Standardisation of river classifications:

Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive

Contract No: EVK1-CT 2001-00089
11th deliverable, due 31/12/04, entitled:

Matrix of possible class boundaries of grades of 'Ecological Status' associated with different methods and stressors

"Contribution of the STAR Project to the European CIS Intercalibration process"

(Paper version)

Compiled by Andrea Buffagni[8], Sebastian Birk[2], Stefania Erba[8], Marcello Cazzola[8], Christian Feld[2], John Murray-Bligh[11], Mike T. Furse[1], Ralph Clarke[1], Thomas Ofenböck[3], Daniel Hering[2], Hania Soszka[19]

Partners 8 (CNR-IRSA, Italy), 2 (UDE, Germany), 11 (EA, United Kingdom), 1 (CEH, United Kingdom), 3 (BOKU, Austria) and 19 (IOEP, Poland)

A project under the 5th Framework Programme
Energy, Environment and Sustainable Development
Key Action 1: Sustainable Management and Quality of Water

Index

1. Introduction

The actual title of the present Deliverable “Matrix of possible class boundaries of grades of ‘Ecological Status’ associated with different methods and stressors” can be complemented by the following sub-title:

“Contribution of the STAR Project to the European CIS Intercalibration process”.

“The intercalibration process is aimed at consistency and comparability of the classification results of the monitoring systems operated by each Member State for the ecological quality elements. The intercalibration exercise must establish values for the boundary between the classes of high and good status, and for the boundary between good and moderate status, which are consistent with the normative definitions of those class boundaries given in Annex V of the WFD.” (ECOSTAT WG 2.A, 2004)

The European Commission and the Member States agreed upon a programme of co-operation to develop a shared approach on the technical issues for implementing the Water Framework Directive (WFD, European Commission, 2000). Such programme is named the Common Implementation Strategy (CIS, European Commission, 2003a). The European Commission, by means of the activity of the Directorate-Generale and its technical institutions, is leading, managing and co-ordinating a number of specific issues aimed at putting the WFD into practice. At this stage of the CIS action, the Intercalibration (IC) process (European Commission, 2003b) is a primary issue to be addressed. This is intended to be a pan-European activity to set harmonized values for class boundaries of National classification systems (European Commission, 2004). Among the contributors to the CIS (e.g. the Commission, the Member States, candidate countries, stakeholders, etc.), a relevant role is devolved to scientific institutions and experts, who should bring the results of scientific research into feasible, pragmatic solutions to the urgent problems linked to the WFD application. To support the Water Framework Directive implementation and strengthen the scientific basis of future biomonitoring and classification of European water bodies, the European Commission co-funded some Europe-wide research projects, such as AQEM (Hering et al., 2004), FAME (Schmutz et al., 2004), STAR (Furse et al., 2004), REBECCA (Rekolainen et al., 2004), etc. The present paper represents a contribution from the STAR and AQEM projects to the delineation of a procedure to perform the Intercalibration process for European rivers.

As a general tendency in the U.S.A. (e.g. Karr et al., 1986; Barbour at al., 1996), and now in Europe (e.g. AQEM Consortium, 2002; Hering et al., 2004), multimetric assessment systems have been applied in a variety of circumstances. This is due to their scientifically sound performance, cost effectiveness and easiness of interpretation (e.g. Thorne & Williams, 1997; Milner & Oswood, 2000). This led European Community delegates, scientists and CIS Working Groups members to initiate the development of Intercalibration Common Metrics to be calculated for river sites within or among GIGs, Member States and stream types (European Commission, 2004).

The STAR and AQEM projects results can support a number of different analytical approaches dealing with the EU “Intercalibration” process. Possible procedures to harmonize European class boundaries based on STAR/AQEM data, for aquatic macroinvertebrates are here provided.

According to Köhl et al. (2000), "harmonization is based on existing concepts which should be brought together in a way to be more easy to compare".

A few definitions

Harmonization

The process by which the class boundaries of MS National methods should be adjusted to be consistent with a common trans-National benchmarking. It must be preformed for High/Good and Good/Moderate status borders.

Note: The harmonization is intended among the results of biological assessment methods only.

Class boundary

The EQR value representing the threshold between two quality classes.

Note: Estimations of uncertainty are not considered in the present paper.

1.1 Objectives of the Deliverable

Among other important aims, the STAR Project worked to make stream assessment methods and results in all of Europe comparable in order to achieve equivalent river quality in future. The present Deliverable deals with the inter-calibration of assessment methods in terms of harminization of their resulting classification (i.e. class boundaries).

In considering future standardisation and harmonization of methods, the 'real world' and the very different situation and traditions in the European countries must be taken into consideration:

- It is unlikely that proven assessment methods will be changed, e.g. RIVPACS in Great Britain, IBGN in France, Saprobic Systems in Austria and Germany, EBEOSWA in The Netherlands and IBE in Italy, at least in the short period available to run the IC exercise. Existing national standards are not likely to be changed. Hence, comparability of results can only be achieved through an inter-calibration. In addition, it is worth mentioning how the data being used for the IC process will largely be already existing data, thus requiring the respective collection and calculation methods to be considered.
- Many existing assessment methods, which will continue to be used in some countries, are not entirely fulfilling the demands of the EU Water Framework Directive. These methods need some adaptation and, in particular, development of procedures for converting results into the series of degradation classes demanded by the EU Water Framework Directive. The results obtained also need to be related to reference conditions. It is crucial that this step is done in a comparable way for all the methods that will be applied in future. This is a central point to be considered in any procedure for the IC process. No simple 'averaging' of existing class boundaries should be considered for the European intercalibration, at least until all MSs assessment systems will be proved to be fully WFD compliant.

- Compared to invertebrates, assessment methods and systems are less developed and far less data are available for stream assessment with fish, macrophytes, phytoplankton and phytobenthos. However, the Water Framework Directive requires methods that take account of these groups. On a European scale, field methods for monitoring fish and phytobenthos are being worked on and committee drafts and draft standards are already in existence. However, in standard water management, these groups are less commonly used than macroinvertebrates and few widespread methods exist for calculating valid indices and converting the results into degradation classes. In reality, it is unlikely that fish and aquatic flora will be applied as frequently as macroinvertebrates in future stream assessment. However, in order to combine the information content of all sources of ecological data, defined and standardised methods are needed to integrate and inter-calibrate the results obtained from different organism groups. Nevertheless, <u>the present Deliverable will mainly deal with invertebrate data</u>, which can be looked at as the most abundant and most complex data. If examples and approaches can be provided and tested for such organisms, it might be comparatively easier to check the appropriateness of IC options for other BQEs later on.

In the AQEM and STAR Projects, a multimetric approach for assessing river quality based on biological indicators was jointly adopted, together with the need to integrate ecological assessments, at a higher level, with quality evaluations based on water chemistry and hydromorphological information. The same approach, even if simplified to make it compatible with the timing and scopes of the IC process, will be considered in this Deliverable.

Indirectly, we will deal with the problem of defining reference conditions and we'll provide a broad scale overview of monitoring datasets available across Europe. The final cross-validation of the results of different MSs' assessment methods will finally depend on the adequacy of the protocol used to derive reference conditions (i.e. to accept/refuse sites as reference sites). After this step will be fully completed, assessment methods can be standardized and the definition of class boundaries between the individual quality classes mutually agreed.

The definition of class boundaries is a necessary step for implementing the Water Framework Directive. It will ultimately be the task of the European Commission to set the class boundaries for what stream conditions are regarded as of 'High', 'Good', 'Moderate', 'Poor' and 'Bad'. This process will need to involve political and ecological considerations. Ecological judgements will need to be based on a variety of messages emanating from a variety of different taxonomic groups and hydromorphological conditions. At present there is no sound scientific basis for integrating these different sources of information. It was the intention of STAR to provide the background science needed to link classes defined by the use of different organism groups and to advice the European Commission how this information may be used, in conjunction with political considerations, to assist the process of defining and delimiting the five grades of ecological status.

The present report is especially aimed at illustrating some possibilities for comparing and harmonizing the MSs classification results by setting comparable boundaries to quality classes, mainly based on invertebrate data. The Deliverable is not focused on intercalibrating biological methods or monitoring systems (this is discussed in STAR Deliverable 8). The example harmonization of the national class boundaries presented here are intended to demonstrate the possibility of identifying and eliminating possible differences by means of different approaches.

The main aims of the present Deliverable can be summarized as follows:

- To illustrate some of the possible procedures to perform the IC exercise across Europe, among those proposed within the STAR Consortium and preliminarily discussed at various CIS WG 2A ECOSTAT meetings.
- To give some examples of their potential applicability across a range of European stream types and GIGs .
- To briefly argue on the results, with the idea of providing a general framework for discussion to people involved in the formal Intercalibration process, to be performed during the next two years.
- To provide general information on potentially suitable metrics at large scale, for an example area in Europe (Central Europe).
- To outline the overall differences among the test datasets and countries in terms of distance from each other, from average conditions or from a tentative benchmarking system, which is supposed to fully satisfy WFD requirements.
- To provide a few full examples of application of some of the considered harmonization procedures.

Issues which are not within the scopes of the present Deliverable can be summarized as follows:

- To define new assessment systems (i.e. the proposed common approaches are explicitly dadicated to the IC exercise and do not represent a proposal for common European assessment systems).
- To propose final options for the IC process.
- To define methods or examples helpful to cover the whole gamut of Water Body Types, Stream Types and Biologcal Quality Eements to be intercalibrated for the WFD implementation.
- To select any final options for technical choices within the single steps of the illustrated procedures (i.e. it is expected that additional and better data will be available during the IC process to support e.g. a robust boundary setting protocol).
- To provide harmonized boundaries for MSs assessment systems (this will be the result of the EU CIS IC process).
- To combine scientific evaluations with socio-economic or political aspects, which should be stressed elsewhere, in the due scene.

1.2 Suitability of the proposed procedures for the three IC Options presented in ECOSTAT WG 2.A, 2004

A general outline of the different options actually considered for the IC process, has been recently presented within the ECOSTAT WG, in the form of a Guidance for the IC process (European Commission, 2004d). Three different Options are present in the Guidance, where their respective advantages and disadvantages are listed and briefly discussed. The following flow-charts are taken from the Guidance, for the three Options.

1.2.1 Option 1: Member States in a GIG area are using the same WFD assessment method

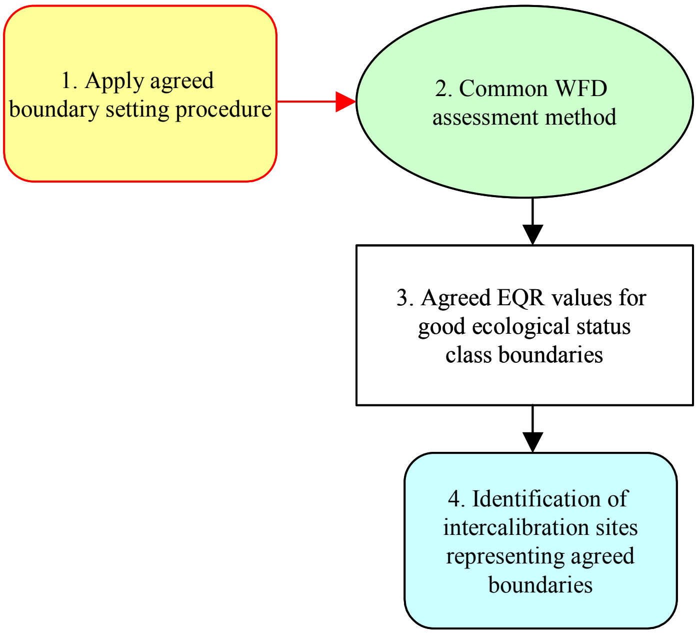


Figure 1.1 Example of how the application of Option 1 might take place (from the IC Guidance, European Commission, 2004).

1.2.2 Option 2: Use of a common metric(s) method identified specifically for the purposes of the intercalibration exercise

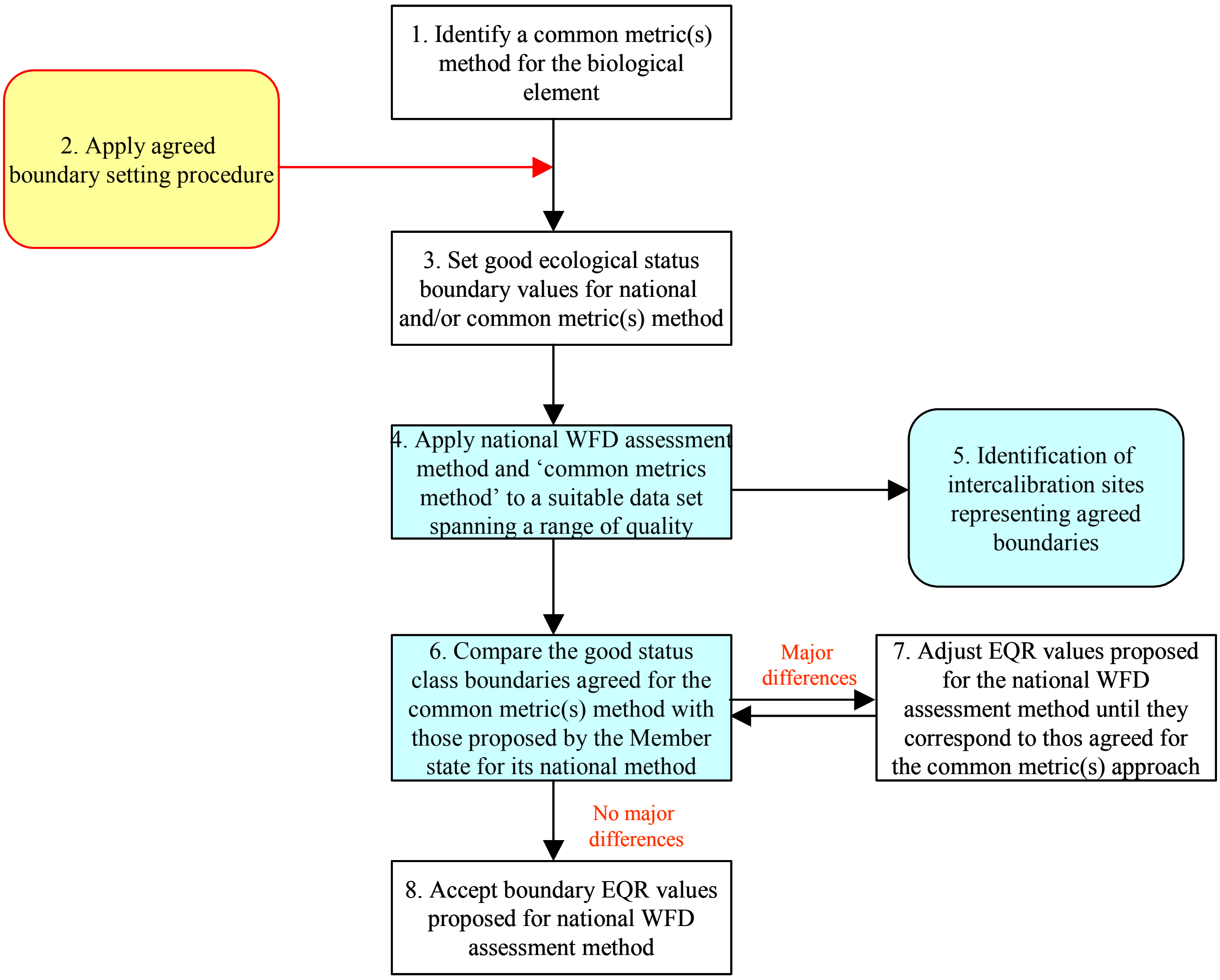


Figure 1.2 Example of how the application of Option 2 might take place (from the IC Guidance, European Commission, 2004).

1.2.3 Option 3: Direct comparison of national methods at intercalibration sites

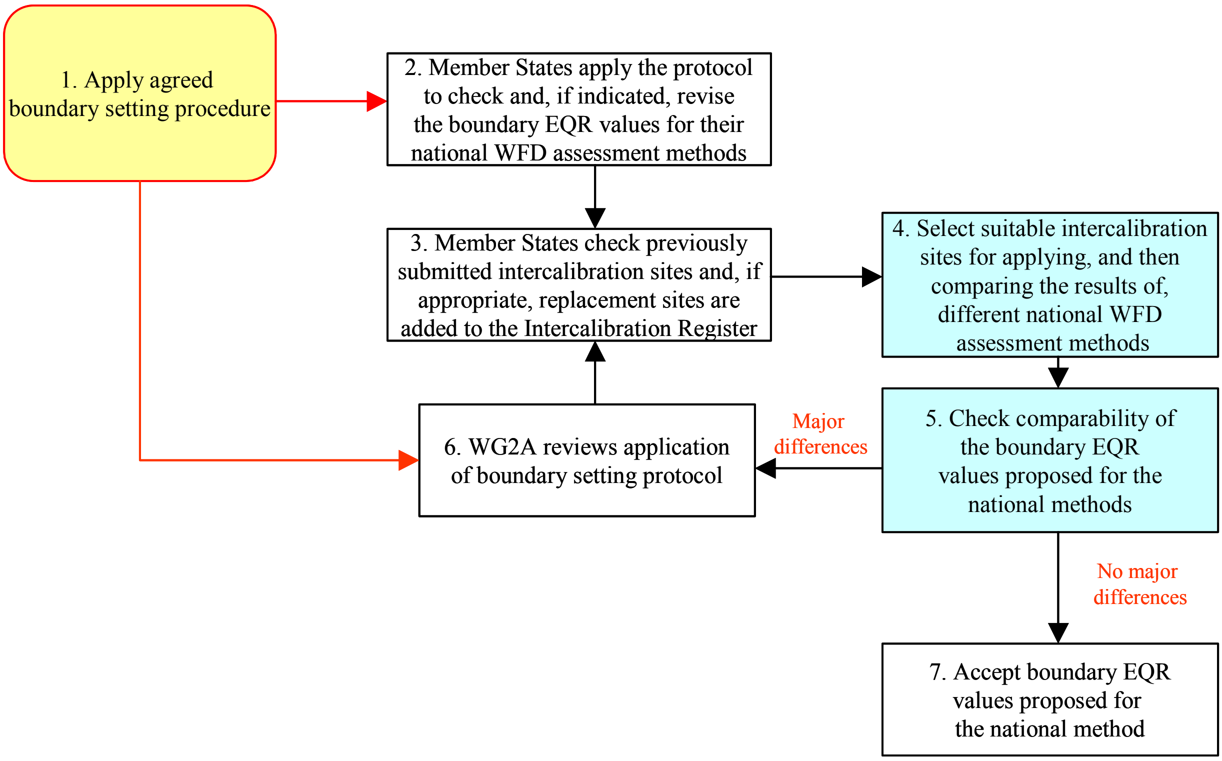


Figure 1.3 Example of how the application of Option 3 might take place (from the IC Guidance, European Commission, 2004).

The use of Option 1, while ideal, will be possible only at the local scale, for a limited number of European countries and stream types. It can adequately support a high degree of comparability among countries as well as consistency with the WFD definitions. This last point must be guaranteed prior to applying any of the options, also because it is the basis for the acceptance/refuse of sites as 'reference' sites. The possible steps of a 'boundary setting protocol' have been outlined in the IC Guidance and individual GIGs are assumed to develop sharp protocols adapted to geographic areas and main degradation factors acting.
Given the central point of consistency with normative definitions, Option 2 put the emphasis on looking for a clear comparability of European class boundaries and assessment systems. To delineate this Option, an Intercalibration Common Metrics (ICMs) approach was proposed (Buffagni & Erba, 2004). In addition, a full application of Option 2 assumes that a trans-National, benchmarking system is adopted. In general terms, it means that the data provided by single MSs should be matched up to International data so that all datasets are compared to the same benchmarking dataset (e.g. within a GIG or, when possible, across GIGs).
Option 3 assumes that the data are compared between countries directly, in the format they are collected by each MS. This Option show clear scientific limitations (e.g. different areas of Europe show quite distinct faunas, different methods were designed to detect the impact of different degradation factors). Moreover, a reasonably acceptable and scientifically sound application of Option 1 would require very detailed data to be provided and jointly examined and/or large field activities not planned at present. If not, there is the authentic danger that the application of this Option will result in a 'political agreement', which is beyond the aims of the present activity.

An important difference between the options is whether the action for its application is done at Member State level, at the GIG level or at a pan-European level (when possible). Another important feature to be considered is the sole use of national metrics (option 3) or the use of Intercalibration Common Metrics (ICMs approach: option 1 and 2).

1.2.4 Hybrid Options

Quite a high number of hybrid options might be conceived, combining single elements of the three main Options. In the IC Guidance, two of them are indicated:
a) To select a ICM index (see Option 2) to underpin the development of the boundary setting procedure, but to follow Option 3 for the application of the procedure to each MSs' data to establish EQR values for relevant boundaries.
b) Boundary values are first established with national classification assessment methods (as in Option 3)(this assumes that compliance to WFD requirements has been demonstrated). The subsequent comparison of the boundary values could then be done with the help of a ICMi approach (as in Option 2).

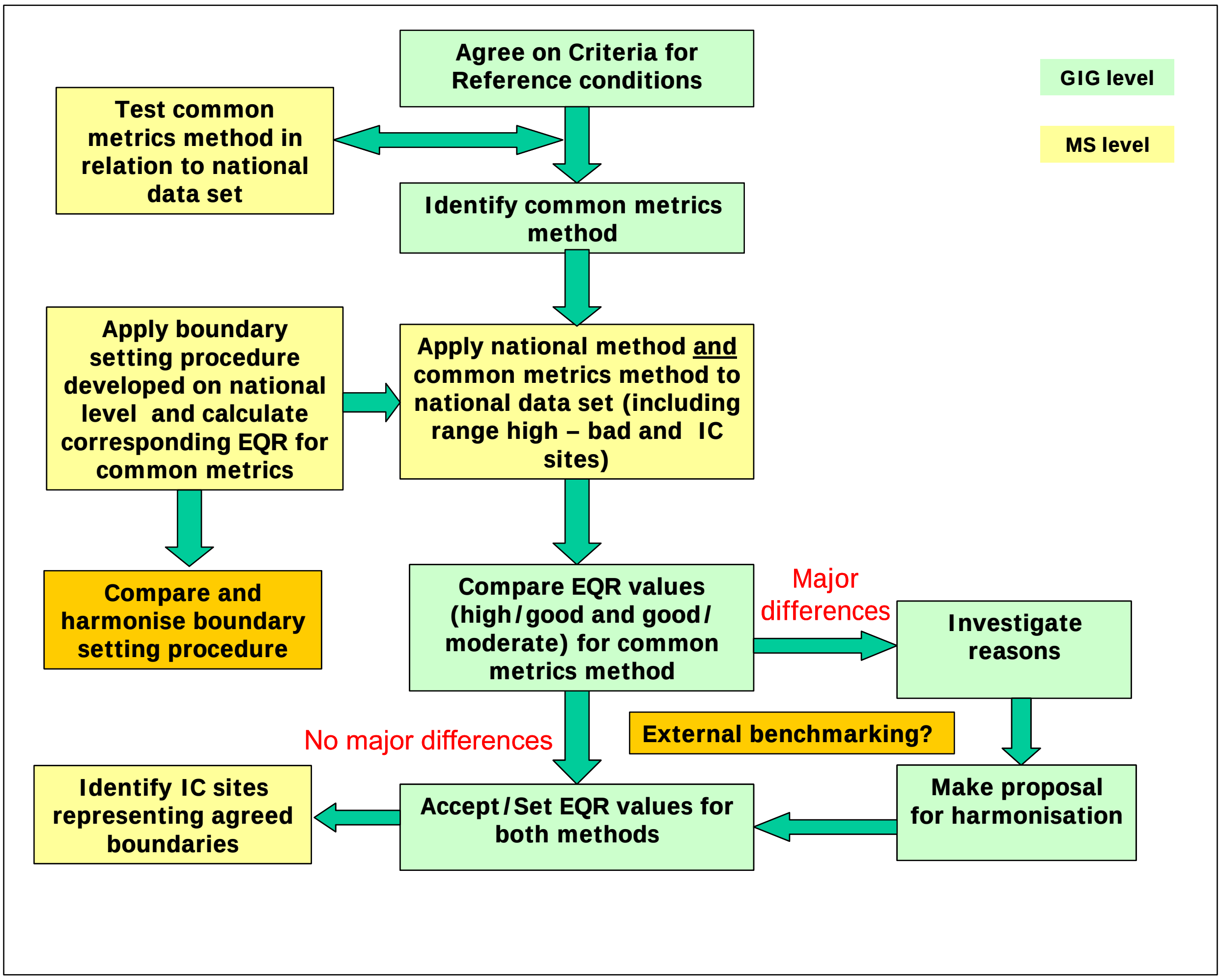


Figure 1.4 Example of a hybrid intercalibration approach, combining elements of Options 2 and 3 (from the IC Guidance, European Commission, 2004).

Some examples for Options 2, 3 and hybrids are given in the present Deliverable, referring to different European areas (GIGs) and stream types.

In the Deliverable, we discuss some possible procedures for the determination of European class boundaries of Ecological Status. In doing so, we are aware that class boundaries are likely to be set by the European Union themselves. We therefore envisage that our results might serve as a proposal for the EC and must be capable of being re-calculated when the final class boundaries are set. We are aware of the large range of possibilities and options to perform such an important and potentially difficult task like the IC process of European water bodies. It is not the intention of this Deliverable and of the whole STAR Consortium to push one option or another. The general applicability of the approaches – with the focus on rivers and aquatic invertebrates – is being evaluated, through a scattered application to datasets provided by STAR partners, other scientific institutions and environment agencies or Environmental Ministries from around Europe. A relevant part of the data presented and processed here (see Chapter 4) were provided as a part of the ongoing pilot IC exercises. In particular, most test datasets refers to the Central GIG countries and activities, with notable exceptions from the South of Europe (e.g. France and Italy), where preliminary actions for the pilot started quite early during 2004 and lead to an important improvement in the delineation of IC Options.

1.3 Participating institutions and countries

In this section, a list of the Institutions and countries that contributed, by directly writing or providing data, to the present Deliverable is reported.

STAR partners who participated to the compilation of Deliverable 11

- Consiglio Nazionale delle Ricerche…………………………..Italy (coordination)
- University of Duisburg-Essen…………………………......... Germany
- Environment Agency, Bristol..…..……………………………United Kingdom
- Centre for Ecology and Hydrology …………………………. United Kingdom
- BOKU – University of Agricultural Sciences……………….. Austria
- Intitute of Environmental Protection, Warsaw………………...Poland

Institutions that provided test datasets
STAR partners

- Consiglio Nazionale delle Ricerche…………………………..Italy
- Environment Agency, Bristol..…..……………………………United Kingdom
- Ministry of the Environment, Warsaw……………………......Poland
- Universitity of Duisburg-Essen…………………………….....Germany
- National Environmental Research Institute………………….. Denmark

Non-STAR partners

- ARPA Lombardia (Parabiago Dep.)……………...…………..Italy
- Estonian Agricultural University……………………………...Estonia
- University of Vigo………….. …………………..………….Spain
- Flemish Environment Agency………………...……….……Belgium

Institutions that provided benchmark datasets
STAR partners

- Consiglio Nazionale delle Ricerche…………………………..Italy
- University of Duisburg-Essen…………………………......... Germany
- Centre for Ecology and Hydrology …………………………. United Kingdom
- BOKU – University of Agricultural Sciences……………….. Austria
- Masaryk University Brno……………………………………. Czech Republic

Non-STAR partners

- CEMAGREF, Lyon……………………………………………France

Institutions that provided test datasets, which were not used in the calculations.

- CEMAGREF, Lyon……………………………………………France
- Royal Haskoning…..The Netherlands

2. Procedure and general topics

2.1 Summary of the STAR ICMi intercalibration procedure for macro-invertebrates – Comparison phase

2.1.1 General statement

- The summary procedure for Intercalibration presented here is a technical supplement to the ECOSTAT WG 2.A discussion paper distributed in February 2004 (Buffagni & Erba, 2004).
- The details provided can be considered as a complement to the description of Option 2 (and hybrids): *Use of a common metric(s) method identified specifically for the purpose of the intercalibration exercise* (ECOSTAT WG 2.A, 2004).

For intercalibration aims, the direct comparison of classification results from different countries is not possible, due to the natural river variability and faunal variation. Even if the IC is run on (broad) river types, the fauna can differ for biogeographical reasons even in similar physical contexts. In addition, the existing methods have different sampling strategies and laboratory procedures, and are also based on different concepts. This is why an intermediate step such as the Intercalibration Common Metrics index (ICMi) is needed (see Buffagni & Erba, 2004; ECOSTAT WG 2.A, 2004).
The example provided here refers to river invertebrates but the procedure can be applied to all Biological Quality Elements, if enough data are available, as well as to other water body types.
The European WFD intercalibration process should compare National assessment methods whose consistency to the normative definitions is demonstrated.

2.1.2 Aim
The aim of the presented procedure is to compare biological WFD class boundaries for rivers across the whole of Europe, despite differences in sampling, analytical and computational methods used by different national monitoring and classification schemes. It is likely to be supplemented by more precise bilateral and multilateral intercalibration between national methods that are similar and which may be based on more detailed taxonomic resolutions.

2.1.3 Overview of intercalibration via ICMi
Intercalibration involves 2 main steps:

1. Comparison of existing national class boundaries
2. Harmonisation (adjustment) of boundaries

Provided the consistency of each of the assessment methods to the normative definitions, harmonisation will be necessary only if the existing class boundaries differ significantly.
Ideally, the class boundaries of the National method for each Member State should be adjusted to correspond to European, trans-national boundaries, e.g. set on the basis of an International, WFD compliant benchmarking system. Alternatively, when the option above is not applicable, the harmonization might be performed through a 'bilateral' comparison of two national methods (but this will hardly guarantee a complete European comparability). For both, the way of making the MSs quality classifications comparable, is to calculate a set of agreed Intercalibration Common Metrics and to combine them into an ICM index.

2.1.5 Summary of the concept of comparison

For comparison, a range of general metrics relating to tolerance, abundance/habitat and richness/diversity are calculated and combined into an Intercalibration Common Metric index (ICMi)(Buffagni & Erba, 2004). The ICMi and the national classification metrics are converted to EQRs by normalization, i.e. dividing them by the value for the reference state for the particular IC River type. This reference state is determined by a specified procedure (see below). The relationship between ICMi and the national classification metric is determined by simple regression. The class boundaries are converted from values of the national classification to values of ICMi for comparison with boundaries of other countries' national systems.

2.1.6 Summary of the concept of harmonization

The concept of harmonization deals with the common understanding of ecological status, expecially for what can be considered good and what moderate. In the present Deliverable the harmonization is carried out shifting boundaries, High/Good and Good/Moderate in order to reduce/eliminate difference among different datasets and methods. The basis for the harmonization is the calculation of a common intercalibration index derived from the combination of a pool of selected metrics. The options presented for the harmonization follow comparisons *via* ICMi. In one case, an example of harmonization based on the selection of median boundaries values derived from the comparison of test datasets is presented (see chapter 7).

Differently, in another example, the procedure involves the comparison of test datasets to WFD-compliant, trans-national datasets (benchmark datasets) for which a Best Available Classification is provided (i.e. STAR/AQEM, see chapter 8). In this last example the ICMi calculated for STAR/AQEM biological classes (benchmark dataset) were compared to the values observed for the corresponding National system classes (test dataset) by means of the Mann-Whitney U test (Helsel & Hirsch, 1992).

2.1.7 Summary description of the harmonization procedure(indirect comparison via ICMi)

A dataset assembled for the purposes of the WFD (benchmark dataset), including quality classification of sites, is identified (e.g. STAR/AQEM), which should be independent form National monitoring datasets.

The relationships between the environmental quality (e.g. water pollution, habitat degradation, acidification, etc.) and the biological response are examined for such dataset, to properly interprete the observed range of e.g. metric values and check the proposed ecological classification criteria. For each site, the "Best Available Classification" (BAC) is provided/derived, which fulfils the WFD requirements.

A statistical comparison is executed between the ICM index values found in the benchmark dataset and the same observed in the test dataset, firstly considering Good status class.

If the ICM index values based on the two classification schemes significantly differ, the class boundary Good/Moderate for the National dataset is shifted in order to eliminate the differences. After the adjustment of the G/M boundary (corresponding to no significant differences according to the two classification systems), the boundary High/Good is considered. The procedure of statistical comparison between high status classes, as it was carried out for Good status, is repeated. The new, harmonized boundaries for the National classification system are thus set for High/Good and Good/Moderate classes.

2.1.8 Criteria used for ICMs selection

The Intercalibration Common Metrics selected and presented here showed a high correlation with the quality classification of the considered sites and stream types. The analysis for the selection of metrics was done considering intra type datasets (e.g. M1, C1) and cross type datasets (e.g. AQEM and STAR datasets), as well as considering recent metric selection experiences (e.g. AQEM Consortium, 2002; Buffagni *et al.*, 2004; Pinto *et al.*, 2004). The potential applicability of the metrics over a wide geographical scale was taken into account. The identification level used for the calculation of these metrics is family. In Annex 2 the selected metrics are reported.

The metrics reported here are tentative. Almost certainly, some changes should occur in the metric composition of the ICM indices to be used for the intercalibration exercise, e.g. in different GIGs. In particular, the metric 1-GOLD might result as not properly describing the quality gradients across Europe. Abundance-based metrics are here included to fit the normative (WFD) definitions. The applicability and suitability of this metric category for different datasets and stream types across Europe will be checked within GIGs.

2.1.9 Normalization options

Why to normalize the invertebrate data?
The invertebrate samples to be compared across Europe for the IC process are often collected with obviously different field procedures. The sampling procedures performed can vary widely, in terms of technique (e.g. net type and proportionality of the sample to different habitat) and sampled area (i.e. quantitative, semi-quantitative or qualitative samples). Also, the calculation formulae and the classification criteria show broad differences. In addition, the range of river types to be compared across Europe greatly differ in fauna for natural reasons (e.g. zoogeographic, climatic, hydrological).
In order to gain comparability among the datasets, the ICMs calculated on the datasets are normalized, i.e. the score of each ICM is divided by a reference value. Some possible options were considered to define this value (Chapter 3.3).

2.1.10 ICM index

After the normalization of the metrics they were combined into a Intercalibration Common Metric index (ICMi)(Chapter 3). Metrics are grouped into 3 groups Tolerance, Abundance/Habitat and Richness/Diversity (see Chapter 3/Annex II). To obtain the final multimetric score a different weight was attributed to the metrics within each group (see Chapter 3/Annex II), giving higher importance to the metrics based on the whole community (Buffagni et al., 2004). The same weigh is attributed to each metric group (0.333).

2.1.11 References

AQEM Consortium, 2002. Manual for the application of the AQEM system. A comprehensive method to assess European streams using benthic macroinvertebrates, developed for the purpose of the Water Framework Directive. Version 1.0, February 2002. www.aqem.de.

Armitage, P. D., D. Moss, J. F. Wright & M. T. Furse, 1983. The performance of a new biological water quality scores system based on macroinvertebrates over a wide range of unpolluted running-water sites. Wat. Res. 17: 333-347.

Böhmer,J., C. Rawer-Jost & A. Zenker, 2004. Multimetric assessment of data provided by water managers from Germany: assessment of several different types of stressors with macrozoobenthos communities. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe. Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 215-228.

Buffagni A. & S. Erba, 2004. A simple procedure to harmonize class boundaries of European assessment systems. Discussion paper for the Intercalibration process – WFD CIS WG 2.A ECOSTAT, 6 February 2004, 21pp.

Buffagni A., S. Erba, M. Cazzola & J. L. Kemp, 2004. The AQEM multimetric system for the southern Italian Apennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe. Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 313-329.

ECOSTAT WG 2.A, 2004. Guidance on the Intercalibration Process. Final Draft 4.1. October, 14th 2004.

Helsel, D. R. & R. M. Hirsch, 1992. Statistical Methods in Water Resources. Elsevier, Amsterdam, 522 pp.

Hering D., O. Moog, L. Sandin & P. F. M. Verdonschot, 2004. Overview and application of the AQEM assessment system. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 1–20.

Ofenböck, T., O. Moog, J. Gerritsen & M. Barbour, 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 251-268.

Pinto, P., J. Rosado, M. Morais & I. Antunes, 2004. Assessment methodology for southern siliceous basins in Portugal. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 191-214.

2.2 Identification level

2.2.1 Taxonomic resolution in aquatic biomonitoring and reserch

The issue of taxonomic resolution has been discussed manifold (e.g. Resh & McElravy 1993, Stubauer & Moog 2000, Schmidt-Kloiber & Nijboer 2004). Practical considerations (e.g. lack of taxonomic expertise, unavailability of autecological information) and administrative needs (e.g. cost efficiency, lack of time and human resources) give reasons for identification to higher taxonomic levels.
The level at which scientists identify freshwater macroinvertebrate varies due to the high number of species of different orders that compose the benthic community and to different available knowledge. Sometimes the informations to go to species are restricted to experts and not always available to end-users. The scientific community it's used to call the attitude of reaching different taxonomic resolution depending on the final aim: best available taxonomy or lowest practicable level. Generally the identification level should be chosen depending on the purpose of the different studies, on the data analysis techniques that are used and on the taxonomical groups studied (Resh and McElravy 1993). The central question is to define

when is necessary to identify at the lowest practicable level and if this options is obtainable in terms of human resources and available time. The lowest practicable level that is possible to reach depends on the technical taxonomical expertise on each different taxon that constitutes the benthic community, and the availability of the needed time to reach a lower identification than family level. Higher precision in bioindication is reached with data on species level. As illustrated in the niche concept each species has evolved special abilities to exploit resources and to cope with the heterogeneity of its habitat. The occurrence of specific species assemblages is therefore a result of the present environmental conditions. In case autecological requirements of characteristic species associations are well established they provide useful evaluation criteria for the structural and functional quality of freshwater ecosystems making them powerful bioindicators for the ecological status of aquatic habitats. Because of autecological differences among related species that form the higher levels, the use of higher taxonomic levels may result in a loss of information relevant for bioindication. Lack of species level information may reduce the ability to detect more subtle changes in ecological quality. On the opposite hand it has to be considered as a genus/species identification level may add variability due to the increased possibility to make mistake on the individual's identification, a situation that can reduce the objectivity of biological assessment data and analysis.

During the EU funded project STAR (www.eu-star.at) it was tested the error made on sorting and taxonomical identification. Some preliminar results seem to indicate as, in some cases, the percentage of misclassification can be 20% even at family level. The final purpose of improving the available ecological information reaching lower level of identification, potentially useful for better assess the quality status of rivers, can be strongly affected by identification errors. Apart of the taxonomic resolution that is been chosen, emerge the call for an evaluation of the accuracy and the precision of taxonomic analysis. Stribling (2003) point out that often a key source of errors is due to the human factor. The employ of different taxonomists and the ability of each researcher could affect the quality of the result, due usually to problems regarding the lack of time and the limited experience or training of the taxonomist. The Environment Canada (1993) provides some recommendations: the identifications should be verified by an expert in the taxonomic group of interest, persons who carry out the identifications should be named, with appropriate details of qualifications and literature and taxonomic keys used for benthos identification should be referenced.

Meanwhile bioassessment methods are rapidly growing and evolving during the last 20 years there is a decline of research taxonomist despite an increased demand for taxonomist expertise (Stribling, 2003). Large funding resources have been diverted to other ecological fields while taxonomic, faunistic and autoecological investigations, essential as a basis for any applied ecological study, have been almost entirely abandoned. Research is needed now on benthic macroinvertebrate taxonomy and distribution to improve bioassessment as a water resource management tool (Buffagni et al., 2001). In some areas identification levels lower than family or genus are difficult for a lack of basic knowledge about taxonomic and ecologic composition of the benthic fauna. In several south European areas, new species have probably still to be recorded for the first time or described (e.g., Belfiore & Buffagni unpublished data; Pinto & Puig pers. comm.; Rossaro pers. comm.; Valle pers. comm.). This is expected for some major macroinvertebrate groups (e.g., Ephemeroptera, Trichoptera and Diptera). For instance, with regard to the mayflies in Italy, a number of studies revealed that comprehensive data on the taxonomy, distribution and ecology of most species are not available (e.g., Buffagni & Belfiore, 1994). In recent years, endemic species have been described (e.g., Belfiore 1995; Belfiore *et al*. 1997) and many others have been reported for the first time (e.g., Belfiore & Buffagni 1994; Belfiore & Desio 1995; Buffagni 1997; 1998; Buffagni & Desio 1999), but information is still restricted to specialist journals and identification keys are not up-to-date. This lack of basic information and the general difficulty to correctly identify

individuals are the major problems for choosing a low taxonomic resolution. In aim to preserve some of the species-level information, without the necessity to identify to species the collected organisms, Buffagni (1997) proposed to achieve this problem fixing an intermediate identification level between genera/family and species. To approximate the specific composition of the community a definition of benthic groups with fixed identification level could be employed. These groups, either taxonomic or morphotaxonomic, are named Operational Units (OU). These groups are created following the philosophy of join together species by similar autoecological features and whenever it's possible was performed on the basis of the most easily visible characters that could be singled out, so as not to complicate identification. The OU allows reducing the identification's errors, if compared to species level and, at the same time, offers a more detailed ecological information. In Italy was proposed the following aggregation of the Italian mayfly species at different levels: genus level (24 OU), morpho-taxonomic group level (11 OU) and species level (*Baetis rhodani*). A linear regression analysis between species numbers and OU numbers on 150 samples collected in Northern Italy was conducted. A correlation coefficient equal to 0.98 (OU = 0.18 + 0.91s; $p<0.001$) was found, showing OU number can be considered a good approximation of species number.
It can be summarised as there is the need for improving taxonomical expertise. These information should constitute the basis for the development of sensitive metrics and assessment systems possibly based on higher taxonomic resolution in order to reduce the identification effort.

2.2.2 Identification level used for monitoring in Europe and WFD requirements

The Water Framework Directive (EU, 2000) establishes new European standard for assessing river quality. Three main types of monitoring are indicated by the Water Framework Directive: surveillance, operational and investigative (E.U. 2000). Each of them has different aims and frequency of application, focusing on different information to be obtained for a river site (WFD, Annex V, 1.3.1/3). The WFD leaves it to the Member States to "identify the appropriate taxonomic level required to achieve adequate confidence and precision in the classification of the quality elements" (Annex V, WFD). Especially in investigative and operative monitoring (Annex V, WFD) pressure specific assessment is required to assess the impact of different stressors, in order to guide future management. Better taxonomical resolution allows for more detailed ecological interpretation of monitoring data to detect the cause of degradation in investigative monitoring. To measure the exact biological response to a pressure in operative monitoring species level data are, in most cases, inevitable.
Every European country has historically developed different bioassessment methods. The methods use different techniques for sampling, sorting, use different taxonomic resolution and analysis of the data. All these features contribute to make difficult to compare the results obtained from the various national methods. Regarding the identification level, the family level is used in different national assessment methods all over Europe in indices such as BMWP, ASPT (Armitage *et al.*, 1983), IBGN (Vernaux *et al.*, 1983)etc. Nevertheless, a more detailed identification level are already used in national methods such as: IBE in Italy (genus and family level, APAT-IRSA/CNR, 2004), Saprobic index in Germany and Austria (species level, BMLF, 1999; Friedrich & Herbst, 2004).
In those parts of Europe, where organic pollution is still the overwhelming stressor affecting running waters, assessment systems based on family level are sometimes sufficient. The family level, easier to be identify, seems to be appropriate whenever there is the need to determine large differences between sites, with the result of a coarse or primary values, that in case of bioassessment means a preliminary rough classification. In areas taxonomically poor,

where it's known that the number of species and genera it's similar to the number of families, the family level is appropriate and enable a significant saving of resources, in terms of time spent and money invested . Lenat and Resh (2001) suggest that situations where a lower identification is advised is the detection of small differences between sites or between date, and in conservation studies. This studies need to be done with an especially accurate method because of the presence of rare species. Particularly, steep pollution gradients can easily be assessed with a large number of invertebrate-based assessment systems, such as ASPT, BMWP or IBE (family based), Belgian Biotic Index or Danish Stream Fauna Index (mixed taxonomic level) and Saprobic Systems (species level). The results of these assessment systems are in many cases comparable. As soon as organic pollutions vanishes or the pollution gradients are less steep, the above mentioned systems must be replaced or supplemented by other assessment methods. In Central European countries (e.g. Germany, Austria), where organic pollution is nowadays a side-problem in river management, assessment systems focussed on the detection of organic pollution give almost everywhere the same results. The dominant stressors, which are affecting Central European rivers today (e.g. hydromorphological degradation, catchment land use, eutrophication, pesticides) are acting in a much more subtle way. However, they still have effects on the biocoenosis, which still can be detected. There are indicator taxa for certain habitats, while others reflect catchment integrity or hydromorphological structures. In most cases, these are species level information (Feld & Hering, submitted).

2.2.3 Taxonomic requirements of the ICM index

The ICM index for benthic invertebrates is calculated with taxonomic data on family level. This allows for integrating monitoring data of various countries to gain intercalibration across a GIG. The monitoring programmes of several countries in Europe operate on species level (e.g. Austria, Czech Republic, Germany, Netherlands) or mixed taxonomic level (e.g. Belgium, Denmark, Sweden, Italy). For the intercalibration exercise outlined in this deliverable, these data have to be adjusted to higher taxonomic level, i.e. summing up the abundances of all species and genera within one family. This procedure sets common ground for the intercalibration via the ICM index and serves solely the purpose of comparison of national assessment methods. Thus, the ICM index is calculated using family level data. These results are then correlated with classification results of the national assessment system which have been obtained by application of the taxonomic resolution in use for the national monitoring programme. Even if the ICM approach does not imply any recommendations for the adequate taxonomic level needed in biomonitoring, it can be seen as the different metrics calculated at family level show, in most datasets, high correlation with ecological quality gradient. In chapter 6.3, different metrics calculated at different taxonomic level are compared. Even when compared to metrics based on species level, the metrics based on family level have comparable correlation coefficient. This confirms that the use of an ICMi based on family level can be a good solution for representing sites quality.

2.2.4 Level of taxonomic identification needed for bilateral comparison

Bilateral comparison of national assessment results, as a direct way of intercalibration in transboundary catchments, is based on species level data. In relation to the ICM approach, which aims at a coarse comparison of assessment systems on a Europe-wide scale, this direct intercalibration technique focuses on class boundary adjustment between two countries.

Since the ICM index is designed to respond to general degradation, its application to intercalibrate stressor-specific assessment methods (e.g. Saprobic Systems) might reveal inadequate. Bilateral comparison based on species level data offers a possibility to integrate these systems in intercalibration without losing the precision of low taxonomic levels used in stressor-specific assessment.

References

Bailey R C, Norris R H and Reynoldson T B (2001) Taxonomic resolution of benthic macroinvertebrate communities in bioassessments. Journal of the North American Benthological Society, 20, 280-286.

Barton, David (1996): The use of Percent Model Affinity to assess the effects of agriculture on benthic invertebrate communities in headwater streams of southern Ontario, Canada. Freshwater Biology 36 (2), 397-410.

ENVIRONMENT CANADA: Guidelines For Monitoring Benthos In Freshwater Environments

Feld, C.K. & D. Hering: Linking macroinvertebrate taxa and derived ecological metrics to hydromorphology and land use at different spatial scales in Central European lowland rivers: a multivariate approach. Submitted to Freshwater Biology.

Fore, L.S., K. Paulsen, and K. O’Laughlin. 2001. Assessing the performance of volunteers in monitoring streams. Freshwater Biology 46: 109-123

GAYRAUD, BERNHARD STATZNER, PIERRE BADY, ARNE HAYBACHP, FRANZ SCHO¨ LL, PHILIPPE USSEGLIO-POLATERA, ‡AND MICHEL BACCHI: Invertebrate traits for the biomonitoring of large European rivers: an initial assessment of alternative metrics, Freshwater Biology (2003) 48, 2045–2064

Ian R. Waite, Alan T. Herlihy, David P. Larsen, N. Scott Urquhart and Donald J. Klemm

The effects of macroinvertebrate taxonomic resolution in large landscape bioassessments: an example from the Mid-Atlantic Highlands, U.S.A. Freshwater Biology, Volume 49, Issue 4, Page 474-489, April 2004

Kilgour B.W. & Barton D.R. (1999) Associations between stream fish and benthos across environmental gradients in southern Ontario, Canada. Freshwater Biology, 41, 553566.

King, R.S., and C. J. Richardson. 2002. Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment. Journal of the North American Benthological Society 21 (1):150-171.

Lenat, D.R. and V.H. Resh. 2001. Taxonomy and stream ecology– The benefits of genus and species-level identification. Journal North American Benthological Society, 20, 287-298.

Leska S. Fore*, Kit Paulsen and Kate O'Laughlin. Assessing the performance of volunteers in monitoring streams, Volume 46 Issue 1 Page 109 - January 2001, Freshwater Biology

Moulton S.R., Stribling J.B., and Lester G.T: Research vs. production taxonomy: defining the quality of taxonomic data necessary for bioassessments. J. N. Am. Benthol. Soc., 2003, 22(4):621–631

Nielsen, R. J. Shiel, F. J. Smith, Ecology versus taxonomy: is there a middle ground?, Hydrobiologia, Volume 387-388, Issue 0, 1998, Pages 451 – 457

Pik Anthony J. , J. M. Dangerfield, Roger A. Bramble, Craig Angus, David A. Nipperess, The Use of Invertebrates to Detect Small-scale Habitat Heterogeneity and its Application to Restoration Practices, Environmental Monitoring and Assessment, Volume 75, Issue 2, April 2002, Pages 179 – 199

Rabeni Charles F. , Ning Wang, Bioassessment of Streams using Macroinvertebrates: Are the Chironomidae Necessary?, Environmental Monitoring and Assessment, Volume 71, Issue 2, October 2001, Pages 177 – 185

Resh, V. H. & E. P. McElravy (1993): Contemporary quantitative approaches to biomonitoring using benthic macroinvertebrates. Freshwater biomonitoring and benthic macroinvertebrates. D. M. Rosenberg and V. H. Resh. New York, Chapman & Hall**:** 159-194.

Ryan S. King 1 And Curtis J. Richardson (2002): Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment J. N. Am. Benthol. Soc., 2002, 21(1):150–171

Schmidt-Kloiber, A. & R. C. Nijboer (2004): The effect of taxonomic resolution on the assessment of ecological water quality classes. Hydrobiologia **516**(1): 269-283.

Stubauer, I. & O. Moog (2000): Taxonomic sufficiency versus need for information - comments based on Austrian experience in biological water quality monitoring. Verh. Internat. Verein. Limnol. **27**: 1-5.

3. Intercalibration Common Metrics (ICM)

Based on previous experience with similar sets of metrics (e.g. AQEM Consortium, 2002; Buffagni *et al.*, 2003; Pinto *et al.*, 2003), six metrics were selected to test the procedure of intercalibration and to give examples of a possible harmonization of European class boundaries. From the initially proposed metrics (Buffagni & Erba, 2004), the selected metrics result from a wide discussion that took place within GIG meetings (especially Mediterranean and Central) and AQEM/STAR consortium.

The main criteria used for the metrics selection are:

- their consistency with WFD definitions, i.e. they have to deal with the three main aspects outlined for aquatic invertebrates in the WFD (tolerance, richness and abundance)
- their ability in describing degradation gradients and discriminating different quality classes, i.e. based on existing literature and AQEM/STAR projects experience
- the possibility of calculating them from a wide range of geographical contexts, i.e. where different effort is placed on the monitoring exercise and different expertise is available for taxonomic identification.

Looking at single metrics' behaviour, the following criteria were followed:

- o The single metrics - and their combination into an ICM index (see the box below) - should follow well the degradation gradient described by most biological assessment systems of European MSs
- o The variability of the metrics at reference sites should preferably be low
- o As most invertebrate methods used up to now in Europe do not require a quantitative sample to be collected, the use of logarithmic transformation for abundance metrics has to be preferred (to derive broad abundance categories).

The selected metrics, here termed Intercalibration Common Metrics: ICMs (Buffagni & Erba, 2004; Table 3.1), have been calculated for all the samples from each of the considered test and benchmark datasets (see Chapters 4 and 5). They can be clustered in two groups: qualitative metrics, only using qualitative information; quantitative metrics, based on abundance estimates.
The identification level initially proposed when the ICMs concept was presented (Buffagni & Erba, 2004) corresponded to the Sistematic Units in use in Italy for the application of the IBE method (APAT-IRSA, 2004). Later on, after a joint discussion at the ECOSTAT and Intercalibration meetings on data availability around Europe, they were set to be at the Family level (Erba et al., 2004). The identification level here adopted for the calculation of these metrics is family.

A few definitions

(Biological) Metric

A metric is a calculated index, which represents some aspects of the biological population's structure, function or other measurable characteristic that changes in a predictable way with increased human influence (Barbour *et al.*, 1999).

Intercalibration Common Metric (ICM)

A biological metric widely applicable within a GIG or across GIGs, which can be used to derive comparable information among different countries/stream types

Notes: (a) Different GIGs may adopt different sets of ICMs, according to the quality of the available data (e.g. identification level) and sampling procedure adopted (e.g. qualitative *vs* quantitative). Nevertheless, a set of ICMs applicable across Europe would ensure a full comparability at the pan-European scale. (b) Whenever an accepted and well-performing assessment method is available for a given stream type, the ICM index should not be considered as a tool for classification beyond the scopes of the IC process. (c) The metrics used in the present Deliverable are example metrics which, while showing an overall applicability, might be profitably substituted or integrated by others at the GIG scale.

Intercalibration Common Metric Index (ICMi)

The combination of the values obtained by ICMs into a single multi-metric index.

Notes: As multi-metric systems are better suitable than single metrics to assess ecological quality and to describe biological communities, more than one metric should preferably be considered when comparing class boundaries. Such metrics can be combined into a simple ICM index (e.g. by averaging the single metrics score) for a straightforward comparison across MSs.

Tab. 3.1 The Intercalibration Common Metrics (ICMs) used in the analysis and comparisons shown in the present Deliverable.

Intercalibration Common Metrics (ICMs) used in the STAR Deliverable 11 for the Intercalibration procedure

Information type	Metric type	Metric name	Taxa considered in the metric	Literature reference	weight
Tolerance	Index	ASPT	Whole community (Family level)	e.g. Armitage *et al.*, 1983	0.333
Abundance/Habitat	Abundance	Log_{10} (Sel_EPTD +1)	Log(sum of Heptageniidae, Ephemeridae, Leptophlebiidae, Brachycentridae, Goeridae, Polycentropodidae, Limnephilidae, Odontoceridae, Dolichopodidae, Stratyomidae, Dixidae, Empididae, Athericidae & Nemouridae)	Buffagni *et al.*, 2004; Buffagni & Erba, 2004	0.266
	Abundance	1-GOLD	1 - (relative abundance of Gastropoda, Oligochaeta and Diptera)	Pinto *et al.*, 2004	0.067
Richness and Diversity	Taxa number	Total number of Families	Sum of all Families present at the site	e.g. Ofenboch et al., 2004	0.167
	Taxa number	number of EPT Families	Sum of Ephemeroptera, Plecoptera and Trichoptera taxa	e.g. Ofenboch et al., 2004; Böhmer et al., 2004.	0.083
	Diversity index	Shannon-Wiener diversity index	$D_{S-W} = -\sum_{i=1}^{s} \left(\frac{n_i}{A}\right) \cdot \ln\left(\frac{n_i}{A}\right)$	e.g. Hering et al., 2004; Böhmer et al., 2004.	0.083

3.1 Why to use ICMs?

Some simple and apparent concerns make the use of the ICMs approach advantageous. Some of them are reported below.
The use of Intercalibration Common Metrics for the IC process can be adopted because:

- They support the translation of MSs' assessment systems results into a single kind of information, which makes different methods comparable (see Chapter 6)
- They can be used to simply compare MSs assessment systems results as well as to harmonize class boundaries at the GIG or pan-European scale (see Chapter 7)
- They make the quality judgement express by MSs systems WFD compliant, in terms of tolerance, richness/diversity and abundance information
- These three broad categories support the detection of a variety of impact types, which concur in determining the general quality of the site in the large majority of European rivers
- Their use allows to go back to the original information collected by each MS (i.e. the invertebrate samples) with their own method, thus supporting the use of large existing datasets all over Europe
- As they can be selected to be metrics of quite general applicability, they can support a large scale comparison of European streams and rivers
- The use of ICMs is encouraged by the European Commission Intercalibration Guidance (2004)

3.2 Weights of the ICMs in the calculation

The metrics have been weighted according to the conceptual group they belong. Into each group, more weight is given to more robust metrics (eg. metrics taking in account the all community). Each ICM normalized value is multiplicated by its weight (see also Table 3.1).
The selection of the weights to be used followed an analysis of the correlation of the ICM Index resulting by different combinations with some example tast datasets (especially C1, M1, M2). In the following table (3.2) some of the results of such comparison are reported.

The weights of the six ICMs finally adopted are reported below:

- ASPT * 0.333
- Log_{10}(sel_EPTD+1) * 0.266
- 1-GOLD * 0.067
- N-taxa * 0.167
- EPT * 0.083
- Shannon-Weiner * 0.083

The ICMi value is calculated by sum of all the ICMs.

3.3 Scaling and Normalizing EQR values: a central point in the WFD Intercalibration process

3.3.1 Introduction

The Water Framework Directive demands for numerical values used to describe ecological quality to be expressed as Ecological Quality Ratios (EQRs). This means that those values must be related to a 0 to 1 scale, where 0 corresponds to the lowest obtainable value (i.e. lowest quality) while 1 is for the highest achivable condition (i.e. highest quality, reference conditions). An additional statement to derive EQRs, is that all observed values must be related to a previously set reference value for each biological metric, resource, etc. in the form of an Observed/Expected ratio.
The two concepts, i.e. to relate to a 0-1 scale and to refer to reference values, are both central in the Intercalibration process. The combination of the two items leads to the setting of the class boundaries (at least of the High/Good status boundary).
A clear definition of the normalization option is crucial when directly comparing National methods as well as when comparing them via ICMs and ICMi.

A definition

EQR setting criteria

.
The calculation options used to define the range of variation of EQRs, i.e. how to set the highest (EQR=1) and lowest (EQR=0) benchmarking (upper and lower anchors), and to derive class boundaries.

3.3.2 Setting the Reference condition value to normalize data

The calculation of an anchor value for reference conditions, i.e. that value used to derive the Observed/Expected EQR value, should be performed after a strict protocol to accept/refuse a site as a 'reference site' was applied (see the ECOSTAT Intercalibration Guidance, EU 2004a). <u>If this so-called 'boundary setting protocol', which should entirely assure the WFD compliance, is applied correctly and thoroughly</u>, each test site will be indisputably assigned (or not assigned) to the pool of reference sites. Those sites only – and the correspondent biological metrics – will be used to confidently calculate the anchor value for reference conditions, for a given water body type, season, etc. Thus, if the pool of reference sites was established according to the (agreed) boundary setting protocol, the High/Good boudary being derived will offer a high degree of confidence (see Table 3.1). <u>The median value can then be used as the most robust measure for setting the reference condition, to be used in the EQRs calculation</u>.

Table 3.1

Protocol to accept/refuse reference sites - Boundary setting protocol	**Use of the protocol**	**Expected confidence of the existing High/Good boundary for the WFD aims**	**Number of reference sites available for the water body type**	**Option to calculate the anchor value for reference conditions**			
				Median	**Maximum**	**75th %ile**	**90th %ile**
available	applied	high	high	encouraged (best)	not suited	encouraged	possible
			low	encouraged (best)	possible	encouraged	not suited
	not applied	low	high	discouraged	discouraged	encouraged	possible
			low	discouraged	possible	encouraged	not suited
not available	not applied	low	high	discouraged	discouraged	encouraged	possible
			low	discouraged	possible	encouraged	not suited

If the boundary setting protocol is not available or not applied, the confidence that the nationally derived High/Good status boundary fits the WFD requirements and attitude will be low. The use of the median value as the reference anchor value might here bias the information towards poorer quality conditions. To estimate the entity of the bias will not be possible until the boundary setting protocol is applied. In these circumstances, the use of the maximum observed value can be considered, especially if the number of reference sites included in the dataset is very low. Nevertheless, the observed maximum can greatly vary according to the number of observatons and e.g. natural variability. When a larger number of sites will be included into the dataset, the observed maximum will highly presumably increase.To partly deal with this tendency, which is not acceptable from a statistical point of view, a fixed number of sites to be considered for the calculation of the maximum can be fixed. As this option is especially suitable for small datasets, the number of sites/samples to be considered can be e.g. 12. If a dataset contains more than 12 samples, the additional information available can be saved by using an electronic resampling technique (e.g. by bootstrapping) to extract 12 samples for each resampling. This will support a more robust estimation of the maximum value, calculated as the average of the resampled maximum values after extracting n times (e.g. 1000) 12 samples from the dataset.

The option of using the median or the maximum value, if the boundary setting protocol is not available or not applied, are both unsatisfactory, for different reasons. To partly cope with the limitations of such approaches, the use of the 75^{th} %ile of the High status sites defined according to the existing National boundaries can be proposed for this preliminary phase of the IC process. By using this percentile as the anchor value, the possible bias of being pushed down towards a poorer quality by the eventual presence of samples classified as High status but not being acceptable reference sites can be partly overcomed. In the meantime, a 'correction' to exclude possible outliers and to reduce the extreme values due to a potentially high natural variabilty will be provided.

From the statistical point of view, 95th %ile have been often used when defining boundaries for biological methods. In this same Deliverable two examples of its use are provided (see Chapter 6.1). In the first example, this %ile is calculated from pre-classified 'reference sites'. The main problem with this procedure is that a relatively high number of sites/samples are needed to estimate it properly (e.g. compared to the median or to the 75%ile), while the scarce availability of data from reference sites is a common problem all over Europe. The second example, here provided for macrophytes, uses this %ile as calculated on samples belonging to all quality classes together. This support the statistical evaluation, but strongly depend on the distribution of samples in the classes and on the effective presence of reference sites (i.e. again, on the major weakness of European datasets).

While all MSs will have to deal with agreed criteria for accepting reference sites in the future to properly apply the WFD, at present it seems unrealistic that all MSs can be provide all the supporting data needed to check the suitability of the adopted protocols. Thus, it will be problematical that – for all MSs, GIGs, IC stream types, IC network sites - the supporting data can be provided in due time for the IC process. This will result in an incomplete, jeopardized scenario, where some MSs will be able to provide the needed data and others will not. To aid comparison and in the respect of the WFD requirements, not too much confidence should then be placed on the class boundaries set by individual MSs. In turn, together with the scarce amount of data expected from 'true' reference sites, this supports the use of the 75th %ile as an anchor for reference condition, within the scopes of the IC process (at least in this pilot phase).

The same option (75th %ile) can be suitably used as well when the WFD compliant boundary setting protocol is followed. For this reason, if not differently specified, the data presented and discussed in this Deliverable have been normalized on the basis of the 75th %ile of the High status or reference sites/samples. This will support an easier comparison of results across Europe, GIGs and stream types, for both test and benchmark datasets (see Chapters 4 and 5).

IMPORTANT WARNING

If calculated on the basis of MSs biological protocols only, the simple agreement on the use of any statistical values (e.g. median, 75th %ile) as an anchor value for Reference conditions, is not acceptable for the formal IC process, because it would not guarantee conformity to the WFD.

Even if a MS has a WFD-compliant assessment system, the use of the National biological method of classification to set the upper anchor value would result in the benchmarking of High status sites, and NOT of WFD-compliant Reference sites. The latter must be derived by integrating biological data with physico-chemical and hydromorphological information (i.e. pressures data).

The use of the 75th %ile in the present Deliverable has been adopted because for most MSs WFD-compliant systems and criteria for setting reference conditions are presently unavailable. The need for comparison requires a common value to be set to normalize data, based on existing datasets.

It is here assumed that the biological metrics considered in the assessment systems of MSs for the purposes of ecological quality classification and the Intercalibration Common Metrics (ICMs) used here to aid the illustration of possible options for the IC process can get values higher than one. This means that, e.g. after equating the reference value to the 75th %ile or to the median, some of the observed values for any metrics might get a value highe than one. To keep this possibility – i.e. to avoid equating all the obtained values higher than one to one – will decrease the uncertainty of the resulting classification.

Quite obviously, each of the three discussed options to normalize data for further comparison do need the availability of at least a few sites/samples that have been classified as reference sites and fullfil the prerequisite of the agreed boundary setting protocol.

3.3.3 The scaling factor and boundary setting option

Accordingly with the WFD requirements, the final scaling of the EQRs must be on a 0-1 scale. This means that the final step in the EQRs calculation will be to rescale single metrics or multi-metric indices to make them fit into the 0-1 scale, indifferently to their potential attitude to show a higher or lower variability in e.g. the High status class. This final re-scaling will potentially lead to a disomogeneous positioning of the boundaries along the quality gradient in mathematical terms i.e. the High/Good boundary might result in apparently different values (e.g. 0.8 vs 0.72) according to two different assessment systems and MSs. This is not incompatible with the WFD and introduces no seriuos problems for its implementation. Nevertheless, it will reduce the possibility of directly comparing the assessment systems and classification results from the various countries, which was one of the main aims for which the EQR concept was introduced in the WFD. To set one of the WFD-relevant boundaries (i.e. High/Good and Good/Moderate) equal to ove would highly increase the direct comparability of classification results across Europe.
Different European countries are actually employing different options to scale the values used to describe the quality gradient for classification purposes. In some MSs, e.g. France, the median value of High status samples is equated to one and the 25th %ile of High status samples is set as the boundary between High and Good status. The other boundaries are calculated by dividing the remaining range into equal classes. In the U.K., the value of 1 is attibuted to the Good/Moderate status boundary, so that it becomes immediately obviuos if a site has to be restored/enhanced or not, for the aims of the WFD. In both countries, type or site specific calculation of reference conditions is actually provided or being defined. Elsewhere, e.g. in Italy, a fixed value – not yet converted into a normalized scale – is set for the all the class boundaries, independently from any stream type-specific reference condition assumption. The idea behind the boundaries setting is quite different from the WFD type-specific principle and assumes that a single index value is suitable to adequately discriminate between e.g. Moderate and Good or Good and High status sites for any stream type. This approach clearly reflects the knowedge available in the period when the assessment method (IBE) was developed and an evident upgrade is expected for the WFD implementation in Italy. The three examples are useful to depict how a standardization of the normalization option across Europe is needed to support an effective comparability of results.

4. Test datasets

Test datasets contain data from national monitoring networks, scientific national projects, exercises among Environmental Agencies etc. In some cases the National legislations and, consequently, the collected data often do not fulfil WFD requirements.

A definition

Test data
Data derived by standard monitoring activities according to MS legislation and tradition.

Notes: (a) The data going to be presumably the basis for the IC process. (b) They can correspond, totally or partially, to the data provided by MSs for the sites included in the formal network of IC sites. (c) For their use and testing, they must be attributed to a GIG stream type.

4.1 Needed characteristics for test data

The presented data were collected during the parallel activity jointly performed by STAR and GIG delegates to collate data useful for the pilot IC exercises. The information reported thus refers to data as well as dataset features.

In general terms, the characteristic for each test dataset are:
- taxalist to family level
- taxalist must include at least an estimation of abundance for each taxon
- sites have to be classified according to assessment method to test
- the boundaries between classes according to such assessment method must be known
- preferably the samplig area should be known
- high status samples must be present
- a wide quality gradient has to be present in the dataset
- criteria to classify high status sites must be indicated. E.g. sites classified according to the MS standard biological method only or considering also other elements (pressures, etc).

4.2 Features describing each test dataset and dataset presentation

- Institution that collected the data (e.g. EPA, EA) and property (Regional Authority, etc.)
- aim of the collection
- how many sites are considered
- how many samples/sites/seasons
- how wide is the quality gradient (e.g. form High to Moderate, from Good to Bad)
- river type
- ancillary data (pressure, chemicals, RHS derived indices, morphological classification, etc.)
- method of classification, including information on class boundaries, min and max values
- type of sampling method (qualitative, quantitative, semiquantitative)
- calculation formulae
- final classification (BAC, MS's) for the presented data

Test database presentation

The present paragraph contains the description of the considered test datasets. These are presented by groups of the same Intercalibration type (according to ECOSTAT WG 2.A, 2004).
The information provided are:

General features
Very brief overall description of the area and characteristics fitting with the IC type requirements, such as catchment area and altitude. Indication on sites distribution (i.e., if sites are spread in a large area or not. A useful datum is an estimation of the maximum distance between two sites).

Aim of collection, number of samples
The Institute who collected the data and/or make the data available, together with a contact person is reported. The aim of the collection, the number of sites and samples, the period of the collection of the data are also reported.

Degradation factor
Information on the main degradation causes and the quality gradient covered. Information on available support data such as chemical variables or other pressures.

National method: sampling and sorting
Description of sampling and sorting method used, usually (even if not always) corresponding to the national member state method. Include information about: sampling surface (real or estimated), sorting semi/quantitative/qualitative, identification level.

National method: criteria for abundance registration
Indication on criteria for abundance registration.

National method: sites' classification
Description of the technique of sites' classification (calculation formulae, two entries table etc.). Maximum and minimum values (possible and observed) are reported. The boundaries between classes are indicated. The boundaries represent the starting step for the following comparison and the harmonization (see chapters 6 and 7) .

Notes on classification
Number of 'high status' sites according to the national method and, if available, according to a Best Available Classification. Best Available Classification (BAC) is the biological classification obtained by applying a WFD compliant procedure and all the available, relevant information on a site. Depending on the main kind of pressure acting, it may result from integrating biological, physico-chemical and hydromorphological information. It is based on detailed community analysis (e.g. by multivariate analysis on one or more BQEs) and not on the standard National methods of classification.

Comparison between the ICMi and MS method's EQRs
For all datasets collected, all Intercalibration Common Metrics and test methods have been recalculated according to 75th percentile observed in 'high status' sites according to test method, in order to uniform the criteria and to make possible the comparison. Thus, the conversion formulae between test method and ICMi, as well as the regression coefficient may

differ from the original calculation provided by each institute. In few cases this normalization option has not been followed, see explanations in the single dataset.
The conversion of the class boundary values for the MS method from the original boundaries to ICMi values, and the linear regressions between the ICMi and the MS methods, and the single ICM and the MS method (with MS method on y axis) are reported through a table.

General remarks/comments
Comments, problems encountered during the treatment of the data.

References

ECOSTAT WG 2.A, 2004. Overview of common Intercalibration types. Version 4.0. February, 26th 2004.

4.3 IC type C1 (small lowland streams dominated by sandy substrates)

4.3.1 Belgium C1

General features
The sites enclosed in this dataset have an altitude lower than 200m and catchment area is comprised between 10 and 100km^2.; they belong to the Flemish river types 'small brooks' or 'small brooks from the Kempen region'. The sites are randomly distributed throughout Flanders. The total area of Flanders is approximately 13 500 km².

Aim of collection, number of samples
Data are provided by Mrs. Gaby Verhaegen of the Flemish Environment Agency (VMM). Data were collected within the monitoring network of the Flemish Environment Agency (Flemish region of Belgium).
The data set includes 208 samples. Collection was performed in three years (2000-2002).

Degradation factor
The sites are affected by general degradation. The quality gradient covers all the quality classes according to both, the currently used regional method BBI and the revised method: the Multimetric Index Flanders (MIF), from 'high' to 'bad' status. No support data are available.

Used regional method: sampling and sorting
Samples have been qualitatively sampled using a hand net. All accessible habitats have been explored for a limited period of time (3 min. effective sampling exceptions in time can be made when substrate exists out of stones). The total sampling area is approximately 20 metres (rough estimation). More than one specimen per taxon has to be present to be considered valid.
The identification is performed to genus/family level.

Used regional method: criteria for abundance registration
The number of specimens is recorded as abundance classes. Such classes have been converted in numbers in order to allow calculation of abundance metrics.

Used regionall method: sites' classification
The classification provided by Mrs. Gaby Verhaegen refers to a Belgian Multimetic Index recently developed (Multimetric Index Flanders - MIF). This is a revised version (Gabriels, 2004) of the index in current usage in Belgium, the Belgian Biotic Index (BBI, De Pauw & Van Hooren, 1983). For the calculation, AQEM rapid assessment program was used for the metrics ASPT, number of families, EPT and Shannon-Wiener.
The determination of the BBI is based on two metrics, combined basing on a two entries table: the faunistic group and the number of systematic units. Assessment is undertaken in 5 quality classes. Values of the index vary from 0 to 10 and boundaries between classes are: high-good: 9; good-moderate: 7; moderate-poor: 5; poor-bad: 3.
The MIF is proposed as a new type specific procedure for index development, in which expert knowledge is incorporated into the existing system. The result of this new procedure is a series of multimetric indices, all consisting of the same five metrics, which are transformed into one index value by means of a scoring system that differs according to the water type. These metrics are total number of taxa, total number of EPT taxa, total number of other sensitive taxa, Shannon-Wiener index and Mean Tolerance Score. The final index is a value

within the interval 0-1, which is equally divided into five quality classes (.high: >=0.8; good: >=0.6; moderate: >=0.4; poor: >=0.2; bad: < 0.2). Because the calculation method differs for each water type, the water type should always be indicated when index results are displayed (Gabriels, 2004).

Notes on classification
MIF: 10 samples on 208 are classified as 'high status' according to MIF assessment method.
BAC or pressures based: No other classification available (no Best Available Classification).

Comparison between the ICMi an MIF classification EQRs, single ICM and national classification EQRs
For the calculation of the ICMi, metrics was normalized according to 75° percentile of high status samples according to MIF method (see explanations in previous chapters). Final ICMi is re-normalized according to 75° percentile value. The minimum and maximum observed values for ICMi (in EQR) are 0 and 1,115. Also the values of the MIF are transformed in EQR through a normalization according to the high status samples' 75° percentile.
Figures below represent the linear regression between ICMi and the MIF
Regression coefficient found are: ICMi vs MIF: 0.74
Results on linear regression between single ICM and the BBI are shown in figure 1 B-G.
The conversion of the class boundary values for the MIF method from the original boundaries to ICMi values is done according to Table 1.

	MIF score	MIF EQR	ICMi EQR
Limit high-good	0,8	0,889	0,836
Limit good-moderate	0,6	0,667	0,621
Limit moderate-poor	0,4	0,444	0,405
Limit poor-bad	0,2	0,222	0,189
ICMi EQR = 0,9698 * MIF EQR - 0,0258			
R^2 =0,7405			

Table 1: MIF class boundaries conversion

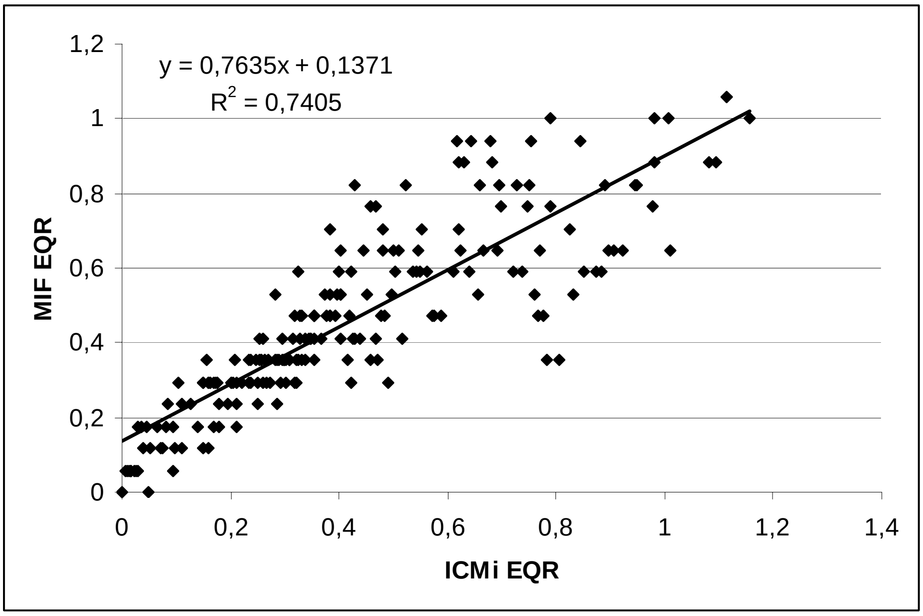


Figure 1A : ICMi vs MIF - R^2 = 0,74; p<0.001

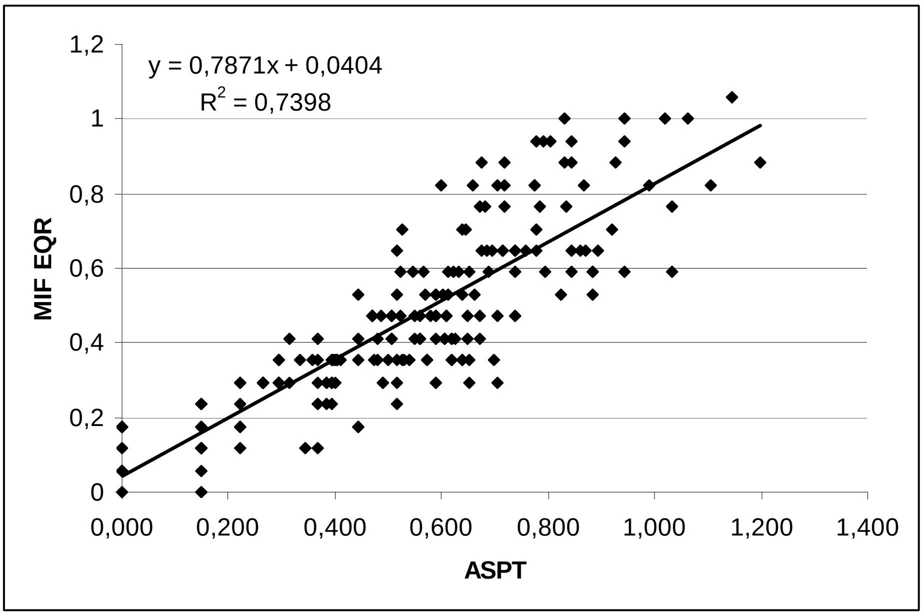


Figure 1B: ASPT - $R^2 = 0.74$; $p<0.001$

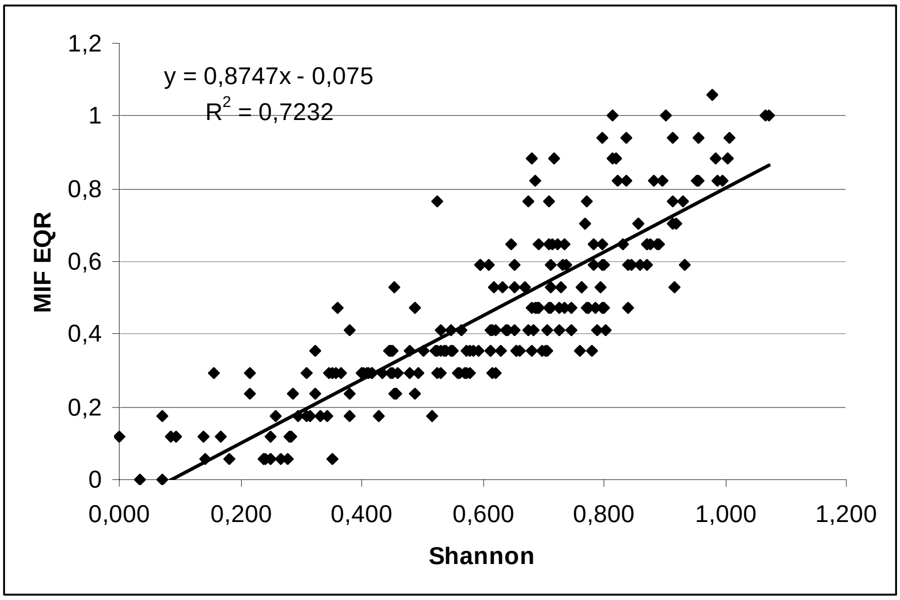


Figure 1C: Shannon - $R^2 = 0.72$; $p<0.001$

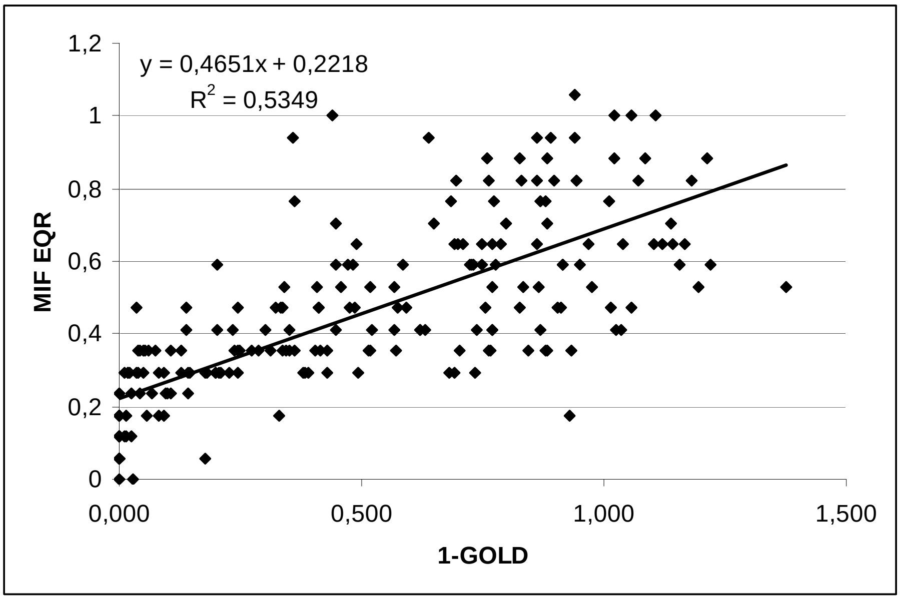


Figure 1D: 1-GOLD - $R^2 = 0.53$; $p<0.001$

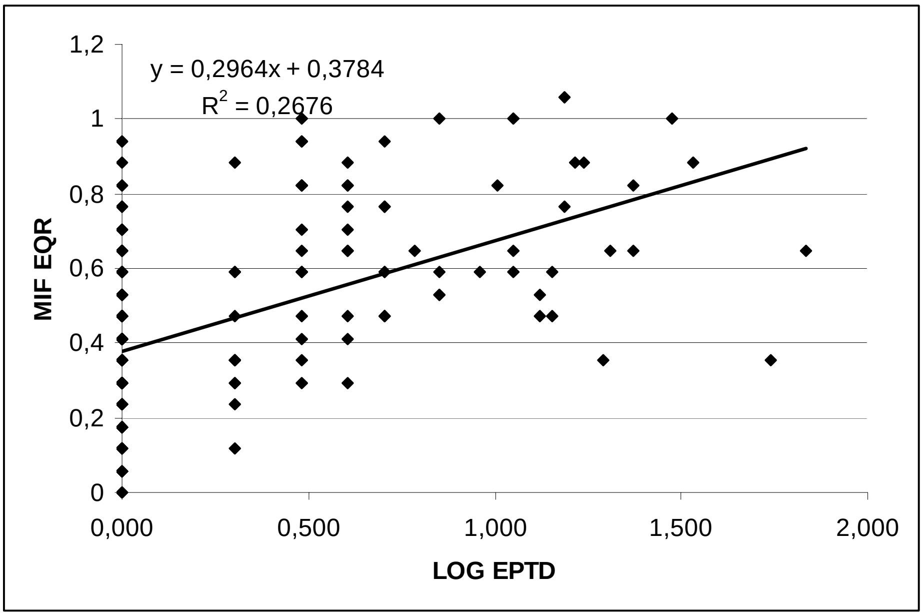


Figure 1E: Log EPTD - $R^2 = 0.27$; $p<0.001$

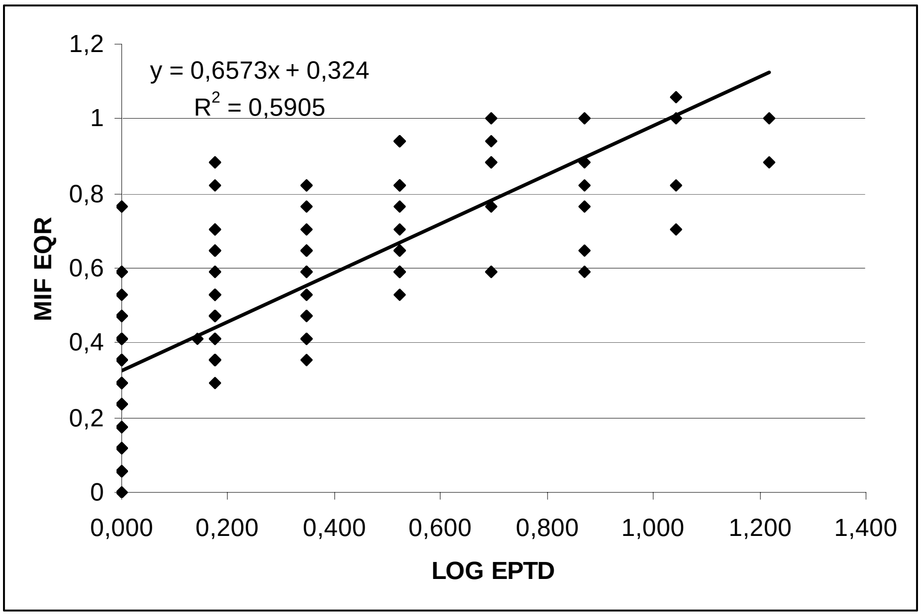


Figure 1F: EPT - R^2 = 0.59; p<0.001

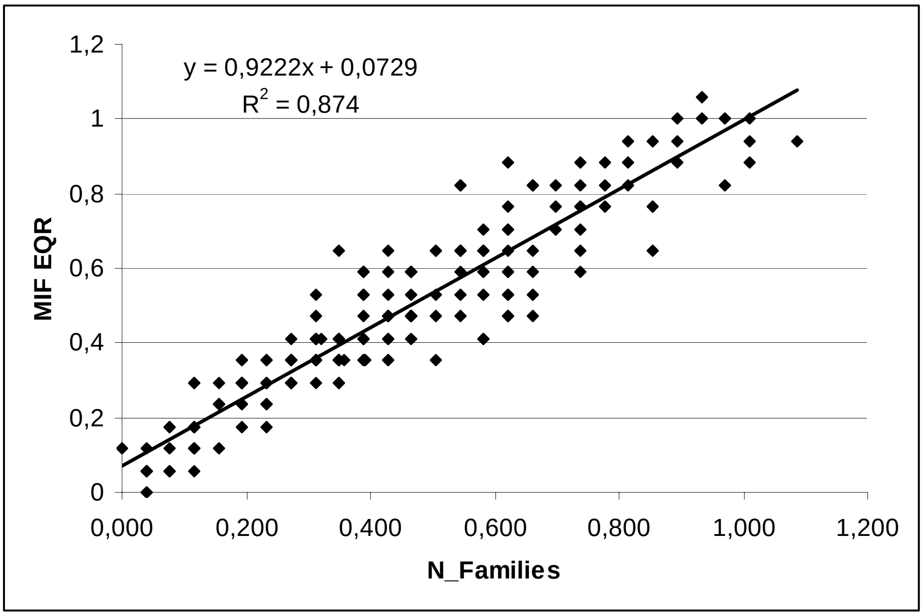


Figure 1G: N families - R^2 = 0.87; p<0.001

General remarks/ comments

The correlation between ICMi and the Multimetric Index Flanders shows better results (R^2 = 0.80) when using results considering the maximum values reached in the high status sites. Nevertheless, the data here presented, related to the MMIF method interested to the IC exercise, give acceptable results in terms of overall ICMi. Also, the single metrics show correlations higher than 0.50, except for the metric Log_EPTD (0.27), probably due to the fact that selected taxa could not represent the quality gradient.

References related to the present dataset

De Pauw, N, & G. Van Hooren, 1983. Method for biological quality assessment of watercourses in Belgium. Hydrobiologia 100: 153-168.

Gabriels, W., Goethals, P., Adriaenssens, V. & De Pauw, N. (2004). Application of different biological assessment systems on Flemish potential intercalibration locations according to the European Water Framework Directive, partim benthic invertebrate fauna. Final Report (in Dutch). Laboratory of Enivironmental Toxicology and Aquatic Ecology, Ghent University, Belgium. 59 p. + appendices.

4.3.2 Denmark C1

General features

Streams with moderate alcalinity can be found only in western parts of Jutland. Here, the landscape is flat and dominated sandy soils. The streams therefore have low slopes and are dominated by sand. Macrophytes are typically covering a major part of the stream bottom. Many C1 streams are regulated because of the use of the land for intensive agriculture (Skriver, 2004).

Aim of collection, number of samples

Data of this dataset were collected by regional Danish authorities (counties) and provided by Dr. Jens Skriver from NERI.

Data have been selected from the National Monitoring Programme (selected catchments and catchment areas between 15 and 100 km^2). Data from the STAR project have been supplemented. Because the number of sites are relatively low, data from all years have been used (typically 1998-2003) (Skriver, 2004). Total number of samples is 346.

Degradation factor

General degradation can be stated for these samples. The quality gradient covers all the quality classes according to the national method, from ‘high’ to ‘bad’ status. But the quality classes poor and bad are only found in a limited number of sites because these streams generally are only slightly polluted with organic matter. Other support data available include a physical description (substrate types, current velocity etc.). The main degradation factors being physical degradation (weed-cutting) and ochre pollution (because of drainage activities in the catchment). Information on water quality only exist from selected sites. There are no microbiological information available from the sites.

National method: sampling and sorting

Macroinvertebrate samples have been collected in spring by kick sampling using a handnet with a mesh size of 0,5 mm (Skriver et al., 2000). Total sampling area is about 1.25 m^2.

Guidelines on sampling, sorting and taxonomic identification have been produced by the Danish Environmental Protection Agency (DEPA, 1998). The Danish Stream Fauna Index (DSFI) is used to express the ecological quality.

The national sorting and identification instructions are minimal guidelines. Samples do not necessarily have to be sorted completely but all “selected” taxa have to be found if they are in the sample (“selected” taxa are defined in the guidelines). (Skriver, 2004). Identification only has to be performed to the genus or family level.

National method: criteria for abundance registration

An estimation of abundance is sufficient for the index calculation. These minimal guidelines are followed by most counties, but some counties have decided to produce macroinvertebrate lists based on complete sorting as well as species identification (Skriver, 2004).

National method: sites’ classification

The index is calculated using a two entries matrix with indicator groups and diversity groups.

The index have values from 1 to 7 were the maximum value expresses a minimal impacted macroinvertebrate community. Classification is performed in 5 quality classes. Index values vary by entire numbers, this can introduce problems during the harmonization.

Notes on classification

National: About 21% (72 on 346) of the samples are classified as ‘high status’ according to national assessment method.
“High status” sites are based on an expert judgement including information on the macroinvertebrate community (species composition), catchment use, water quality data if they are available, point sources, information on weed-cutting, regulation etc. The national classification method DSFI have not been used as a criterion. The DSFI value will typically be 7 for “high status” sites but in a number of cases DSFI 7 can be found in streams that are only believed to represent good status.

Comparison between the ICMi and DSFI EQRs, single ICM and DSFI EQRs
For the calculation of the ICMi, metrics was normalized according to 75° percentile observed in the ‘high status’ samples (see explanations in previous chapters). ICMi was renormalized according to the 75° percentile. The minimum and maximum observed values for ICMi (in EQR) have been 0.30 and 1.09. Between ICMi and DSFI, a regression coefficient of 0.48 was found (see figure 1A).
Results on linear regression between single ICM and DSFI are shown in figure 1 B-G.
The scores of IBE in the graphs are expressed in EQR values, calculated dividing the DSFI score for each sample by the 75° observed in the high status samples.
The conversion of the class boundary values for the DSFI method from the original boundaries to ICMi values is done according to Table 1.

	DSFI score	DSFI EQR	ICM index
Limit high-good	6	0.857	0.888
Limit good-moderate	5	0.714	0.790
Limit moderate-poor	4	0.571	0.692
Limit poor-bad	3	0.429	0.595
ICM index = DSFI EQR * 0.683 +0.3021			
R^2=0.52; p<0.001			

Table 1: DSFI class boundaries conversion

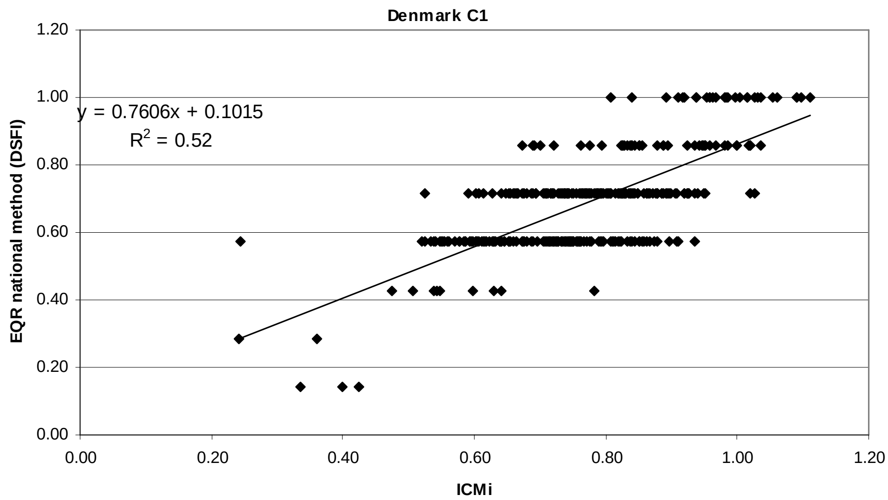


Figure 1: linear regression between ICMs and DSFI in dataset Denmark C1
Figure 1A: ICMi - R^2 = 0.52; p<0.001

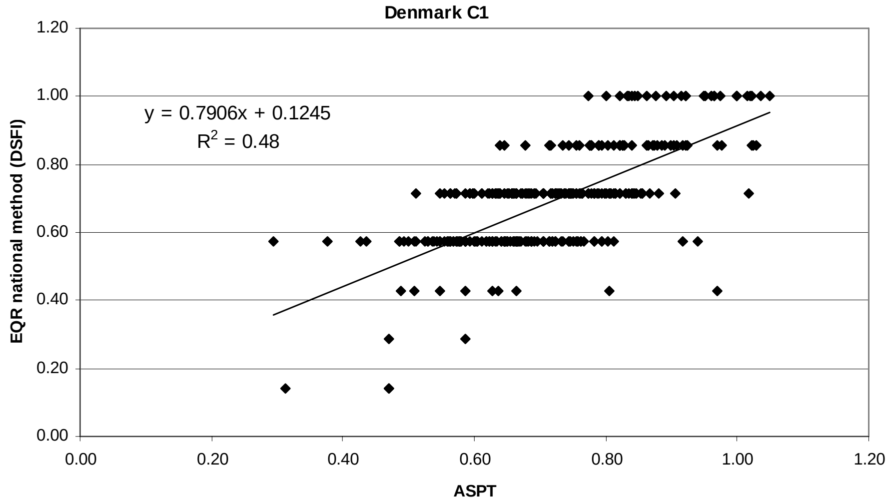


Figure 1B: ASPT index - $R^2 = 0.48$; $p<0.001$

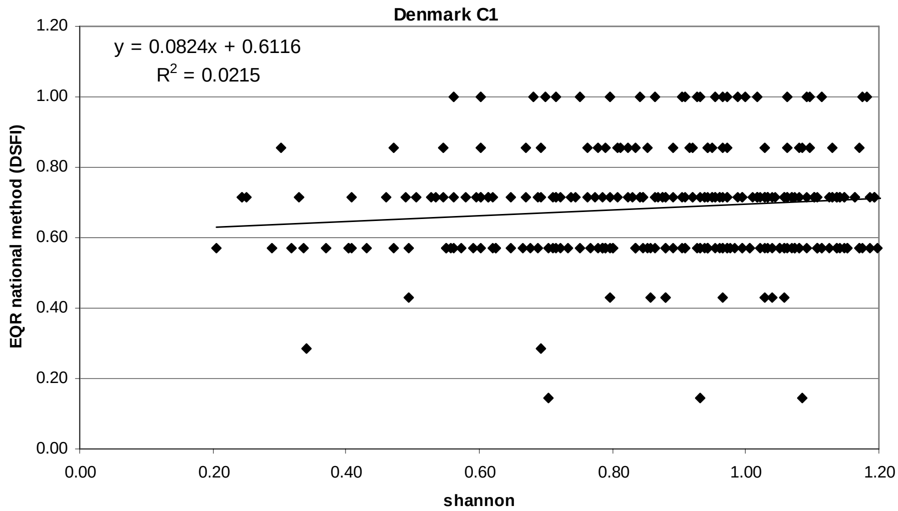


Figure 1C: Shannon index - $R^2 = 0.02$; $p<0.001$

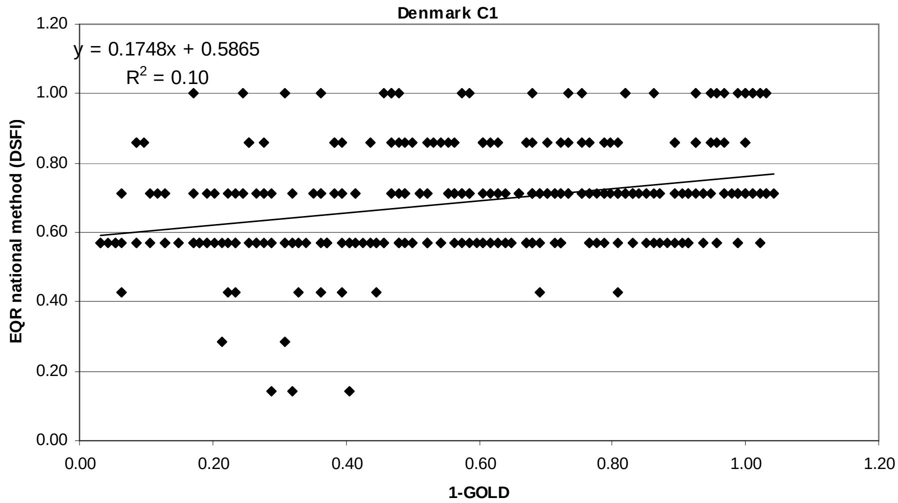


Figure 1D: 1-GOLD - $R^2 = 0.10$; $p<0.001$

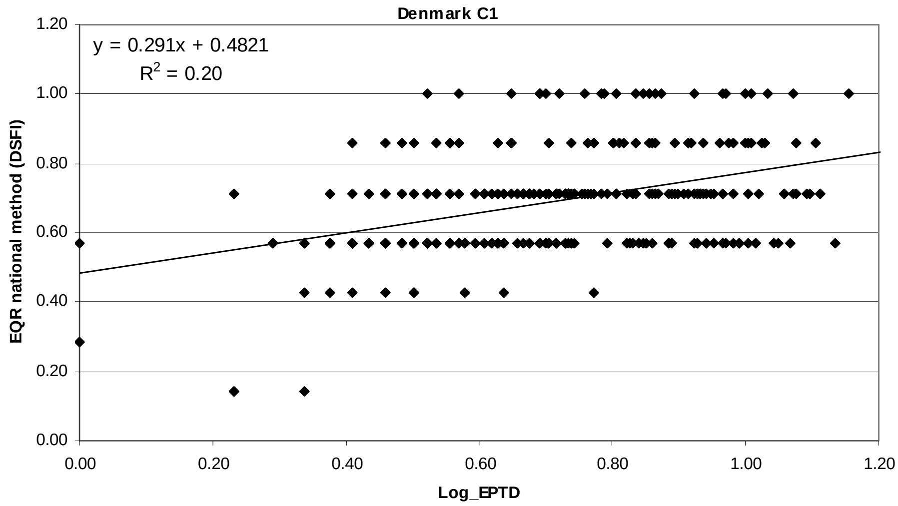


Figure 1E: Log EPTD - $R^2 = 0.20$; p<0.001

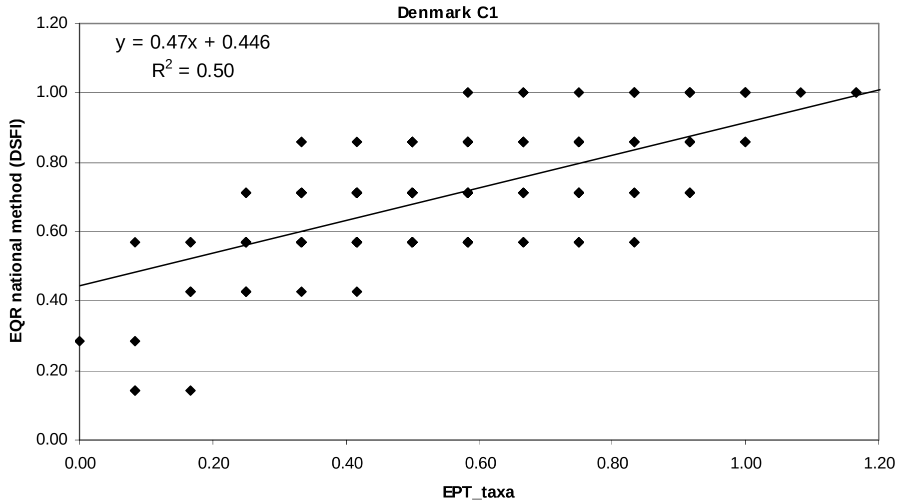


Figure 1F: EPT taxa - $R^2 = 0.50$; p<0.001

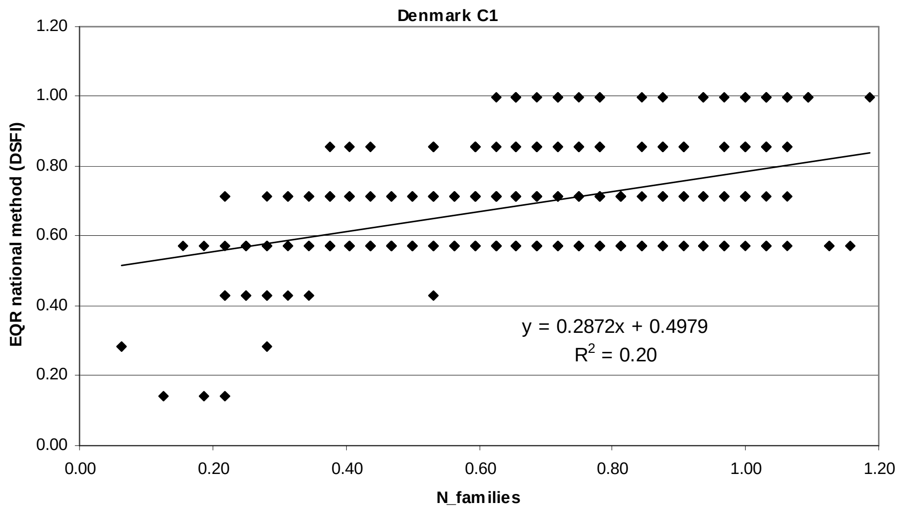


Figure 1G: Number of families - $R^2 = 0.20$; $p<0.001$

General remarks, comments

Skriver (2004) comments: the regression may be influenced by different sorting and identification procedures as described earlier. But also some of the individual metrics values change substantially between years and between sites without any indication of change in ecological state (in high sites as well as in impacted sites). Looking at the single metrics, very low correlations can be observed for Shannon and 1-GOLD. About this result, Skriver (2004) affirms that the Shannon Wiener diversity index may have very low values at some sites that are believed to be only minor impacted (also judged from R-C4 and R-C6 sites). This is also the case for the 1-GOLD metric. And some of the selected families in the Log_{10} (Sel_EPTD+1) metric looks problematic for Northern Europe (Limnephilidae and Nemouridae should be excluded and some other Plecoptera families could be introduced. Some of the Diptera families in this metric only occurs very rarely and in very low numbers in Danish samples).

Notes on dataset description

The content of the present description was verified by Dr. Jens Skriver of NERI who provided the data.

References related to the presented dataset

DEPA, 1998. Danish Environmental Protection Agency. Biological assessment of watercourse quality. Guidelines no. 5. – Ministry of Environment and Energy, Copenhagen. 39pp (in Danish).

Skriver, J., 2004. European intercalibration: Stream type R-C1 in Denmark. Pilot exercise report. 4pp. November 2004.

Skriver, J., N. Friberg & J. Kierkegaard, 2000. Biological assessment of running waters in Denmark: introduction of the Danish Stream Fauna Index (DSFI). Verh. Internat. Verein. Limnol. 27: 1822-1830.

4.3.3 Estonia C1

General features
The most water bodies of Estonia are situated lower than 200 m above sea level. The baserock consists of limestones (northern part), or sandstones (southern part). For the current intercalibration, samples from stony and/or gravelly bottom (sometimes also with sandy areas) were chosen, with velocity > 0.2 m/s. The catchment area for the sites was characterised by the distance to the stream source (4 -72 km), or by Strahler order (2 - 4). The upper limit of the catchment area did presumably not exceed 1000 sq. km, although for the smallest streams it was fairly less than 100. The sites are typical for Estonian lowlands and moraine hills (Timm, 2004)

Aim of collection, number of samples
Data of this dataset were collected and provided by Dr. Henn Timm from Estonian Agricultural University, Institute of Zoology and Botany. The sites are included in the national Estonian database. In most cases, sampling time was April-May (later than the common highwater period, but just before the most intensive emergence of insects). 23 samples are included, only one sample was chosen from each stream (Timm, 2004).

Degradation factor
In general, agricultural or urban pollution (sometimes accompanied by channelisation) was the main impairment type at the stressed samples. The direct influence of impoundments was avoided. Hydrochemical data are available for few samples only, and are almost missing for sites with catchment area <100 km^2 (Timm, 2004). In this dataset the quality classes according to the tested method, range mainly from 'high' to 'moderate' status (three classes equally represented, with about 10% in 'poor' status).

National method: sampling and sorting
Sampling was conducted according to Swedish examples (Johnson, 1999; Medin *et al*., 2001). A single sample consisted of five 1 m-long kicks from the most typical hard bottom of the site, and of one qualitative, unstandardized collection from all habitats available. The handnet´s edge was 25 cm long, and mesh size 0.5 mm. Dimension of each replicate is 0,25 m^2 All five replications, as well as the qualitative sample were fixed in separate jars in the field, and analysed separately later (Timm, 2004). Identification were undertaken at species level where possible, except some particular groups (chironomids, oligochaetes, sphaeriids, water mites).

National method: criteria for abundance registration
The absolute abundance (or its relatives, such as diversity indices) cannot be given for all taxa, because some taxa may originate from the qualitative search only. Therefore, when a taxon occurred in qualitative sample, its "abundance" was always considered 1 and that was added to the "correct" abundance from semi-quantitative samples (Timm, 2004).

National method: sites' classification
Quality classes for sites were established, using British ASPT (Armitage *et al*., 1983). Such classification has to be consider preliminary, anyway the ASPT index is currently used in regular biological monitoring of Estonian streams. Used boundaries between classes are: HG, 6.1; GM, 5.1; MP, 4.1; PB, 3,1. In this database, the minimum and maximum observed values are 3.43 and 7.10.

Notes on classification
National: About 39% (9 on 23) of the samples are classified as 'high status' according to national assessment method.
BAC or pressures based: No other classification available (no Best Available Classification).

Comparison between the ICMi and ASPT EQRs, single ICM and ASPT EQRs
For the calculation of the ICMi, metrics was normalized according to 75° percentile observed in the 'high status' samples (see explanations in previous chapters). The minimum and maximum observed values for ICMi (in EQR) have been 0.16 and 1.16. Between ICMi and ASPT, a regression coefficient of 0.76 was found (see figure 1A).
Results on linear regression between single ICM and ASPT are shown in figure 1 B-G.
The scores of ASPT in the graphs are expressed in EQR values, calculated dividing the ASPT score for each sample by the 75° observed in the high status samples.
The conversion of the class boundary values for the ASPT method from the original boundaries to ICMi values is done according to Table 1.

	estASPT score	estASPT EQR	ICM index
Limit high-good	6.1	0.927	0.892
Limit good-moderate	5.1	0.775	0.678
Limit moderate-poor	4.1	0.623	0.464
Limit poor-bad	3.1	0.471	0.249
ICM index = estASPT EQR * 1.4102 -0.4151 R^2=0.76; p<0.001			

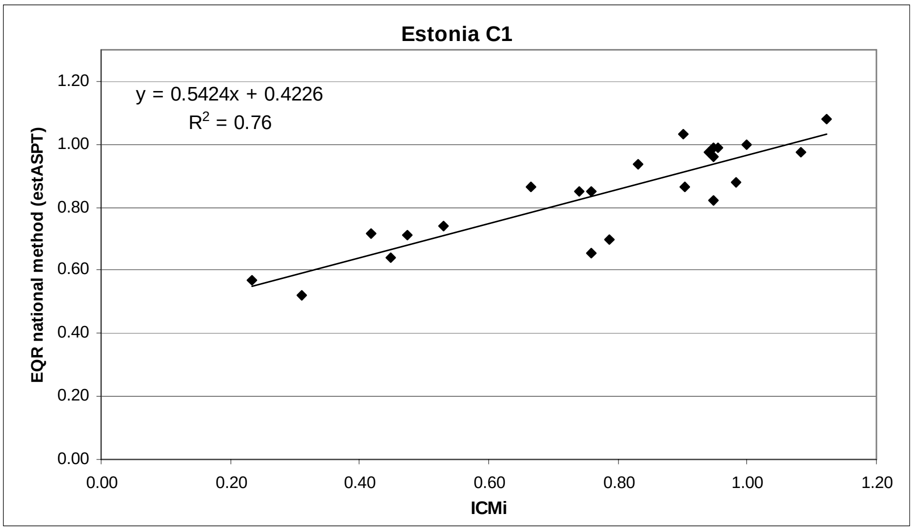


Figure 1: linear regression between ICMs and ASPT in dataset Estonia C1
Figure 1A: ICMi - R^2 = 0.76; p<0.001

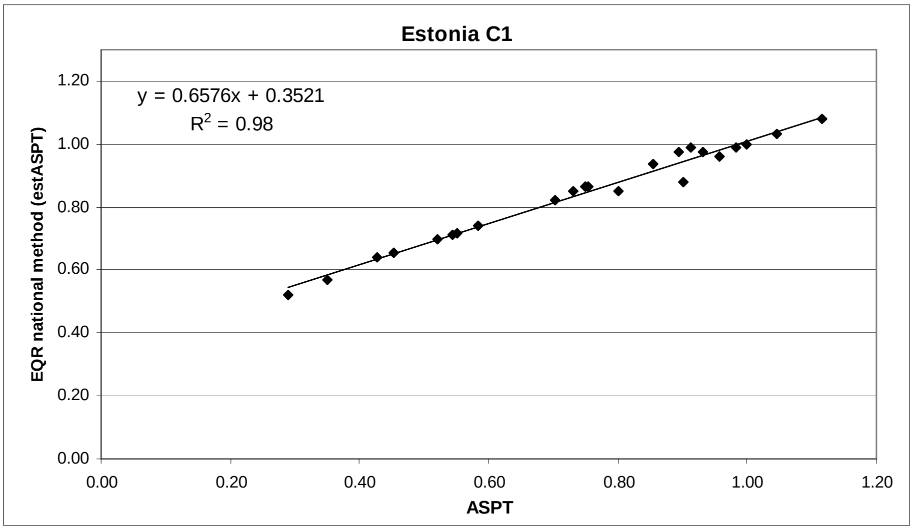


Figure 1B: ASPT index - $R^2 = 0.98$; $p<0.001$

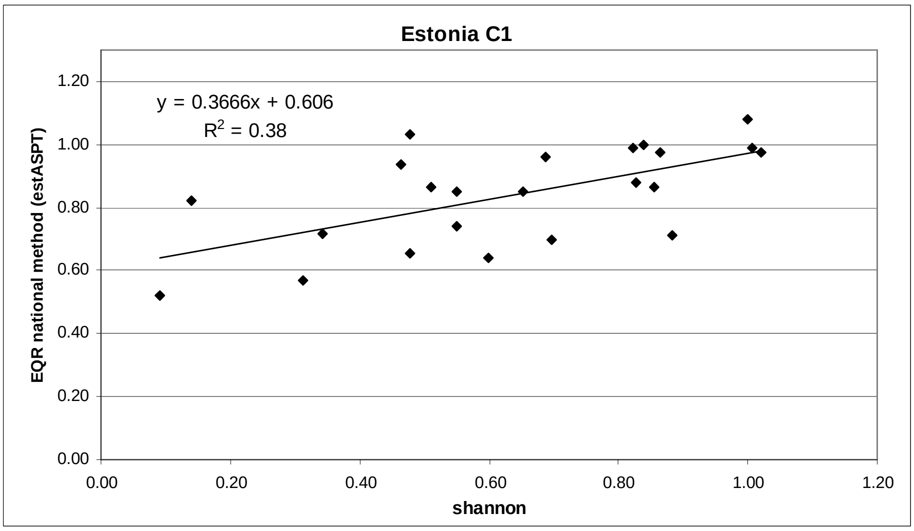


Figure 1C: Shannon index - $R^2 = 0.38$; $p<0.001$

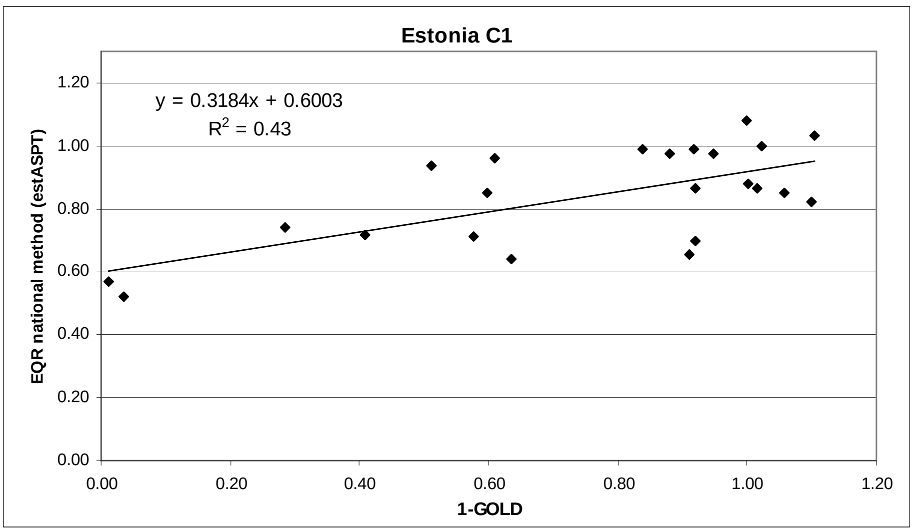


Figure 1D: 1-GOLD - $R^2 = 0.43$; $p<0.001$

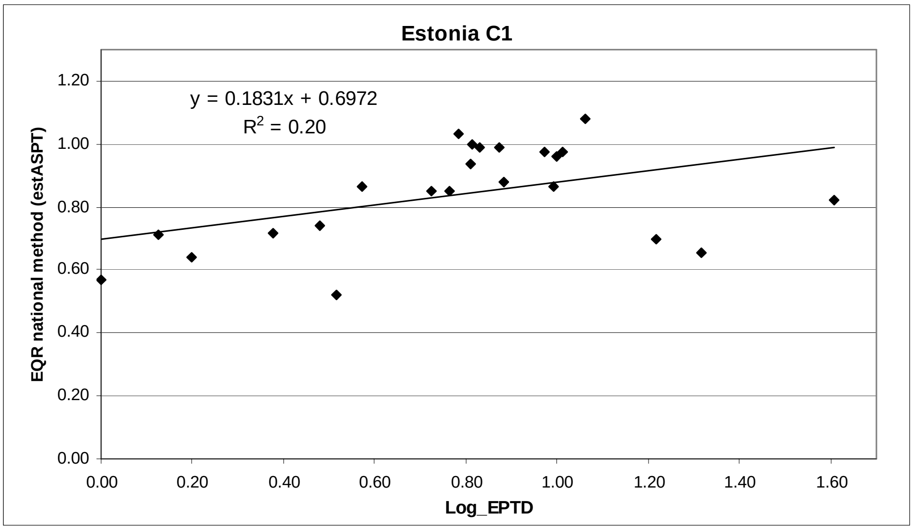


Figure 1E: Log EPTD - $R^2 = 0.20$; p<0.001

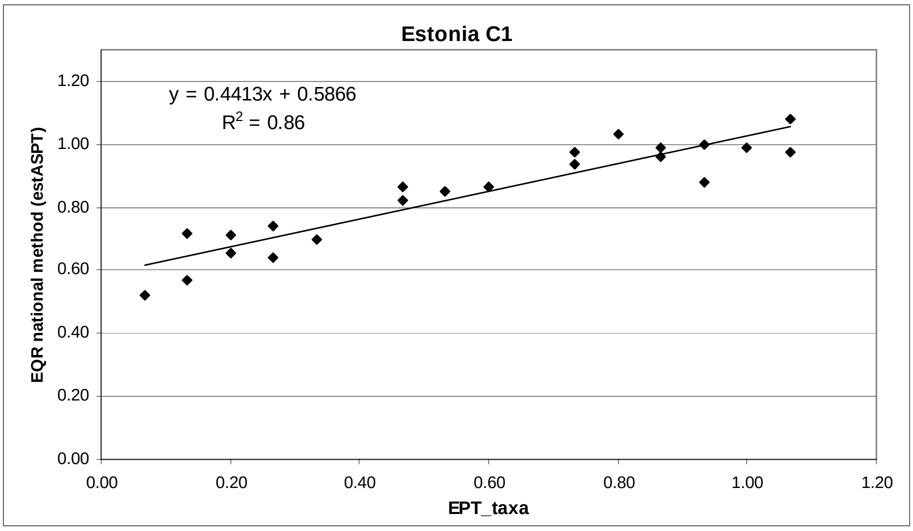


Figure 1F: EPT taxa - $R^2 = 0.86$; p<0.001

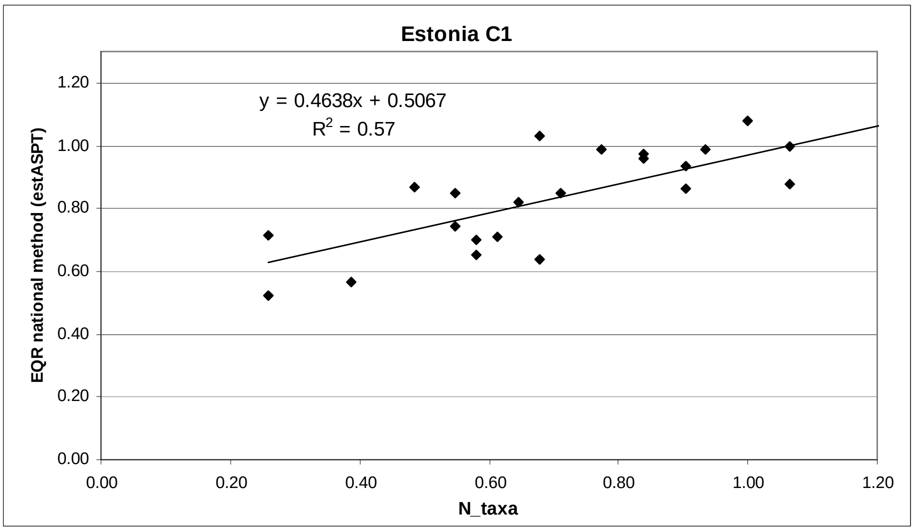


Figure 1G: Number of families - $R^2 = 0.57; p<0.001$

General remarks

-

Notes on dataset description

The content of the present description was verified by Dr. Henn Timm from Estonian Agricultural University, Institute of Zoology and Botany. who provided the data.

References related to the presented dataset

Johnson, R.K., 1999. Benthic macroinvertebrates. In: Bedömningsgrunder för miljökvalitet. Sjöar och vattendrag. Bakgrundsrapport 2. Biologiska parametrar (Ed. by Torgny Wiederholm). Naturvårdsverket Förlag 85-166.

Medin, M., U. Ericsson, C. Nilsson, I. Sundberg & P. A. Nilsson, 2001. Bedömningsgrunder för bottenfaunaundersökningar. Medins Sjö- och Åbiologi AB. Mölnlycke, 12 pp.

Timm, H., 2004. Comment to IC pilot exercise Estonian data. 3pp. November 2004.

4.3.4 France C1

General features
Sites belong to the hydro-ecoregion “Landes” (HER 13) of the French typology. Altitude is for all the sites enclosed in this dataset lower than 100m and catchment area is comprised between 10 and 300km^2. Correspond to the small streams. Geology is high siliceous with a lot of sand. Climatic conditions are oceanic.

Aim of collection, number of samples
Data collection was performed by the Direction Régionale de l’Environment. The database is organized by Lyon Cemagref and has been provided by Dr. Jean Gabriel Wasson. The sites are included in the national monitoring network and regularly investigated for quality assessment.
The total number of sites included is 20. In this dataset, the samples collected from 1992 to 2002 are included. Data collection was performed in several seasons per year (number of seasons not specified). Total number of samples is 139.

Degradation factor
General degradation is the main factor of alteration. The quality gradient covers all the quality classes according to the national method, from ‘high’ to ‘bad’ status. The support data are available from the National monitoring network. The type of data available is not specified.

National method: sampling and sorting
The method of classification is the official French monitoring method IBGN (Indice Biologique Global Noramalisé, Vernaux *et al.*, 1982). Sampling is carried out taking a number of 8 samples with a Surber sampler (base surface 1/20 m^2). These samples are characterized by different fixed couple of substrate dimensions and flow velocity. The total sampling area is 0.4 m^2.
Identification is performed to family level.

National method: criteria for abundance registration
IBGN method is semiquantitative. To be considered as valid, a single taxon has to be present with a minimum number of 3 specimens (or 10 specimens for few taxa). Nevertheless, in the present dataset the number of specimens is recorded as real abundance.

National method: sites’ classification
For the final classification, two metrics are considered: the Faunistic Indicator Group (GFI) whose values range from 1 to 9 and the number of collected families (taxonomic variety, VT) divided into 14 classes. The final IBGN value is obtained by the sum of these two metrics. Values of the index can vary from 0 to 20; boundaries between quality classes can have different values according to the stream type. For the C1 boundaries are: reference-high, 16; high-good, 14; good-moderate, 12. The boundaries moderate-poor and poor-bad are not defined for the C1 stream type. The transformation in EQR is done according to type.
In this database, the minimum and maximum observed values for IBGN are 1 and 18.

Notes on classification
National: 50 samples on 139 (about 36%) are classified as ‘high status’ according to national assessment method.
BAC or pressures based: 24 samples classified as ‘reference’.
Reference sites are selected on the basis of very low anthropic pressures, independently of the biological values in a first approach. The distribution of biological data is then calculated for

all samples of the reference dataset, and the outliers samples are checked. Dubious sites are eliminated, but low biological values are accepted if they come from validated reference sites. The procedure combine both spatial and temporal variability of a given stream type. The Reference Conditions (RC) are defined as the range of variability of a given biological element (index or metric) observed at reference sites. However, the calculation of EQR needs to define a Reference Value (RV) for the normalization of the samples. Due to the small number of reference sites generally observed for most types, the most robust and stable statistic is chosen as RV. For all our calculations, following the recommendation of the REFCOND guidance, the ***median*** was used as Reference Value.
The general approach and Reference Values for each type are described in a work paper (Wasson et al., October 2003, in French) and a summary (in English) will be available soon.
Reference sites were first selected from the monitoring network and other complementary sites in using two independent methods :

- "expert selection" by the field hydrobiologists (DIREN teams), on the basis of a detailed questionnaire combining all the possible pressures at the basin, reach and site scale.
- "GIS selection" run by Cemagref on the basis of known point source pollution discharges (from water agencies), and land use (CORINE), at the scale of hydrologic units (sub-basins ca. 100 km^2). However, this selection eliminates impacted basins where reference sites could be found upstream of pollution discharge (Wasson et al., August 2004, in French)

The IBGN values observed in these two selections of sites were compared to the values calculated from reference sites selected and sampled by the ***Cemagref*** hydrobiologists. The reference value for a given stream type was accepted only if the IBGN values observed in the three datasets were in good concordance. If not, a checking procedure was run and dubious sites were eliminated.
Since December 2004, the boundaries of the IBGN classes are redefined according to this definition of reference samples. In particular the boundary High good is set at the 25° percentiles of the reference samples.

Comparison between the ICMi and IBGN EQRs, single ICM and IBGN EQRs
For the calculation of the ICMi, metrics was normalized according to the 75° observed in the High status and Reference status samples (see explanations in previous chapters). This normalization option is suitable only for this IC exercise purpose. It will not be used in France for WFD implementation.
Final ICMi is re-normalized according to its 75° percentiles. The minimum and maximum observed values for ICMi (in EQR) have been 0 and 1.18. Between ICMi and IBGN, a regression coefficient of 0.83 was found (see figure 1A).
Results on linear regression between single ICM and IBGN are shown in figure 1 B-G.
The scores of IBGN in the graphs are expressed in EQR values, calculated dividing the IBGN score for each sample by the 75° observed in the high status samples.
The conversion of the class boundary values for the IBGN method from the original boundaries to ICMi values is done according to Table 1.

	IBGN score	IBGN EQR	ICM index
Limit ref-high	16	0.941176471	0.935
Limit high-good	14	0.823529412	0.822
Limit good-moderate	12	0.705882353	0.709
Limit moderate-poor	nd	nd	nd
Limit poor-bad	nd	nd	nd
ICM index = IBGN EQR * 0.9574 + 0.0336			
R^2=0.83; p<0.001			

Table 1: IBGN class boundaries conversion for C1 dataset

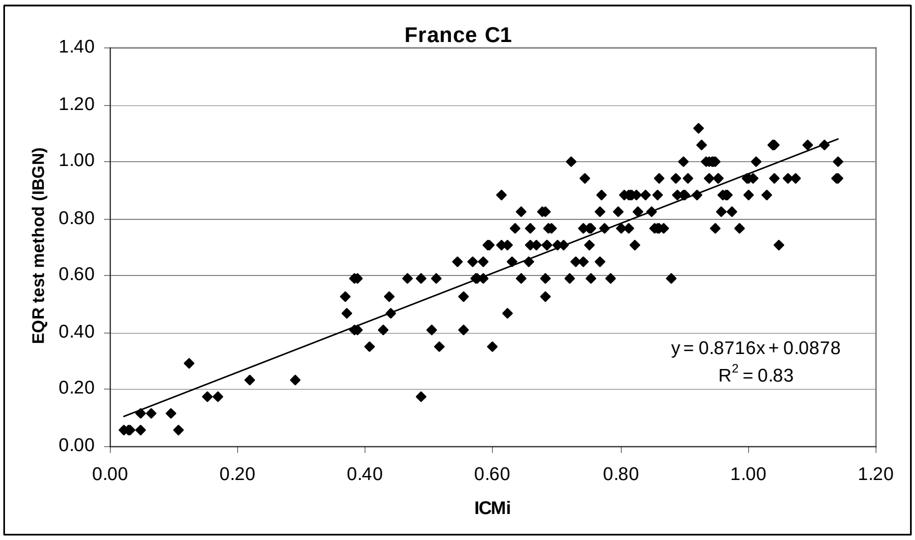


Figure 1: linear regression between ICMs and IBGN in dataset France C1
Figure1A:ICMi - R^2 = 0.83; p<0.001

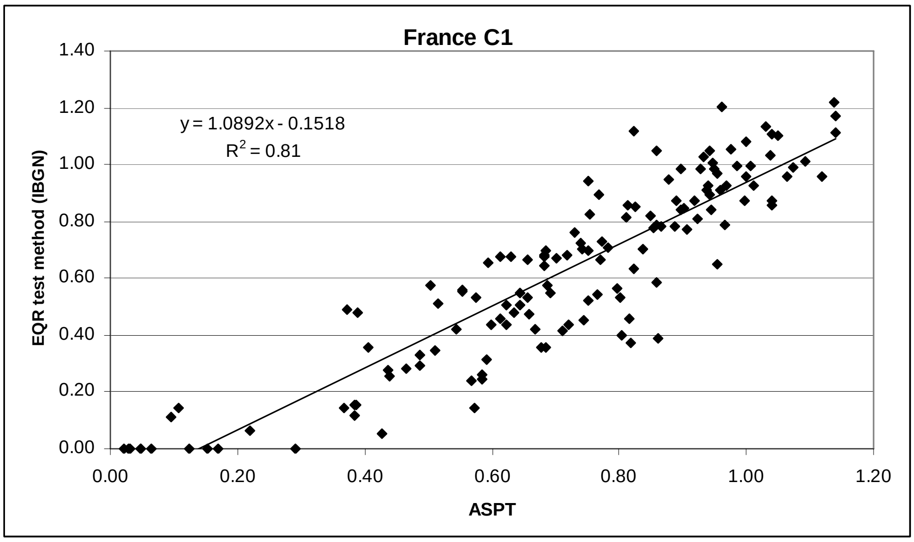


Figure 1B: ASPT index - R^2 = 0.81; p<0.001

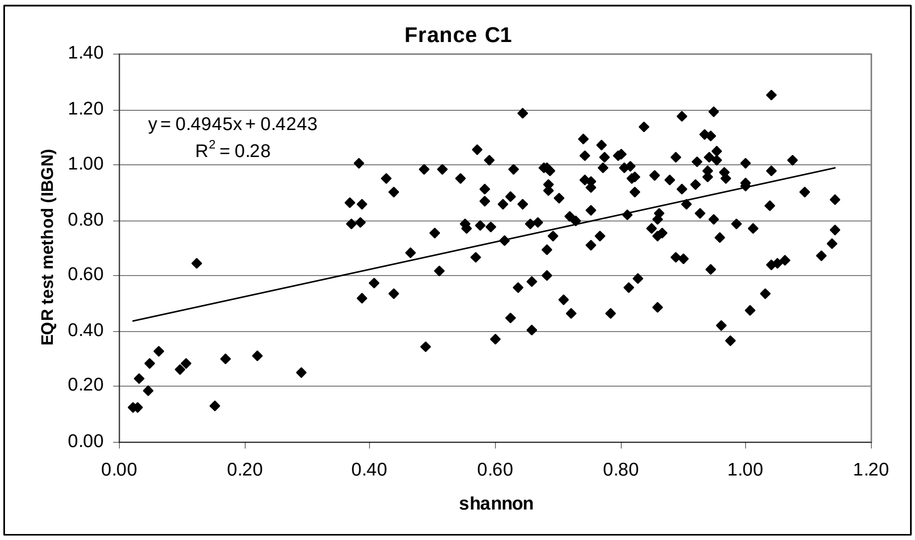


Figure 1C: Shannon index - $R^2 = 0.28$; $p<0.001$

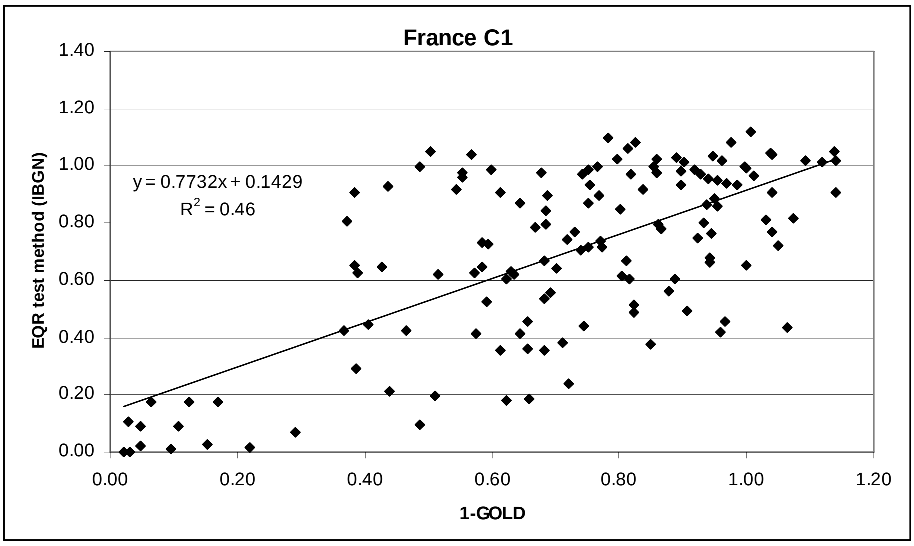


Figure 1D: 1-GOLD - $R^2 = 0.46$; $p<0.001$

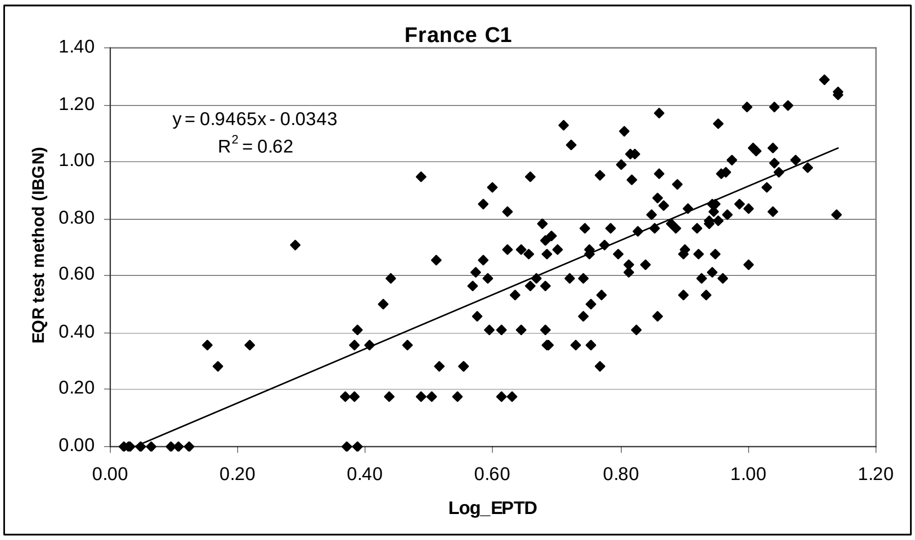


Figure 1E: Log EPTD - $R^2 = 0.62$; $p<0.001$

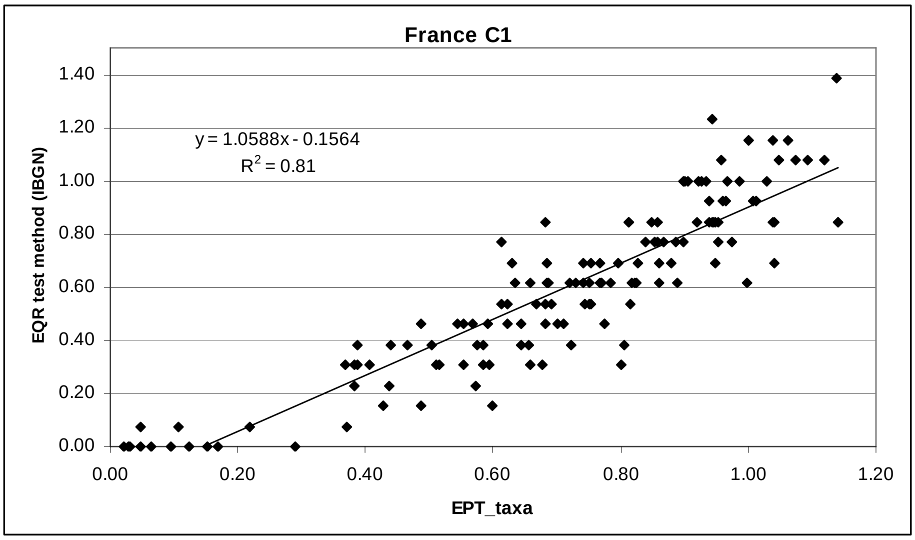


Figure 1F: EPT taxa - $R^2 = 0.71$; $p<0.001$

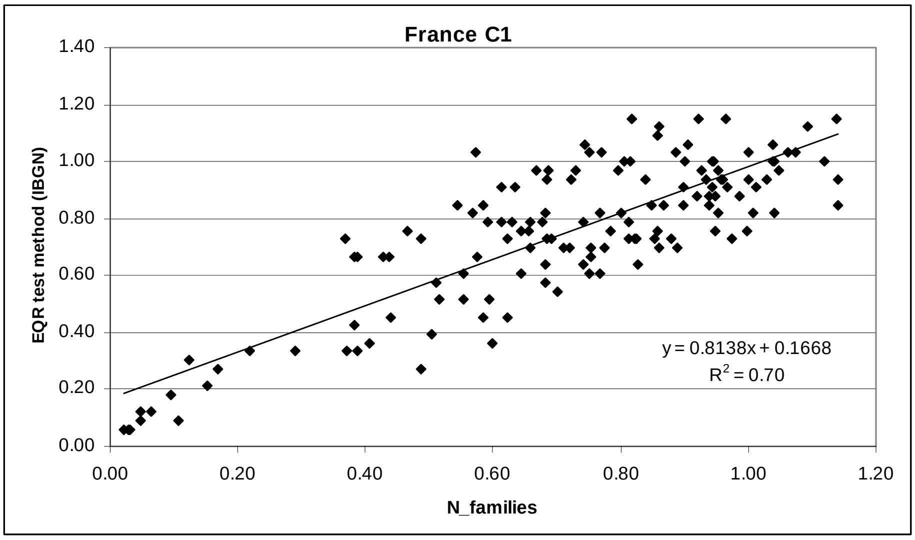


Figure 1G: Number of families - $R^2 = 0.70$; $p<0.001$

Notes on dataset description
The content of the present description was verified by Dr. Jean Gabriel Wasson from Lyon CEMAGREF who provided the data.

References related to the presented dataset

Vernaux, J. P., P. Galmiche, F. Janier & A. Monnot, 1982. Une nouvelle methode pratique d’evaluation de la qualité des eaux courantes: un indice biologique de qualité générale (I.B.G.). Annales Scientifiques de l’Université de Franche-Comté Besançon, Biologie animale 3: 11–21.

4.3.5 Germany C1

General features
The sites enclosed in this dataset have an altitude lower than 200m and catchment area ranges between 10 and 100km^2. Sampling sites are located in the German Lowlands, covering the federal states North Rhine-Westphalia, Lower Saxony, Schleswig-Holstein, Mecklenburg-Western Pomerania and Brandenburg. The maximum distance between two sites is about 450 km.

Aim of collection, number of samples
Data of this dataset were collected by various regional German authorities and are owned by Umweltbundesamt and LAWA. The dataset has been provided by Sebastian Birk from University of Duisburg-Essen. The sites are included in the federal monitoring networks.
In this dataset, 38 sites are included and data refer to several years of collection. Data collection was usually performed in 3 seasons per year (spring, summer and autumn). Total number of samples is 91.

Degradation factor
A 'general degradation' can be observed. In this dataset, according to national method samples are classified from 'good' to 'bad' status, only one sample is classified as 'high' status. No additional data are available.
This range of quality classes results from the overall ecological classification. It reflects the problem of the German lowland stream sites, none of which are in reference condition. The problem arising by this is the definition of a 75th percentile of high status sites – for German R-C1 only one site has high status. And even if you only regard SI(DE), only two sites are of high status.

National method: sampling and sorting
Sampling has been carried out at sites representative for the reach to be assessed, i.e. the sample has to represent the characteristic benthos community of the reach (DIN 38410, 2003). Each habitat exceeding 5 percent coverage is sampled according to its proportion.
Sorting method is semiquantitative.

National method: criteria for abundance registration
As sorting is semiquantitative, absolute abundances are not recorded. In the present dataset number of individuals are estimated using the mean value of each class as absolute abundance.

National method: sites' classification
The German 'ecological classification of benthic fauna in rivers' comprises two assessment modules to evaluate 'general degradation' (multimetric index, named GD (DE)) and 'organic pollution' (saprobic index, named SI (DE)).
The multimetric index for R-C1 includes the metrics 'abundance of EPT species', 'German Fauna Index Type 14', 'Shannon-Wiener diversity', 'number of Plecoptera species', 'percentage of rheophilous species' and 'percentage of shredders'. Single metrics are normalised against reference values and combined by averaging.
The saprobic index is the weighted averaging of the saprobic value and abundance of the present taxa. Identification is undertaken to species level.
Overall ecological quality is derived by the worst class of either module.

Notes on classification

National: 1 out of 91 samples is classified as 'high status' according to national assessment method.
BAC or pressures based: No other classification available (no Best Available Classification).

Comparison between the ICMi and GD (DE) and SI (DE) EQRs, single ICM and GD (DE) and SI (DE) EQRs
The normalization for the ICMs was not undertaken considering the 75° percentile of High status, as in the others dataset. For Germany C1 dataset reference values for the normalisation of ICMs have been obtained by correlation and regression of the German assessment module "General Degradation" against each ICM. ICM values corresponding to a German index value of 1.0 have been taken as reference values (Birk, 2004). The normalization of the index SI(DE) has been modelled on the basis of regression analysis against GD_abs (1.0 = reference). GD(DE) index was considered as absolute value (not normalized).

The minimum and maximum observed values for ICMi normalized are 0.16 and 0.98. Between ICMi and SI(DE), a regression coefficient of 0.32 was found (see figure 1A).
Results on linear regression between single ICM and SI(DE)mod are shown in figure 1 B-G.
The regression coefficient between ICMi and GD(DE) is 0.38 (see figure 2a). For regression of the single ICM and the GD(DE) see figure 2 B-G
The conversion of the class boundary values for the SI(DE) method and GD(DE) from the original boundaries to ICMi values is done according to Table 1 and Table 2.

	SI(DE) score	SI(DE) EQR	ICM index
Limit high-good	1.7	0.848	0.846
Limit good-moderate	2.2	0.664	0.577
Limit moderate-poor	2.8	0.443	0.255
Limit poor-bad	3.4	0.221	-0.067
ICM index = SI(DE) EQR * 1.456 -0.3895 R^2=0.32; p<0.001			

Table 1: SI(DE) class boundaries conversion

	GD(DE) score	GD(DE) EQR	ICM index
Limit high-good	0.8	-	0.884
Limit good-moderate	0.6	-	0.766
Limit moderate-poor	0.4	-	0.648
Limit poor-bad	0.2	-	0.531
ICM index = GD(DE) EQR * 0.5894 +0.4127 R^2=0.38; p<0.001			

Table : GD(DE) class boundaries conversion

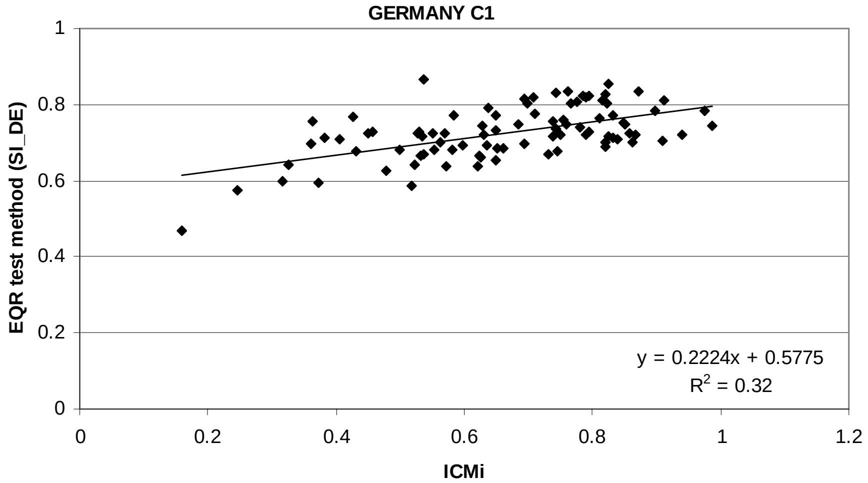


Figure 1: linear regression between ICMs and SI(DE) in dataset Germany C1
Figure 1A: ICMi - R^2 = 0.32; p<0.001

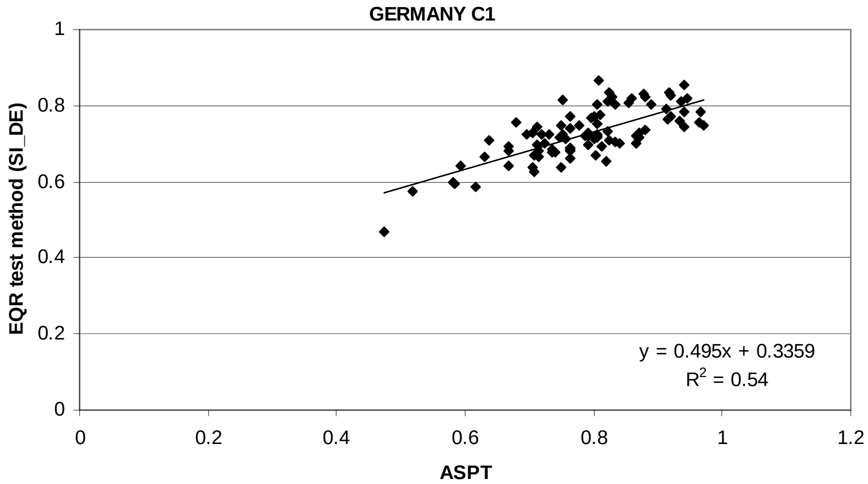


Figure 1B: ASPT index/SI(DE) - R^2 = 0.54; p<0.001

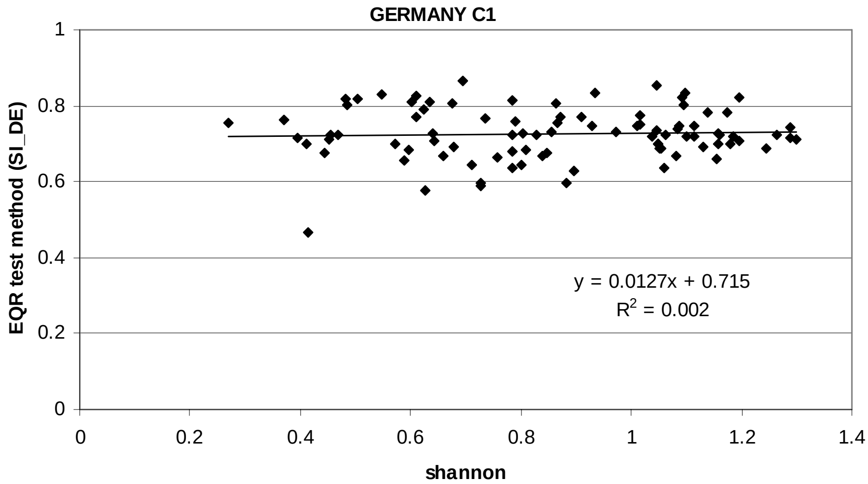


Figure 1C: Shannon index/SI(DE) - R^2 = 0.002; p=0.648

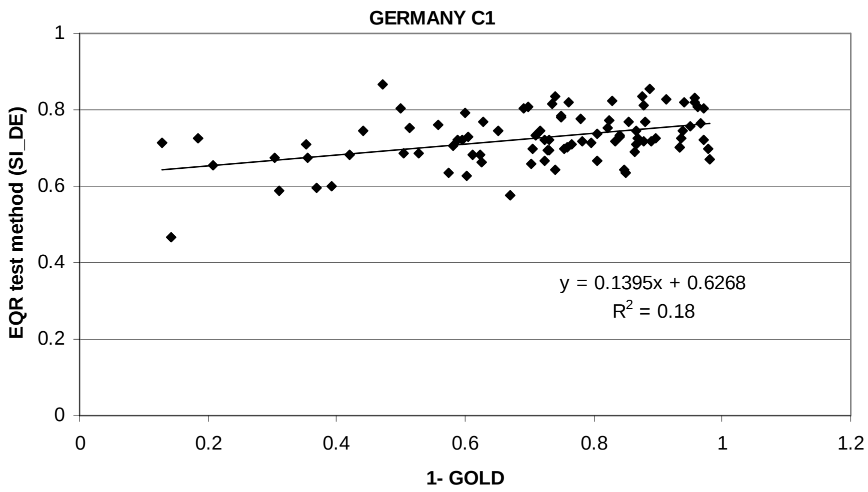


Figure 1D: 1-GOLD/SI(DE) - $R^2 = 0.18$; $p<0.001$

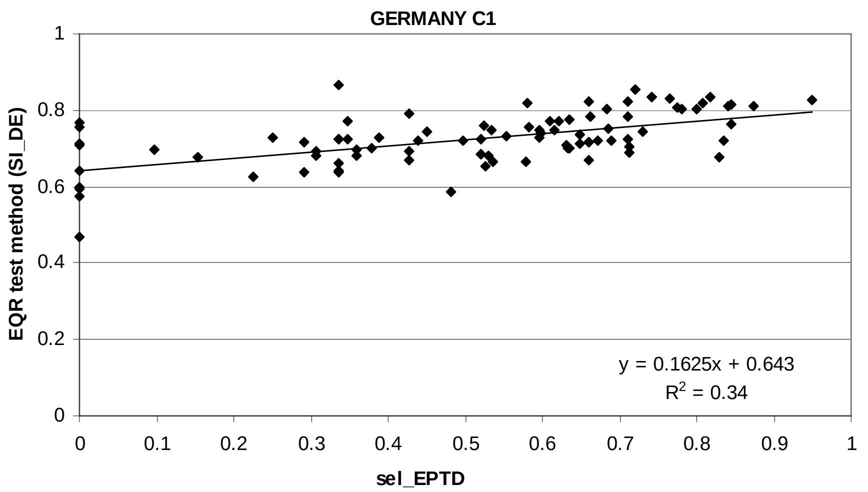


Figure 1E: Log EPTD/SI(DE) - $R^2 = 0.34$; $p<0.001$

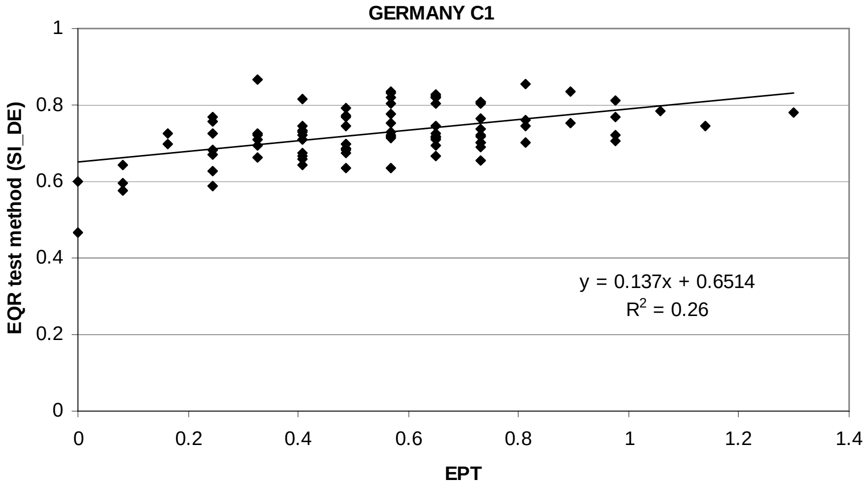


Figure 1F: EPT taxa/SI(DE) - $R^2 = 0.26$; $p<0.001$

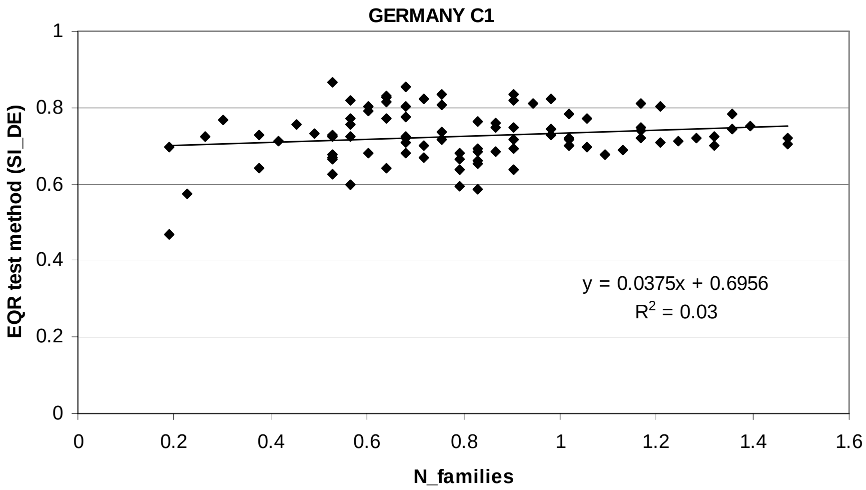


Figure 1G: Number of families/SI(DE) - $R^2 = 0.03$; $p=0.121$

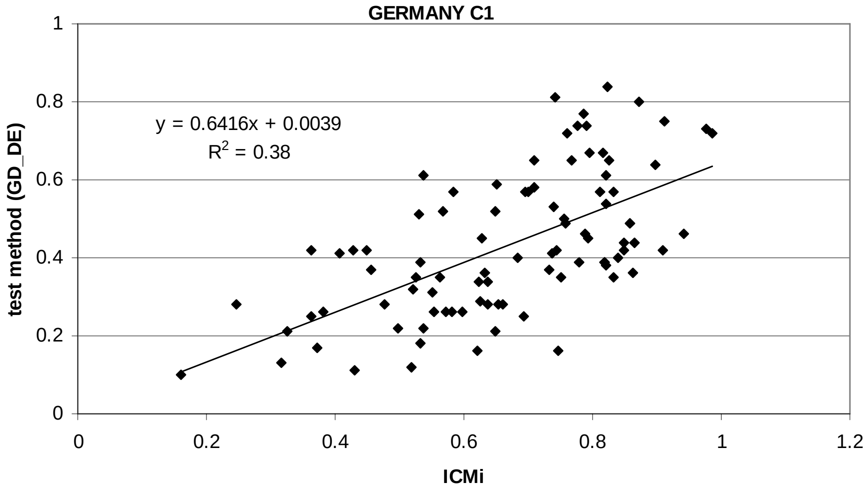


Figure 2: linear regression between ICMs and GD(DE) in dataset Germany C1
Figure 2A: ICMi - R^2 = 0.32; p<0.001

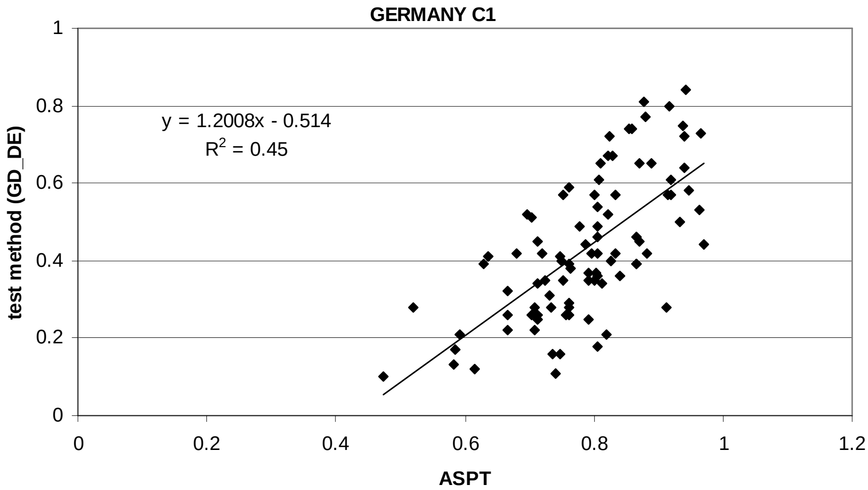


Figure 2B: ASPT index/GD(DE) - R^2 = 0.45; p<0.001

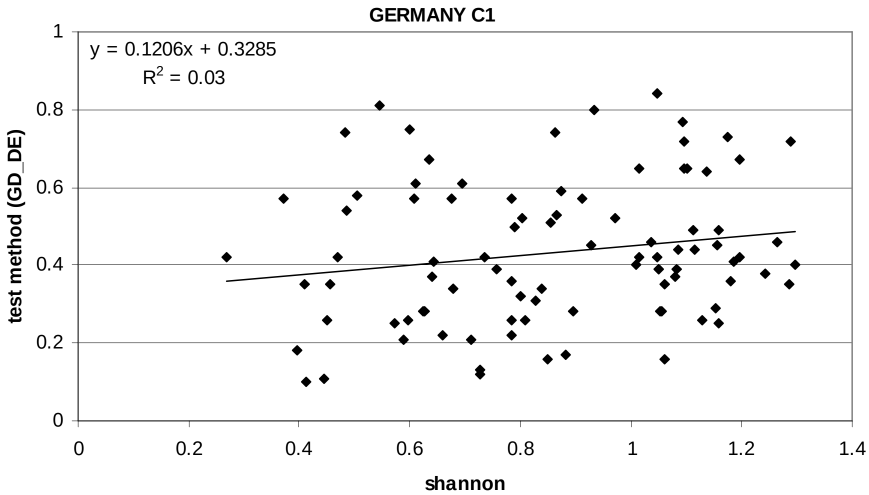


Figure 2C: Shannon index/GD(DE) - R^2 = 0.03; p=0.10

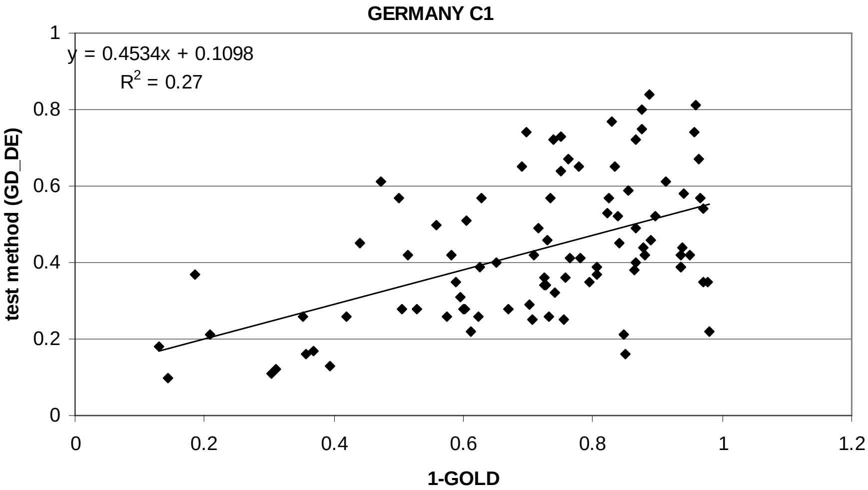


Figure 2D: 1-GOLD/GD(DE) - R^2 = 0.27; p<0.001

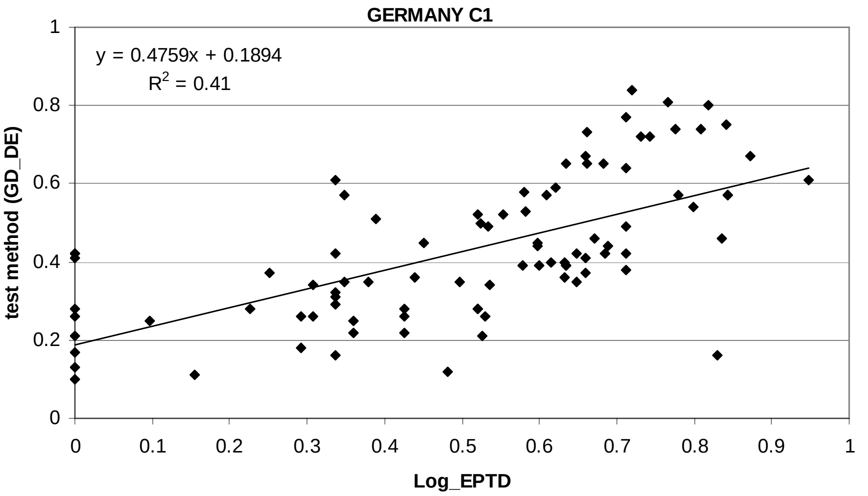


Figure 2E: Log EPTD/GD(DE) - R^2 = 0.41; p<0.001

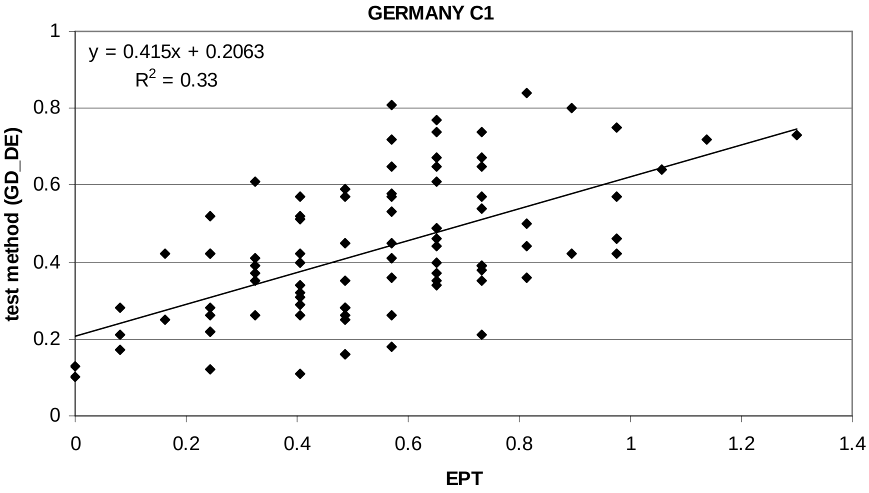


Figure 2F: EPT taxa/GD(DE) - R^2 = 0.33; p<0.001

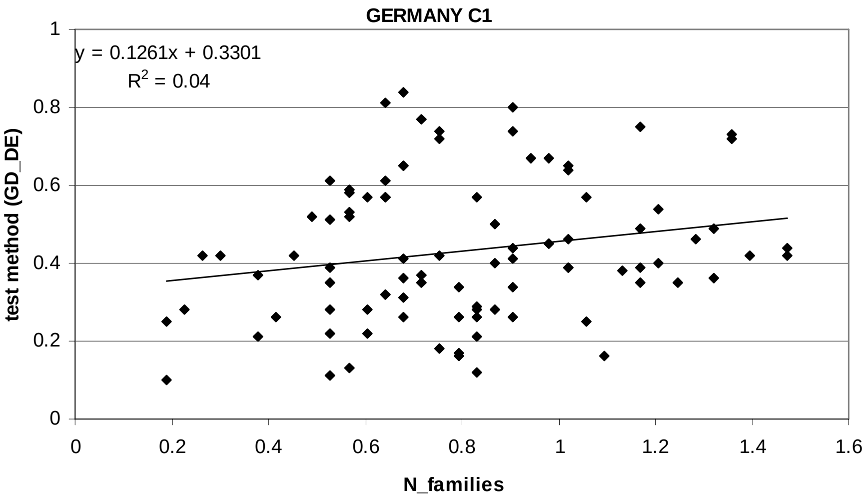


Figure 2G: Number of families/GD(DE) - R^2 = 0.04; p=0.05

General remarks, comments

The correlation with some ICMs is not significant, i.e.: SI(DE) vs Shannon p=0.648, SI(DE) vs Number of families p=0.121 and GD(DE) vs Shannon p=0.10.

The boundary of the poor-bad class transformed in ICMi has negative value. This can be due to the absence of 'bad quality' samples and to an overall low regression between ICMi and national method.

Possible hypothesis to be considered for low correlations ICMi vs National method in German dataset:

- In Germany, strong attention is paid to degradation in stream morphology.
- The identification level is species: different variability for family level data.
- On a total of 91 samples, only one is classified as High status and 8 are classified as Bad status. The dataset has a short gradient, with most of the sites in the 'central' classes.
- The multimetric GD(DE) (general degradation module) index is based on the principle of 'one out, all out' among different subindices that consider different alteration

factors. This can determine a low class for the class, e.g. if only the morphological quality is low. It is important to verify if quality gradients of different stressors are the same.

Notes on dataset description

The content of the present description is verified by Dr. Sebastian Birk from University of Duisburg-Essen who provided the data.

References related to the presented dataset

Birk, S., 2004. Description of how stream type-specific reference conditions using macrozoobenthos have been derived in Germany. 2pp. Essen, 15 November 2004.

DIN 38410, 2003. Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung - Biologisch-ökologische Gewässeruntersuchung (Gruppe M1) - Bestimmung des Saprobienindex in Fließgewässern (M1).

Friederich, G. & V. Herbst, 2004. Another revision of the saprobic index - why and what for? Acta hydrochim. hydrobiol. 32: 61-74. (in German with English abstracts).

4.3.6 Italy C1

Italy C1

General features

This dataset contains samples from typical Northern Italian spring fed streams in the lowland of Po river named ‘fontanili’ (see AQEM Consortium, 2002 for further description of this Italian type). Altitude is for all the sites lower than 200m and catchment area is very little (lower than 100km^2).
All the sites are located in region Lombardia, province of Milan. The sites are enclosed in a small area. Maximum distance between two sites is about 60 km.

Aim of collection, number of samples

Data of this dataset were collected and provided by Dr. Pietro Genoni from ARPA Lombardia (Regional Environmental Protection Agency) and are owned by ARPA Lombardia. Sites have been sampled during different sampling surveys with different aims such as monitoring, methodology testing, EA internal activities etc. (Genoni, unpublished data). Some sites are included in an intercalibration exercise on national assessment method (IBE), performed among different Environmental Agency’s working groups (Genoni, 2003; Genoni *et al.*, 1997; 1998).
In this dataset, 39 sites are included and data refer to 6 years of collection from 1994 to 2000. Data collection was performed in 4-6 sampling surveys seasons per year. Since not all sites were investigated in all the years (and seasons), total number of samples is 361.

Degradation factor

Streams belonging to this Italian stream type have usually managed banks and channel and are located in rural areas. The main degradation factor is not clearly discernible and can include morphological alteration, organic pollution, pesticides or other toxic substances. For these reasons, it’s possible to state here a ‘general degradation’ factor. In this dataset the quality gradient covers all the quality classes according to national method, from ‘high’ to ‘bad’ status. For most of these samples (not all), other support data are available such as main physical, chemicals and microbiological variables.

National method: sampling and sorting

The classification method used is the official national assessment method IBE (Indice Biotico Esteso, APAT-IRSA/CNR, 2003). According to this method, the sampling is performed along a transect between the two banks of the river in a riffle area and the number of replicates varies according to water width and general habitat diversification. The total area sampled is thus not fixed; in this databases it has been considered approximately 0.9m^2. The sorting is semiquantitative (a minimum number of specimens for each taxon has to be considered).
The identification is undertaken at genus and family level.

National method: criteria for abundance registration

As the sorting is semiquantitative, no precise indication of the real number of present specimens is given. Except for taxa present with less than 10 individuals, for which usually a real count is undertaken, the Italian EA operators use to give an indication of the relative abundance of the collected taxa by means of codified symbols, such as I for ‘present’ L for ‘abundant’ and U for ‘dominant’. For the use in this exercise, after consultation with the data collector, the symbols have been converted in numbers, according to the following criteria: 20 for ‘present’ taxa, 60 for ‘abundant’ and 180 for ‘dominant’.

National method: sites' classification
The final index score is obtained via a two-entry table, by comparison of two metrics: the total number of taxa collected and the Faunistic Group (ordered by an increasing scale of tolerance). Values of the index can vary from 0 to 14. In this database, the minimum and maximum observed values are 2.4 and 13.

Notes on classification
National: About 25% (94 on 361) of the samples are classified as 'high status' according to national assessment method.
BAC or pressures based: No other classification available (no Best Available Classification).

Comparison between the ICMi and IBE EQRs, single ICM and IBE EQRs
For the calculation of the ICMi, metrics was normalized according to 75° percentile observed in the 'high status' samples (see explanations in previous chapters). The minimum and maximum observed values for ICMi (in EQR) have been 0.17 and 1.09. Between ICMi and IBE, a regression coefficient of 0.72 was found (see figure 1A).
Results on linear regression between single ICM and IBE are shown in figure 1 B-G.
The scores of IBE in the graphs are expressed in EQR values, calculated dividing the IBE score for each sample by the 75° observed in the high status samples.
The conversion of the class boundary values for the IBE method from the original boundaries to ICMi values is done according to Table 1.

	IBE score	IBE EQR	ICM index
Limit high-good	9.6	0.906	0.837
Limit good-moderate	7.6	0.717	0.631
Limit moderate-poor	5.6	0.528	0.426
Limit poor-bad	3.6	0.340	0.220
ICM index = IBE EQR * 1.0911 + 0.1509 R^2=0.72; p<0.001			

Table 1: IBE class boundaries conversion

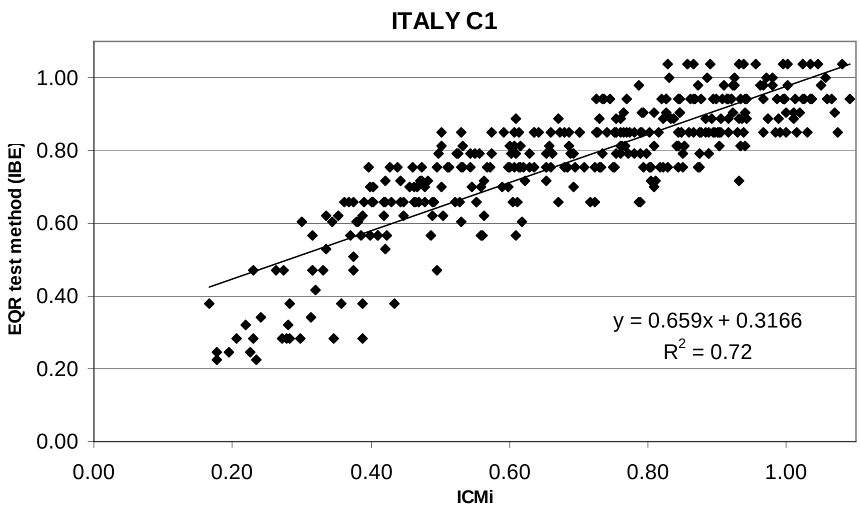


Figure 1: linear regression between ICMs and IBE in dataset Italy C1
Figure 1A: ICMi - R^2 = 0.72; p<0.001

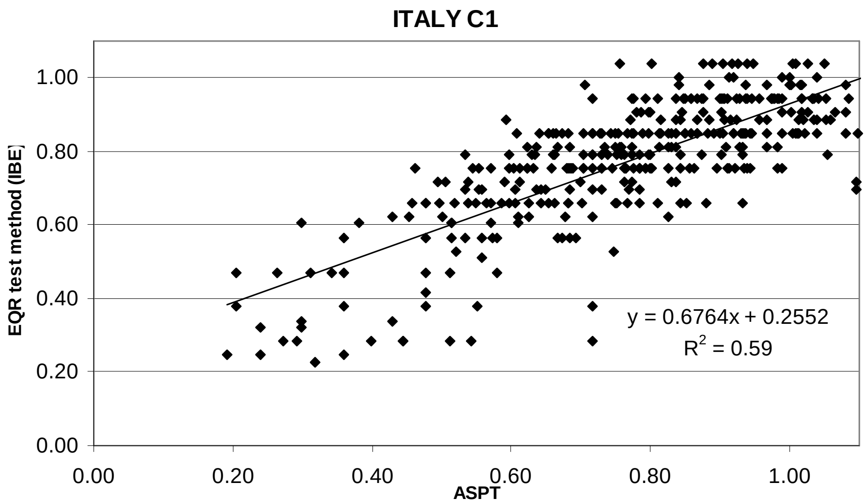


Figure 1B: ASPT index - $R^2 = 0.59$; p<0.001

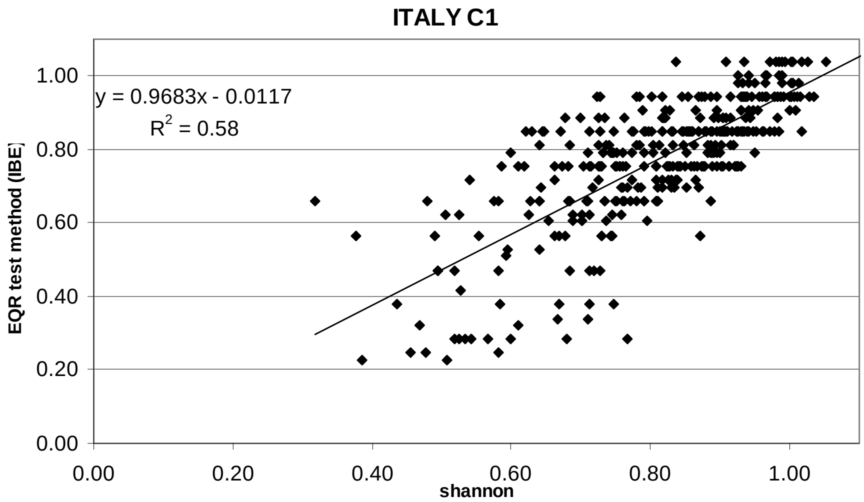


Figure 1C: Shannon index - $R^2 = 0.58$; p<0.001

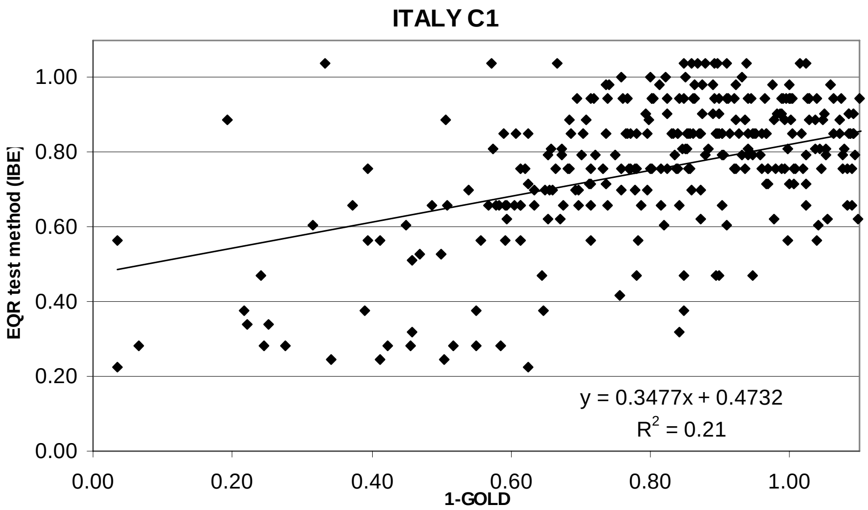


Figure 1D: 1-GOLD - $R^2 = 0.21$; p<0.001

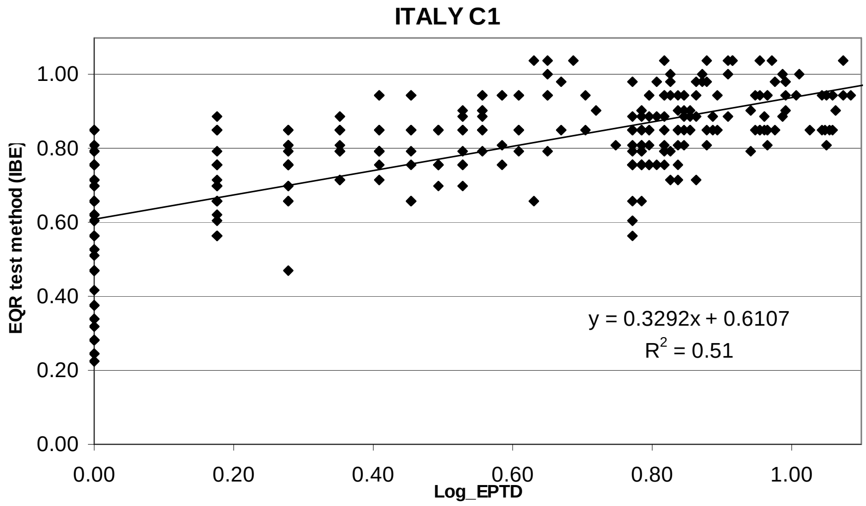


Figure 1E: Log EPTD - $R^2 = 0.51$; p<0.001

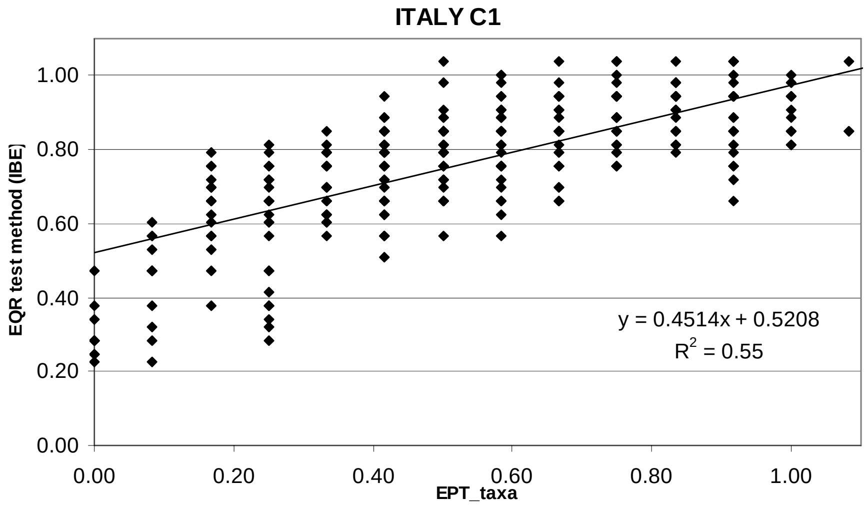


Figure 1F: EPT taxa - $R^2 = 0.55$; p<0.001

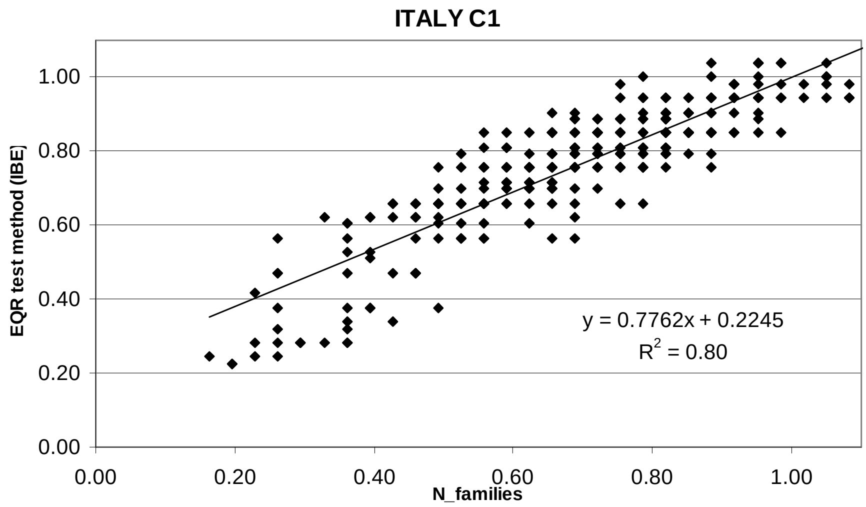


Figure 1G: Number of families - $R^2 = 0.80$; p<0.001

General remarks

In about 20 years, the calculation of IBE index encountered several updates (Ghetti, 1986; 1995; 1997; APAT-IRSA/CNR, 2004), especially in relation to the minimum number of specimens to be considered.
Since the samples refer to a period of 4 years, the calculation of the IBE index was originally performed following different IBE 'versions'.
In this dataset, the IBE values of all the samples have been recalculated according to the most updated version of the index, i.e.: APAT-IRSA/CNR, 2004.
This dataset is used in the example of calculation reported in the enclosed CD-ROM.

Notes on dataset description

The content of the present description is verified by Dr. Pietro Genoni from ARPA Lombardia (Regional Environmental Protection Agency), who collected and provided the data.

References related to the presented dataset

APAT-IRSA/CNR, 2004. Indice Biotico Esteso (I.B.E). In: APAT, Manuali e linee guida 29/2003. APAT-IRSA/CNR, Metodi analitici per il controllo della qualità delle acque 3: 1115-1136.

AQEM Consortium, 2002. Manual for the application of the AQEM system. Version 1.0 February 2002.

Genoni, P., P. Beati, F. Buzzi, P. Casarini, M. Girami E. Gozio, V. Mafessoni, P. Roella & A. Sarzilla, 1997. Intercalibrazione del metodo Indice Biotico Esteso I.B.E. (IRSA-CNR, 1995) per la valutazione della qualità dei corsi d'acqua. Regione Lombardia Direzione Generale Sanità Servizio Prevenzione Sanitaria. 62 pp.

Genoni, P., P. Beati, F. Buzzi, P. Casarini, M. Girami E. Gozio, V. Mafessoni, P. Roella & A. Sarzilla, 1998. Intercalibrazione del metodo Indice Biotico Esteso I.B.E. (IRSA-CNR, 1995) per la valutazione della qualità dei corsi d'acqua. Regione Lombardia Direzione Generale Sanità Servizio Prevenzione Sanitaria. 34 pp.

Genoni P., 2003. Influenza di alcuni fattori ambientali sulla composizione delle cenosi macrobentoniche dei corsi d'acqua planiziali minori. Biologia Ambientale 17 (1): 9-16.

Ghetti, P. F., 1986. I macroinvertebrati nell'analisi di qualità dei corsi d'acqua. Provincia Autonoma di Trento, Trento: 105 pp.

Ghetti, P. F., 1995. Indice Biotico Esteso (I.B.E.). Notiziario dei Metodi Analitici. IRSA - CNR, 7 luglio 1995, Roma: 1-24.

Ghetti, P. F., 1997. Indice Biotico Esteso (I.B.E.). I macroinvertebrati nel controllo della qualità degli ambienti di acque correnti. Provincia Autonoma di Trento: 222 pp.

4.3.7 Poland C1

General features
The sites enclosed in this dataset have an altitude lower than 200m and catchment area is comprised between 10 and 100km^2. The sites are quite evenly distributed across lowland part of Polish territory. The width of investigated river stretches is generally 2-5 m, reaching sporadically 8 m. Bottom substrate constitutes in most cases sand, sometimes with gravel or stones. On sites representing high/good status macrophytes are rather rare. Sites representing worse status are characterised by high abundance of macrophytes and filamentous algae.

Aim of collection, number of samples
Data were collected and provided by dr. Hanna Soszka and Malgorzata Golub from the Institute of Environmental Protection in Warsaw. Most samples were taken by voivodship inspectorates of environmental protection and were included in the pilot monitoring project (Kownacki et al., 2002). Significant part of the samples were collected also by the Institute of Environmental Protection for the intercalibration purposes.
Set of data provided in November 2004 for the present pilot exercise purposes comprises overall 49 samples.

Degradation factor
The sites are affected mainly by eutrophication. The quality gradient covers all the quality classes according to the national method, from 'high' to 'bad' status. Support data are available on water chemistry and characterization of site and catchment.

National method: sampling and sorting
Data were collected according to Polish Protocol. At each sampling occasion 5 samples are taken. Four of them are quantitative (from dominating substrate using Surber net or Ekman-Birge grab) and one is qualitative (from all habitats present at the site) to expand the list of taxa (Kownacki & Soszka 2004).
The abundance of fauna was recalculated to 1 m^2 Macroinverterbrates were identified to the family level.

National method: sites' classification
The method of assessment is based on 2 components: BMWP score adapted to Polish conditions (BMWP-PL) and modified Margalef's diversity index (Kownacki et al., 2004).
The BMWP (Armitage *et al.*, 1983) assigns a score to each collected taxon, decreasing according to its tolerance. The total sum gives the BMWP value of the site.
In this dataset, a modified standard BMWP table is used (BMWP-PL), in order to better represent the ecological gradient in Polish rivers. These modifications include:
- verification of usefulness of taxa scored in the original British system in Polish conditions,
- supplementing the list of families with several taxa not occurring in Great Britain due to zoogeographical isolation, but present in Poland and having a role as indicators of water quality,
- change of score assigned to several taxa (in comparison with the original BMWP)

If the values of both assessment elements differ by one class, the final classification is based on the the worst value. If the values of assessment elements differ by two classes (very rare situation), the average value is taken.

The minimum value for BMWP-PL is 0, the maximum is open end. Boundaries between classes are: high-good 100; good-moderate 70; moderate-poor 40; poor-bad10. In the present dataset maximum observed value is 158, the minimum is 5.

The second index is the modified Margalef diversity index (D), calculated as follows:

D = S/LOGN

S = NUMBER OF FAMILIES

N = TOTAL ABUNDANCE

VALUES ARE FROM 0 TO AN OPEN END. BOUNDARIES BETWEEN CLASSES ARE: HIGH-GOOD 5.5; GOOD-MODERATE4; MODERATE-POOR 2.5; POOR-BAD 1. MAXIMUM OBSERVED VALUE IS 11.75, THE MINIMUM IS 0.74.

NOTES ON CLASSIFICATION

BAC or pressures based: no other classification available (no Best Available Classification).

Comparison between the ICMi an national classification EQRs, single ICM and national classification EQRs

For the calculation of the ICMi, metrics were normalized according to 75° percentile of high status samples provided and classified using national method (see explanations in previous chapters). Final ICMi is re-normalized according to 75° percentile value. The minimum and maximum observed values for ICMi (in EQR) are 0.02 and 1.1. Also the two components of Polish assessment method, BMWP-PL and Margalef index, are transformed in EQR through a normalization according to the high status samples' 75° percentile.

The classification is undertaken following the concept 'one out all out' between the two indices. In the present dataset the index BMWP-PL gives always the worst classification, when non consistence is observed. Thus, it has been decided to undertake the harmonization on this index only.

Regression coefficient found is : ICMi vs BMWP-PL: 0.74, ICMi vs Margalef: 0.40 (see figure 1A, a, b).

Results on linear regression between single ICM and the BMWP-PL are shown in figure 1 B-G.

The conversion of the class boundary values for the method from the original boundaries to ICMi values is done according to Table 1.

	BMWP score	BMWP EQR	ICM index
Limit high-good	100	0.775	0.827
Limit good-moderate	70	0.543	0.612
Limit moderate-poor	40	0.310	0.398
Limit poor-bad	10	0.078	0.183
ICM index = BMWP EQR * 0.9227 +0.1116			
R^2=0.74; p<0.001			

Table 1: Poland index class boundaries conversion

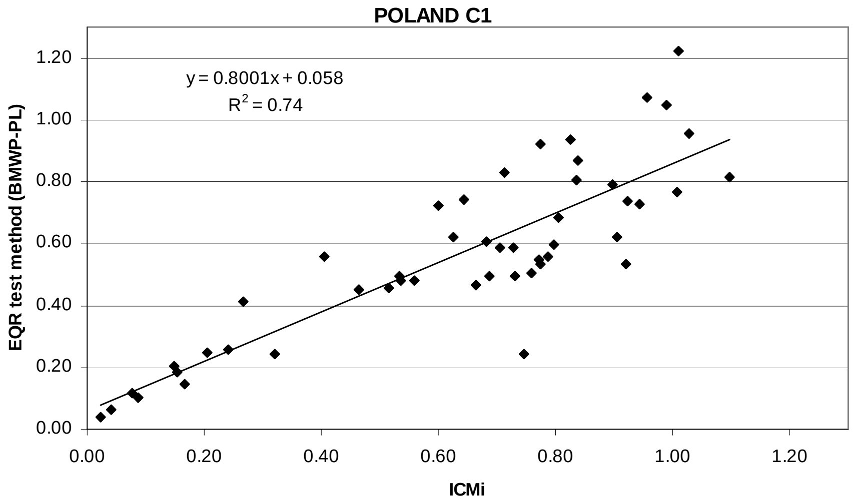


Figure 1Aa : ICMi vs BMWP-PL - $R^2 = 0.74$; $p<0.001$

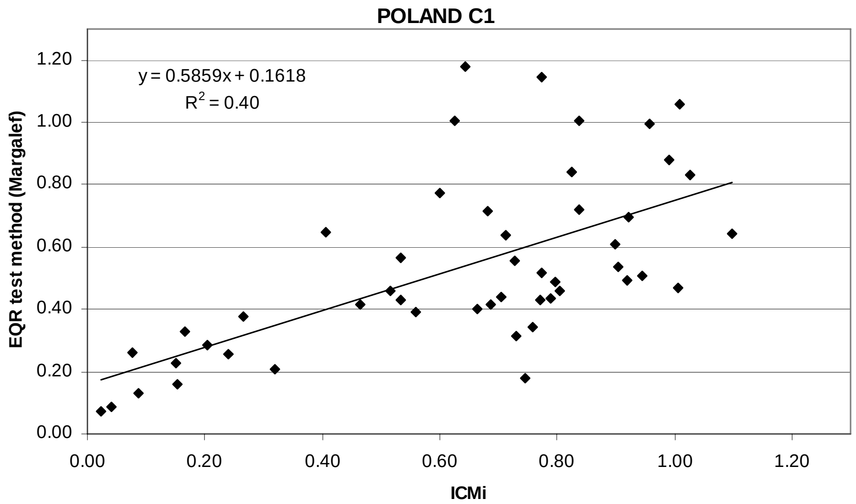


Figure 1Ab: ICMi vs Margalef - $R^2 = 0.40$; $p<0.001$

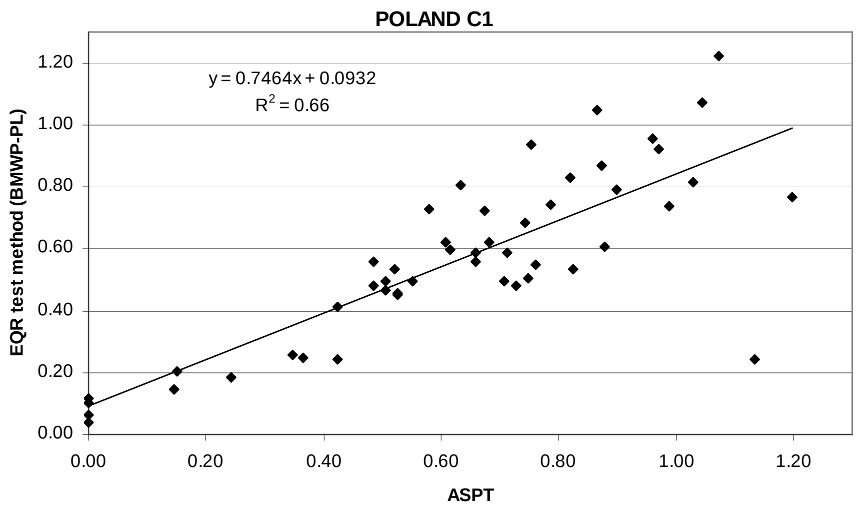


Figure 1B: ASPT vs BMWP-PL - $R^2 = 0.66$; $p<0.001$

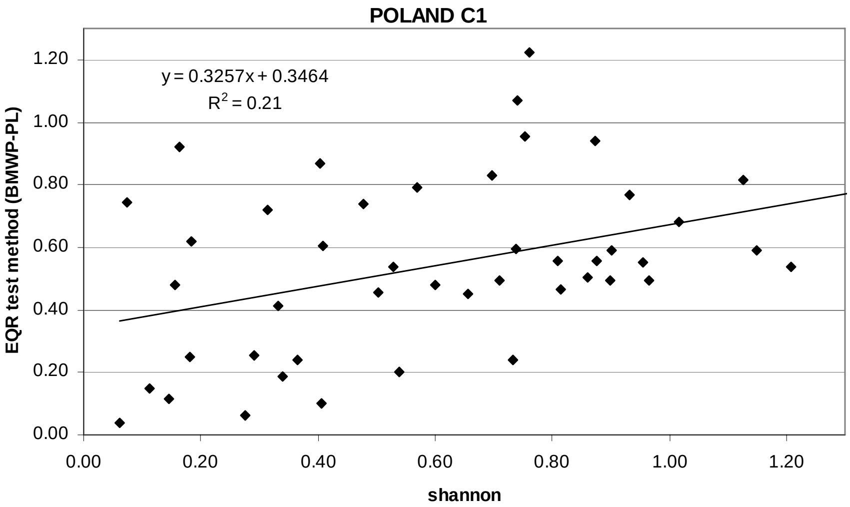


Figure 1C: Shannon - $R^2 = 0.21$; p<0.001

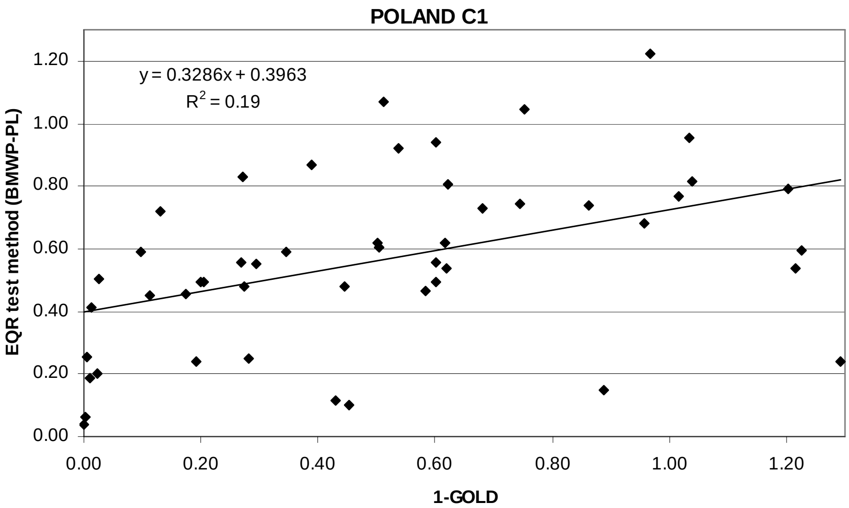


Figure 1D: 1-GOLD - $R^2 = 0.19$; p<0.001

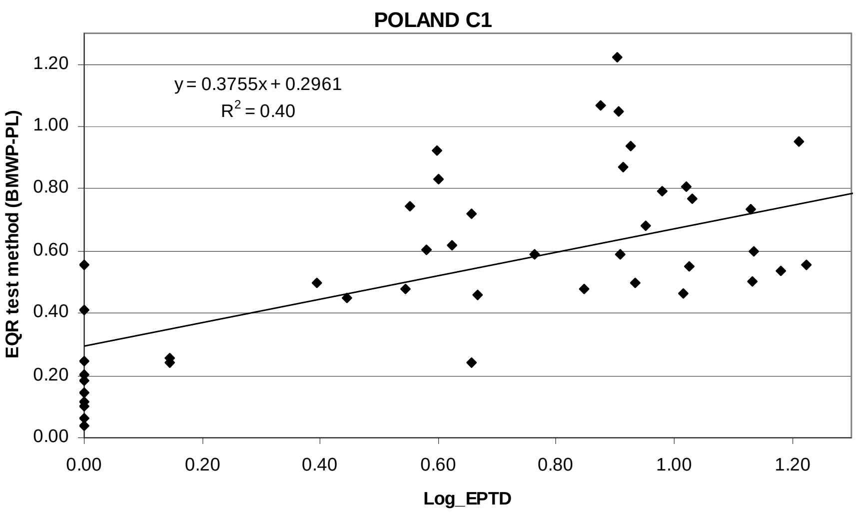


Figure 1E: Log EPTD - $R^2 = 0.40$; p<0.001

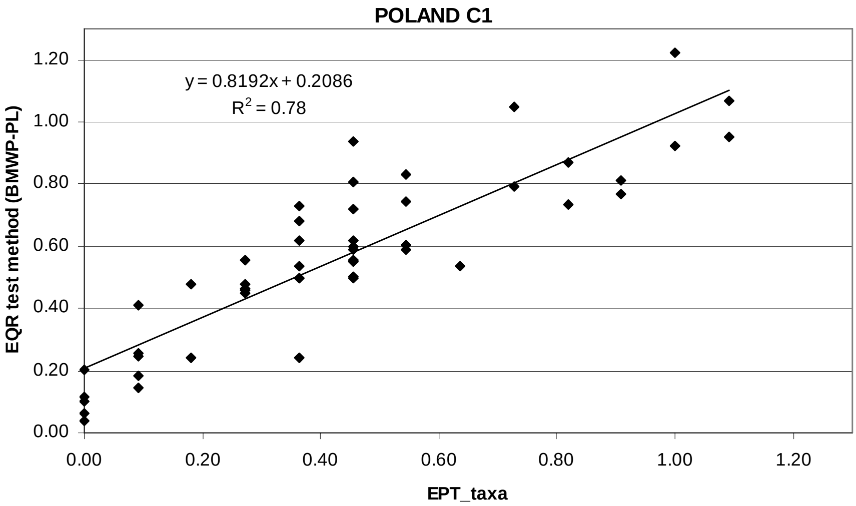


Figure 1F: EPT - $R^2 = 0.78$; p<0.001

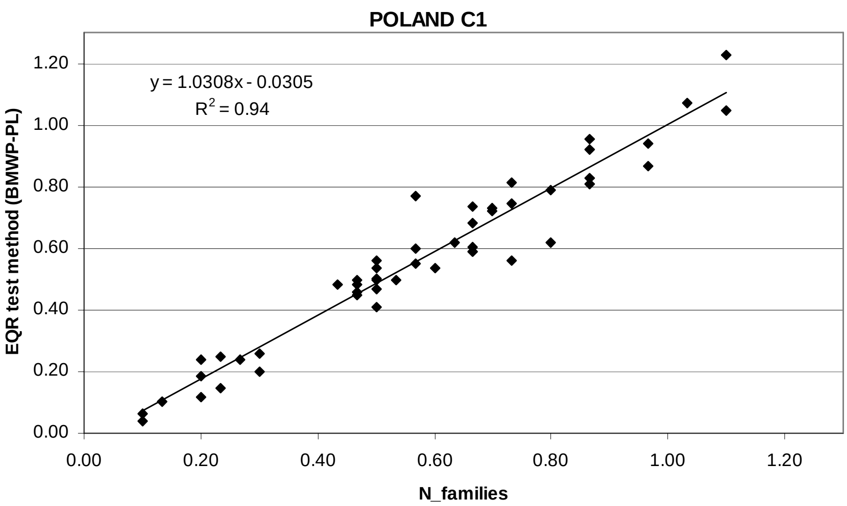


Figure 1G: N families - $R^2 = 0.94$; p<0.001

General remarks

-

Notes on dataset description

The content of the present description will be verified by Dr. Hania Soszka and Malgorzata Golub from the Institute of Environmental Protection in Warsaw, who provided the data.

References related to the presented dataset

Armitage, P. D., D. Moss, J. F. Wright & M. T. Furse, 1983. The performance of a new biological water quality scores system based on macroinvertebrates over a wide range of unpolluted running-water sites. Wat. Res. 17: 333–347.

Kownacki, A., H. Soszka, T. Fleituch & D. Kudelska (eds.), 2002. River biomonitoring and benthic invertebrate communities. Institute of Environmental Protection, Karol Starmach Institute of Freshwater Biology PAS, Warszawa-Krakow, 88 pp.

Kownacki, A., H. Soszka, D. Kudelska & T. Fleituch, 2004. Bioassessment of Polish rivers based on macroinvertebrates. Proceedings of the international 11th Magdeburg Seminar on Waters in Central and Eastern Europe: Assessment, Protection, Management. 18-22 October 2004, UFZ Leipzig (W. Geller et al. (Eds.): 250-251.

Kownacki A. & H. Soszka 2004. Guidelines on the river assessment based on macroinvertebrates for voivodship inspectorates of environmental protection. Chief Inspectorate of Environmental Protection, Warsaw (unpublished, in Polish).

4.3.8 UK C1

General features

The sites have a small catchment surfaces and are located in lowland area.

Aim of collection, number of samples

Data are collected and owned by the Environment Agency. A small amount of the data was collected and is owned by the Scottish Environmental Protection Agency (SEPA). The dataset has been provided by John Murray Bligh from EA. The sites are included in the national monitoring network for the program of Environmental Protection. The data is available on the River biology Monitoring System that can be downloaded from http://www.cies.staffs.ac.uk/. The total number of sites included is 789. The year of collection is 1995. Sites were sampled in two seasons, spring and autumn. Each sample is derived from the combination of the biological samples of the two seasons. Total number of samples is 789.

Degradation factor

As main factor of alteration, a general degradation can be stated. Actually, the sites are affected by different alterations, probably mostly organic. The quality gradient covers all the quality classes according to the national method, from 'high' to 'bad' status. The support data for all the sites regard chemical monitoring data and pressures (perceived stressed) data.

National method: sampling and sorting

The sampling method is the one applied for the RIVPACS method (Wright, 1995; Murray-Bligh, 1999). Sampling is carried out taking two samples of 3 minutes each plus a search of 1 minute. The total sampling area is not specified.
Identification is performed to family level.

National method: sites' classification

The final classification is undertaken through the combination of two indices: EQI ASPT and EQI N-taxa. ASPT is the value of BMWP divided by the total number of collected taxa. The BMWP (Armitage *et al.*, 1983) assigns a score to each collected taxon, decreasing according to its tolerance. The EQI ASPT (and the EQI N-taxa) corresponds to the observed ASPT (or Number of families) for combined spring and autumn sample, divided by the RIVPACS prediction for the same combination. Each of the two indices give a classification, the poorest class indicated by either EQI ASPT or EQI N-taxa is the overall quality class for a site.
Minimum and maximum values can vary according to the considered dataset. In the present set the values of the ICMi vary from 0.16 to 1.20 The transformation in EQR is done according to type.

National method: criteria for abundance registration

Only logarithmic abundance classes were recorded. Actual abundances were simulated:
1-9 = 4
10-99 = 40
100-999 = 400
1000-9999 = 4000

Notes on classification

National: The preliminary UK class boundaries were agreed in a meeting in Edinburgh 20 August 04. They are the 5M classification scheme boundaries used in UK from 1990-94. These were first published in The Scottish Office (1992). The current classification scheme (GQA, EA, 1997) differs from the one proposed for the WFD. The boundaries of the latter is

used in the present exercise. About 36% (299 on 789) of the samples are classified as 'high status' according to such assessment method.
BAC or pressures based: no other classification available (no Best Available Classification).

Comparison between the ICMi and National method, single ICM and National method
For the calculation of the ICMi, metrics was normalized according to 75° percentiles of High status samples (see explanations in previous chapters). Final ICMi is re-normalized according to 75° percentile. In the present set the values of the ICMi vary from 0.1 to 1.1.
Between ICMi and ASPT-EQI a regression coefficient of 0.82 was found (see figure 1A).
Results on linear regression between single ICM ASPT-EQI are shown in figure 1 B-G.
The regression coefficient between ICMi and NFAM-EQI is 0.71 (see figure 2a). For regression of the single ICM and the NFAM-EQI see figure 2 B-G
The conversion of the class boundary values for the ASPT-EQI method and NFAM-EQI from the original boundaries to ICMi values is done according to Table 1 and Table 2.

	ASPT-EQI score	ASPT-EQI_EQR	ICM index
Limit high-good	1	0.943	0.864
Limit good-moderate	0.88	0.830	0.693
Limit moderate-poor	0.76	0.717	0.521
Limit poor-bad	0.65	0.613	0.363
ICM index = combUK EQI_EQR * 1.5169 -0.5667			
R2=0.82; p<0.001			

Table 1: ASPT-EQI class boundaries conversion

	NFAM-EQI score	NFAM-EQI_EQR	ICM index
Limit high-good	1	0.826	0.826
Limit good-moderate	0.78	0.645	0.665
Limit moderate-poor	0.57	0.471	0.511
Limit poor-bad	0.36	0.298	0.357
ICM index = combUK EQI_EQR * 0.8872 +0.0926			
R2=0.71; p<0.001			

Table : NFAM-EQI class boundaries conversion

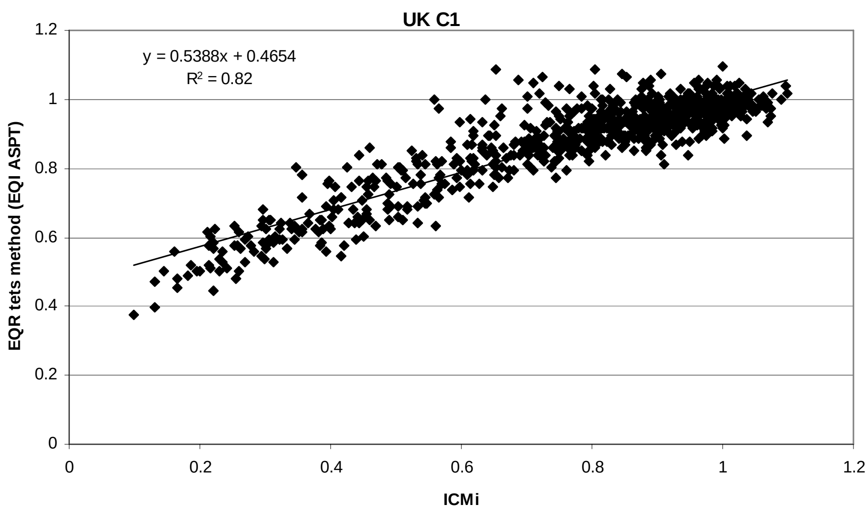


Figure 1: linear regression between ICMs and ASPT-EQI in dataset UK C1
Figure 1A: ICMi - R^2 = 0.82; p<0.001

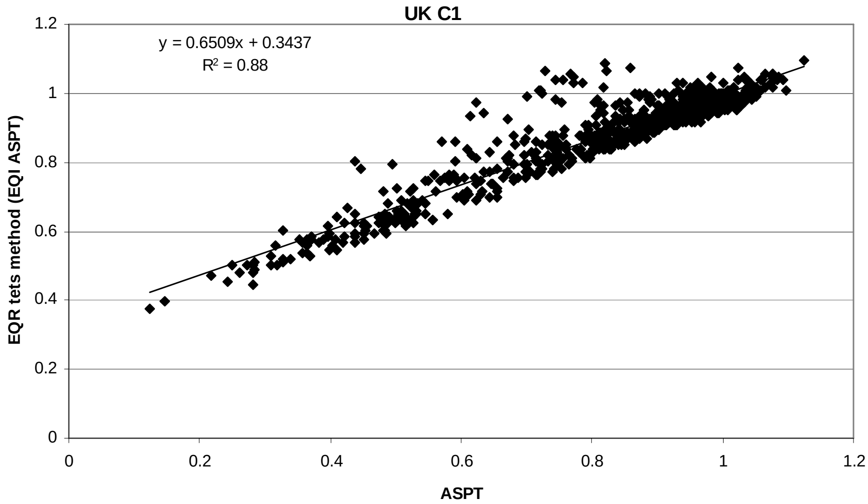


Figure 1B: ASPT index/ ASPT-EQI - R^2 = 0.88; p<0.001

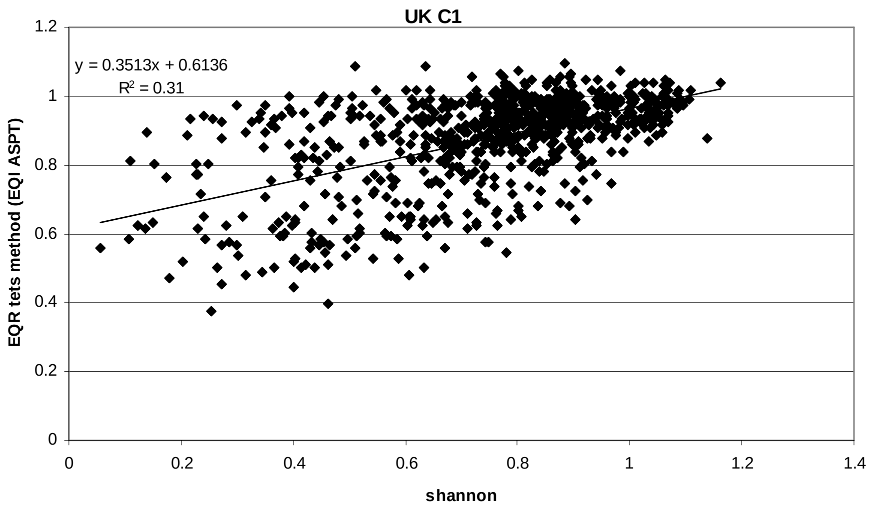


Figure 1C: Shannon index/ ASPT-EQI - R^2 = 0.31; p<0.648

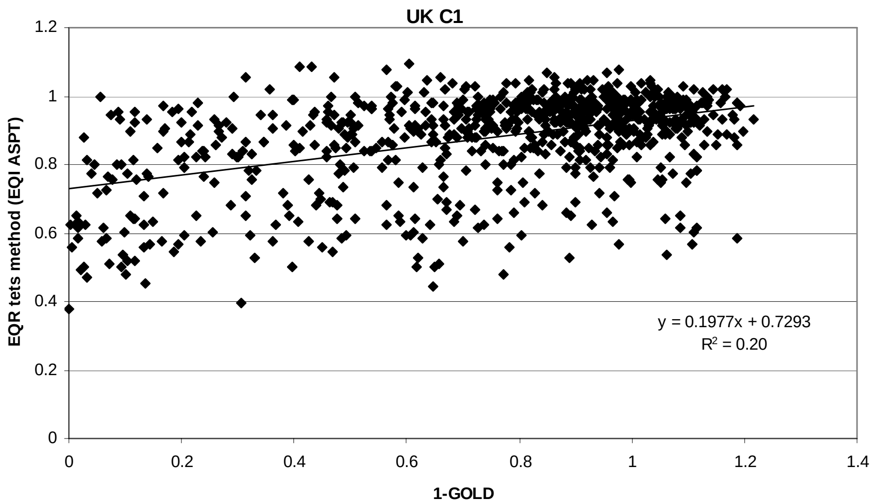


Figure 1D: 1-GOLD/ ASPT-EQI - R^2 = 0.20; p<0.001

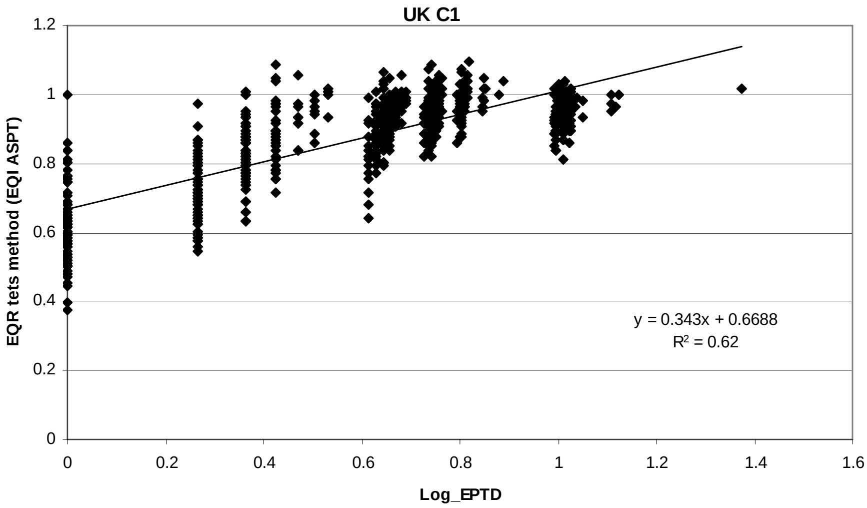


Figure 1E: Log EPTD/ ASPT-EQI - R^2 = 0.62; p<0.001

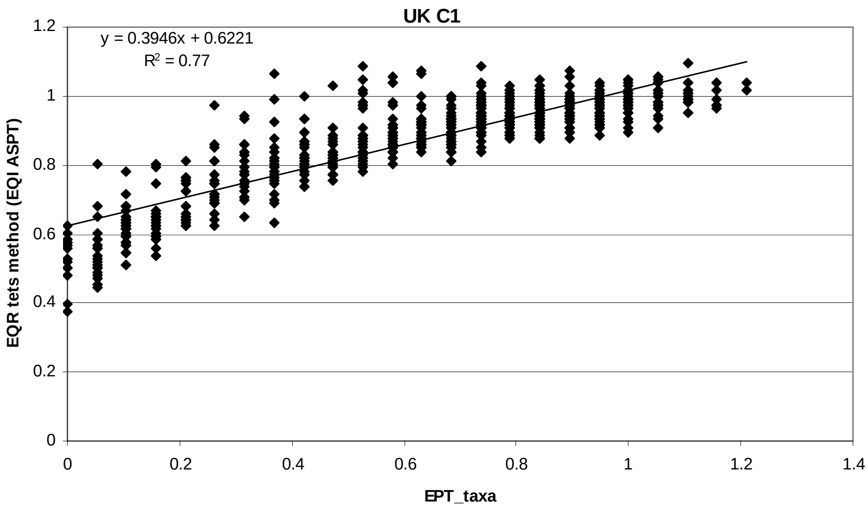


Figure 1F: EPT taxa/ ASPT-EQI - R^2 = 0.77; p<0.001

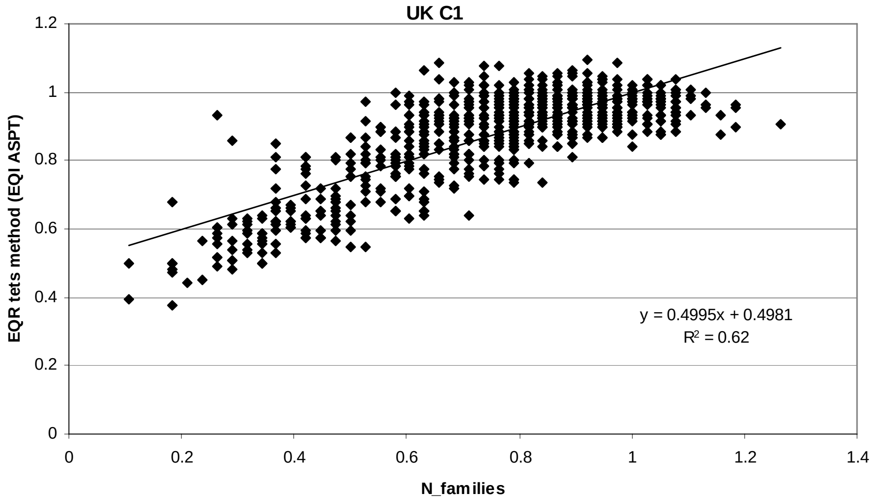


Figure 1G: Number of families/ ASPT-EQI - R^2 = 0.62; p<0.121

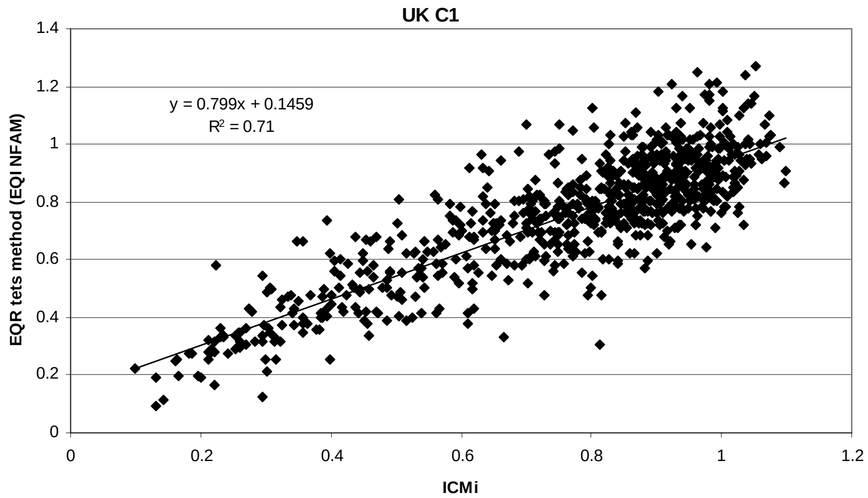


Figure 2: linear regression between ICMs and NFAM-EQI in dataset UK C1
Figure 2A: ICMi - R^2 = 0.71; p<0.001

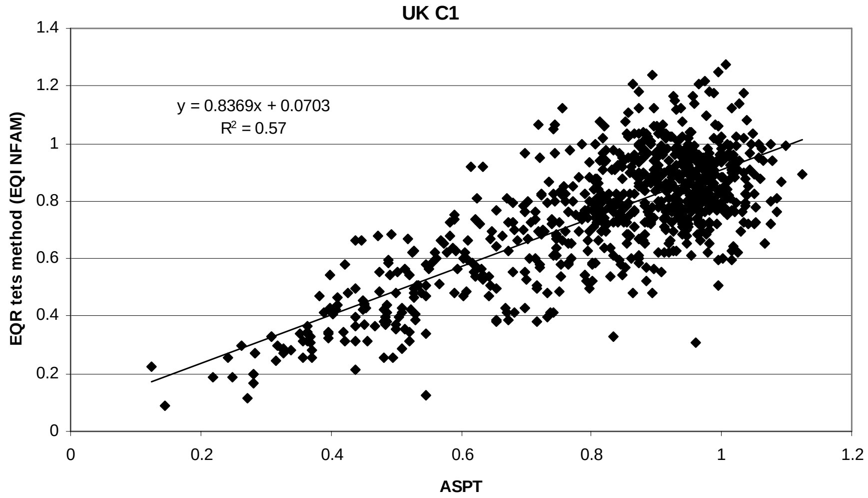


Figure 2B: ASPT index/ NFAM-EQI - $R^2 = 0.57$; $p<0.001$

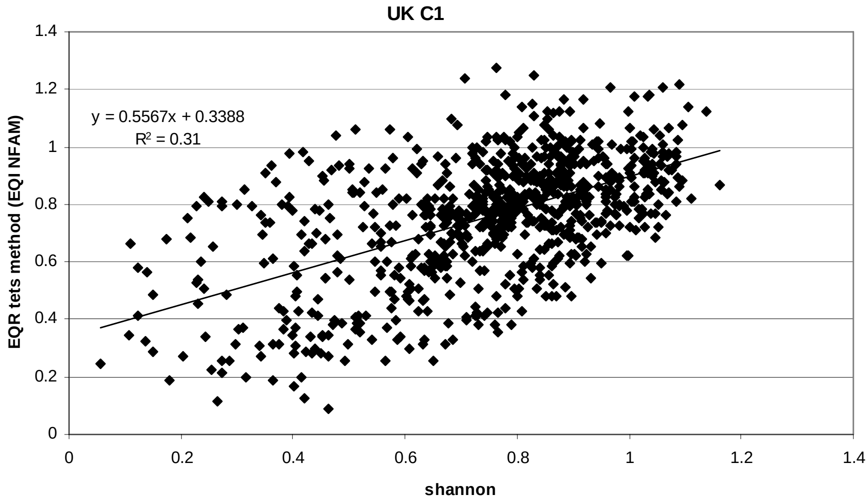


Figure 2C: Shannon index/ NFAM-EQI - $R^2 = 0.31$; $p<0.10$

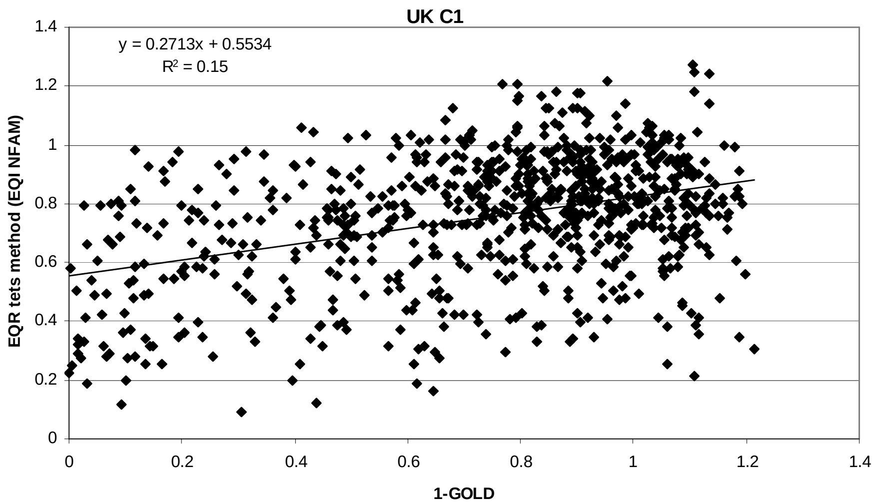


Figure 2D: 1-GOLD/ NFAM-EQI - $R^2 = 0.15$; $p<0.001$

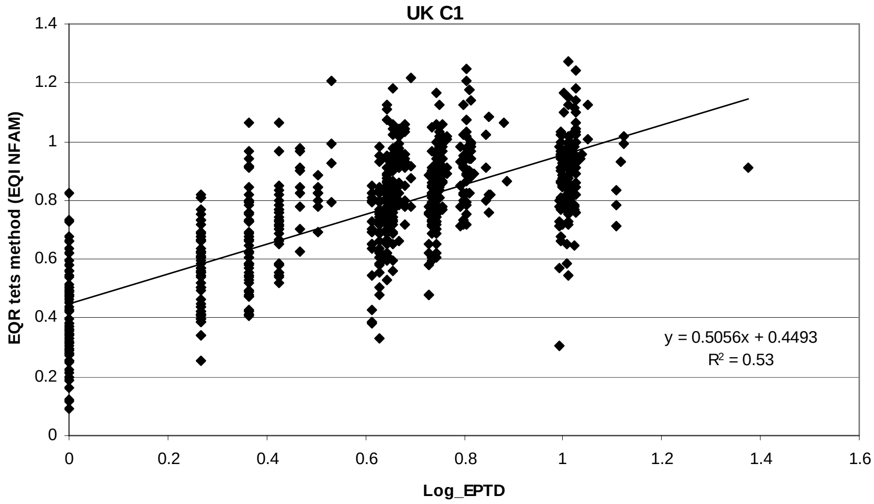


Figure 2E: Log EPTD/ NFAM-EQI - $R^2 = 0.53$; $p<0.001$

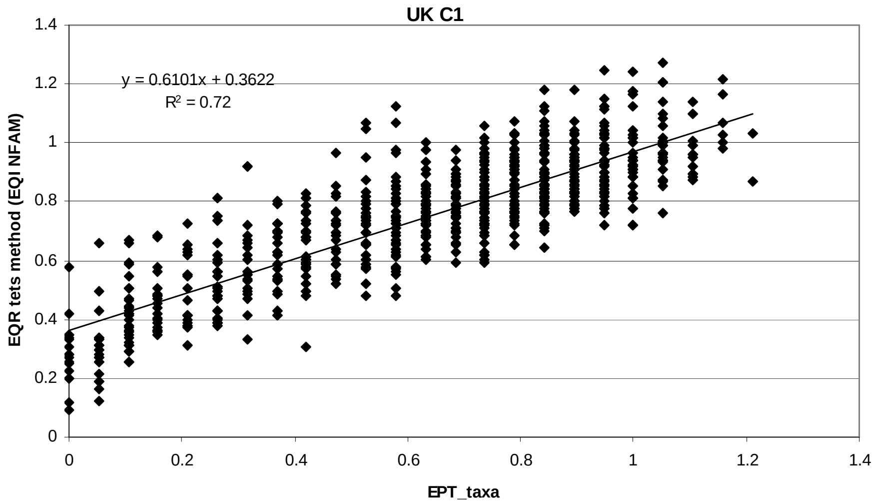


Figure 2F: EPT taxa/ NFAM-EQI - $R^2 = 0.72$; $p<0.001$

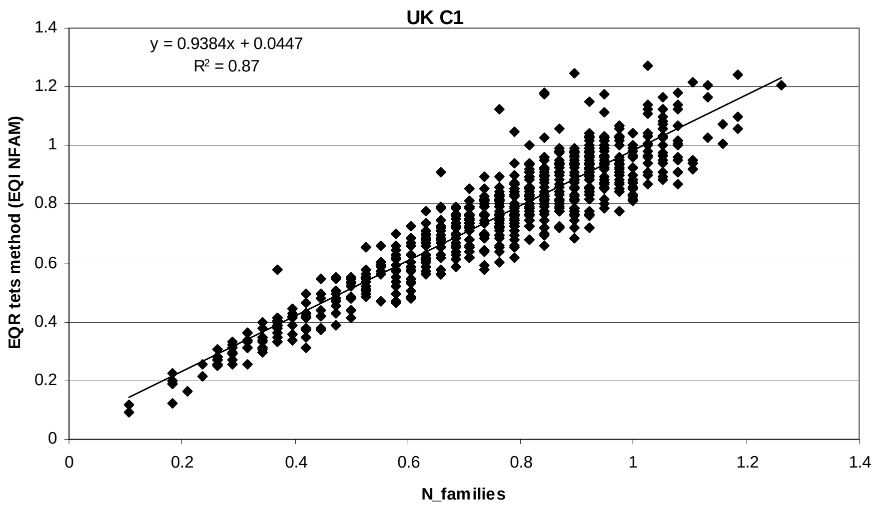


Figure 2G: Number of families/ NFAM-EQI - $R^2 = 0.87$; $p<0.001$

General remarks
Data were normalized according not to the current GQA but to the proposed WFD scheme.

Notes on dataset description
The content of the present description was verified by Dr. John Murray-Bligh of EA, who provided the data.

References related to the presented dataset

EA, 1997. Environment Agency. Assessing Water quality - General Quality Assessment (GQA) scheme for Biology. Fact Sheet. Bristol (Environment Agency).

Murray-Bligh, J. A. D., 1999. Procedure for quality assurance for RIVPACS compatible macro-invertebrate samples analysed to the taxonomic level needed for the BMWP-score system. Quality Management Systems for Environmental monitoring: Biological Techniques, BT003. (Version 1.0, 3 August 1999) Bristol, Environment Agency.

The Scottish Office, 1992. Water Quality survey of Scotland 1990. Tee Scottish Office, Edinburgh ISBN0 7480 0597 8.

Wright, J. F., 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Aus. J. Ecol. 20: 181–198.

4.3.9 France C2

General features

Sites belong to the hydro-ecoregion “Massif Armoricain” (HER 12) of the French typology. Altitude is for all the sites enclosed in this dataset lower than 150m and catchment area is comprised between 10 and 200km^2. Correspond to the small streams. Geology is siliceous with rocky substrates. Climatic conditions are oceanic.

Aim of collection, number of samples

Data collection was performed by the Direction Régionale de l’Environment. The database is organized by Lyon Cemagref and has been provided by Dr. Jean Gabriel Wasson. The sites are included in the national monitoring network and regularly investigated for quality assessment.

The total number of sites included is 38. In this dataset, the samples collected from 1992 to 2002 are included. Data collection was performed in several seasons per year (number of seasons not specified). Total number of samples is 143.

Degradation factor

General degradation is the main factor of alteration. The quality gradient covers all the quality classes according to the national method, from ‘high’ to ‘bad’ status. The support data are available from the National monitoring network. The type of data available is not specified.

National method: sampling and sorting

The method of classification is the official French monitoring method IBGN (Indice Biologique Global Noramalisé, Vernaux et al., 1982). Sampling is carried out taking a number of 8 samples with a Surber sampler (base surface 1/20 m^2). These samples are characterized by different fixed couple of substrate dimensions and flow velocity. The total sampling area is 0.4 m^2. To be considered as valid, a single taxaon has to be present with a minimum number of 3 specimens (or 10 specimens for few taxa). Identification is performed to family level.

National method: sites’ classification

For the final classification, two metrics are considered: the Faunistic Indicator Group (GFI) whose values range from 1 to 9 and the number of collected families (taxonomic variety, VT) divided into 14 classes. The final IBGN value is obtained by the sum of these two metrics. Values of the index can vary from 0 to 20; boundaries between quality classes can have different values according to the stream type. For the C2 boundaries are: reference-high, 16; high-good, 14; good-moderate, 12. The boundaries moderate-poor and poor-bad are not defined for the C2 stream type. The transformation in EQR is done according to type.

In this database, the minimum and maximum observed values for IBGN are 3 and 19.

Notes on classification

National: 73 samples on 143 (about 50%) are classified as ‘high status’ according to national assessment method.

BAC or pressures based: 27 samples classified as ‘reference’.

Reference sites are selected on the basis of very low anthropic pressures, independently of the biological values in a first approach. The distribution of biological data is then calculated for all samples of the reference dataset, and the outliers samples are checked. Dubious sites are eliminated, but low biological values are accepted if they come from validated reference sites.

The procedure combine both spatial and temporal variability of a given stream type. The Reference Conditions (RC) are defined as the range of variability of a given biological element (index or metric) observed at reference sites. However, the calculation of EQR needs to define a Reference Value (RV) for the normalization of the samples. Due to the small number of reference sites generally observed for most types, the most robust and stable statistic is chosen as RV. For all our calculations, following the recommendation of the REFCOND guidance, the median was used as Reference Value.
The general approach and Reference Values for each type are described in a work paper (Wasson et al., October 2003, in French) and a summary (in English) will be available soon.
Reference sites were first selected from the monitoring network and other complementary sites in using two independent methods :
“expert selection” by the field hydrobiologists (DIREN teams), on the basis of a detailed questionnaire combining all the possible pressures at the basin, reach and site scale.
“GIS selection” run by Cemagref on the basis of known point source pollution discharges (from water agencies), and land use (CORINE), at the scale of hydrologic units (sub-basins ca. 100 km^2). However, this selection eliminates impacted basins where reference sites could be found upstream of pollution discharge (Wasson et al., August 2004, in French)
The IBGN values observed in these two selections of sites were compared to the values calculated from reference sites selected and sampled by the Cemagref hydrobiologists. The reference value for a given stream type was accepted only if the IBGN values observed in the three datasets were in good concordance. If not, a checking procedure was run and dubious sites were eliminated.
Since December 2004, the boundaries of the IBGN classes are redefined according to this definition of reference samples. In particular the boundary High good is set at the 25° percentiles of the reference samples.

Comparison between the ICMi and IBGN EQRs, single ICM and IBGN EQRs
For the calculation of the ICMi, metrics was normalized according to the 75° observed in the high status and reference status samples according to (see explanations in previous chapters). This normalization option is suitable only for this IC exercise purpose. It will not be used in France for WFD implementation.
Final ICMi is re-normalized according to its 75° percentiles. The minimum and maximum observed values for ICMi (in EQR) have been 0 and 1.18. Between ICMi and IBGN, a regression coefficient of 0.85 was found (see figure 1A).
Results on linear regression between single ICM and IBGN are shown in figure 1 B-G.
The scores of IBGN in the graphs are expressed in EQR values, calculated dividing the IBGN score for each sample by the 75° observed in the high status samples.
The conversion of the class boundary values for the IBGN method from the original boundaries to ICMi values is done according to Table 1.

	IBGN score	IBGN EQR	ICM index
Limit ref-high	16	0.941176471	0.906
Limit high-good	14	0.823529412	0.794
Limit good-moderate	12	0.705882353	0.681
Limit moderate-poor	nd	nd	nd
Limit poor-bad	nd	nd	nd
ICM index = IBGN EQR * 0.9585 + 0.0043 R^2=0.85; p<0.001			

Table 1: IBGN class boundaries conversion for C2 dataset

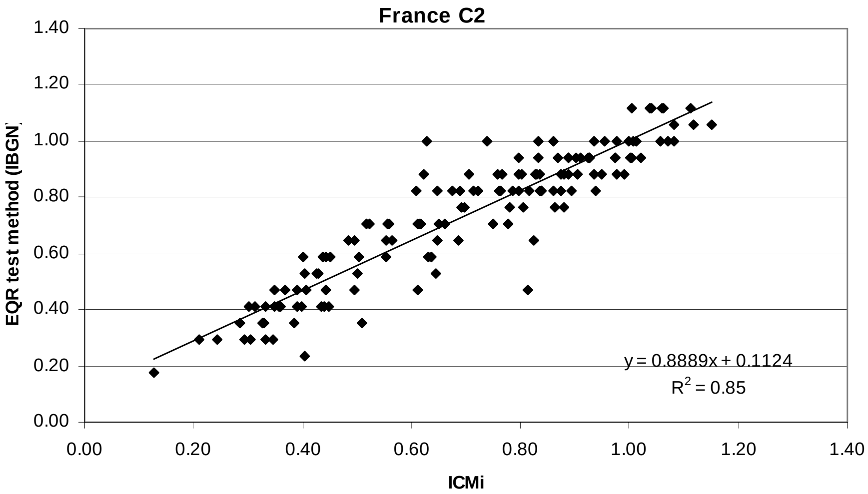


Figure 1: linear regression between ICMs and IBGN in dataset FranceC2
Figure1A:ICMi - R^2 = 0.85; p<0.001

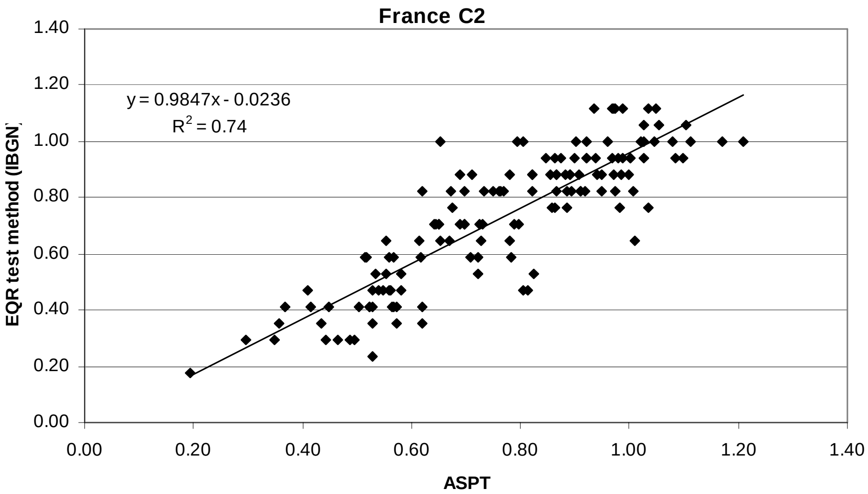


Figure 1B: ASPT index - R^2 = 0.74; p<0.001

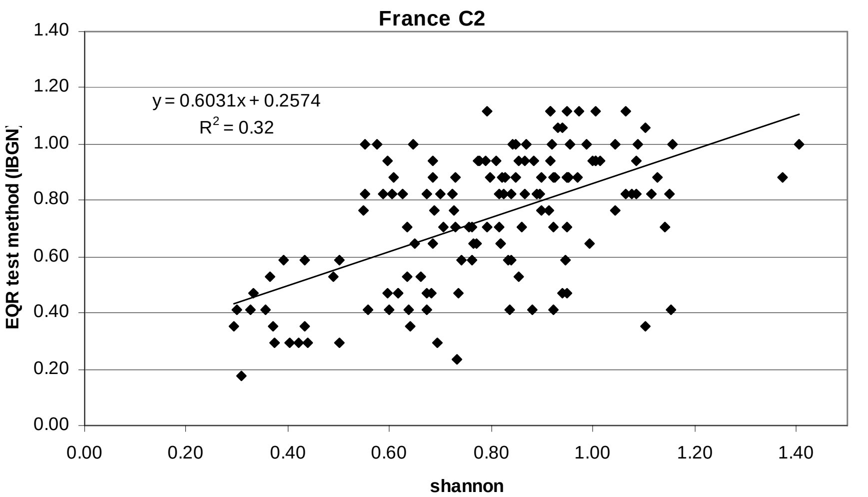


Figure 1C: Shannon index - R^2 = 0.32; p<0.001

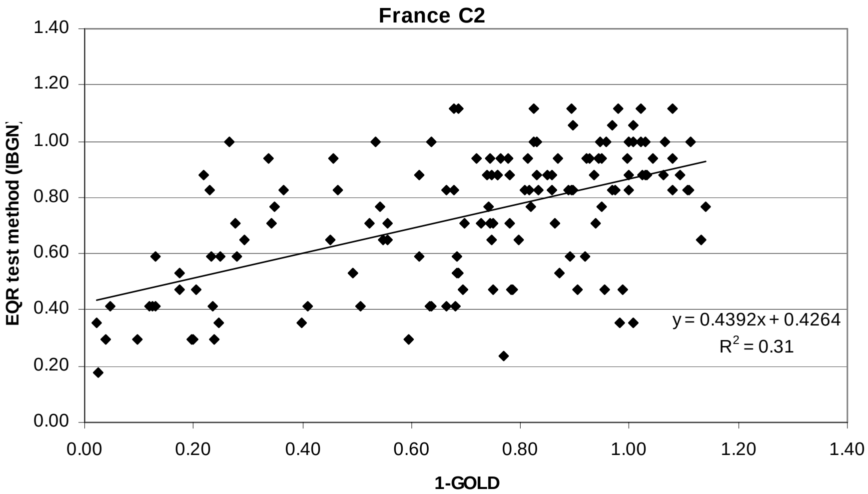


Figure 1D: 1-GOLD - $R^2 = 0.31$; p<0.001

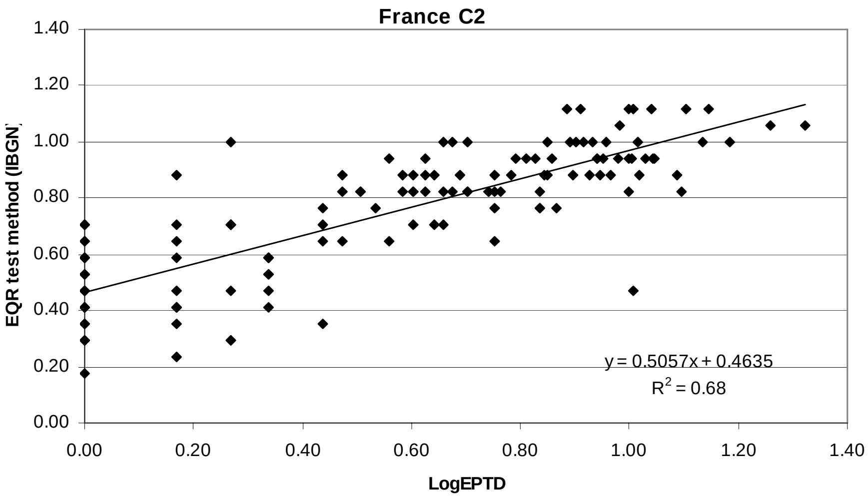


Figure 1E: Log EPTD - $R^2 = 0.68$; p<0.001

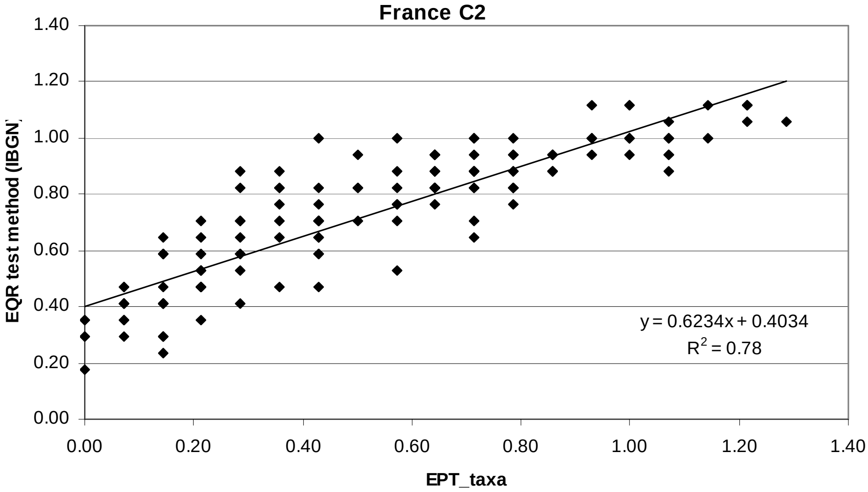


Figure 1F: EPT taxa - R^2 = 0.78; p<0.001

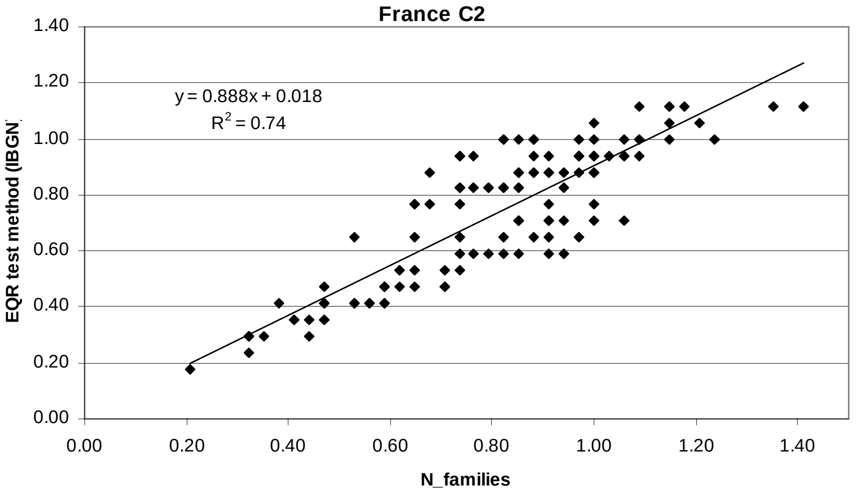


Figure 1G: Number of families - R^2 = 0.74; p<0.001

General remarks

Remarks on the original calculation: normalization performed according to the median value of the 24 reference sites indicted by the data provider (JGW). For the metric ASPT the minimum considered value was 0 and not 2. The values of the metric 1-GOLD vary from 0 to circa 90.

Data was recalculated in accord to all other datasets, thus: normalization performed according to 75° percentile observed in national method High status sites (both for nat meth and ICMs), minimum ASPT value: 2, metric 1-GOLD varying from 0 to 1.

Notes on dataset description

The content of the present description was verified by Dr. Jean Gabriel Wasson from Lyon CEMAGREF who provided the data.

References related to the presented dataset

Vernaux, J. P., P. Galmiche, F. Janier & A. Monnot, 1982. Une nouvelle methode pratique d'evaluation de la qualité des eaux courantes: un indice biologique de qualité générale (I.B.G.). Annales Scientifiques de l'Université de Franche-Comté Besançon, Biologie animale 3: 11–21.

4.3.10 Spain C2

General features

The size class for all the sites is 10-100 km^2 and altitude is lower than 200m. The samples have been collected in coastal river systems throughout the North and Northwest of Spain. From Navarra (West of Pyrenees, 1 sample), to Asturias (2 samples), towards the West of Spain in Galicia (3 samples in Lugo, 10 in Coruña, and 30 in Pontevedra provinces). The samples maximum dispersion is approximately 800 km.
The RC-2 type in Spain is characterised by stony siliceous substrates, mostly dominated by granite blocks and stones, with gravel underneath. They have a high frequency of riffles alternating with small pool areas. Riparian corridors are composed by alder, ash and oak trees accompanied by ferns.

Aim of collection, number of samples

Data of this dataset were collected and provided by Dr. Isabel Pardo from University of Vigo, various are the owners: University of Vigo, the Water Authorities Aguas de Galicia and Confederacion Hidrográfica del Norte. The sites are part of a research and monitoring program.
In this dataset, 25 sites are included. The periods of collection are: 1997 in winter, spring, summer and autumn season; 2002 and 2003 in summer. Total number of samples is 46.

Degradation factor

Sites are mainly affected by two kind of alterations: organic pollution and increase in concentration of nutrients.
According to the test classification method, the quality gradient covers all the quality classes, from 'high' to 'bad' status. The support data available comprise physical and chemicals data, hydromorfological informations and diatoms' community samples, the latter not available for data from 1997.

National method: sampling and sorting

Two sampling method has been performed. For sample collected in 1997 (28 samples), a 3 minutes kick sampling in proportion to habitats present has been carried out. In all the other samples, the sampling technique has been a 20 replicates multiple habitat approach (Barbour, 1999). In the later case the sampling surface is 2.5 m^2. The two different groups are considered as the same dataset, since sampling and classification method provided statistically comparable results in these small streams (Pardo, 2003).

National method: criteria for abundance registration

The number of specimens is recorded as real abundance.

National method: sites' classification

The final index results from a sum of 9 metrics (Spanish MMI). A multiple regression analysis has been performed in order to select metric combination. The resulting metrics significantly predict a specific pressure gradient. In this database, values of the index vary from 2.03 to 6.42 (in EQRs 1.07 and 0.19); boundaries between quality classes in EQRs are: high-good, 0.972; good-moderate, 0.729; moderate-poor, 0.486; poor-bad, 0.243.

Notes on classification

National: the samples classified as 'high status' according to Spanish MMI method are 7 on 46 total samples.
BAC or pressures based: considering pressures data the sites classified as reference are 3.

Comparison between the ICMi and EQRs, single ICM and EQRs
For the calculation of the ICMi, metrics was normalized according to 75° percentile of the high status samples according to MMI classification. (see explanations in previous chapters). The minimum and maximum observed values for ICMi (in EQR) are 0.01 and 1. Between ICMi and Spanish MMI, a regression coefficient of 0.91 is observed (see figure 1A). Results on linear regression between single ICM and Spanish MMI, also normalized according to 75° percentile, are shown in figure 1 B-G.
The conversion of the class boundary values for the MMI Spanish from the original boundaries to ICMi values is done according to Table 1.

	MMI score	MMI 75° EQR	ICM index
Limit high-good	0.97	0.933	0.915
Limit good-moderate	0.73	0.702	0.624
Limit moderate-poor	0.49	0.471	0.334
Limit poor-bad	0.24	0.231	0.032
ICM index = MMI 75° EQR * 1.2585 - 0.2589			
R^2=0.91; p<0.001			

Table 1: Spanish MMI class boundaries conversion for C2 dataset

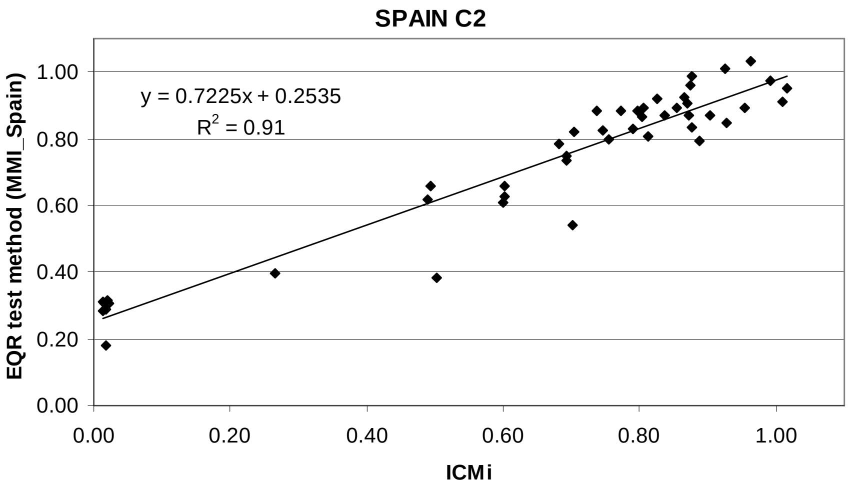


Figure 1: linear regression between ICMs and Spanish MMI in dataset Spain C2
Figure1A:ICMi - R^2 = 0.91; p<0.001

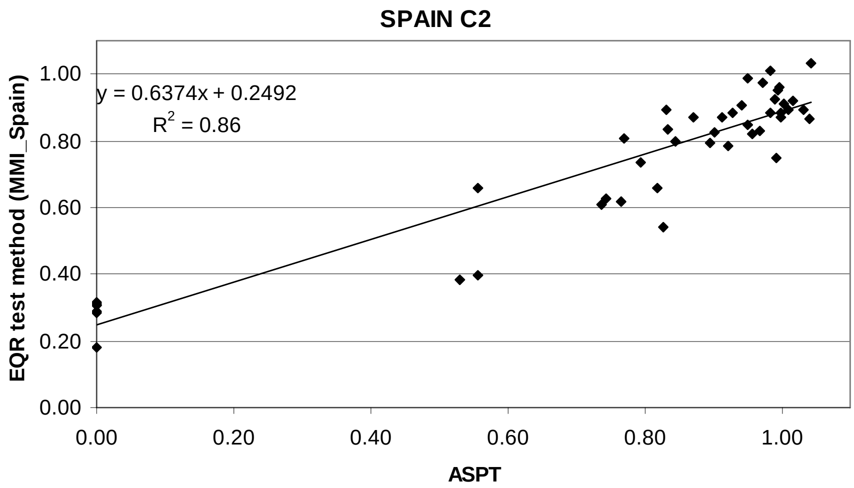


Figure 1B: ASPT index - R^2 = 0.86; p<0.001

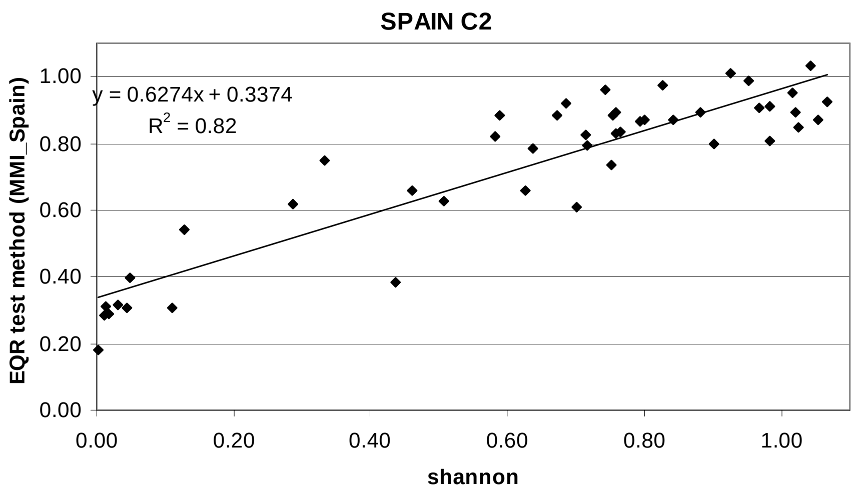


Figure 1C: Shannon index - R^2 = 0.82; p<0.001

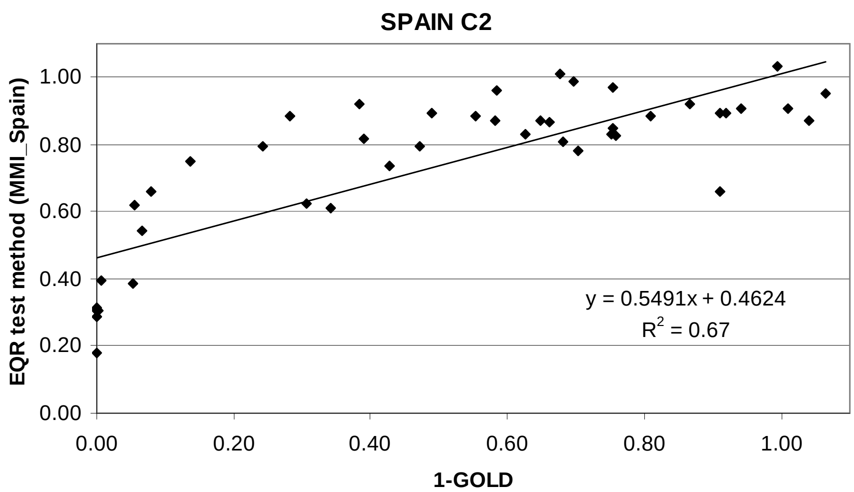


Figure 1D: 1-GOLD - R^2 = 0.67; p<0.001

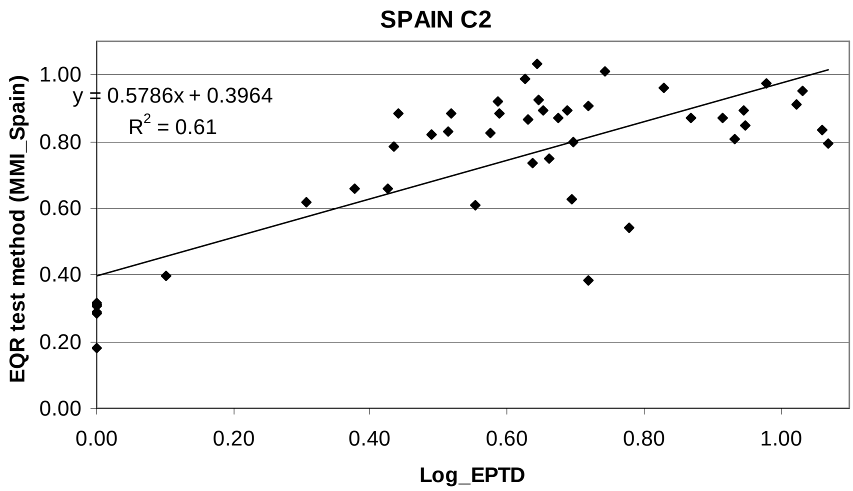


Figure 1E: Log EPTD - $R^2 = 0.61$; $p<0.001$

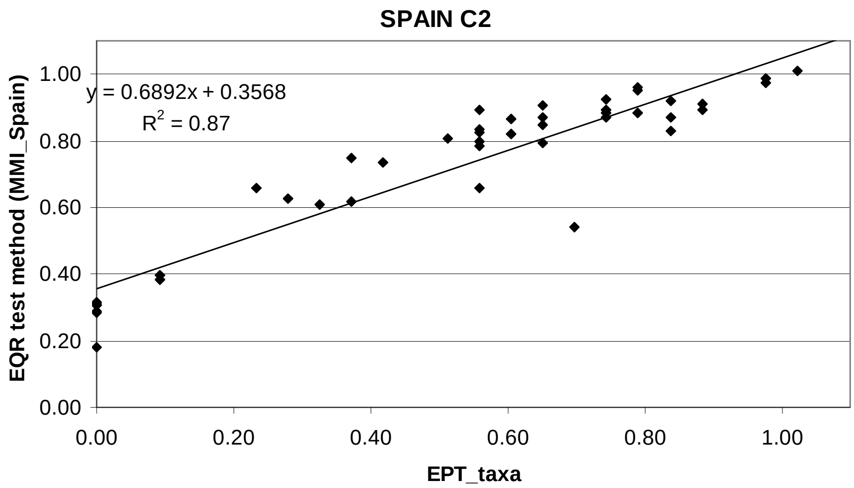


Figure 1F: EPT taxa - $R^2 = 0.87$; $p<0.001$

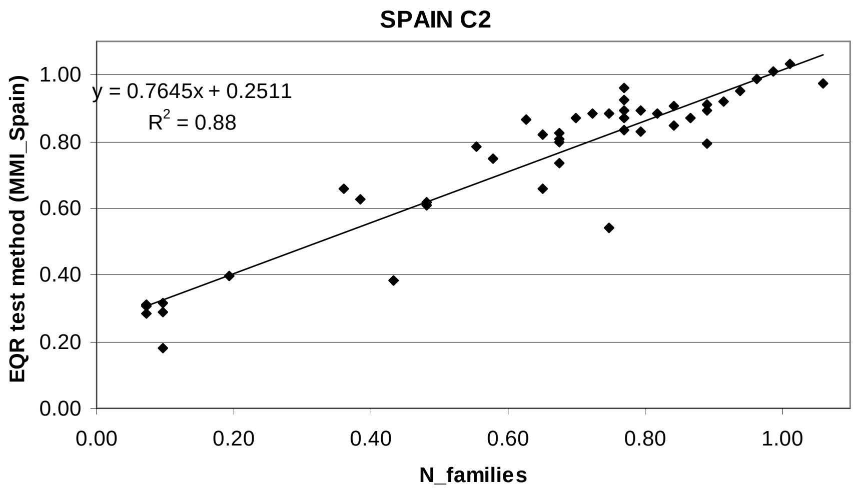


Figure 1G: Number of families - $R^2 = 0.88$; $p<0.001$

Notes on dataset description

The content of the present description was verified by Dr. Isabel Pardo from University of Vigo, who collected and provided the data.

References related to the presented dataset

Barbour, M. T., J. Gerritsen, B. D. Snyder, & J. B. Stribling, 1999. Rapid Bioassessment Protocols for Use in streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency, Office of Water, Washington D.C.

Pardo, I. 2003. Absolute reference conditions for evaluating ecological status of Galician streams and rivers (NW Spain) applying the EU Water Framework Directive. Bull. NABS 20(1): 249.

Pardo, I., N. Felpeto, E. S. Lopez, Fernández, C. & C. Cillero. 1998. Estudio de caracterización ambiental del río Louro (Pontevedra). Saneamiento general de la cuenca del río Louro. Informe Técnico. Dirección General de Obras Públicas y calidad de las Aguas. Ministerio de Obras públicas, Transportes y medio Ambiente. Technical Report.

4.3.11 France M1
test dataset n. FTM101

General features
Sites belong to the hydro-ecoregion "Méditerrannée" (HER 6) of the French typology. Hydrologic seasonality is high, but the streams are not regularly intermittent. Altitude ranges from 0 to 600m, comparable in term of climatic conditions with the range 200-800 m of more southern Mediterranean countries (Spain, Portugal, Italy). Catchment area is small and comprised between 10 and 100 km^2.
Correspond to the small streams (Strahler order 1 to 3). The geology is mixed, with predominance of sedimentary formations. The mean of daily maximum temperature in July is about 29°C. High seasonality, and violent storm events (10 years daily rainfall > 110mm).

Aim of collection, number of samples
Data collection was performed by the Direction Régionale de l'Environment. The database were organized by Lyon Cemagref and has been provided by Dr. Jean Gabriel Wasson. The sites are included in the national monitoring network and regularly investigated for quality assessment.
The total number of sites included is 32. Samples correspond to the years 1992 – 2001; they are representative of the whole hydrologic cycle, with an equal number of samples in late winter and spring (February to June), and in summer and early fall (July to November). Total number of samples is 77.

Degradation factor
General degradation is the main factor of alteration. The dataset covers all the range of ecological status, from 'high' to 'bad' status according to the national method. Data from CORINE Land Cover are available for all the sites. On the basis of land use, pressures of the sites could be further evaluated.

National method: sampling and sorting
The method of classification is the official French monitoring method IBGN (Indice Biologique Global Noramalisé, Vernaux *et al.*, 1982). Sampling is carried out taking a number of 8 samples with a Surber sampler (base surface 1/20 m^2). These samples are characterized by different fixed couple of substrate dimensions and flow velocity. The total sampling area is 0.4 m^2. To be considered as valid, a single taxon has to be present with a minimum number of 3 specimens (or 10 specimens for few taxa). The identification is undertaken at family level.
National method: criteria for abundance registration
IBGN method is semiquantitative. To be considered as valid, a single taxon has to be present with a minimum number of 3 specimens (or 10 specimens for few taxa). Nevertheless, in the present dataset the number of specimens is recorded as real abundance.
All taxa are considered since the first individual, but indicator taxa require a minimum number of individuals (3 or 10) to be taken into account.

National method: sites' classification
For the final classification, two metrics are considered: the Faunistic Indicator Group (GFI) whose values range from 1 to 9 and the number of collected families (taxonomic variety, VT) divided into 14 classes. The final IBGN value is obtained by the sum of these two metrics. Values of the index can vary from 0 to 20. The transformation in EQR is done according to type. For the small Mediterranean streams here presented, the IBGN class boundaries are the

following reference, 17; high-good, 15; good-moderate, 13. In this database, the minimum and maximum observed values for IBGN are 2 and 19.

Notes on classification
National: For this pilot exercise, the calculation of the Reference Values of the ICMs is carried out using the sites with High and Reference Ecological Status according to the IBGN value. These samples are 28 on 77 total (36% circa).

Reference sites are selected on the basis of very low anthropic pressures, independently of the biological values in a first approach. The distribution of biological data is then calculated for all samples of the reference dataset, and the outliers samples are checked. Dubious sites are eliminated, but low biological values are accepted if they come from validated reference sites. The procedure combine both spatial and temporal variability of a given stream type. The Reference Conditions (RC) are defined as the range of variability of a given biological element (index or metric) observed at reference sites. However, the calculation of EQR needs to define a Reference Value (RV) for the normalization of the samples. Due to the small number of reference sites generally observed for most types, the most robust and stable statistic is chosen as RV. For all our calculations, following the recommendation of the REFCOND guidance, the median was used as Reference Value.
The general approach and Reference Values for each type are described in a work paper (Wasson et al., October 2003, in French) and a summary (in English) will be available soon.
Reference sites were first selected from the monitoring network and other complementary sites in using two independent methods :

- "expert selection" by the field hydrobiologists (DIREN teams), on the basis of a detailed questionnaire combining all the possible pressures at the basin, reach and site scale.
- "GIS selection" run by Cemagref on the basis of known point source pollution discharges (from water agencies), and land use (CORINE), at the scale of hydrologic units (sub-basins ca. 100 km^2). However, this selection eliminates impacted basins where reference sites could be found upstream of pollution discharge (Wasson et al., August 2004, in French)

The IBGN values observed in these two selections of sites were compared to the values calculated from reference sites selected and sampled by the ***Cemagref*** hydrobiologists. The reference value for a given stream type was accepted only if the IBGN values observed in the three datasets were in good concordance. If not, a checking procedure was run and dubious sites were eliminated.
Since December 2004, the boundaries of the IBGN classes are redefined according to this definition of reference samples. In particular the boundary High good is set at the 25° percentiles of the reference samples.

Comparison between the ICMi and IBGN EQRs, single ICM and IBGN EQRs
For the calculation of the ICMi, metrics was normalized according to 75° percentile of High status samples and Reference status samples (see explanations in previous chapters). This normalization option is suitable only for this IC exercise purpose. It will not be used in France for WFD implementation.
Final ICMi is re-normalized according to 75° percentile value. The minimum and maximum observed values for ICMi (in EQR) are 0.19 and 1.09. Between ICMi and IBGN, a regression coefficient of 0.86 was found (see figure 1A).
Results on linear regression between single ICM and IBGN are shown in figure 1 B-G.
The scores of IBGN in the graphs are expressed in EQR values, calculated dividing the IBGN score for each sample by the 75° observed in the high status samples.

The conversion of the class boundary values for the IBGN method from the original boundaries to ICMi values is done according to Table 1.

	IBGN score	IBGN EQR	ICM index
Limit reference-high	17	0.986	0.977
Limit high-good	15	0.870	0.865
Limit good-moderate	13	0.754	0.754
Limit moderate-poor	nd	nd	nd
Limit poor-bad	nd	nd	nd
ICM index = IBGN EQR * 0.9614 + 0.0292 R^2=0.86; p<0.001			

Table 1: IBGN class boundaries conversion for M1 dataset

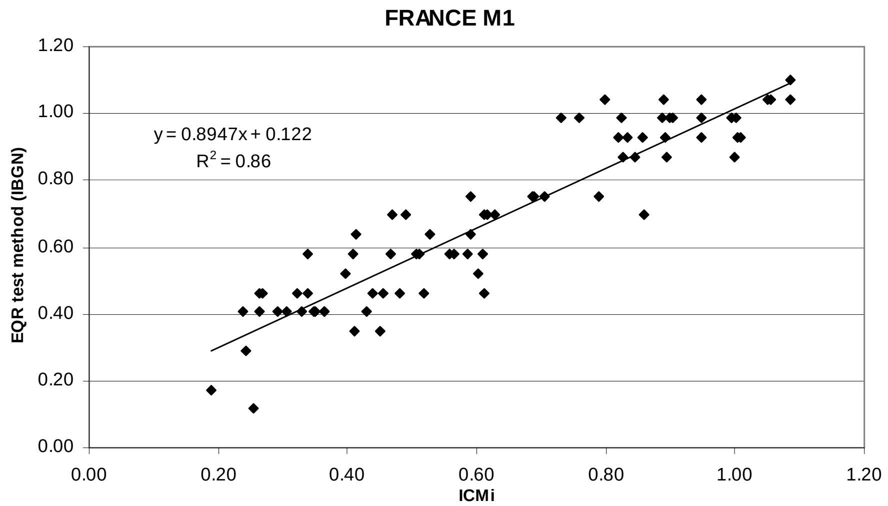


Figure 1: linear regression between ICMs and IBGN in dataset France C1
Figure1A:ICMi - R^2 = 0.86; p<0.001

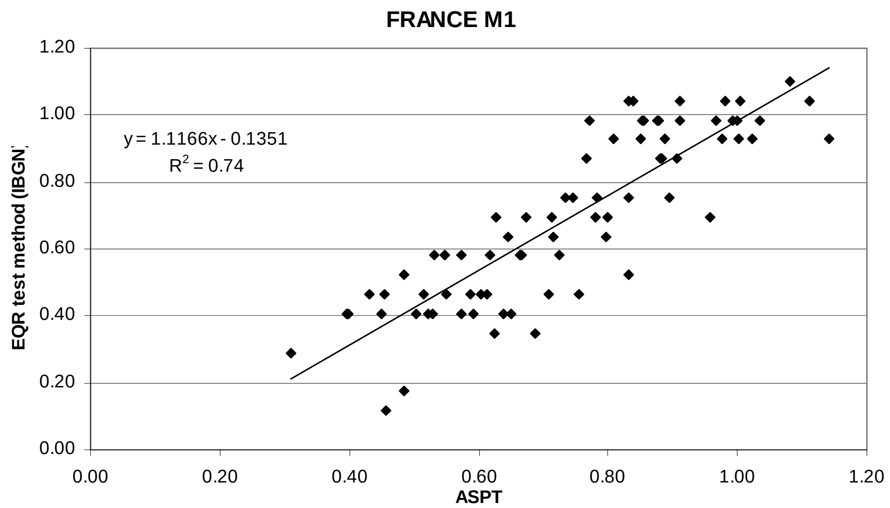


Figure 1B: ASPT index - R^2 = 0.74; p<0.001

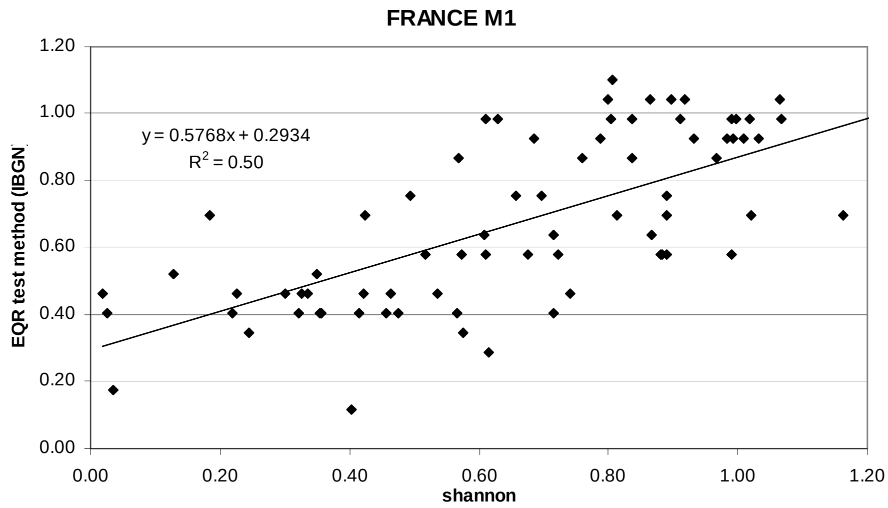


Figure 1C: Shannon index - R^2 = 0.50; p<0.001

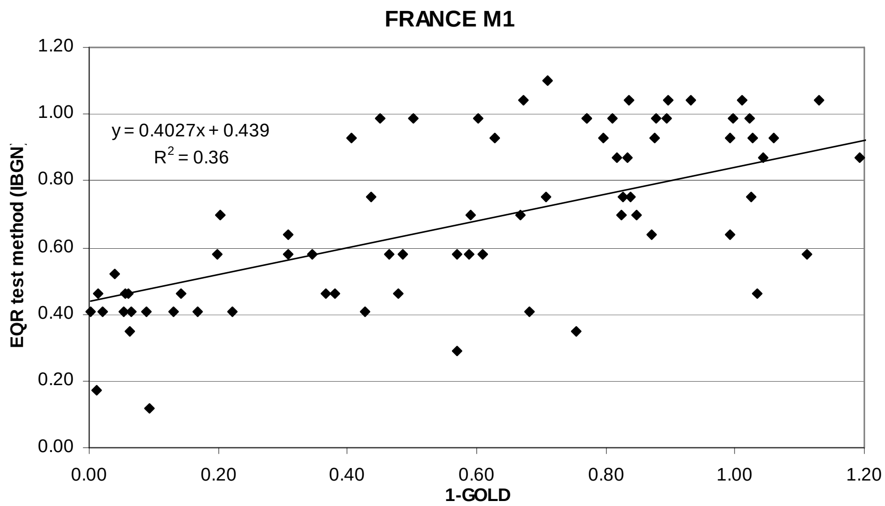


Figure 1D: 1-GOLD - R^2 = 0.36; p<0.001

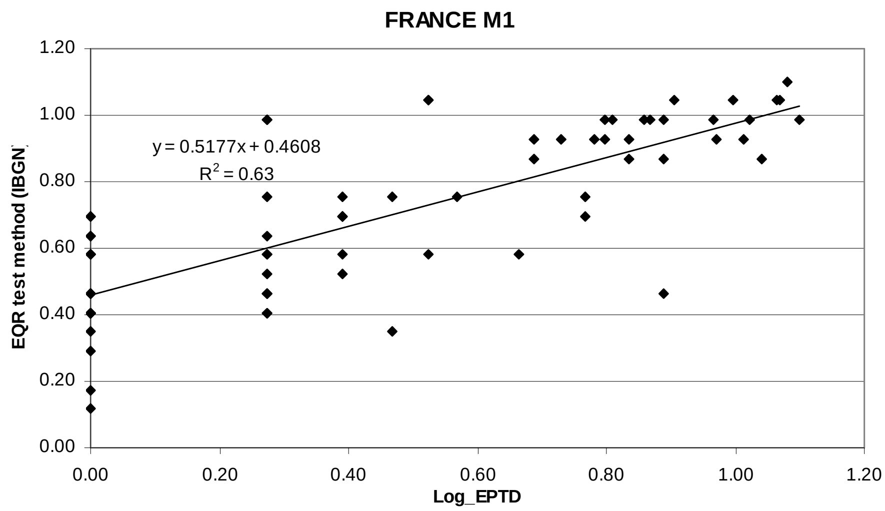


Figure 1E: Log EPTD - R^2 = 0.63; p<0.001

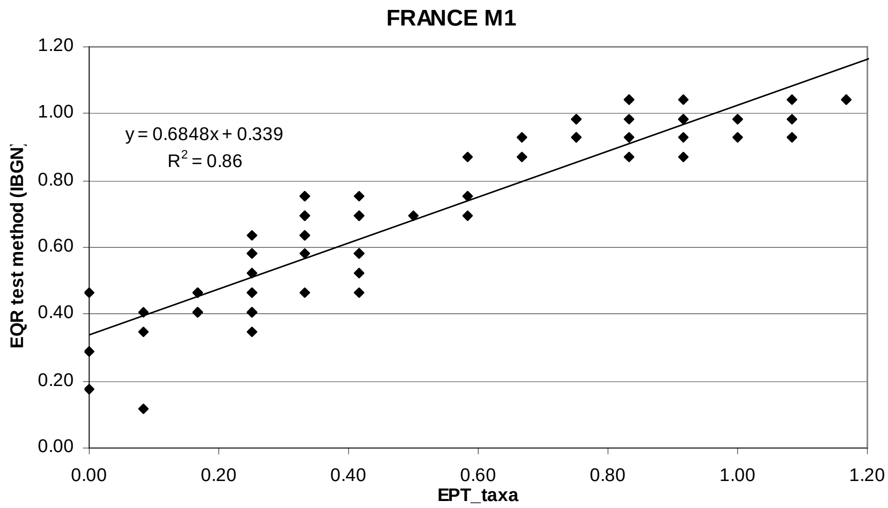


Figure 1F: EPT taxa - R^2 = 0.86; p<0.001

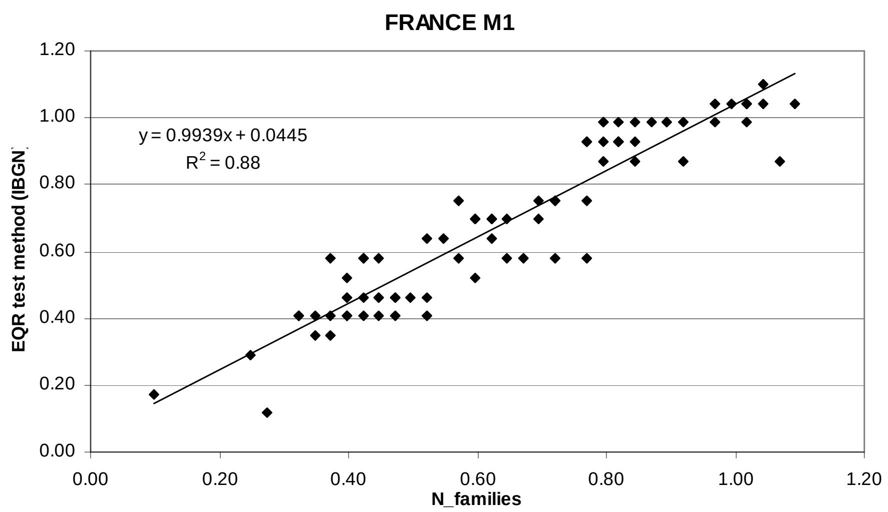


Figure 1G: Number of families - R^2 = 0.88; p<0.001

General remarks
The characteristics of this dataset are achieved from Wasson, 2004, a work paper provided for the First Mediterranean GIG Intercalibration meeting, Evora May 2004.

Notes on dataset description
The content of the present description is verified by Dr. Jean Gabriel Wasson from Lyon CEMAGREF who provided the data.

References related to the presented dataset

Vernaux, J. P., P. Galmiche, F. Janier & A. Monnot, 1982. Une nouvelle methode pratique d’evaluation de la qualité des eaux courantes: un indice biologique de qualité générale (I.B.G.). Annales Scientifiques de l’Université de Franche-Comté Besançon, Biologie animale 3: 11–21.

Wasson, J. G., 2004. Comparison of the French IBGN index with Intercalibration Common Metrics. First Mediterranean GIG Intercalibration meeting, Evora 19-21May 2004. Work paper.

4.3.12 Italy M1

test dataset n. ITM101

General features
The sites are located in Southern Apennines (region Campania, see AQEM Consortium, 2002; Buffagni *et al.*, 2004; Balestrini *et al.*, 2004, for further description) and in Tuscany (Central Italy). Even if streams are not intermittent, high seasonal variations of flow regime can be observed. Sites are small-sized (catchment area lower than 100km^2 except for two site), and have an altitude range of 200–800 m. The two areas are about 400 km distant, in each area maximum distance among sites is about 50 km.

Aim of collection, number of samples
Data were collected and provided by Dr. Andrea Buffagni from CNR-IRSA. This institute is the data owner.
Sites in Tuscany have been investigated for the Project EU-STAR. The aim is to provide a standardization of ecological quality classification in streams and rivers all over Europe (Furse, 2001; Hering & Strackbein, 2002). Sites in Campania have been investigated to test first application of the assessment method developed for South Apennine Italian stream type during the AQEM Project and to provide a comparison with the national method IBE (Indice Biotico Esteso, APAT-IRSA/CNR, 2003).
11 sites have been investigate in Tuscany for three seasons: summer 2002, winter 2003 and spring 2003. For 6 out of 11 sites, two replicates of the sampling method were undertaken.
In Southern Apennines, 11 sites have been investigate once in one year: autumn 2003. For 1 site, two replicates of the sampling method were undertaken. Total number of samples is 63.

Degradation factor
Stressor observed is mainly organic pollution often associated with degradation of stream morphology (Buffagni *et al.*, 2001). Other kinds of water pollution can be present (such as impact from farming activities, trace metals and presence of livestock). According to the national classification method performed, the quality gradient covers all the quality classes according to from ‘high’ to ‘bad’ status. Additional data available are for all samples main physical, chemical and microbiological variables. Data from the following environmental indices are also available: Habitat Modification Score, Habitat Quality Assessment (HMS and HQA, Raven *et al.*, 1998, Buffagni & Kemp, 2001), Index of Fluvial Functioning (IFF, Siligardi *et al.*, 2000, Balestrini *et al.*, 2004)

National method: sampling and sorting
The classification method used is the official national assessment method IBE (Indice Biotico Esteso, APAT-IRSA/CNR, 2003). According to this method, the sampling is performed along a transect between the two banks of the river in a riffle area and the number of replicates varies according to water width and general habitat diversification. The total area sampled is thus not fixed; in this databases it has been considered approximately 0.9m^2. The sorting is semiquantitative (a minimum number of specimens for each taxon has to be considered). The identification is undertaken at genus and family level.

National method: criteria for abundance registration
The sorting is semiquantitative, and usually an indication of the relative abundance of the collected taxa by means of codified symbols is given. In the present dataset, an estimation of the absolute abundance is provided.

National method: sites' classification

The final index score is obtained via a two-entry table, by comparison of two metrics: the total number of taxa collected and the Faunistic Group (ordered by an increasing scale of tolerance). Values of the index can vary from 0 to 14;

In this database, the minimum and maximum observed values are 2 and 11.6.

Notes on classification

National: About 33% (21 on 63) of the samples are classified as 'high status' according to national assessment method.

BAC or pressures based: a BAC is available. Reference sites are defined according to ecological breakpoints along the multivariate axis that explains the main degradation factor. Remaining classes equally spaced. 12 reference sites are present according to BAC.

Comparison between the ICMi and IBE EQRs, single ICM and IBE EQRs

For the calculation of the ICMi, metrics was normalized according to 75° percentile observed in the 'high status' samples (see explanations in previous chapters). The minimum and maximum observed values for ICMi (in EQR) have been 0.17 and 1.09. Between ICMi and IBE, a regression coefficient of 0.72 was found (see figure 1A). The conversion of the class boundary values for the IBE method from the original boundaries to ICMi values is done according to Table 1.

Results on linear regression between single ICM and IBE are shown in figure 1 B-G.

	IBE score	IBE EQR	ICM index
Limit high-good	9.6	0.881	0.901
Limit good-moderate	7.6	0.697	0.722
Limit moderate-poor	5.6	0.514	0.543
Limit poor-bad	3.6	0.330	0.364
ICM index = IBE EQR * 0.9756 + 0.0419 R^2=0.75; p<0.001			

Table 1: IBE class boundaries conversion

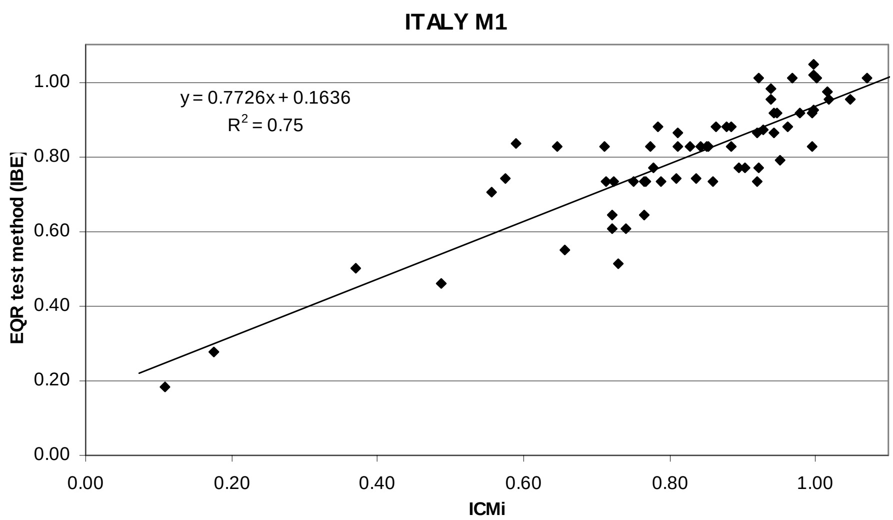


Figure 1: linear regression between ICMs and IBE in dataset ITM101

Figure 1A: ICMi - $R^2 = 0.75$; $p<0.001$

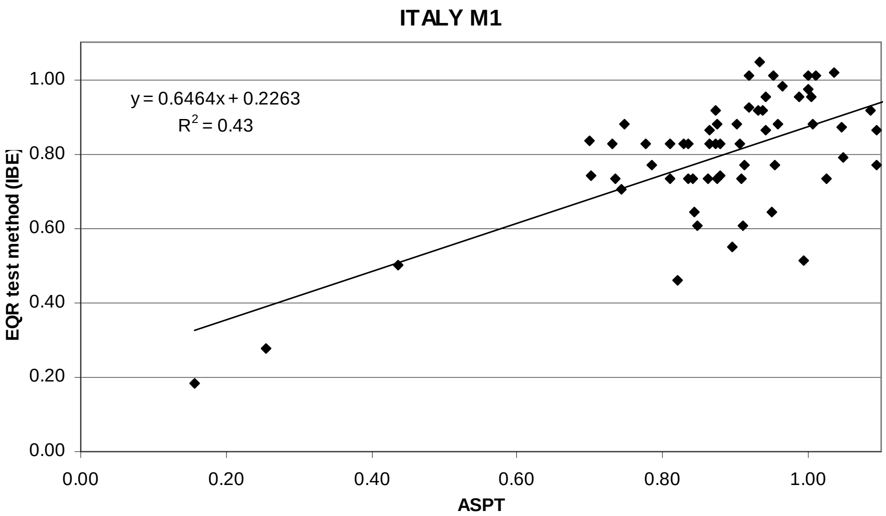


Figure 1B: ASPT index - $R^2 = 0.43$; $p<0.001$

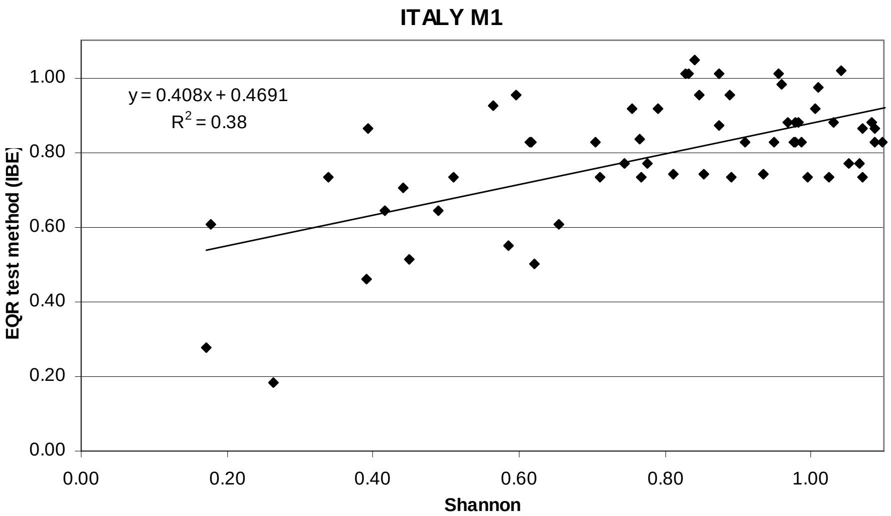


Figure 1C: Shannon index - $R^2 = 0.38$; $p<0.001$

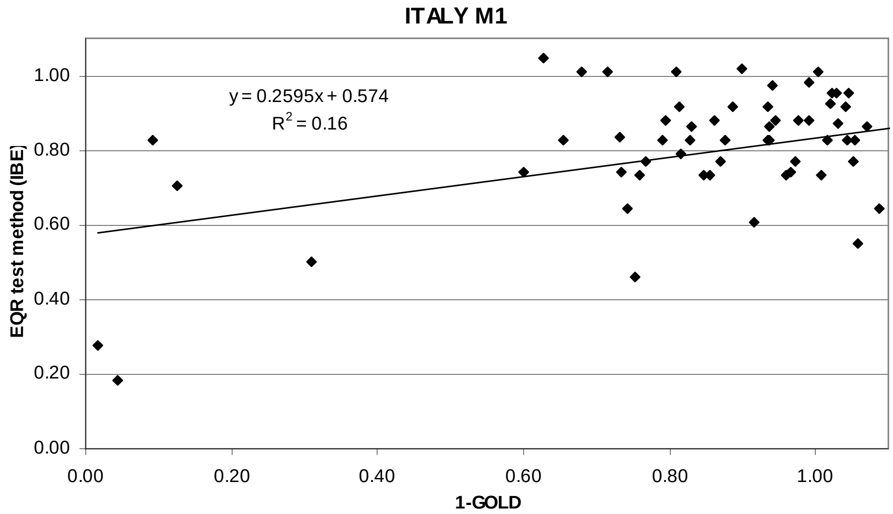


Figure 1D: 1-GOLD - R^2 = 0.16; p=0.001

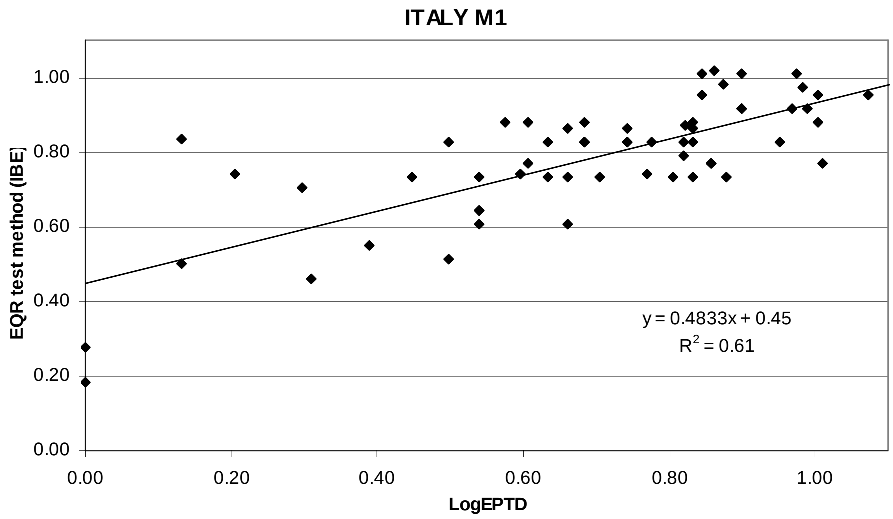


Figure 1E: Log EPTD - R^2 = 0.61; $p<0.001$

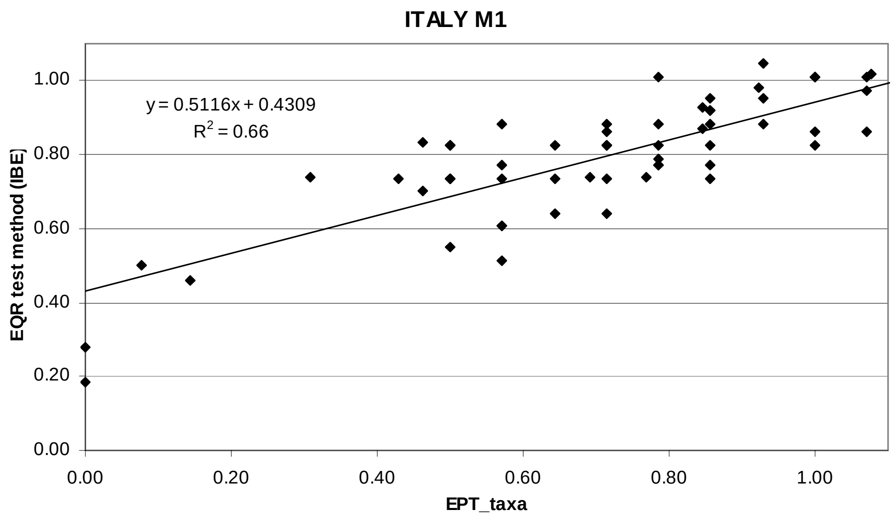


Figure 1F: EPT taxa - $R^2 = 0.66$; $p<0.001$

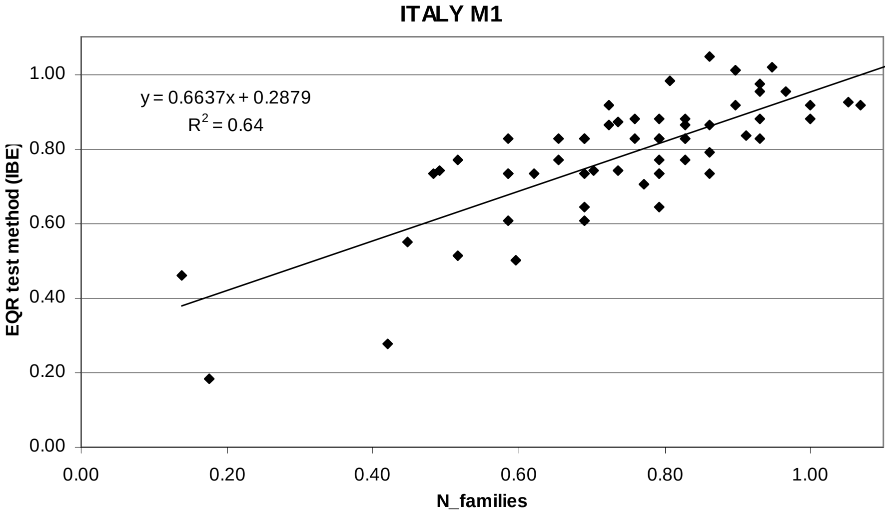


Figure 1G: Number of families - $R^2 = 0.64$; $p<0.001$

General remarks

In about 20 years, the calculation of IBE index encountered several updates (Ghetti, 1986; 1995; 1997; APAT-IRSA/CNR, 2004), especially in relation to the minimum number of specimens to be considered.

In this dataset, the IBE values of all the samples have been calculated according to the most updated version of the index, i.e.: APAT-IRSA/CNR, 2004.

References

APAT-IRSA/CNR, 2004. Indice Biotico Esteso (I.B.E). In: APAT, Manuali e linee guida 29/2003. APAT-IRSA/CNR, Metodi analitici per il controllo della qualità delle acque 3: 1115-1136.

AQEM Consortium, 2002. Manual for the application of the AQEM system. Version 1.0 February 2002.

Balestrini, R., M. Cazzola & A. Buffagni, 2004. Characterizing hydromorphological features of selected Italian rivers: a comparative application of environmental indices. Hydrobiologia 516: 367–381.

Buffagni, A., J. L. Kemp, S. Erba, C. Belfiore, D. Hering & O.Moog, 2001. A Europe wide system for assessing the quality of rivers using macroinvertebrates: the AQEM project and its importance for southern Europe (with special emphasis in Italy). J. Limnol. 60 (suppl. 1): 39–48.

Buffagni, A., S. Erba, M. Cazzola & J. L. Kemp, 2004. The AQEM multimetric system for the southern Italian Apennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. Hydrobiologia 516: 313–329.

Buffagni, A. & J. L. Kemp, 2002. Looking beyond the shores of the United Kingdom: *addenda* for the application of River Habitat Survey in South European rivers. J. Limnol. 61 (2): 199–214.

Furse, M., 2001. “Standardisation of River Classifications: Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive" STAR Project, Part B - EU Contract number: EVK1-CT 2001-00089.

Ghetti, P. F., 1986. I macroinvertebrati nell'analisi di qualità dei corsi d'acqua. Provincia Autonoma di Trento, Trento: 105 pp.

Ghetti, P. F., 1995. Indice Biotico Esteso (I.B.E.). Notiziario dei Metodi Analitici. IRSA - CNR, 7 luglio 1995, Roma: 1-24.

Ghetti, P. F., 1997. Indice Biotico Esteso (I.B.E.). I macroinvertebrati nel controllo della qualità degli ambienti di acque correnti. Provincia Autonoma di Trento: 222 pp.

Hering, D. & J. Strackbein, 2002. Standardisation of river classifications: Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive. 1st deliverable - STAR stream types and sampling sites. 30/06/02.

Raven, P. J., T. H. Holmes, F. H. Dawson, P. J. A. Fox, M. Everard, I. R. Fozzard & K. J. Rouen, 1998, River Habitat Survey, the physical character of rivers and streams in the UK and Isle of man. River Habitat Survey No. 2, May 1998. The Environment Agency, Bristol, 86 pp.

Siligardi, M., S. Bernabei, C. Cappelletti, E. Chierici, F. Ciutti, F. Egaddi, A. Franceschini, B. Maiolini, L. Mancini, M. R. Minciardi, C. Monauni, G. L. Rossi, G. Sansoni, R. Spaggiari & M. Zanetti, 2000. I.F.F. Indice di funzionalità fluviale. Manuale ANPA/2000: 223 pp.

4.3.13 Italy M5

test dataset n. ITM501

General features

The sites are located in three areas of the region Sardinia (Buffagni *et al.*, 2004). Maximum distance between two sites is about 300km. In all the streams high seasonal variations of flow regime can be observed; moreover, for most of them, large part of the channel can run dry during summer season. Sites have catchment area lower than $100km^2$ except for two sites, and an altitude range of 100–450 m.

Aim of collection, number of samples

Data were collected and provided by Dr. Andrea Buffagni from CNR-IRSA. This institute is the data owner.

Sites are included in a national research Project named MICARI funded by Italian Ministry of Instruction, University and Research. The aim is the improvement of carrying capacity of streams and, in particular for this area, the development of a quality assessment method for temporary streams.

From 11 to 13 sites have been investigated in three months of 2004: February, June and August. A total number of 37 samples were collected.

Degradation factor

Stressor observed is mainly organic pollution often associated with degradation of stream morphology (Buffagni *et al.*, 2004). According to the national classification method performed, the quality classes range from 'high' to 'moderate' status, with only one site in 'poor' status. Additional data available are for all samples main physical, chemical and microbiological variables. Data from the following environmental indices are also available: Habitat Modification Score, Habitat Quality Assessment (HMS and HQA, Raven *et al.*, 1998, Buffagni & Kemp, 2001), Index of Fluvial Functioning (IFF, Siligardi *et al.*, 2000, Balestrini *et al.*, 2004)

National method: sampling and sorting

The classification method used is the official national assessment method IBE (Indice Biotico Esteso, APAT-IRSA/CNR, 2003). According to this method, the sampling is performed along a transect between the two banks of the river in a riffle area and the number of replicates varies according to water width and general habitat diversification. The total area sampled is thus not fixed; in this databases it has been considered approximately $0.9m^2$. The sorting is semiquantitative (a minimum number of specimens for each taxon has to be considered).

National method: criteria for abundance registration

For method IBE, as the sorting is semiquantitive, no precise indication of the real number of specimens present is given. In this dataset, for the taxa present with less than 10 individuals, a count of the real abundance has been undertaken. For all other taxa, an estimation of the specimens present in the whole sample has been carried out by steps of 10 individuals.

National method: sites' classification

The final index score is obtained via a two-entry table, by comparison of two metrics: the total number of taxa collected and the Faunistic Group (ordered by an increasing scale of tolerance). Values of the index can vary from 0 to 14. In this database, the minimum and maximum observed values are 5 and 10.4.

Notes on classification
National: only one site is classified as 'high status' according to national assessment method.
BAC or pressures based: a PCA analysis has been carried out on biological data, in order to highlight the main variation axes. To explain the meaning of the axes, correlations with environmental variables have been considered. A sites classification along the quality axis was performed and 8 samples was classified as reference. Boundaries between classes were performed basing on the selection of the ecological breakpoint.

Comparison between the ICMi and IBE EQRs, single ICM and IBE EQRs
For the calculation of the ICMi, metrics was normalized according to the value of the only high status site according to IBE. The minimum and maximum observed values for ICMi (in EQR) have been 0.24 and 1. Between ICMi and IBE, a regression coefficient of 0.46 was found (see figure 1A).
Results on linear regression between single ICM and IBE are shown in figure 1 B-G.
The conversion of the class boundary values for the IBE method from the original boundaries to ICMi values is done according to Table 1.

	IBE score	IBE EQR	ICM index
Limit high-good	9.6	0.923	0.914
Limit good-moderate	7.6	0.731	0.717
Limit moderate-poor	5.6	0.538	0.521
Limit poor-bad	3.6	0.346	0.324
ICM index = IBE EQR * 1.0223 - 0.0298			
R^2=0.46; p<0.001			

Table 1: IBE class boundaries conversion

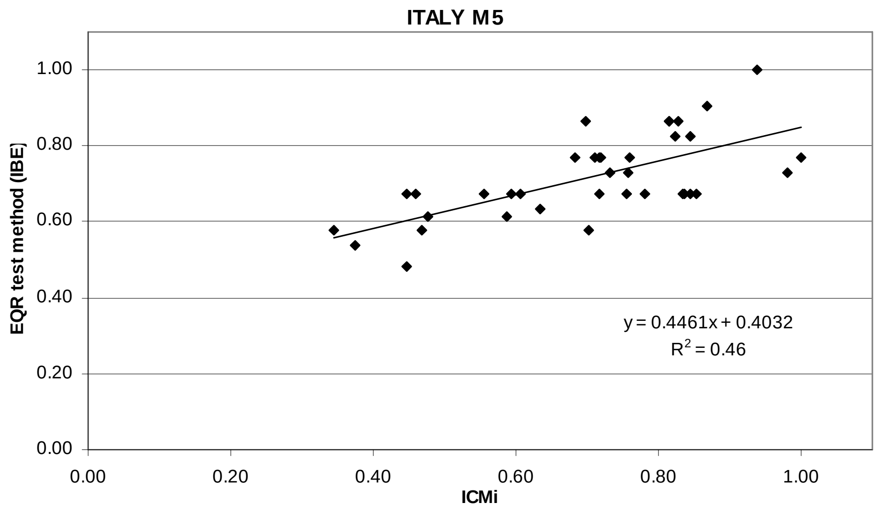


Figure 1: linear regression between ICMs and IBE in dataset ITM101
Figure 1A: ICMi - R^2 = 0.46; p<0.001

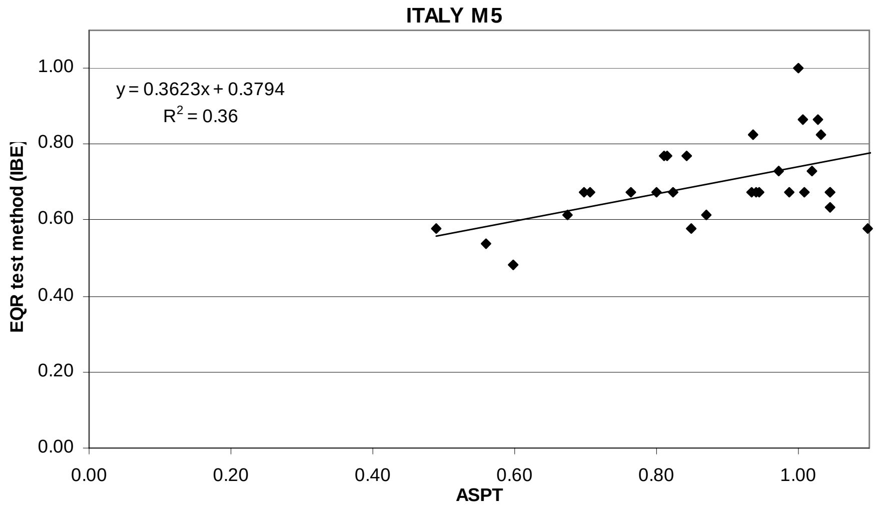


Figure 1B: ASPT index - $R^2 = 0.36; p<0.001$

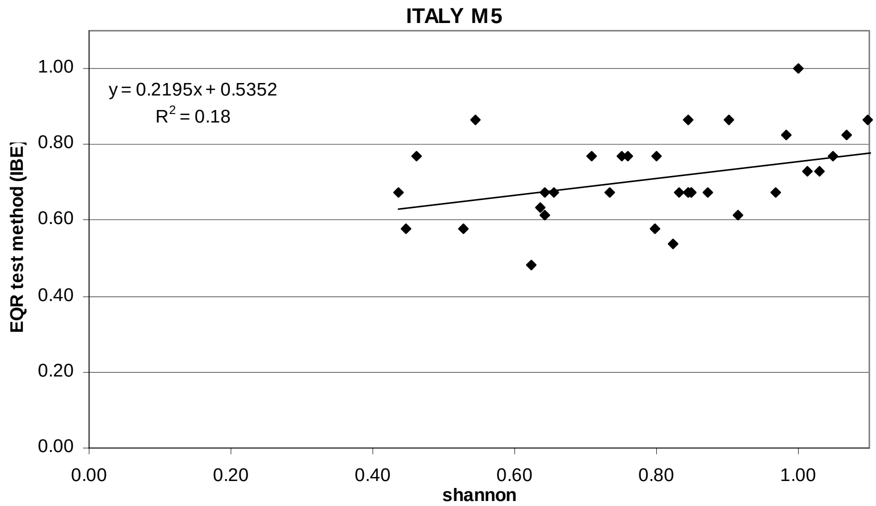


Figure 1C: Shannon index - $R^2 = 0.18; p=0.009$

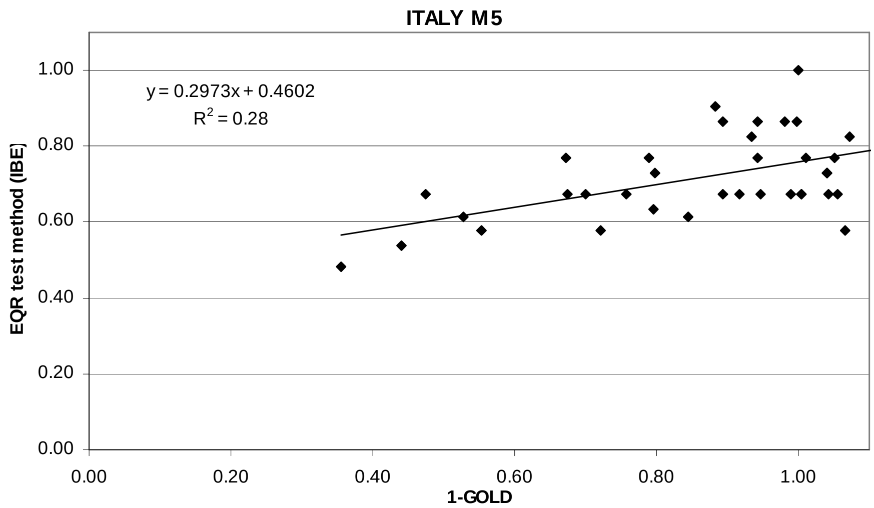


Figure 1D: 1-GOLD - $R^2 = 0.28; p<0.001$

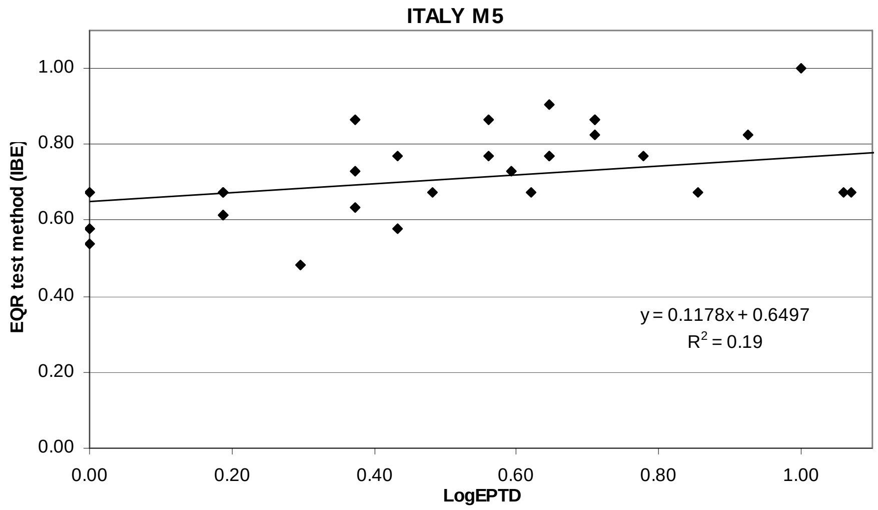


Figure 1E: Log EPTD - $R^2 = 0.19$; p=0.008

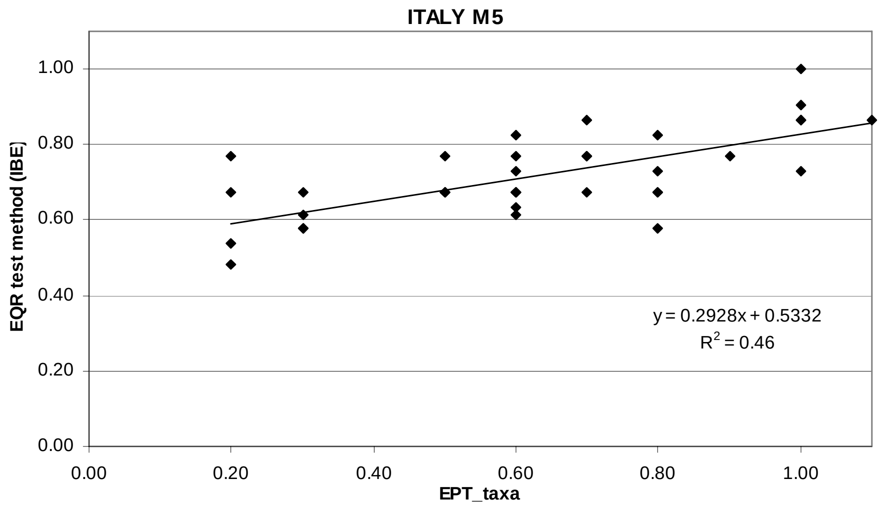


Figure 1F: EPT taxa - $R^2 = 0.46$; p<0.001

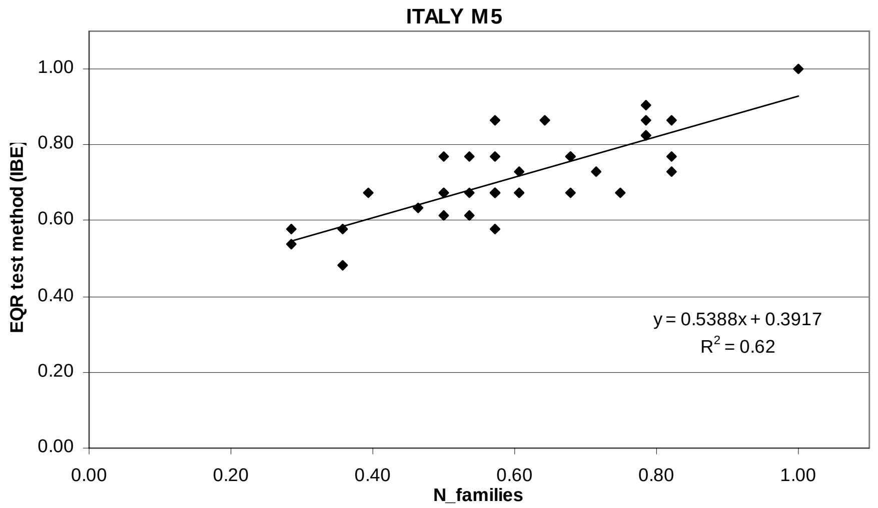


Figure 1G: Number of families - $R^2 = 0.62$; p<0.001

General remarks

In about 20 years, the calculation of IBE index encountered several updates (Ghetti, 1986; 1995; 1997; APAT-IRSA/CNR, 2003), especially in relation to the minimum number of specimens to be considered.

In this dataset, the IBE values of all the samples have been calculated according to the most updated version of the index, i.e.: APAT-IRSA/CNR, 2003.

Low correlations can be due to the weakness of the IBE in describing the quality gradient in temporary rivers (Buffagni *et al.*, 2004).

References

APAT-IRSA/CNR, 2003. Indice Biotico Esteso (I.B.E). In: APAT, Manuali e linee guida 29/2003. APAT-IRSA/CNR, Metodi analitici per il controllo della qualità delle acque 3: 1115-1136.

Balestrini, R., M. Cazzola & A. Buffagni, 2004. Characterizing hydromorphological features of selected Italian rivers: a comparative application of environmental indices. Hydrobiologia 516: 367–381.

Buffagni, A., S. Erba & R. Pagnotta, 2004. Carattere lentico-lotico dei fiumi mediterranei e classificazione biologica di qualità. Quad. Ist. Ric. Acque, 122: 165-189.

Buffagni, A. & J. L. Kemp, 2002. Looking beyond the shores of the United Kingdom: *addenda* for the application of River Habitat Survey in South European rivers. J. Limnol. 61 (2): 199–214.

Ghetti, P. F., 1986. I macroinvertebrati nell'analisi di qualità dei corsi d'acqua. Provincia Autonoma di Trento, Trento: 105 pp.

Ghetti, P. F., 1995. Indice Biotico Esteso (I.B.E.). Notiziario dei Metodi Analitici. IRSA - CNR, 7 luglio 1995, Roma: 1-24.

Ghetti, P. F., 1997. Indice Biotico Esteso (I.B.E.). I macroinvertebrati nel controllo della qualità degli ambienti di acque correnti. Provincia Autonoma di Trento: 222 pp.

Raven, P. J., T. H. Holmes, F. H. Dawson, P. J. A. Fox, M. Everard, I. R. Fozzard & K. J. Rouen, 1998, River Habitat Survey, the physical character of rivers and streams in the UK and Isle of man. River Habitat Survey No. 2, May 1998. The Environment Agency, Bristol, 86 pp.

Siligardi, M., S. Bernabei, C. Cappelletti, E. Chierici, F. Ciutti, F. Egaddi, A. Franceschini, B. Maiolini, L. Mancini, M. R. Minciardi, C. Monauni, G. L. Rossi, G. Sansoni, R. Spaggiari & M. Zanetti, 2000. I.F.F. Indice di funzionalità fluviale. Manuale ANPA/2000: 223 pp.

4.4 Summary tables for testdatasets

In table 4.4.1 a summary of the characteristics of the presented test datasets is reported. Information is taken from the general descriptions presented in the previous pages and from the data provided.

dataset name (country and stream type)	size class km2	altitude class m	data collector	contact person	data owner	aim of the collection	number of sites	number of year/season	number of samples	number of high status samples (according to MS method)	number of high status samples Classification) (according to Best Available
Belgium C1	10-100	<200	Flemish Environment Agency	G. Verhaegen Flemish Environment Agency	Flemish Environment Agency	national monitoring program	70 ca.	3 years (2000-2002)	208	10	nd
Denmark C1	15-100	<200	regional Danish authorities (counties)	J. Skriver - NERI	regional Danish authorities	national monitoring program	135 ca.	6 years, various seasons	346	72	17 high status sites according to expert judjement, inverts community and abiotic data
Estonia C1	<100-<1000	<200	Estonian Agricultural University	H. Timm - EAU	Estonian Agricultural University	national Estonian database	23	1 year, 1 season usually April-May	23	9	nd
France C1	10-300	<100	Direction Regionale de l'Environment	J. G. Wasson - Cemagref-Lyon	database: Cemagref-Lyon	national monitoring network	20	10 years	139	59	24 reference sites according to pressures data
Germany C1	10-100	<200	various regional authorities	S. Birk - UNI Essen	Umweltbundesamt and LAWA	quality monitoring	38	different not specified years, 3 sesons	91	1	nd
Italy C1	10-100	<200	ARPA Lombardia	P. Genoni - ARPA Lombardia	ARPA Lombardia	various ARPA surveys (monitoring, internal IC, method testing)	39	6 years, 4 seasons	361	94	nd
Poland C1	10-100	<200	Institute Environmet Protection, Warsaw	H. Soszka - Institute Environmet Protection, Warsaw	Voivodships Inspectorates of Environmental Protection in Poland, Warsaw University	various, pilot monitoring; scientific project	49	1year, 1season	49	11	nd
UK C1	small	lowland	Environment Agency	J. Murray Bligh - Environment Agency	Environment Agency	Environment Protection	789	1 year, 2 season combined data	789	202	nd
totale	**C1**						**1165**		**1897**	**115**	

Table 4.4.1 Test datasets features

dataset name (country and stream type)	main degradation factor	quality gradient	support data	classification method	min/max possible values	sampling method	sampling area m2	calculation formulae
Belgium C1	general	from high to bad	not available	Multimtric Index Flanders	0/1	3 minutes sampling from all available microhabitats	not specified	Multimetric Index (combination of 5 metrics)
Denmark C1	general	from high to moderate (only 14 samples in poor and bad status)	physical description, water quality data	DSFI	1/7	kick sampling from all microhabitats of the site across 3 transctes	about 1.25	two entries table (2 metrics: indicator group and diversity group)
Estonia C1	general	from high to poor	hydrochemical data for few sites	British ASPT	1/10	5 kicks from most tipical substrate + 1 qualitative sample from all habitats	1.25	BMWP divided number of families
France C1	general	from high to bad	national PC monitoring network	IBGN	1/20, transformed in EQR	8 habitat samples charaterized by substrate dimension and flow velocity	0.4	two entries table (2 metrics: n° of family and Faunistic Indicator Group)
Germany C1	morphology (general)	from high to bad	none	German Official System	SI(DE): 5/1 indicative (min=high quality); GD(DE): 0.01/0.84	proportional to microhabitats presence, semiquantitative (DIN 38410)	not specified	two indices: multimetric and weighted averaging
Italy C1	general	from high to bad	main physical, chemicals, microbiological variables (not available for all samples)	IBE (national method)	1/14	riffle only, semiquantitative	0.9 (estimated)	two entries table (2 metrics: n° of taxa and Faunistic Group)
Poland C1	eutrophication	from high to bad	Waterchemistry data	BMWP-POL and Margalef div. Index	BMWP-POL: 0/>100 Margalef: <1/>5,50	4 quant replicates from dominating substrates + 1 qualitative from all habitats	1	worst classification between BMWP-POL and Margalef div. index
UK C1	organic	from high to bad	chemical monitoring data and pressures	National GQA classification	0/>1	RIVPACS	not specified	EQI ASPT (observed ASPT/RIVPACS predicted ASPT)

Table 4.4.1 Test datasets features (continued)

dataset name (country and stream type)	size class km2	altitude class m	data collector	contact person	data owner	aim of the collection	number of sites	number of year/season	number of samples	number of high status samples (according to MS method)	number of high status samples (according to Best Available Classification)
France C2	10-200	<150	Direction Regionale de l'Environment	J. G. Wasson Cemagref-Lyon	database: Cemagref-Lyon	national monitoring network	38	10 years	143	73	27 reference sites according to pressures data
Spain C2	10-100	<200	UNI Vigo	I. Pardo - UNI Vigo	various: UNI Vigo, Aguas de Galicia, Confederacion Hidrografica del Norte	research/ monitoring	25	3 years, 4 seasons	46	7	7
totale	**C2**						**25**		**46**	**7**	
France M1	10-100	200-800	Direction Regionale de l'Environment	J. G. Wasson Cemagref-Lyon	database: Cemagref-Lyon	national monitoring network	32	6 years	77	28	nd
Italy M1	10-100	200-800	CNR-IRSA	A. Buffagni - CNR-IRSA	CNR-IRSA	EU STAR Project sites (51 samples) and test AQEM method (12)	23	3/1 seasons	63	21	12
totale	**M1**						**23**		**63**	**21**	
Italy M5	10-400	100-150	CNR-IRSA	A. Buffagni - CNR-IRSA	CNR-IRSA	National research samples	12	3 seasons	37	1	8
totale	**M5**						**12**		**37**	**1**	

Table 4.4.1 (continued): test datasets features

dataset name (country and stream type)	main degradation factor	quality gradient	support data	classification method	min/max possible values	sampling method	sampling area m2	calculation formulae
France C2	general	from high to bad	National PC monitoring network	IBGN	1/20, transformed in EQR	8 habitat samples charaterized by substrate dimension and flow velocity	0.4	two entries table (2 metrics: n° of family and Faunistic Indicator Group)
Spain C2	organic - nutrients	from high to bad	Physico-chemistry, hydromorfological, diatoms (not for 1 year data)	Multimetric index. Multiple regression analysis to select metric combinations	2.03/6.45	multihabitat sampling proportional 20 kick (most samples) 3 min. kick proportional habitat (1 year samples)	2.5	sum of 9 metrics
total	C2							
France M1	general	from high to bad	National PC monitoring network	IBGN	1/20, transformed in EQR	8 habitat samples charaterized by substrate dimension and flow velocity	0.4	two entries table (2 metrics: n° of family and Faunistic Indicator Group)
Italy M1	general	from high to bad	main physical, chemicals, microbiological variables. Environmental indices: HMS, HQA, IFF	IBE (national method)	1/14	riffle only, semiquantitative	0.9 (estimated)	two entries table (2 metrics: n° of taxa and Faunistic Group)
total	**M1**							
Italy M5	general	from high to bad	main physical, chemicals, microbiological variables. Environmental indices: HMS, HQA, IFF	IBE (national method)	1/14	riffle only, semiquantitative	0.9 (estimated)	two entries table (2 metrics: n° of taxa and Faunistic Group)
totla	**M5**							

Table 4.4.1 (continued): test datasets features

4.5 Summary of the biological assessment methods tested

Table 4.5.1 gathers the main features of the considered assessment methods and contains information about sampling and sorting method, identification level, criteria for abundance registration, calculation formulae etc.

method name (acronym)	method full name (English translation)	country	main degradation factor investigated	sampling method (in the present exercise)	sampling area m2	abundance recording
MIF	Multimetric Index Flanders	Belgium	General degradation	3 min. sampling from all available microhabitats	not fixed	abundance classes
DSFI	Danish Stream Fauna Index	Denmark	Organic Pollution, General Degradation (stressor not specified);	sampling of all microhabitats at the site, 12 kick samples along three transect	not fixed	abundance classes
sampling: Swedish example, classification: ASPT	sampling: Swedish example, classification: ASPT	Estonia	General degradation (stressor not specified)	five 1 m-long kicks from the most typical hard bottom of the site, and of one qualitative, unstandardized collection from all habitats available. Swedish example	not fixed	semiquantitative for the five kicks, presence recorded for qualitative samples (0 or 1)
IBGN	Indice Biologique Global Noramalisé (Global Biological Index Normalized)	France	General degradation (stressor not specified)	8 habitat samples charaterized by substrate dimension and flow velocity, semiquantiative	0.4	abundance not recorded
DIN 38 410 - Determination of Saprobic Index of Running Waters	German Official System	Germany	morphology (general)	Different sampling tools and techniques. Sampling of each habitat exceeding 5% coverage	not fixed	abundance classes
IBE	Indice Biotico Esteso (Extended Biotic Index)	Italy	General degradation (stressor not specified)	not fixed N. of replicates, possibly across a representative transect in *riffle* area.	not fixed (estimated 0.9)	semiquantitative sorting. Relative abundance estimation in three class
Polish Assessement method	BMWP-Polish version & Margalef index	Poland	Organic pollution	4 quantitative sampling + 1 qualitative sampling	not fixed	N° of individuals
Spanish MMI	Spanish Multimetric Index	Spain	General degradation (stressor not specified)	proportional sampling according to microhabitats presence. 20 replicates (18 samples) or 3 min. sampling (28 samples)	2.5 for 20 replicate samples	real abundances
GQA	General Quality Assessment	UK	General degradation (stressor not specified)	3 min. sampling + 1 min. search. All habitats sampled in proportion to their cover, both in *riffle* and *pool*	not fixed	abundance classes, number of individuals

Table 4.5.1: considered assessment methods

method name (acronym)	ID level	calculation formulae	min/max values	class boundaries	literature references
MFI	Genus/Family	Mulitmetric index (sum of 5 metrics)	0/1	H-G: 0.8; G-M: 0,6; M-P: 0.4; P-B: 0.2	Gabriels *et al.*, 2004
DSFI	Genus or Family	matrix with 6 indicator groups along one axis and 4 diversity groups along another axis. 7 quality classes	1/7 The calculation result directly delivers the quality class	H-G: 7 and 6	Skriver *et al.*, 2000
sampling: Swedish example, classification: ASPT	Family	Brirtish *average BMWP-score per taxon* (ASPT).	1/10	In this exercise for Estonia: H-G: 6.1; G-M: 5.1; M-P: 4.1; P-B: 3.1.	for sampling: Johnson R.K., 1999; Medin et al., 2001; for ASPT: Armitage *et al.*, 1983.
IBGN	Family	two entries table (2 metrics: n° of family and Faunistic Indicator Group). 5 quality classes	1/20	H-G: 17; G-M: 13; M-P: 9; P-B: 5	Vernaux *et al.*, 1982
DIN 38 410 - Determination of Saprobic Index of Running Waters	species, species groups, genus	Saprobic Index. 5 quality classes	4/0 (highest value, worst class	H-G: 1.7; G-M: 2.2; M-P: 2.8; P-B: 3.4	DIN 38410, 2003; Friedrich & Herbst, 2004
IBE	Genus/Family	two entries table (2 metrics: n° of taxa and Faunistic Group). 5 quality classes	0/14	H-G: 9,6; G-M: 7,6; M-P: 5,6; P-B: 3,6	APAT/IRSA, 2003; Ghetti, 1997
Polish Assessement method	Family	combination of two indices: *BMWP* scores, modified according to Polish river and *Margalef diversity index*. 5 quality classes	BMWP: 0/open end (usually more than 100); Margalef. 0/not fixed (usually more than 6)	For BMWP-POL, H-G: 100; G-M: 70; M-P: 40; P-B: 10. For Margalef, H-G: 5.5; G-M: 4; M-P: 2.5; P-B: 1.	Armitage *et al.*, 1983
Spanish MMI	spacies/Genus	sum of 9 metrics	0/1	H-G: 0.97; G-M: 0.73; M-P: 0.49; P-B: 0.24	Barbour et al., 1999; Pardo, 2003
GQA	Family	combination of two indices: the *average BMWP-score per taxon* (ASPT) and the *number of scoring taxa.* Comparison with expected value in unpolluted site. The resulting EQI values are asigned to 6 quality classes	EQR	For EQI-ASPT, H-G: 1; G-M: 0.89; M-P: 0.77; P-B: 0.66, B: 0.50. For EQI-N_taxa, H-G: 0.85; G-M: 0.70; M-P: 0.55; P-B: 0.45, B: 0.30.	Wright *et al.*, 2000; EA, 1997; Armitage *et al.*, 1983

Table 4.5.1 (continued): considered assessment methods

5 Benchmark dataset

A definition

Benchmark data

Data fulfilling the WFD demands (e.g. stream type specific, reference conditions established, EQRs calculated, five quality classes considered), including biological, phisico-chemical and general pressure data.

Notes: (a) Such data should provide evidence of a high degree of comparability among countries and can be used to derive trans-National information and benchmarking. (b) Example of potential benchmark datasets already existing (derived by E.U. co-funded projects): AQEM (invertebrates), FAME (fishes), STAR (invertebrates, diatoms, macrophytes, fishes, hydromorphology).

5.1 Benchmark dataset

According to the procedure described in the previous paragraph, 11 datasets from 6 different European countries are included in the benchmark. Datasets are described here below, divided in AQEM Project datasets, STAR Project datasets and extra AQEM/STAR datasets.

5.2 AQEM Project datasets and STAR Project datasets

At sites investigated for the Project EU-AQEM (Hering *et al.*, 2003), samples were collected with the aim of developing and testing macroinvertebrate based assessments method, which satisfiy WFD requirements. Following the principles of the AQEM Project, the STAR Project (Furse, 2001) aims to develop a framework method for calibrating different biological survey results against ecological quality classifications following the indication of the Water Framework Directive.

For the collection of all invertebrate data for both projects, the sampling procedure followed a multihabitat approach (derived from Barbour *et al.*, 1999; see also Hering *et al.*, 2004), in which 20 replicate are collected proportionally from the observed microhabitat. Additional data are available for all samples and sites, concerning the main physical, chemical and microbiological variables. Also, the AQEM/STAR site protocol provides information on environmental variables such as morphological features, degree of general degradation, measures of discharge, land use in floodplain and catchment area, etc. (for a detailed description see AQEM consortium, 2002). Additionally, extra environmental data can be available for single datasets.
Every set of data collected for the AQEM and STAR Project includes a set of Reference sites, selected according the demands of the WFD. Criteria for the selection of the Reference sites are specified in Hering *et al.* (2003).
Samples have been classified according three classification method: firstly, a pre-classification was provided to have a rough idea of the sites degree of degradation; after data collection, a post-classification was undertaken, based on the analysis of the benthic community and to its response to the stressors identified by means of multivariate techniques

or others statistical methods; at last, the final-classification based on the definition of a multimetric index well fitting with the post-classification (see Hering *et al.*, 2004). The BAC corresponds, depending on the countries, to the post or the final classification, depending on which of the two better represent the quality gradient of the sites described by macroinvertebrates' community. For the STAR datasets, the analysis are still in progress, thus the BAC is provided according to the analysis provided till this moment.

A few more definitions

Best Available Classification (BAC)

The biological classification obtained by applying a WFD compliant procedure and all the available, relevant information on a site. E.g. depending on the kind of the main pressures acting, it may result from the integration of biological, physico-chemical and hydromorphological information. It must be based on detailed community analysis (e.g. by multivariate analysis) and not simply on the standard National method of classification. Agreed BACs will be produced based on the criteria outlined in the Guidance on the Intercalibration process (Annexes I and II: Outline protocol for comparing Member States' class boundaries).

Notes: (a) A BAC classification, which is provided at this early stage of the WFD implementation for IC purposes, should correspond to the classification we would obtain by fully applying a WFD compliant classification system. The main difference with such classification is that a BAC refers to a single BQE, because the biological intercalibration is being performed at the BQE level (i.e. not at the final classification stage). (b) It refers to a river site or sample. (c) A benchmark classification (i.e. a preliminary surrogate for a final, agreed BAC) for a number of European stream types and sites is provided by the AQEM and STAR projects, expressly co-funded by the E.C. to support the WFD implementation across Europe. A part of the data produced by the two projects has been used for the comparison and harmonization exercises presented in this STAR Deliverable.

National Standard Classification

The biological classification obtained by applying the current MS quality classification scheme for each BQE.

Notes: (a) Each MS has its National legislation regulating the quality classification of rivers/river sites. In many cases, the procedure applied up to the present time by MSs for classifying sites does not satisfy, or only partially, the WFD requirements. (b) It refers to a river site or sample.

5.3 Acceptable criteria to derive a BAC classification

- Evaluation of tolerance to pollution included
- Richness/Diversity considered
- Abundance considered
- Type specific classification → the used calssification system must be stream type adapted (i.e. type or site specific reference conditions)
- Pressure analysis is combined with biological information → nor abiotic or biotic classification only are acceptable (e.g. multimetric systems alone: no BAC;

morphological or chemical classification alone inadequate; no pressures-based classification only)
- o Classification based on multivariate analysis acceptable
- o Sample-level classification (not site-level classification → one site can be a 'reference' site in one season and not in others, e.g. according to an identified seasonal disturbance)

It is beyond of the scope of the present Deliverable to argue on the procedure and protocols to be used to derive Best Available Classifications, which will be defined in the proper circumstances (e.g. European Commission, 2004).

During the AQEM project, different partners adopted distinct kinds of analysis to derive the BAC, including multivariate techniques (ordination and classification), k-means analysis, ecological breakpoint identification, expert judgement, etc.

For Italian benchmark datasets, multivariate analysis (PCA) was performed for each area on invertebrate samples data to describe the main biological gradients and relate them to the environmental variables (Buffagni et al., 2004). The PCA sites' scores relative to the multivariate axis expressing environmental quality were utilized to classify sites into five quality classes. The resulting classification was further checked by directly looking at pressures, especially to accept/refuse reference sites/samples, thus deriving the Best Available Classification. To set class boundaries along the multivariate axis the approach used outlined ecological breakpoints between reference and good quality sites, following a method close to the k-means analysis. Equally spaced classes were then selected to set the other thresholds (see Buffagni et al., 2004 for details). This BAC is thus based on the whole available information from the benthic community and the environmental variables investigated, which included water chemistry, hydromorphology, catchment characteristics, etc. (AQEM Consortium, 2002).

5.4 Needed characteristics for benchmark data

The presented data were collected during the AQEM and STAR projects activities. An additional dataset, included into the general Benchmark dataset used to run the statistical testing across datasets and countires, was provided by Jean-Gabriel Wasson (CEMAGREF, Lyon, France).

In general terms, the characteristic for each dataset are:
- taxalist to family level
- taxalist must include abundance for each taxon (at least estimated)
- preferably the samplig area should be known
- high status samples must be present
- a good quality gradient has to be represented
- criteria to classify refrence conditions must be indicated. E.g. sites classified according to direct/indirect multivariate analysis on invertebrate taxa abundances and pressures, etc.

5.5 Features describing each benchmark dataset

- Institution that collected the data (e.g. CEH, CNR-IRSA) and property (Ministry of Environment, etc.)
- aim of the collection
- how many sites are considered

- how many samples/sites/seasons
- how wide is the quality gradient (e.g. form High to Moderate, from Good to Bad)
- river type
- ancillary data available (pressure, chemicals, RHS derived indices, morphological classification, etc.)
- method of classification, including information on class boundaries, min and max values (if defined)
- type of sampling method (qualitative, quantitative, semiquantitative)
- calculation formulae (not statutory)
- final classification (pre-classification, post-classification, BAC, MS's) for the presented data.

5.6 AQEM Project datasets

5.6.1 Austrian Benchmark dataset
Austria ABC101 (A04)

Sites' classification: Best Available Classification
For the Austrian benchmark set, BAC correspond to the multimetric classification developed for AQEM Project (final-classification). Thus, BAC correspond to what in AQEM has been called final classification. Class boundaries were set using the 25% percentile of references and the 75% percentile of bad sites. This range was divided by three.

Number of samples
Data were collected and provided by Dr. Otto Moog from BOKU, Wien. 12 sites have been investigate for two seasons. Total number of samples is 24, 5 of which are classified as reference sites.

General features, stream type
The stream type is named 'Mid-sized streams in the Bohemian Massif'; description can be found in AQEM Consortium, 2002 and Ofenböck *et al.*, 2004.
Sites have catchment area comprised between 100 and 1000 km2, prevalent geology is siliceous and the altitude is 200-800m.

Degradation factor
The main stressor observed is degradation of stream morphology: impoundment measures are the main source of degradation (Ofenböck *et al.*, 2004). According to the Best Available Classification performed, the quality gradient quality classes from 'high' to 'bad' status.

Sampling and sorting notes
The samples include all the 20 replicates proportionally sampled. Sorting is quantitative and sample size is approximately 1.25 m^2.

5.6.2 Czech Benchmark datasets
Czech Republic CB01 (C01)
Czech Republic CB03 (C03)

Sites' classification: Best Available Classification
The BAC in Czech benchmark dataset consists on the post-classification, based on combination of community structure and threshold of saprobic index. Brabec *et al.*, (2004) describe the development of the multimetric index in AQEM Czech stream types on the basis of such post classification.

Czech Republic CB01 (C01)

Number of samples
Data were collected and provided by Dr. Karel Brabec from Masaryk University, Brno. 12 sites have been investigated for two seasons. Total number of samples is 24, with 2 reference sites.

General features, stream type

Sites belong to the stream type ‘Medium sized streams in the central sub-alpine mountains’. Streams are permanent, with maximum discharge in spring. (see AQEM Consortium, 2002 and Brabec *et al.*, 2004 for further description)
Sites have catchment area comprised between 100 and 1000 km^2, siliceous geology and altitude of 200-500m.

Degradation factor
The main stressor observed is organic pollution. According to the Best Available Classification performed, the quality gradient quality classes from ‘high’ to ‘bad’ status.

Sampling and sorting notes
The samples include all the 20 replicates proportionally sampled. Sorting is quantitative and sample size is approximately 1.25 m^2.

Czech Republic CB03 (C03)

Number of samples
Data were collected and provided by Dr. Karel Brabec from Masaryk University, Brno. 12 sites have been investigated for two seasons. Total number of samples is 22, with 7 reference sites.

General features, stream type
Sites belong to the stream type ‘Mid-sized streams in the Carpathians’. Streams have braided channels under natural conditions (see AQEM Consortium, 2002 and Brabec *et al.*, 2004 for further description).
Sites have catchment area comprised between 100 and 1000 km^2, flysch geology is dominated by flysch and altitude of 200-500m.

Degradation factor
The main stressor observed is organic pollution. According to the Best Available Classification performed, the quality gradient ranges from ‘high’ to ‘bad’ status.

Sampling and sorting notes
The samples include all the 20 replicates proportionally sampled. Sorting is quantitative and sample size is approximately 1.25 m^2.

5.6.3 German Benchmark dataset

Sites’ classification: Best Available Classification
In German benchmark dataset the Best Available Classification is the post-classification. Descriptions can be found in (..)

Germany DB04 (D04)

Number of samples
Data were collected and provided by Dr. Daniel Hering from University of Essen. 38 sites have been investigated for two seasons (spring and summer 2000). Total number of samples is 58. Reference sites are 12.

General features, stream type

Sites belong to the stream type 'Small streams in lower mountainous areas of central Europe'. Streams are characterized by sinuating channel with anabranched sections and temporarily connected side arms. The floodplain is completely covered with woody vegetation. (descriptions in AQEM Consortium, 2002 and Lorentz *et al.*, 2004).
Sites have catchment area comprised between 10 and 100 km^2, siliceous geology and altitude of 200-500m.

Degradation factor
The main stressor observed is degradation in stream morphology, associated with organic pollution (AQEM Consortium, 2002). According to the Best Available Classification performed, the quality gradient quality classes from 'high' to 'bad' status.

Sampling and sorting notes
The samples include all the 20 replicates proportionally sampled. Sorting is quantitative and sample size is approximately 1.25 m^2.

5.6.4 Italian Benchmark datasets
Italy IBM101 (IO2)
Italy IBM102 (IO3)
Italy IBC101 (IO4)

Sites' classification: Best Available Classification
For all Italian benchmark datasets, the BAC correspond to the post-classification, performed through a multivariate analysis. A PCA analysis was applied to the samples. The ordination axes were correlated to environmental and water quality data in order to clarify the observed gradients. The scores along the PCA axis interpreted as an environmental quality gradient, considering the different degradation factor in each dataset, is used to classify the sites.
The classification, is based on the selection of the ecological breakpoint for the separation between reference and good sites. The remaining classes are, whenever possible, equally spaced. Buffagni *et al.* (2004) describe in detail the assessment module for Southern Apennines Italian stream type (I02) and the selection of the PCA based classes (BAC).

Additional information
Data from the following environmental indices are also available: Habitat Modification Score, Habitat Quality Assessment (HMS and HQA, Raven *et al.*, 1998, Buffagni & Kemp, 2001), Index of Fluvial Functioning (IFF, Siligardi *et al.*, 2000, Balestrini *et al.*, 2004).

Number of samples
Data were collected and provided by Dr. Andrea Buffagni from CNR-IRSA, Brugherio. For each of the three Italian AQEM areas, 11 sites have been investigate for three seasons: spring and autumn 2000 and winter 2001. Total number of samples is 33, 9 of which are classified as reference sites. Thus, AQEM Italian dataset included in the benchmark comprehend 99 samples with 27 reference sites.

Italy IBM101 (IO2)
General features, stream type
Samples of this dataset belong to non intermittent rivers located in Southern Apennines (region Campania, see AQEM Consortium, 2002; Buffagni *et al.*, 2004; Balestrini *et al.*, 2004, for further description). Sites are small-sized (catchment area lower than 100km^2

except for one site), calcareous and have an altitude range of 200–800 m. Maximum distance between two sites is about 100 km.

Degradation factor
The main stressor observed is organic pollution, often associated with degradation of stream morphology (Buffagni *et al.*, 2001). It's thus possible to consider a 'general degradation' factor. According to the Best Available Classification performed, the quality gradient covers all the quality classes from 'high' to 'bad' status.

Sampling and sorting notes
The samples consists of the 10 pool replicates, since the assessment system is developed on this area (Buffagni *et al.*, 2004). The 10 pool sample resulted more representative of the quality gradient (Buffagni *et al.*, op. cit.). Sorting is quantitative and sample size is approximately 0.5 m^2.

Italy IBM102 (IO3)
General features, stream type
Samples of this dataset belongs to rivers located in Northern Apennines (region Emilia Romagna, see AQEM Consortium, 2002; Balestrini *et al.*, 2004, for further description). Sites are medium-sized (catchment area between 100 and 1000 km^2), calcareous and have an altitude range of 200–800 m. Maximum distance between two sites is about 50 km.

Degradation factor
The main stressor observed is degradation of stream morphology (Buffagni *et al.*, 2001). According to the Best Available Classification performed, the quality gradient quality classes from 'high' to 'moderate' status.

Sampling and sorting notes
The samples refer to the 10 pool replicates, The 10 pool sample resulted more representative of the quality gradient. Sorting is quantitative and sample size is approximately 0.5 m^2.

Italy IBC101 (IO4)
General features, stream type
Samples of this dataset belongs to spring fed small streams also named 'fontanili' located in the lowland of the Po river (region Piemonte, see AQEM Consortium, 2002; Balestrini *et al.*, 2004 for further description). Sites are small-sized (catchment area usually lower than 100), calcareous and the altitude is lower than 200m. Maximum distance between two sites is about 40 km.

Degradation factor
The main stressor observed is general degradation: water pollution associated to alteration in stream morphology (Buffagni *et al.*, 2001). According to the Best Available Classification performed, the quality gradient quality classes from 'high' to 'bad' status.

Sampling and sorting notes
The samples include all the 20 replicates proportionally sampled. Sorting is quantitative and sample size is approximately 1 m^2.

5.7 STAR Project datasets

5.7.1 United Kingdom Benchmark datasets

UK 1

Sites' classification: Best Available Classification

The BAC in UK samples is determined according to the RIVPACS method (Wright *et al.*, 2000).

Number of samples

Data were provided by Dr. John Murray Bligh from Environment Agency. 13 sites have been investigated for two seasons (spring and autumn). Total number of samples is 60. Reference samples are 18.

General features, stream type

Sites belong to the stream type 'Small lowland calcareous streams', broadly correspondent to RIVPACS group 32 (Type I sites). Altitude is lower than 200m and catchment area comprised between 10 and 100 km^2. Geology is calcareous ($CaCO_3$ >80mgl-1).

Degradation factor

Main degradation factor is organic pollution. According to the Best Available Classification performed, the quality gradient quality classes from 'high' to 'bad' status.

Sampling and sorting notes

The dataset comprehend two sampling method: the national assessment method RIVPACS (Murray-Bligh, 1999) and the STAR sampling method. It is the Additional stream type investigated in the STAR project (see Hering & Strackbein, 2002).

UK 2

Sites' classification: Best Available Classification

The BAC in UK samples is determined according to the RIVPACS method (Wright *et al.*, 2000).

Number of samples

Data were provided by Dr. John Murray Bligh from Environment Agency. 12 sites have been investigated for two seasons (spring and autumn). Total number of samples is 66. Reference samples are 18.

General features, stream type

Sites belong to the stream type 'Medium sized, deeper, calcareous lowland' sites in RIVPACS Group 20 (Type J sites).
Altitude is lower than 200m and catchment area comprised between 10 and 100 km^2. Geology is calcareous ($CaCO_3$ >80mgl-1).

Degradation factor

Main degradation factor is organic pollution. According to the Best Available Classification performed, the quality gradient quality classes from 'high' to 'bad' status.

Sampling and sorting notes
The dataset comprehend two sampling method: the national assessment method RIVPACS (Murray-Bligh, 1999) and the STAR sampling method. The samples are investigated in the STAR Project as core stream types (see Hering & Strackbein, 2002).

5.7.2 Italian Benchmark datasets

Italy IBM102 (IO6)

Sites' classification: Best Available Classification
As the other Italian datasets, the BAC classification is performed via a multivariate analysis (post-classification) see previous Italian description for further details.

Number of samples
Data were collected and provided by Dr. Andrea Buffagni from Environment Agency. 11 sites have been investigated. For all sites, data from the summer sampling period is included. For few sites also winter an spring season is enclosed. Total number of samples is 16. Reference samples are 2.

General features, stream type
Stream type is 'Small sized calcareous streams in the Central Apennines'. Sites are located in Tuscany region and are characterised by gravel to cobble substrate, and a sinuate channel form in a Ushaped valley. The annual regime is usually permanent, even if under extreme conditions some sites can run dry in summer. Catchment area is 10-100 km^2 and altitude class: 200-800 m. Geology is dominated by calcareous formations.

Degradation factor
Streams are mainly affected by sewage, pasture and agriculture. Some alteration in stream morphology can be observed. Thus a general degradation can be stated. According to BAC, reference, good and moderate status samples are present.

Sampling and sorting notes
The samples refer to the 10 pool replicates. The 10 pool sample resulted more representative of the quality gradient. Sorting is quantitative and sample size is approximately 0.5 m^2.

5.8 Extra AQEM/STAR datasets

5.8.1 France FBM101

Important note
The same dataset with a different normalization (i.e.: according to the 75^{th} percentile of high status samples) is used also as test dataset as France M1 (see description in chapter 4). For the harmonization process to the test dataset France M1, the French benchmark subset here described is excluded.

Sites' classification

The classification method is WFD compliant. Adaptation on the IBGN criteria for abundance registration, originally without considering the abundances, has been performed. In the present dataset the number of specimens is recorded as real abundance.

Jean-Gabriel Wasson comments
Reference sites are selected on the basis of very low anthropic pressures, independently of the biological values in a first approach. The distribution of biological data is then calculated for all samples of the reference dataset, and the outliers samples are checked. Dubious sites are eliminated, but low biological values are accepted if they come from validated reference sites. The procedure combine both spatial and temporal variability of a given stream type. The Reference Conditions (RC) are defined as the range of variability of a given biological element (index or metric) observed at reference sites. However, the calculation of EQR needs to define a Reference Value (RV) for the normalization of the samples. Due to the small number of reference sites generally observed for most types, the most robust and stable statistic is chosen as RV. For all our calculations, following the recommendation of the REFCOND guidance, the median was used as Reference Value. The general approach and Reference Values for each type are described in a work paper (Wasson et al., October 2003, in French) and a summary (in English) will be available soon.

Number of samples (see dataset France M1 description in cap.4)
The total number of sites included is 32. Samples correspond to the years 1992 – 2001; they are representative of the whole hydrologic cycle, with an equal number of samples in late winter and spring (February to June), and in summer and early fall (July to November). Total number of samples is 77, 17 are classified as reference.

General features, stream type (see dataset France M1 description in cap.4)
Sites belong to the hydro-ecoregion “Méditerrannée” (HER 6) of the French typology. Hydrologic seasonality is high, but the streams are not regularly intermittent. Altitude ranges from 0 to 600m, comparable in term of climatic conditions with the range 200-800 m of more southern Mediterranean countries (Spain, Portugal, Italy). Catchment area is small and comprised between 10 and 100 km^2.

Degradation factor(see dataset France M1 description in cap.4)
GENERAL DEGRADATION IS THE MAIN FACTOR OF ALTERATION. THE DATASET COVERS ALL THE RANGE OF ECOLOGICAL STATUS, FROM ‘HIGH’ TO ‘BAD’ STATUS ACCORDING TO THE NATIONAL METHOD. DATA FROM CORINE LAND COVER ARE AVAILABLE FOR ALL THE SITES. ON THE BASIS OF LAND USE, PRESSURES OF THE SITES COULD BE FURTHER EVALUATED.

Sampling and sorting notes
The sampling and sorting method is the French national method IBGN. Adaptation in abundance recording has been performed in order to assure WFD compliancy.

5.9 Summary tables for benchmark datasets

In table 5.1 the selected benchmark datasets are reported, with a synthesis of all the major features, related to samples characteristics, method of classification, etc.

dataset code	country, IC type	AQEM/STAR code	size class km2	altitude class m	data collector and owner	aim of the collection	sites N.	Number of season	total samples N.	reference samples N. (BAC)	sampling method
AB04	Austria, A2	A04	100-1000	200-800	BOKU-Wien O. Moog	EU AQEM Project sites	12	2	24	5	AQEM sampling protocol
CB01	Czech Republic, C	C01	100-1000	200-500	Masaryk University K. Brabec	EU AQEM Project sites	12	2	24	2	AQEM sampling protocol
CB03	Czech Republic, C	C03	100-1000	200-500	Masaryk University K. Brabec	EU AQEM Project sites	11	2	22	7	AQEM sampling protocol
DB04	Germany, C	D04	10-100	200-800	UNI-Essen D. Hering	EU AQEM Project sites	29	2	58	12	AQEM sampling protocol
IBC101	Italy, C1	I04	10-100	<200	CNR-IRSA A. Buffagni	EU AQEM Project sites	11	3	33	9	AQEM sampling protocol
IBM101	Italy, M1	I02	10-100	200-800	CNR-IRSA A. Buffagni	EU AQEM Project sites	11	3	33	8	AQEM sampling protocol
IBM102	Italy, M2	I03	100-1000	200-800	CNR-IRSA A. Buffagni	EU AQEM Project sites	11	3	33	7	AQEM sampling protocol
IBM103	Italy, M1	I06	10-100	200-800	CNR-IRSA A. Buffagni	STAR Project sites	12	1 (3 for 3 sites)	16	2	AQEM sampling protocol
UB01	UK, C	U15	10-100	<200	CEH Dorset	EU STAR Project additional stream type	13	2	60	18	STAR sampling protocol and RIVPACS
UB02	UK, C	U23	10-100	<200	CEH Dorset	EU STAR Project core stream type	12	2	66	18	STAR sampling protocol and RIVPACS
	Total						**134**		**369**	**88**	

Table 5.9.1 Selected benchmark datasets.

dataset code	main degradation factor	quality gradient	support data	Best Available Classification criteria
AB04	Morphology	from high to bad	main phisic, chemicals, microbiological variables. AQEM site protocol	Multimetric classification. Range between 25° percentile of high staus and 75° of bad divided by three
CB01	Organic pollution	from high to poor	main phisic, chemicals, microbiological variables. AQEM site protocol	Post-classification, community structure and theresholds for saprobic value
CB03	Organic pollution	from high to poor	main phisic, chemicals, microbiological variables. AQEM site protocol	Post-classification, community structure and thresholds for saprobic value
DB04	Morphology, Organic pollution	from high to bad	main phisic, chemicals, microbiological variables. AQEM site protocol	Post-classification
IBC101	General degradation	from high to bad	main phisic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between high and good class along multivariate axis. Remaining classes equally spaced
IBM101	General degradation	from high to bad	main phisic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between high and good class along multivariate axis. Remaining classes equally spaced
IBM102	Morphology	from high to moderate	main phisic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between high and good class along multivariate axis. Remaining classes equally spaced
IBM103	General degradation	from high to moderate	main phisic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between high and good class along multivariate axis. Remaining classes equally spaced
UB01	Organic pollution	from high to bad	main phisic, chemicals, microbiological variables. STAR site protocol	RIVPACS classification
UB02	Organic pollution	from high to bad	main phisic, chemicals, microbiological variables. STAR site protocol	RIVPACS classification

Table 5.9.1 (continued) Selected benchmark datasets.

References

AQEM Consortium, 2002. Manual for the application of the AQEM system. Version 1.0 February 2002.

Balestrini, R., M. Cazzola & A. Buffagni, 2004. Characterizing hydromorphological features of selected Italian rivers: a comparative application of environmental indices. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe. Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 367–381.

Barbour, M. T., J. Gerritsen, B. D. Snyder, & J. B. Stribling, 1999. Rapid Bioassessment Protocols for Use in streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency, Office of Water, Washington D.C.

Brabec, K., S. Zahradkova, D. Nemejcova, P. Paril, J. Kokeš & J. Jarkovsky, 2004. Assessment of organic pollution effect considering differences between lotic and lentic stream habitats. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 331–346.

Buffagni, A., J. L. Kemp, S. Erba, C. Belfiore, D. Hering & O.Moog, 2001. A Europe wide system for assessing the quality of rivers using macroinvertebrates: the AQEM project and its importance for southern Europe (with special emphasis in Italy). J. Limnol. 60 (suppl. 1): 39–48.

Buffagni, A. & J. L. Kemp, 2002. Looking beyond the shores of the United Kingdom: *addenda* for the application of River Habitat Survey in South European rivers. J. Limnol. 61 (2): 199–214.

Buffagni, A., S. Erba, M. Cazzola & J. L. Kemp, 2004. The AQEM multimetric system for the southern Italian Apennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 313–329.

Furse, M.T. 2001. “Standardisation of River Classifications: Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive” STAR Project, Part B - EU Contract number: EVK1-CT 2001-00089.

Hering, D. & Strackbein J., 2002. STAR stream types and sampling sites. STAR Project First Deliverable. June 2002.

Hering, D., A. Buffagni, O. Moog, L. Sandin, M. Sommerhäuser, I. Stubauer, C. Feld, R. K. Johnson, P. Pinto, N. Skoulikidis, P. F. M. Verdonschot & S. Zahrádková, 2003. The development of a system to assess the ecological quality of streams based on macroinvertebrates – design of the sampling programme within the AQEM project. Int. Rev. Hydrobiol. 88: 345–361.

Hering, D., O. Moog, L. Sandin & P. F. M. Verdonschot, 2004. Overview and application of the AQEM assessment system. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands* Hydrobiologia 516: 1–20.

Lorentz, A., D. Hering, C. K. Feld & P. Rolauffs, 2004. A new methot for assessing the impact of hydromorfological degradation on the macroinvertebrate fauna of five German stream types. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds),* *Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 107–127.

Murray-Bligh, J. A. D., 1999. Procedure for quality assurance for RIVPACS compatible macro-invertebrate samples analysed to the taxonomic level needed for the BMWP-score system. Quality Management Systems for Environmental monitoring: Biological Techniques, BT003. (Version 1.0, 3n August 1999) Bristol, Environment Agency.

Ofenböck, T., O. Moog, J. Gerritsen & M. Barbour, 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 251-268.

Raven, P. J., T. H. Holmes, F. H. Dawson, P. J. A. Fox, M. Everard, I. R. Fozzard & K. J. Rouen, 1998, River Habitat Survey, the physical character of rivers and streams in the UK and Isle of man. River Habitat Survey No. 2, May 1998. The Environment Agency, Bristol, 86 pp.

Siligardi, M., S. Bernabei, C. Cappelletti, E. Chierici, F. Ciutti, F. Egaddi, A. Franceschini, B. Maiolini, L. Mancini, M. R. Minciardi, C. Monauni, G. L. Rossi, G. Sansoni, R. Spaggiari & M. Zanetti, 2000. I.F.F. Indice di funzionalità fluviale. Manuale ANPA/2000: 223 pp.

Wasson, J.G., Chandesris, A., Pella, H. and Blanc, L. (2003). Typologie des eaux courantes pour la Directive Cadre Européenne sur l'Eau : l'approche par Hydro-écorégion. "5ème séminaire du réseau d'animation REGLIS (Représentation et Gestion de l'Information Spatialisée)", Cemagref UMR Structures et Systèmes Spatiaux, Montpellier, 13-14 novembre, 7 p.

Wright, J. F., Sutcliff, D. W., Furse, M. T. (eds.), 2000. Assessing the biological quality of freshwaters. RIVPACS and similar techniques. Ambleside (Freshwater Biological Association).

6 Common European metrics: ICMs and others

Three examples of possible European metrics are presented in the present Paragraph.
The first set of metrics is represented by the ICMs (Table 3.1), used in the Deliverable to illustrate the different Options for the European IC process. They were especially conceived and selected to be applicable in the short time and simple, also considering a quite 'rough' identification level, such as Family level.
The second and the third sets of metrics were identified with the aim of providing a more scientifically robust selection of metrics, based on STAR and AQEM data, able to effectively describe the degradation gradients observed in two clusters of stream types in Europe. They correspond, respectively, to "Central Lowland" and "Central Mountain" groups of stream types sampled during the two projects.

6.1 Performance of ICMs and ICM index in a range of European test datasets

For the fulfilment of the WFD, all methods consider tolerance and richness of the benthic community. The abundance, even if in some cases not properly considered in the original method, has been always considered in the present exercise. The methods that have defined type specific biological reference conditions are here (tentatively) considered as WFD-compliant.

	ASPT	Shannon	EPT	N° Families	1-GOLD	Log(sel-EPTD)	ICMi
BELC1	0.74	0.72	0.59	0.87	0.53	0.27	0.74
DENC1	0.48	0.02	0.5	0.2	0.1	0.2	0.48
ESTC1	0.98	0.38	0.86	0.57	0.43	0.2	0.76
FRAC1*	0.81	0.28	0.71	0.7	0.46	0.62	0.83
GERC1 SI(DE)*	0.54	0.01	0.26	0.03	0.18	0.34	0.32
GERC1 GD(DE)*	0.45	0.03	0.33	0.04	0.27	0.41	0.32
ITAC1	0.59	0.58	0.55	0.8	0.21	0.51	0.72
POLC1	0.66	0.21	0.78	0.94	0.19	0.4	0.74
UKC1 ASPT-EQI*	0.88	0.31	0.77	0.62	0.2	0.62	0.82
UKC1 NFAM-EQI*	0.57	0.31	0.72	0.87	0.15	0.53	0.71
FRAC2*	0.74	0.32	0.78	0.74	0.31	0.68	0.85
SPAC2*	0.86	0.82	0.87	0.88	0.67	0.61	0.91
FRAM1*	0.74	0.5	0.86	0.88	0.36	0.63	0.86
ITAM1	0.43	0.38	0.66	0.64	0.16	0.61	0.75
ITAM5	0.36	0.18	0.46	0.62	0.28	0.19	0.46
Mean	0.66	0.34	0.65	0.63	0.30	0.45	0.68
Mean WFD compliant	0.70	0.32	0.66	0.60	0.33	0.56	0.70

Table 6.1 R^2 between National assessment methods and ICMs values in test datasets *=WFD compliant.

In test datasets the fit of the ICMi is in general good with a mean value of 0.68. In most of the datasets the R^2 for the ICMi are very good (>0.70). As for the benchmark data, the metrics 1-

GOLD and Shannon show the worst correlation, while the best fits are observed for ASPT, EPT taxa and number of families.

6.2 Performance of ICMs and ICM index in a range of European benchmark datasets

	ASPT	Shannon	EPT	N° Families	1-GOLD	Log(sel-EPTD)	ICMi
AB04*	0.32	0.58	0.73	0.81	0.61	0.78	0.75
CB01*	0.50	0.40	0.48	0.35	0.69	0.52	0.63
CB03*	0.75	0.46	0.71	0.56	0.53	0.73	0.79
DB04	0.20	0.16	0.20	0.25	0.01	0.18	0.25
UB01*	0.80	0.14	0.71	0.51	0.20	0.56	0.74
UB02*	0.76	0.07	0.74	0.51	0.14	0.65	0.74
IBM1*	0.46	0.60	0.48	0.36	0.10	0.56	0.64
IBM202*	0.02	0.46	0.55	0.59	0.15	0.49	0.61
IBC101*	0.83	0.33	0.92	0.65	0.45	0.86	0.86
FBM101*	0.68	0.49	0.81	0.84	0.33	0.59	0.81
DB01	0.24	0.02	0.05	0.01	0.24	0.28	0.13
DB03	0.26	0.04	0.24	0.18	0.08	0.44	0.43
DB05	0.38	0.46	0.55	0.42	0.28	0.51	0.59
Mean	0.48	0.33	0.56	0.48	0.28	0.55	0.62
Mean selected dataset	0.53	0.37	0.63	0.54	0.32	0.59	0.68

Table 6.2: R^2 between BAC and ICMs values in available benchmark dataset.
*=selected dataset.

The relationship observed between the BAC of the benchmark datasets and the values of the ICMs is presented (R^2 values). The mean value of R^2 for the selected datasets is higher than 0.50 in all the ICMs except for Shannon index and 1-GOLD. In general, ICMs showing the best fit are the EPT taxa and Log_selEPTD. The ICMi the regression is good in all selected datasets (mean = 0.68).
Datasets showing a R^2 lower than 0.50 have been excluded from the benchmark. The dataset DB05 has been excluded because no reference sites are present.

The relationship between ICMi and BAC is showed in box and whiskers representations. From such figures it is possible to consider 1) the variability of the ICMi values in each BAC classes and 2) if the quality gradient expressed from the BAC is represented by the ICMi.
Figure 6.1 considers the benchmark datasets of the IC types R-C.

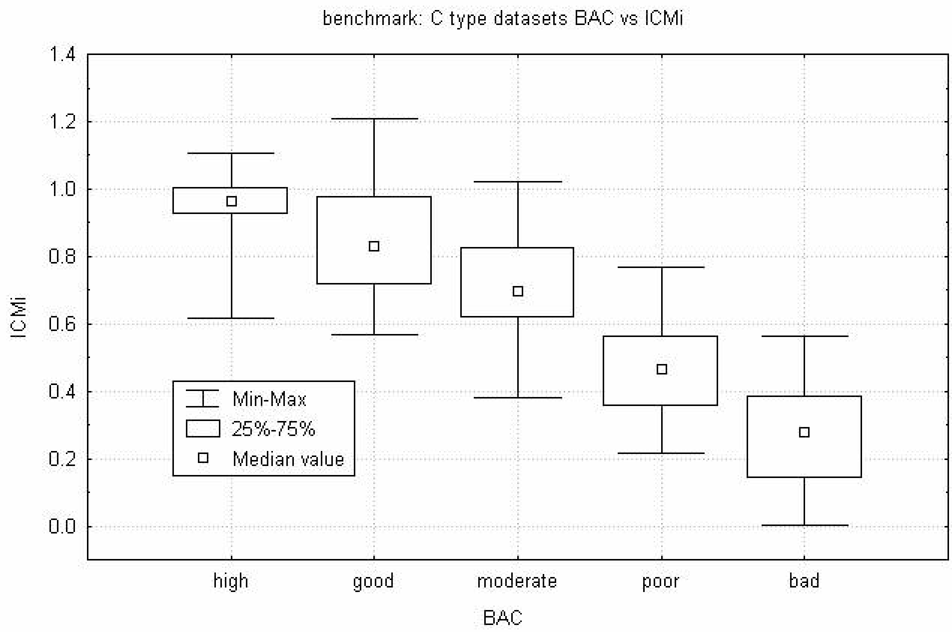


Figure 6.1 Box and whiskers representation for the benchmark datasets selected for R-C types. BAC vs ICMi.

The trend of the median values show a good fit with the quality classes expressed from the BAC. The classes moderate, poor and bad result well separated. A slight overlap of the interquartiles can be observed for class good and moderate and class good and high.

In Figure 6.2, the same representation is provided for the benchmark datasets belonging to IC types R-M.

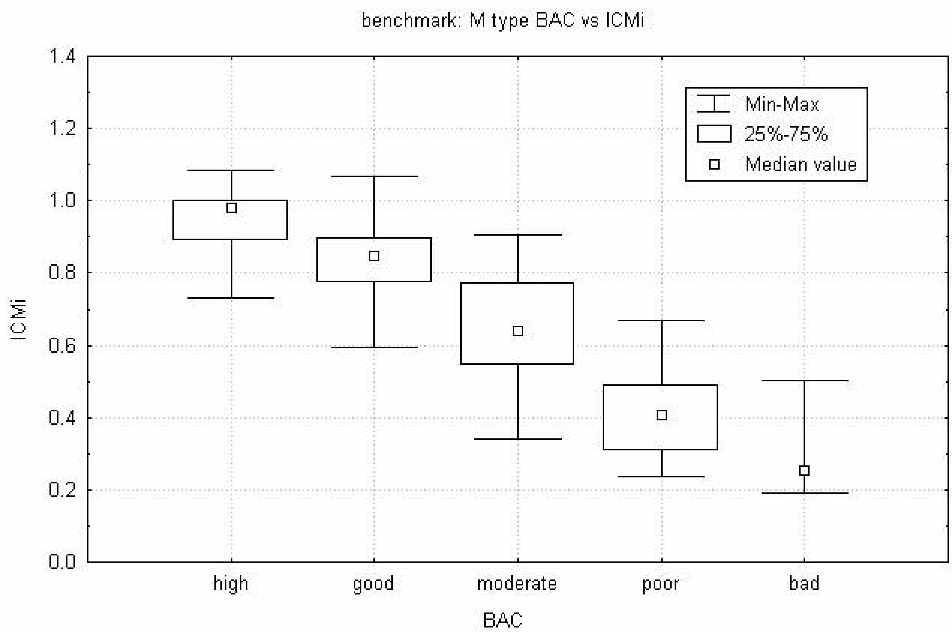


Figure 6.2 Box and whiskers representation for the benchmark datasets selected for R-M types. BAC vs ICMi.

The good trend of the median values of the ICMi related to the BAC quality classes is confirmed also for R-M types. The overlap for all classes is absent or very minor, especially considering the interquartile range.

Figure 6.3 represents the results for all the benchmark datasets selected.

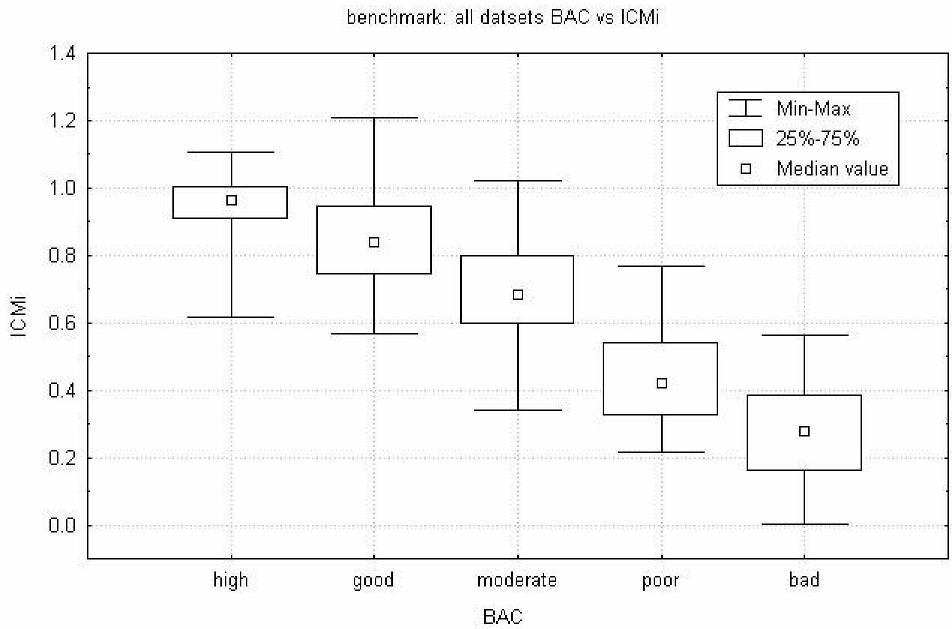


Figure 6.4 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs ICMi

For the complete benchmark dataset, the good trend of the ICMi values in the classes is confirmed. Combining the benchmark sets for R-C and R-M types the overlap is slight between classes good and moderate and class poor and bad.

The following illustration shows the relationship between single ICMs and the BAC for all the benchmark datasets.

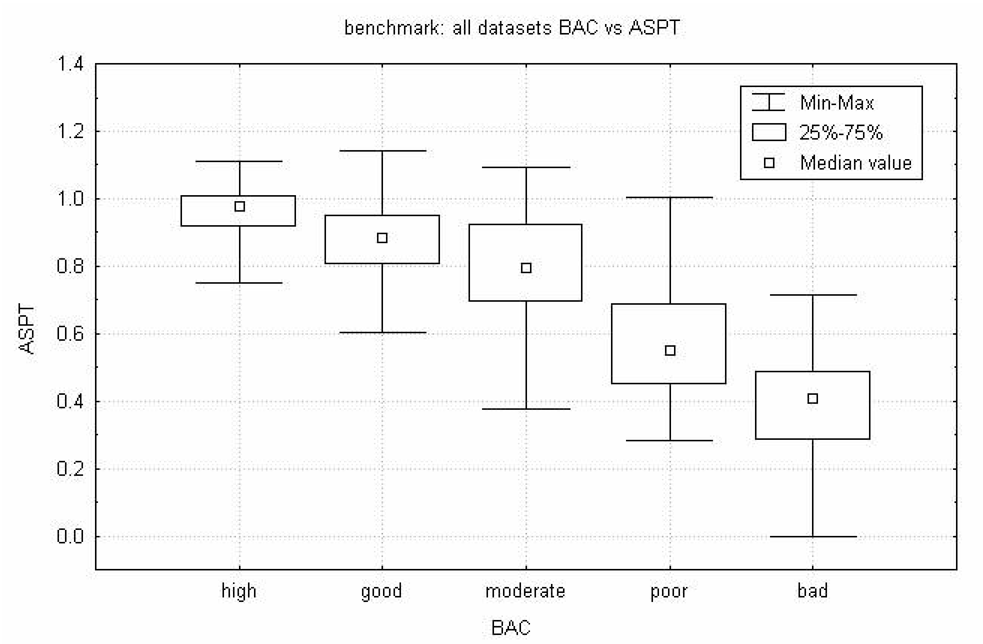


Figure 6.5 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs ASPT.

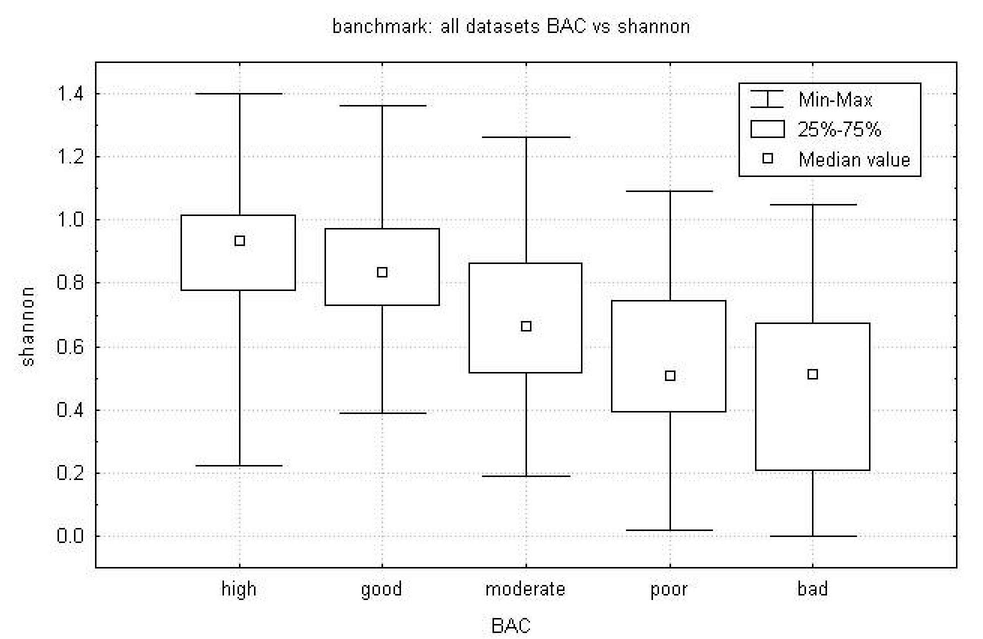


Figure 6.6 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs Shannon

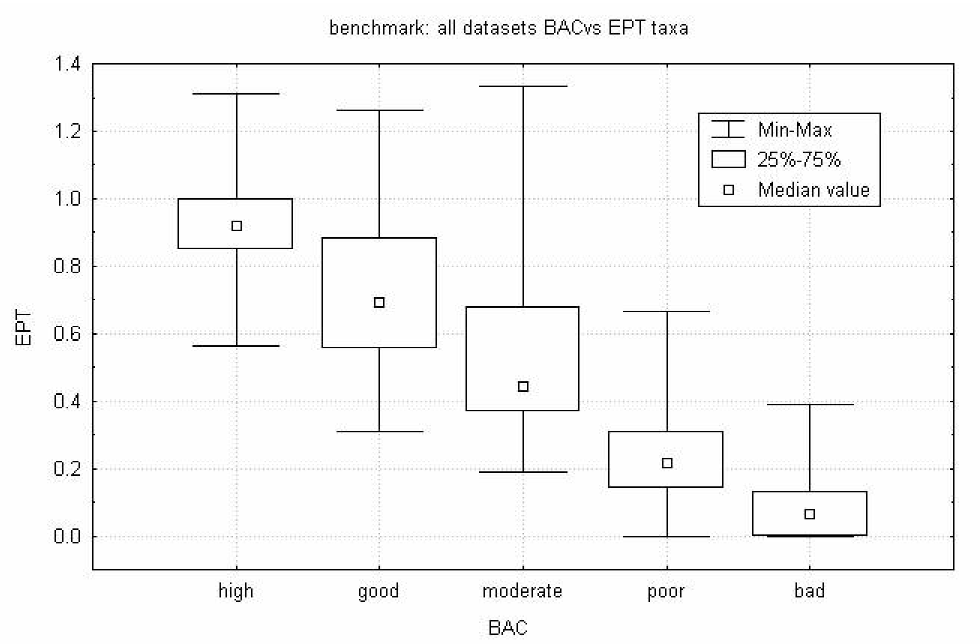


Figure 6.7 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs EPT

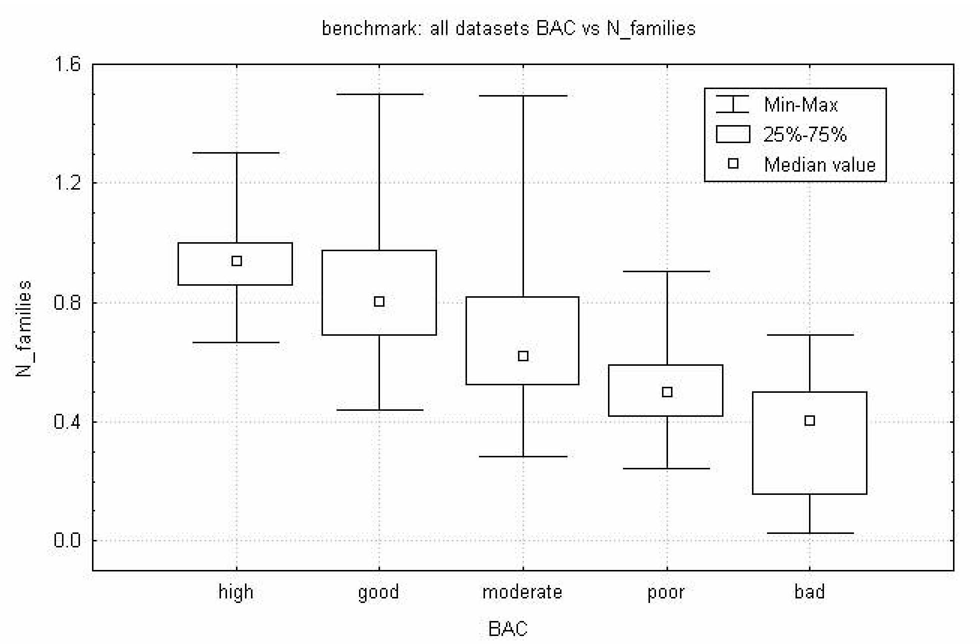


Figure 6.8 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs N_families.

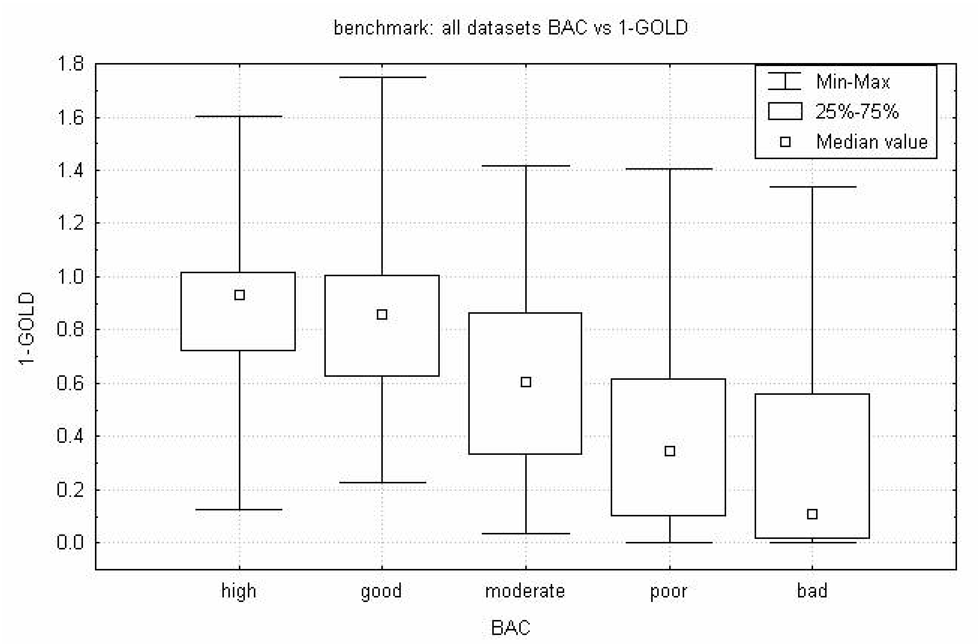


Figure 6.9 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs 1-GOLD.

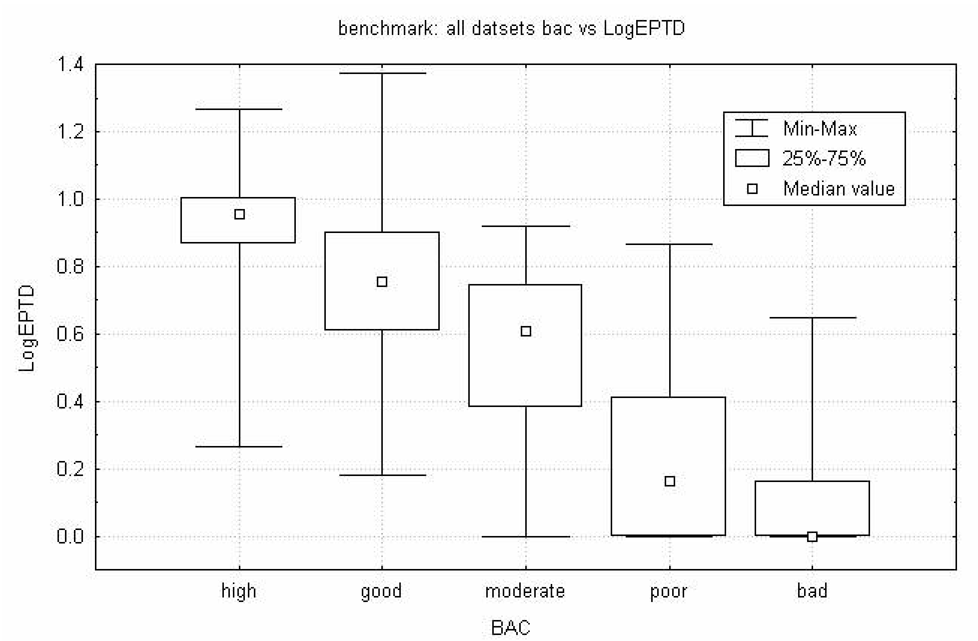


Figure 6.10 Box and whiskers representation for the benchmark datasets selected for all the types. BAC vs Log_EPTD.

The trend of the median values show a good response for all the metrics. For Number of families, EPT taxa, ASPT and LogEPTD the high and good class are well separated. Some ICMs show a general overlap among classes, e.g. 1-GOLD and Shannon.

6.3 The identification of metrics to assess the impact of different environmental stressors in large geographical areas

6.3.1 Introduction

The STAR project covers almost the entire geographical area of Europe, with a north-south extension from Sweden to Greece and an east-west extension from Latvia to Portugal. The project covers more than 20 stream types and the question arose, whether the stream types could be combined into Stream Type Groups (STG), representing streams that are comparable in terms of ecoregion, altitude and size (system A descriptors of the WFD), as well as environmental aspects, such as physical-chemical status and hydromorphological conditions. . In particular for the inter-calibration exercise, water bodies in large geographic areas need to be compared. This comparison should be performed by means of "Common Inter-calibration Metrics", which are suited to assess environmental degradation in a large variety of stream types. Though these metrics give usually not as precise results as metrics specifically selected for an individual stream type, they are suited for comparison purposes. This chapter presents a method to identify suited biological parameters (metrics) to assess the impact of abiotic environmental impacts (stressors).

For the first time, "Common Metrics" are selected which are capable of assessing degradation in broadly defined Stream Type Groups. They should in future be used

- as "Intercalibration Common Metrics" for the EC-inter-calibration exercise – this relates in particular to those metrics acting on a coarse taxonomic level (e. g., family level), see above;
- to compare assessment results within a watershed, which is shared by two or more countries;
- as a preliminary basis to develop (multimetric) assessment systems for those countries, which have not yet developed a system specifically dedicated to the demands of the Water Framework Directive.

6.3.2 Database and methods

Database

The analysis was restricted to the two largest Stream Type Groups defined for STAR, since they represent many stream types and cover a wide geographical area: the "Central Lowland" and "Central Mountain" groups (Table 1). Several stream types that were investigated within the AQEM project (Hering et al., 2004; www.aqem.de) also fit into the two STAR Stream Type Groups and were gained with comparable methods, and, thus, the respective AQEM stream types were added to the database.

Table 1: Stream type groups, group members and designated main stressor of the AQEM and STAR project used for the analysis. Letters of group members indicate the respective country: A = Austria, C = Czech Republic; D = Germany; K = Denmark; N = The Netherlands, S = Sweden; U = United Kingdom, V = Slovakia); Main stressors are indicated by O = Organic pollution, M = Morphological degradation, A = Acidification, G = General degradation.

Stream Type Group	Group members (stream types)	Project	Designated main stressor
STG 1 "Central Lowland"	D01: Small sand bottom lowland streams	AQEM	M
	D02: Organic type lowland brooks	AQEM	M
	D03: Mediumd-sized sand bottom lowland streams	AQEM, STAR	M
	K02: Medium-sized lowland streams	STAR	M
	S05: Medium-sized lowland streams in Southern Sweden	STAR	O
	U23: Medium-sized lowland streams	STAR	O
	N01: Small lowland streams	AQEM	G
	N02: Small hill streams	AQEM	G
STG 2 "Central Mountain"	A04: Medium-sized streams in the Bohemian Massif	AQEM	M
	A05: Small shallow mountain streams	STAR	M
	C04: Small shallow mountain streams	STAR	O
	C05: Small streams in the Central Sub-Alpine Mountains	STAR	M
	C01: Medium-sized streams in the Central Sub-Alpine mountains	AQEM	O
	C15: Small streams in the Carpathian	AQEM	O
	C16: Medium-sized streams in the Carpathian	AQEM	O
	D04: Small shallow mountain streams	AQEM, STAR	M
	D06: Small Buntsandstein streams	STAR	G
	V01:Small calcareous mountain streams in the East Carpathians	STAR	O

Stream Type Group 1 covers a total of eight stream types with 387 samples, Stream Type Group 2 a total of 10 stream types and 369 samples. Each sample comprises i) a taxalist derived from quantitative multi-habitat samples and ii) numerous environmental parameters on different spatial scales, which were derived either from maps or in parallel to macroinvertebrate sampling in the field.

The environmental variables were divided into three groups, representing the supposed main stressors in the datasets (Annex 1): i) physical-chemical measures (organic pollution/eutrophication), ii) hydromorphological parameters (hydromorphological/general

degradation), and iii) land use parameters (organic pollution, general degradation). Table 2 shows the number of environmental variables and samples used for the Stream Type Groups.
Each taxalist was used to calculate nearly 200 metrics, such as richness/diversity measures (e. g. Margalef diversity, # EPT taxa) or functional measures (e. g., feeding types, habitat preferences).
Finally, each sample was represented by environmental variables and biocoenotic metrics which provided the basis for the statistical analysis.

Table 2: Number of environmental variables used for the analysis of the main stressors. A complete list is given in Annex 1.

Stream Type Group	Environmental variable group (possible stressor)	No. of variables (samples)
STG 1	Physical-chemical (organic pollution/eutrophication)	11 (309) for PC1, 8 (387) for PC1a
	Hydromorphology (hydromorphological/general degradation)	41 (367)
	Land use (organic pollution, general degradation)	14 (373)
STG 2	Physical-chemical (organic pollution/eutrophication)	12 (309)
	Hydromorphology (hydromorphological/general degradation)	36 (369)
	Land use (organic pollution, general degradation)	11 (332)

Statistical analysis
Environmental variables and gradients
The statistical analysis aimed at identifying those variables that show the highest relation to certain environmental stressors. In a first step PCA was used to reduce the number of variables by i) calculating hypothetical main gradients of the environmental dataset and ii) identifying redundant (co-correlating) variables. For each environmental variable group a separate PCA was run. Interval-scaled variables were “log (x+1)”-transformed except for pH. Proportional variables (%) were transformed arcsin sqrt x. Variables with a frequency of < 5 samples were excluded from the analysis.

Stream Type Group 1: “Central Lowland”
Physical-chemical variables of this Stream Type Group were analysed twice, since oxygen parameters (dissolved oxygen content, oxygen saturation) were missing for two stream types. The first PCA (PC1) comprised 309 samples for which all parameters were available, the second PCA (PC1a) was run with 387 samples, yet without oxygen parameters. The PCA of hydromorphological (HY1) and land use (LU1) variables were run once each with the number of variables and samples listed in Table 2.

Stream Type Group 2: “Central Mountain”
A PCA was run once for each variable group (PC, HY, and LU) with the number of variables and samples listed in Table 2 and Annex 1.

Biocoenotic metrics

The number of metrics was reduced before statistical analysis in order to eliminate those metrics, which did not provide i) a sensible range of values and ii) provide redundant information due to high inter-relationship. Therefore, box plots were produced for each metric, and those covering only a small range of values (e. g: STG 2: xylophageous feeding preferences: 0-0.44 %) were deleted from the set. A triangular correlation matrix was produced for the remaining metrics. If metrics correlated with $r > 0.85$, those metrics were excluded from further analysis, that showed the lower overall correlation with other metrics. This procedure also identified metrics with a very low frequency in the dataset.

A total of 90 metrics for STG 1 and 102 metrics for STG 2 remained for further analysis. Proportional (%) metrics were transformed using 'arcsin sqrt x', all other variables were 'log (x+1)'-transformed.

Canonical Ordination (RDA) of metrics and environmental gradients

The link of environmental and biotic variables was realized by direct gradient analysis. Detrended Correspondence Analysis (DCA) identified a short biotic (metric) gradient of 1.31. Therefore, Redundance Analysis (RDA) was the appropriate method to directly analyse the environmental and biotic gradients (ter Braak & Smilauer 2002). A RDA was run for each Stream Type Group to identify the individual strength of the environmental gradients. This was followed by a second RDA for which the dataset was limited to samples of sites affected by the same designated stressor. In addition, the physical-chemical gradient was used as a co-variable if hydromorphological degradation was the designated main stressor and the hydromorphological gradient was used as a co-variable for the analysis of organically polluted sites. Hence, the impact of subordinate stressors was partialled out to focus on the identification of stressor-specific metrics.

All multivariate analysis was run with CANOCO 4.51 (ter Braak & Smilauer2003) and correlations were calculated with STATISTICA 6.1 (StatSoft, 2003).

Final metric selection

The final selection of metrics was realized in three steps:

1. The metrics were ordered according to their RDA "species fit", a measure for the contribution of a metric to the multiple regression of metrics on the environmental variables. The selection was limited to the 50 highest ranking metrics.
2. Each metric was correlated (Pearson product moment) to the individual gradients, whereas the respective sites (and samples) were restricted to only those samples previously allocated to the relevant main stressor. Example: If metrics were correlated with the gradient HY1, the dataset was restricted to sites presumed to be mainly impacted by hydromorphological degradation. Those samples allocated to organic pollution or acidification were excluded.
3. Step 2 was repeated, but stream type-specific. Therefore, the analysis was run for each stream type separately, and the mean, minimum and maximum correlation coefficients (r_{mean}, r_{min}, and r_{max}) were calculated (Annex 2 and 3).

The metrics were ordered according to their correlation with the main gradient in the dataset (HY1 for STG 1, PC1 for STG 2), and the final selection was restricted to the 50 highest ranking metrics. Annex 2 and 3 show the final tables with additional correlation results (Spearman rank) of metrics and pre-/post-classifications.

Validation of environmental gradients

Although multivariate analysis provides an effective and time-saving method to identify the inherent multidimensional structure of different kinds of objects, the results may represent artificial patterns and suggest erroneous conclusions. Therefore, the environmental gradients were compared with a pre-/post classification of the respective sample sites which was based on expert judgement of the field researchers having sampled the streams and, if available, additional knowledge derived from previous studies. Each site was assigned to a quality class (reference = 5, good = 4, moderate = 3, poor = 2, or bad =1) referring to the estimated main stressor's degree of impairment. The validation was checked by Spearman correlation of the stressor-specific pre-classification and the respective gradients represented by the PCA axis values (PC1, PC1a, HY, and LU) (Table 3). Therefore, samples were grouped according to their designated main stressor and correlations were calculated only with the respective environmental gradient. For example, if the main stressor was organic pollution, the samples were correlated with the physical-chemical gradient. During the AQEM project, the pre-classification of most sites was corrected after sampling due to additional abiotic data gained during the field work (physical-chemical measures, site protocol parameters of hydromorphological variables). If available, the pre-classification was replaced by the post-classification. No post-classification was available for STAR sites.

In conclusion, the classification applied is mainly coherent to the "best available classification" that is used as a benchmark within the inter-calibration exercise.

Table 3: Correlation coefficients (r; Spearman rank) of PCA gradient values and pre-/post-classification of sites (see text for explanation). p = level of significance; N = number of valid samples in the analysis

	Gradient	r	p	N
Stream Type Group 1: "Central Lowland"	Physical-chemical (PC1)	-0.144	0.402	36*)
	Physical-chemical (PC1a)	0.180	0.201	52
	Hydromorphology (HY1)	-0.893	< 0.001	160
	Land use (LU2)	0.193	< 0.001	157
Stream Type Group 2: "Central Mountain"	Physical-chemical (PC1)	-0.67	< 0.001	146
	Hydromorphology (HY1)	-0.82	< 0.001	121
	Land use (LU1)	-0.25	< 0.001	365

*) Correlation of gradient and pre-classification only possible for stream type U23.

Discussion of gradient validation

The correlation analysis showed a high correlation between the pre/post-classification and the hydromorphological gradient (HY1) for both stream type groups and for those sites designated to be mainly hydromorphologically impacted (Table 3). Hence, the HY1 gradient fits well the expert judgement on the hydromorphological status of the sites, which confirms the capability to "impartially" indicate hydromorphological degradation by measurable hydromorphological parameters. Vice versa, it may confirm the selection of appropriate parameters for the corresponding gradients targeting the detection of hydromorphological impairment.

The correlation of the PC1 gradient of Stream Type Groups 2 with the organic pollution-based pre-/post-classification (r = -0.670) was fairly high, too. Yet, the correlation coefficient was low for Stream Type Group 1 (-0.144 and 0.180 for PC1 and PC1a, respectively). This means that the Central Lowland dataset probably does not adequately reflect a physical-chemical gradient. The gradient may be to short or, as another explanation, the selected physical-chemical parameters may be inappropriate to measure a pollution gradient.

The land use gradients (except LU1 for STG 1) were comparatively weak: r = 0.193 for Stream Type Group 1 and r = -0.250 for Stream Type Group 2. Therefore, land use seems to be of minor importance within the dataset when compared to the other gradients. However, intensive land use (crop land, pasture) may be a good descriptor for eutrophication as shown below for Stream Type Group 2 (Figure 3).

Results

6.3.3 Stream Type Group 1: "Central Lowland"

Environmental gradients (PCA)

Figure 1a and b show the PCA ordination plots for the physical-chemical variables, which are used here exemplarily to visualize the results. The ordination plot shows a main physical-chemical gradient along (PC1) axis 1 that is characterized by and positively related to the N (NO_2, NO_3, NH_4) and P (diss. PO_4) nutrient components, and the biological oxygen demand (BOD_5). A similar gradient was derived from the PCA of physical-chemical parameters without the oxygen components (PC1a, not shown here) and with 'natural' parameters (pH, alkalinity, total hardness) as co-variables. The main hydromorphological gradient (HY1) was positively related to bank and bed modification, stagnation, scouring, and straightening. The gradient was negatively related to 'natural' variables, such as the number of logs and debris dams and % xylal (wooded debris) on the stream bed, the shaded proportion of the stream bed, and the wooded riparian and floodplain area. The main land use gradient (LU1) was positively related to the proportion of forest, wetland and standing water bodies in both, the floodplain and the catchment area. The other end of the gradient was characterized by the proportion of crop land, pasture and urban settlement/industry and, hence, represents the 'impacted' end. Two more land use gradients were identified, which are almost independent of LU1.

LU2 divides the proportion of grass-/bushland (positive) and pasture (negative) in the catchment and floodplain, and LU3 is positively correlated with the proportion of urban settlement/industry at both spatial scales.

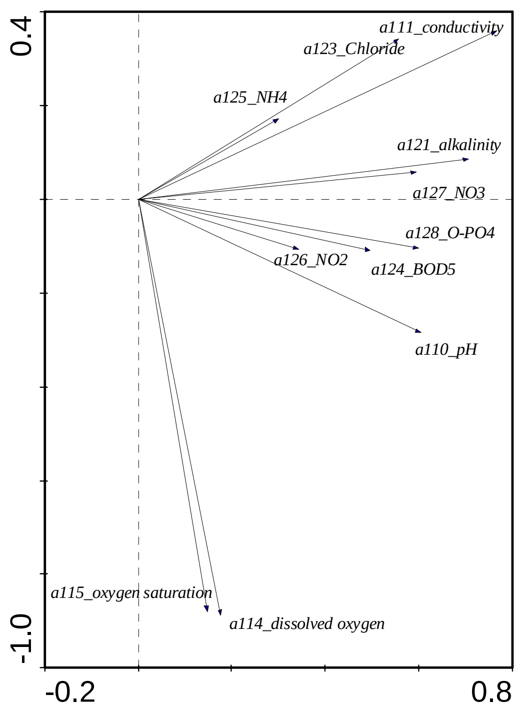


Figure 1a: PCA of eleven physical-chemical variables of Stream Type Group 1, axis 1 *vs*. 2. The PC1 gradient is represented by axis 1.

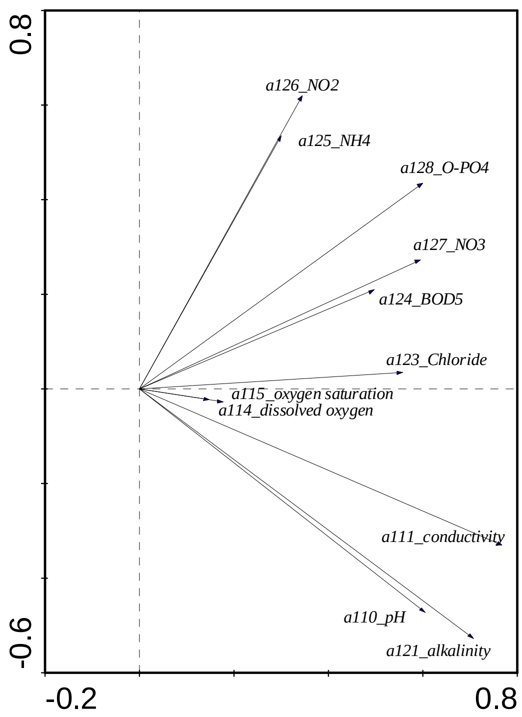


Figure 1b: PCA of eleven physical-chemical variables of Stream Type Group 1, axis 1 *vs*. 3. The PC1 gradient is represented by axis 1.

Linkage of environmental gradients and metrics (RDA)

The RDA was run with 85 metrics, 5 gradients (PC1a, HY1, and LU1-3) and 387 samples. The hydromorphological gradient (HY1) clearly dominated in STG 1 as indicated by high *lambda*-A and F values (Table 4).

Table 4: RDA statistics and results of the forward selection of environmental gradients (Monte Carlo simulation with 1000 permutations). *Lambda*-A is a measure to evaluate the strength of an environmental variable (gradient) in the analysis (ter Braak & Smilauer, 2002).

Gradient		*Lambda*-A	p	F	
HY1 (Hydromorphology)		0.07	0.002	29.69	
LU1 (Land use forest *vs*. crop land)		0.04	0.002	15.89	
LU2 (Land use grass-/bushland *vs*. pasture)					
LU3 (Land use urban settlement/industry)					
PC1a (Physical-chemical)		0.01	0.022	2.13	
	A x e s				
	1	2	3	4	Total variance
Eigenvalues:	0.107	0.023	0.014	0.008	1.000

Species-environment correlations:	0.687	0.544	0.639	0.427	
Cumulative percentage variance					
- of species data:	10.7	13.0	14.4	15.2	
- of species-environment relation:	69.1	83.8	92.7	97.6	
Sum of all eigenvalues					1.000
Sum of all canonical eigenvalues					0,155

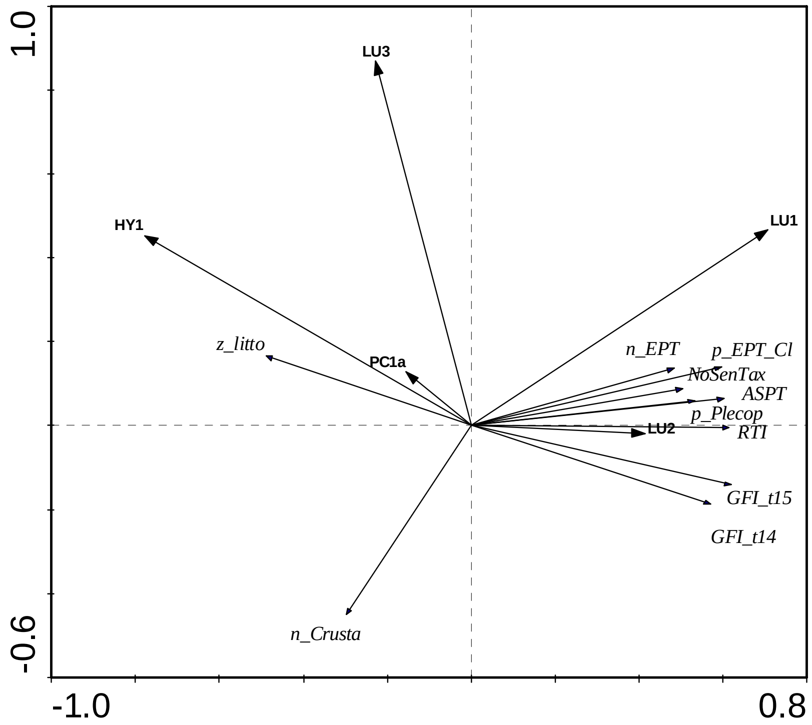


Figure 2: RDA biplot of 85 metrics, 5 gradients, and 387 samples of STG 1. For clarity, the 'species fit' was set to > 25 % in order to show the ten strongest metrics in the analysis. Metric codes: n_EPT = number of Ephemeroptera-Plecoptera-Trichoptera taxa; p_EPT_cl(*) = % EPT based on abundance classes; NoSenTax = number of sensitive taxa; ASPT(*) = Average score per taxon; p_Plecop(*) = % Plecoptera individuals; RTI = Rhithron Typie Index, GFI t14, t15 = German Fauna Index types 14 and 15, n Crusta = number of Crustacea taxa, z_litto = % individuals with littoral preferences. *(* metric is working on family level and, thus, suited for the inter-calibration exercise on a large scale working with existing datasets)*

The German Fauna Indices and the proportion of littoral preferring individuals show the highest relation to the HY1 (Figure 2). These metrics seem to react stressor-specific, whereas the other are also related to the second-strongest LU1 gradient. The Average Score Per Taxon (ASPT), proportion of Plecoptera, Rhithron Typie Index, number of sensitive taxa, and proportion and number of Ephemeroptera-Plecoptera-Trichoptera (EPT) individuals and taxa, respectively. The number of Crustacea taxa was strongly related to LU2.

The 50 highest-scoring metrics for the indication of hydromorphological degradation of the Central Lowland dataset (STG 1) are given in Annex 2.

6.3.4 Stream Type Group 2: "Central Mountains"

Environmental gradients (PCA)

A total of six gradients (PC1, HY1-2, and LU1-3) have been extracted from the PCA gradient analysis. The physical-chemical gradient (PC1) of STG 2 was similar to that of STG 1 and was positively correlated with nutrient components (N, P), electric conductivity, and the biological oxygen demand (BOD_5). The PCA of hydromorphological variables lead to two main gradients: HY1 was positively related to the impact by bed and bank fixation and riparian modification. The other end of the gradient was, for example, connected with the proportion of shaded stream bottom, the number of logs and debris dams, and the width of the wooded riparian vegetation. The degree of flow regulation (stagnation, damming, torrent modification) was positively correlated with HY2. The main land use gradient (LU1) separated between crop land/urban settlement/industry (positive correlation) and forest (negative) for both, catchment and floodplain land use. LU2 divided the dataset into those samples located in catchments/floodplains dominated by pasture and grass-/bushland. The third gradient (LU3) separated the impact of extensive grass-/bushland and crop land.

Linkage of environmental gradients and metrics (RDA)

The RDA was run with 102 metrics, 6 gradients (PC1, HY1-2, and LU1-3) and 295 samples. The physical-chemical gradient (PC1) was clearly dominating in STG 2 which is indicated by very high *lambda*-A and F values (Table 5). In comparison with PC1, the other gradients are fairly weak and reveal the role of the physical-chemical pollution as the main stressor in this Stream Type Group.

Table 5: RDA statistics and results of the forward selection of environmental gradients (Monte Carlo simulation with 1000 permutations). *Lambda*-A is a measure to evaluate the strength of an environmental variable (gradient) in the analysis (ter Braak & Smilauer, 2002).

Variable	*Lambda*-A	p	F
PC1 (Physical-chemical)	0.15	0.002	52.06
HY2 (Stagnation, dams, torrent modification)	0.03	0.002	9.63
HY1 (Bed/bank fixation, riparian, floodplain)	0.02	0.002	8.14
LU1 (Forest *vs.* cropland and urban settlement/industry	0.01	0.002	5.23
LU3 (Grass-/bushland *vs.* crop land)	0.01	0.002	2.76
LU2 (Grass-/bushland vs. pasture and urban settlement/industry	0.01	0.002	2.34

	A x e s				
	1	2	3	4	Total variance
Eigenvalues:	0.08	0.03	0.01	0.01	1.000
Species-environment correlations:	0.57	0.72	0.49	0.45	
Cumulative percentage variance					
- of species data:	7.60	10.80	12.00	12.80	
- of species-environment relation:	57.90	82.30	91.40	97.00	

Sum of all eigenvalues:	1.000
Sum of all canonical eigenvalues:	0.130

The RDA confirms the dominant role of the PC1 gradient in the Central Mountain data (Figure 3). Many metrics were directly related to the gradient, either positive, such as the Hirudinea abundance and the German Saprobic Index new, or negative, such as the proportion and number of EPT individuals and taxa, respectively, the number of Plecoptera taxa, the Average Score Per Taxon, or the German Fauna Indices. Although rather weak in the analysis, the HY1 gradient shows an almost rectangular orientation and, thus, is fairly independent from the PC1 (Figure 3). This is not true for the HY2 and LU1 gradients, which run in nearly the same direction as PC1. Accordingly, higher nutrient concentrations in Central Mountain streams came along with stagnation as well as intensive agricultural land use (crop land). LU2 and LU3 are subordinate.

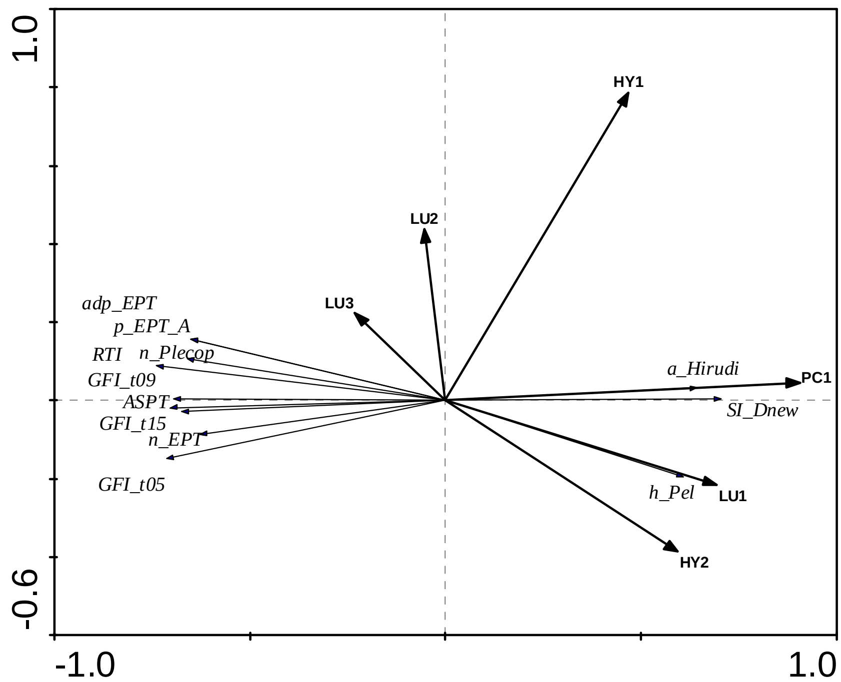


Figure 3: RDA biplot of 102 metrics, 6 gradients, and 295 samples of STG 2. For clarity, the 'species fit' was set to > 40 % in order to show the twelve strongest metrics in the analysis. Metric codes: adp_EPT = number of Ephemeroptera-Plecoptera-Trichoptera taxa, adjusted; p_EPT_A(*) = % EPT Austrian version; n_Plecop = number of Plecoptera taxa; RTI = Rhithron Typie Index; GFI t05, 09, and 14 = German Fauna Index types 05, 09, and 14, respectively; ASPT (*)= Average score per taxon; n_EPT = number of EPT taxa; a_hirudi(*) = Hirudinea abundance; SI_Dnew = Revised German Saprobic Index; h_Pel = % Pelal preferences.
(metric is working on family level and, thus, suited for the inter-calibration exercise on a large scale working with existing datasets)*

The 50 highest-scoring metrics for the indication of organic pollution/eutrophication of the Central Mountain dataset (STG 2) are given in Annex 3.

6.3.5 Conclusion

The data evaluation has proven that it is possible to select "Common Metrics", which are suited to assess environmental degradation within a large geographic area and broadly defined stream types. Furthermore, it is obvious that individual metrics react to different stressors. While most of the metrics strongly correlating to environmental gradients are based on species level, some family-based metrics have a comparatively good performance, too.

For the inter-calibration exercise, "Inter-calibration Common Metrics" (ICM) need to be selected, which will be used for comparing the output of different national assessment systems.

Since the AQEM/STAR dataset is the first pan-European benthic invertebrate data set, we propose to select the ICM for the "Central and Baltic GIG" and, if feasible, also for the "Northern GIG" from the Annexes 1 and 2. These should mainly be restricted to family-based metrics, to allow for comparing data also from those countries, which do not have datasets on species level. Overall, the ICM should include metrics reacting on different stressors.

The species-based metrics, which have proven their ability to detect environmental stress in a large variety of stream types, are a valuable tool for comparing assessment results between countries (restricted to those countries who work on species level). Furthermore, they can be used as a first draft assessment system in countries without a national system.

Cited literature

Armitage P.D., Moss D., Wright J.F. & Furse M.T. (1983) The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. *Water Research*, **17**, 333-347.

Hering D., Moog O., Sandin L. & Verdonschot P.F.M. (2004) Overview and application of the AQEM assessment system. *Hydrobiologia*, **516**, 1-20.

Lorenz A., Hering D., Feld C.K. & Rolauffs P. (2004) A new method for assessing the impact of hydromorphological degradation on the macroinvertebrate fauna of five German stream types. *Hydrobiologia*, **516**, 107-127.

Moog O. (ed) (1995) Fauna Aquatica Austriaca. Wasserwirtschaftskataster, Bundesministerium für Land- und Forstwirtschaft, Wien.

Podraza P., Schuhmacher H. & Sommerhäuser M. (2000) Composition of macroinvertebrate feeding groups as a bioindicator in running waters. *Proceedings of the International Association of Theoretical and Applied Limnology*, **27**, 3066-3069.

Rolauffs P., Stubauer I., Zahrádkova Z., Brabec K. & Moog O. (2004) Integration of the Saprobic System into the Water Framework Directive approach. *Hydrobiologia*, **516**, 285-298.

Schweder H. (1992) Neue Indices für die Bewertung des ökologischen Zustandes von Fließgewässern, abgeleitet aus der Makroinvertebraten-Ernährungstypologie. *Limnologie Aktuell*, **3** (eds Friedrich G. & Lacombe J.), pp. 353-377. Gustav Fischer, Stuttgart-Jena-New York.

StatSoft, Inc. (2003) STATISTICA (data analysis software system), version 6.1.

ter Braak C.J.F. & Smilauer P. (2002) *CANOCO reference manual and CanoDraw for Windows user's guide version 4.5*. Biometris – Plant Research International, Wageningen and České Budějovice.

ter Braak C.J.F. & Smilauer P. (2003) *CANOCO for Windows version 4.51*. Biometris – Plant Research International, Wageningen.

Zelinka M. & Marvan P. (1961) Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer. *Archiv für Hydrobiologie*, **57**, 389-407.

Annex 1: Table of environmental variables used for the different PCA gradient analysis. "+" indicates variable's usage for the Stream Type Groups. Variables are allocated to the variable groups: LU = land use; HY = hydromorphology, PC = physical-chemical.

Environmental variable	Stressor gradient	STG 1	STG 2
a19_91_% Forest catchment	LU	+	+
a19_4_% Wetland catchment	LU	+	
a19_5_% Grass-/bushland catchment	LU	+	+
a19_9_% Standing water bodies catchment	LU	+	
a19_12_% Crop land catchment	LU	+	+
a19_13_% Pasture catchment	LU	+	+
a19_92_% Urban settlement/industry catchment	LU	+	+
a24_1_Permanent flowing (y/n)	HY	+	
a25_Lakes in the stream continuum (y/n)	HY	+	
s26_Floodplain width [m]	HY	+	+
s26_2_Flood prone area width [m]	HY	+	
s26_3_Entrenchment depth [m]	HY	+	
s26_5_Mean depth [m]	HY	+	+
a30_91_% Forest floodplain	LU	+	+
a30_4_% Wetland floodplain	LU	+	+
a30_5_% Grass-/bushland floodplain	LU	+	+
a30_9_% Standing water bodies floodplain	LU	+	
a30_12_% Crop land floodplain	LU	+	+
a30_13_% Pasture floodplain	LU	+	+
a30_92_% Urban settlement/industry floodplain	LU	+	+
a69_Shading at zenith (foliage cover) [%]	HY	+	+
a70_Width wooded riparian vegetation [m]	HY	+	+
a71_Meandering (y/n)	HY	+	+
a71_Sinuate (y/n)	HY	+	+
a71_Constrained (y/n)	HY	+	+
a71_Anabranching (y/n)	HY	+	+
a71_6_Artificially constrained (y/n)	HY	+	+
a73_Standing water bodies in the floodplain (y/n)	HY	+	
a74_No. of debris dams (> 0.3 m³)	HY	+	+
a75_No. of logs (>10 cm diameter)	HY	+	+
a76_Riparin wooded vegetation [%	HY	+	+

Environmental variable	Stressor gradient	STG 1	STG 2
length]			
a77_Dams (y/n)	HY		+
a79_91_Bank fixation concrete [%]	HY	+	+
a79_92_Bank fixation stones [%]	HY	+	+
a79_93_Unfixed banks [%]	HY	+	+
a80_91_Bed fixation concrete [%]	HY	+	+
a80_92_Bed fixation stones [%]	HY	+	+
a80_93_no Bed fixation [%]	HY	+	+
a81_Stagnation (y/n)	HY	+	+
a82_Torrent modification	HY		+
a84_Straightening	HY	+	+
a87_Scouring	HY	+	+
a93_Lack of natural floodplain vegetation (y/n)	HY	+	+
a103_Megalithal [%]	HY	+	+
a103_Macrolithal [%]	HY	+	+
a103_Mesolithal [%]	HY	+	+
a103_Microlithal [%]	HY	+	+
a103_Akal [%]	HY	+	+
a103_Psammal [%]	HY	+	+
a103_Argyllal [%]	HY	+	+
a104_Algae [%]	HY	+	+
a104_Submerged macrophytes [%]	HY	+	+
a104_Emerged macrophytes [%]	HY	+	+
a104_Living parts of terrestrial plants [%]	HY	+	+
a104_Xylal [%]	HY	+	+
a104_CPOM [%]	HY	+	+
a104_FPOM [%]	HY	+	+
a105_Average stream width [m]	HY	+	+
a110_pH	PC	+	+
a111_Electric conductivity [µS/cm]	PC	+	+
a114_Dissolved oxygen [mg/l]	PC	+	+
a115_Oxygen saturation [%]	PC	+	+
a121_Alkalinity [mmol/l]	PC	+	+
a122_Total hardness [mmol/l]	PC	+	+
a123_Chloride [mg/l]	PC	+	+
a124_BOD5 [mg/l]	PC	+	+
a125_NH4 [mg/l]	PC	+	+
a126_NO2 [mg/l]	PC	+	+
a127_NO3 [mg/l]	PC	+	+
a128_Ortho-PO4 [µg/l]	PC	+	+
a129_Total PO4 [mg/l]	PC	+	+

Annex 2: Table of the 50 highest ranking metrics for the identification of the impact of hydromorphological degradation (HY1) in the Central Lowland Stream Type Group (STG 1). The metrics were ranked according to their correlation (Pearson product moment; r) with the main gradient HY1. In addition the metric's correlation with the five-class pre-/post-classification (Spearman rank; r) and the respective significance levels (p) are given. The last three columns list stream type-specific correlations (Pearson product moment) of metric values with the main gradient as mean, maximum, and minimum values of the individual stream types.
Bold metrics work on family level and, thus, are suited for the inter-calibration exercise on a large scale working with existing datasets

			Metric with HY1		Metric with pre-/post-classification		Metric with HY1: stream type-specific		
Order	Metric short	Metric name	r	p	r	p	r mean	r max	r min
1	GFI_T15	German Fauna Index D03 (Lorenz et al., 2004)	-0,80	< 0,001	0,77	< 0,001	-0,78	-0,87	-0,67
2	GFI_T14	German Fauna Index D01 (Lorenz et al., 2004)	-0,80	< 0,001	0,80	< 0,001	-0,83	-0,84	-0,81
3	GFI_T09	German Fauna Index D05 (Lorenz et al., 2004)	-0,60	< 0,001	0,60	< 0,001	-0,61	-0,65	-0,59
4	**ASPT**	Average score per Taxon (Armitage et al., 1983)	-0,58	< 0,001	0,64	< 0,001	-0,65	-0,88	-0,52
5	Z_LITTO	[%] Littoral preferences (Moog, 1995)	0,57	< 0,001	-0,68	< 0,001	0,59	0,75	0,38
6	SI_DNEW	German Saprobic Index new (Rolauffs et al., 2004)	0,56	< 0,001	-0,74	< 0,001	0,66	0,84	0,51
7	C_RP	[%] Rheophilic preferences (Moog, 1995)	-0,56	< 0,001	0,64	< 0,001	-0,51	-0,75	-0,33
8	RTI	Rhithron Typie Index	-0,56	< 0,001	0,71	< 0,001	-0,66	-0,77	-0,45
9	SI_ZM	Saprobic Index (Zelinka & Marvan, 1961)	0,54	< 0,001	-0,69	< 0,001	0,64	0,69	0,58
10	GFI_T05	German Fauna Index D04 Lorenz et al., 2004)	-0,51	< 0,001	0,48	< 0,001	-0,63	-0,78	-0,48
11	SIZM_OLI	[%] Oligosaprobic valences (Moog, 1995)	-0,51	< 0,001	0,65	< 0,001	-0,62	-0,72	-0,45
12	H_AKLIPS	[%] Type Akal + Lithal + Psammal preferences	-0,51	< 0,001	0,47	< 0,001	-0,46	-0,63	-0,18
13	**P_EPT_CL**	[%] EPT (abundance classes)	-0,50	< 0,001	0,55	< 0,001	-0,62	-0,87	-0,42
14	**LOG10_SE**	Log selected taxa [%]	-0,49	< 0,001	0,61	< 0,001	-0,50	-0,55	-0,41
15	SI_CZ	Czech Saprobic Index	0,48	< 0,001	-0,66	< 0,001	0,62	0,79	0,49
16	NOSENTAX	Number of sensitive taxa (Austria)	-0,47	< 0,001	0,65	< 0,001	-0,50	-0,76	-0,31
17	BIOREG_A	Index of Biocoenotic Region (Austria)	0,47	< 0,001	-0,59	< 0,001	0,54	0,73	0,35
18	C_IN	[%] Indifferent current	0,47	< 0,001	-0,61	< 0,001	0,47	0,58	0,29

			Metric with HY1		Metric with pre-/ post-classification		Metric with HY1: stream type-specific		
Order	Metric short	Metric name	r	p	r	p	r mean	r max	r min
		preferences (Moog, 1995)							
19	N_DIPTER	Number of Diptera taxa	-0,44	< 0,001	0,53	< 0,001	-0,43	-0,58	-0,18
20	H_PEL	[%] Pelal preferences (Moog, 1995)	0,43	< 0,001	-0,51	< 0,001	0,40	0,54	0,14
21	H_AKA	[%] Akal preferences (Moog, 1995)	-0,42	< 0,001	0,39	< 0,001	-0,37	-0,50	-0,17
22	Z_MEPOT	[%] Metapotamal preferences (Moog, 1995)	0,42	< 0,001	-0,55	< 0,001	0,46	0,56	0,39
23	GFI_T11	German Fauna Index D02 (Lorenz et al., 2004)	-0,41	< 0,001	0,44	< 0,001	-0,39	-0,45	-0,33
24	BBI	Belgian Biotic Index	-0,41	< 0,001	0,53	< 0,001	-0,43	-0,67	-0,23
25	Z_HYRHIT	[%] Hyporhithral preferences (Moog, 1995)	-0,41	< 0,001	0,55	< 0,001	-0,43	-0,62	-0,15
26	Z_MERHIT	[%] Metarhithral preferences (Moog, 1995)	-0,39	< 0,001	0,57	< 0,001	-0,50	-0,61	-0,31
27	SIZM_BME	[%] Beta-mesosaprobic valences (Moog, 1995)	-0,39	< 0,001	0,44	< 0,001	-0,42	-0,45	-0,36
28	RHEOIND	Rheoindex Banning (abundance)	-0,39	< 0,001	0,43	< 0,001	-0,46	-0,73	-0,22
29	F_GATHCO	[%] Gatherers/collectors (Moog, 1995)	0,39	< 0,001	-0,44	< 0,001	0,30	0,55	0,06
30	N_EPT	Number of EPT taxa	-0,38	< 0,001	0,14	< 0,001	-0,44	-0,75	-0,25
31	**BMWP**	Biological Monitoring Working Party (Armitage et al., 1993)	-0,37	< 0,001	0,49	< 0,001	-0,39	-0,77	-0,11
32	SI_NL	Dutch Saprobic Index	0,36	< 0,001	-0,26	< 0,001	0,15	0,53	-0,18
33	**IBE**	IBE	-0,35	< 0,001	0,57	< 0,001	-0,39	-0,67	-0,12
34	Z_HYPPOT	[%] Hypopotamal preferences (Moog, 1995)	0,35	< 0,001	-0,59	< 0,001	0,41	0,53	0,32
35	**P_EPT**	[%] EPT taxa	-0,35	< 0,001	0,39	< 0,001	-0,48	-0,62	-0,23
36	H_PHY	[%] Phytal preferences (Moog, 1995)	0,34	< 0,001	-0,11	0,172	0,21	0,46	-0,17
37	H_LIT	[%] Lithal preferences (Moog, 1995)	-0,34	< 0,001	0,44	< 0,001	-0,31	-0,56	-0,05
38	**P_TRICHO**	[%]Trichoptera	-0,33	< 0,001	0,38	< 0,001	-0,33	-0,50	-0,12
39	C_LP	[%] Limnophilic preferences (Moog, 1995)	0,32	< 0,001	-0,39	< 0,001	0,38	0,57	0,26
40	N_GASTRO	Number of Gastropoda taxa	0,32	< 0,001	-0,29	< 0,001	0,33	0,43	0,17
41	**N_FAMIL**	Number of Families	-0,30	< 0,001	0,41	< 0,001	-0,32	-0,66	-0,02
42	**P_PLECOP**	[%] Plecoptera	-0,30	< 0,001	0,63	< 0,001	-0,43	-0,51	-0,34

			Metric with HY1		Metric with pre-/post-classification		Metric with HY1: stream type-specific		
Order	Metric short	Metric name	r	p	r	p	r mean	r max	r min
43	F_ACTFIL	[%] Active filter feeders (Moog, 1995)	0,28	< 0,001	-0,42	< 0,001	0,31	0,60	0,03
44	RETI	Rhithron Feeding Type Index (Schweder, 1992; Podraza et al., 2000)	-0,27	< 0,001	0,32	< 0,001	-0,32	-0,35	-0,29
45	NO_TAXA	Number of taxa	-0,27	< 0,001	0,37	< 0,001	-0,28	0,06	-0,64
46	C_RL	[%] Rheo-limnophilic preferences (Moog, 1995)	0,26	< 0,001	-0,26	< 0,001	0,13	0,51	-0,34
47	H_STONES	[%] Stone-dwellers	-0,26	< 0,001	0,47	< 0,001	-0,32	-0,84	0,17
48	**P_BIVALV**	[%] Bivalvia	0,26	< 0,001	-0,41	< 0,001	0,27	0,55	0,04
49	P_INDEX	Portuguese Index	-0,26	< 0,001	0,23	< 0,001	-0,32	-0,41	-0,25
50	H_POM	[%] Particulate Organic Matter preferences (Moog, 1995)	-0,25	< 0,001	0,07	0,394	-0,23	0,42	-0,75

Annex 3: Table of the 50 highest ranking metrics for the identification of the impact of organic pollution/eutrophication in the Central Mountain Stream Type Group (STG 2). The metrics were ranked according to their correlation (Pearson product moment; r) with the main gradient PC1. In addition the metric's correlation with the five-class pre-/post-classification (Spearman rank; r) and the respective significance levels (p) are given. The last three columns list stream type-specific correlations (Pearson product moment) of metric values with the main gradient as mean, maximum, and minimum values of the individual stream types.
Bold metrics work on family level and, thus, are suited for the inter-calibration exercise on a large scale working with existing datasets

			Metric with PC1		Metric with pre-/post-classification		Metric with PC1: stream type-specific		
Order	Metric short	Metric Name	R	p	R	p	R mean	R max	R min
1	GFI_T05	German Fauna Index D04 (Lorenz et al., 2004)	-0,73	< 0.001	0,73	< 0.001	-0,74	-0,81	-0,54
2	SI_DNEW	German Saprobic Index new (Rolauffs et al., 2004)	0,76	< 0.001	-0,77	< 0.001	0,73	0,85	0,56
3	GFI_T09	German Fauna Index D05 (Lorenz et al., 2004)	-0,73	< 0.001	0,65	< 0.001	-0,71	-0,84	-0,54
4	RTI	Rhithron Typie Index	-0,70	< 0.001	0,75	< 0.001	-0,71	-0,81	-0,59
5	GFI_T15	German Fauna Index D03 (Lorenz et al., 2004)	-0,69	< 0.001	0,73	< 0.001	-0,69	-0,76	-0,47
6	**1-GOLD**	1-relative abundance Gastropoda, Oligochaeta, and Diptera	-0,66	< 0.001	0,60	< 0.001	-0,68	-0,74	-0,57
7	H_LIT	[%] Lithal preferences (Moog, 1995)	-0,62	< 0.001	0,67	< 0.001	-0,67	-0,83	-0,51
8	H_STONES	[%] Stone dwellers	-0,62	< 0.001	0,67	< 0.001	-0,67	-0,83	-0,51
9	GFI_T14	German Fauna Index D01 (Lorenz et al., 2004)	-0,64	< 0.001	0,68	< 0.001	-0,66	-0,78	-0,58
10	NOSENTAX	Number of sensitive taxa (Austria)	-0,64	< 0.001	0,76	< 0.001	-0,66	-0,81	-0,45
11	**ASPT**	Average score per Taxon (Armitage et al., 1983)	-0,73	< 0.001	0,68	< 0.001	-0,66	-0,86	-0,29
12	RETI	Rhithron Feeding Type Index (Schweder, 1992; Podraza et al., 2000)	-0,63	< 0.001	0,59	< 0.001	-0,65	-0,75	-0,50
13	**P_EPT**	[%] EPT taxa	-0,68	< 0.001	0,73	< 0.001	-0,65	-0,81	-0,30
14	BBI	Belgian Biotic Index	-0,65	< 0.001	0,61	< 0.001	-0,65	-0,75	-0,49
15	H_AKLIPS	[%] Type Akal + Lithal + Psammal	-0,54	< 0.001	0,56	< 0.001	-0,64	-0,78	-0,51
16	C_RP	[%] Rheophilic preferences (Moog, 1995)	-0,54	< 0.001	0,52	< 0.001	-0,61	-0,71	-0,51
17	IBE	IBE	-0,60	< 0.001	0,61	< 0.001	-0,60	-0,75	-0,53
18	N_EPT	Number of EPT taxa	-0,56	< 0.001	0,69	< 0.001	-0,59	-0,78	-0,34
19	N_COLEOP	Number of Coleoptera taxa	-0,60	< 0.001	0,56	< 0.001	-0,59	-0,72	-0,43
20	SI_NL	Dutch Saprobic Index	-0,54	< 0.001	0,49	< 0.001	-0,58	-0,78	-0,33

Order	Metric short	Metric Name	Metric with PC1 R	Metric with PC1 p	Metric with pre-/post-classification R	Metric with pre-/post-classification p	Metric with PC1: stream type-specific R mean	R max	R min
21	**BMWP**	Biological Monitoring Working Party (Armitage et al., 1993)	-0,58	< 0.001	0,61	< 0.001	-0,58	-0,74	-0,29
22	SIZM_OLI	[%] Oligosaprobic valences (Moog, 1995)	-0,58	< 0.001	0,62	< 0.001	-0,58	-0,70	-0,37
23	N_EPT_DI	Number of EPT / Diptera taxa	-0,44	< 0.001	0,67	< 0.001	-0,57	-0,70	-0,36
24	**LOG10SEL**	Log selected taxa (ICM)	-0,58	< 0.001	0,70	< 0.001	-0,56	-0,81	-0,29
25	**P_EPT**	[%] EPT taxa	-0,48	< 0.001	0,50	< 0.001	-0,55	-0,77	-0,33
26	N_EPHEME	Number of Ephemeroptera taxa	-0,55	< 0.001	0,60	< 0.001	-0,55	-0,73	-0,39
27	N_PLECOP	Number of Plecoptera taxa	-0,49	< 0.001	0,59	< 0.001	-0,55	-0,69	-0,28
28	**P_EP**	[%] Ephemeroptera-Plecoptera	-0,46	< 0.001	0,39	< 0.001	-0,55	-0,84	-0,29
29	RHEOIND	Rheoindex Banning (abundance)	-0,49	< 0.001	0,55	< 0.001	-0,54	-0,75	-0,29
30	Z_HYRHIT	[%] Hyporhithral preferences (Moog, 1995)	-0,49	< 0.001	0,50	< 0.001	-0,54	-0,69	-0,37
31	N_PLETRI	Number of Plecoptera + Trichoptera taxa	-0,49	< 0.001	0,63	< 0.001	-0,52	-0,77	-0,25
32	**N_FAMIL**	Number of Families	-0,46	< 0.001	0,51	< 0.001	-0,49	-0,65	-0,28
33	**P_EPHEME**	[%] Ephemeroptera	-0,39	< 0.001	0,32	< 0.001	-0,49	-0,81	-0,24
34	**P_COLEOP**	[%] Coleoptera	-0,41	< 0.001	0,59	< 0.001	-0,49	-0,59	-0,36
35	**P_PLECOP**	[%] Plecoptera	-0,47	< 0.001	0,60	< 0.001	-0,47	-0,53	-0,43
36	N_TRICHO	Number of Trichoptera taxa	-0,41	< 0.001	0,57	< 0.001	-0,46	-0,73	-0,15
37	Z_EPIRHI	Epirhithral preferences [%] (Moog, 1995)	-0,45	< 0.001	0,61	< 0.001	-0,45	-0,53	-0,23
38	**A_PLECOP**	Abundance Plecoptera	-0,44	< 0.001	0,58	< 0.001	-0,41	-0,52	-0,25
39	SIZM_XEN	[%] Xenosaprobic preferences (Moog, 1995)	-0,48	< 0.001	0,58	< 0.001	-0,39	-0,51	-0,10
40	F_XYSHFI	[%] Xylophageous + shredders + active filterers + passive filterers	-0,48	< 0.001	0,51	< 0.001	-0,36	-0,75	-0,10
41	F_SHRED	[%] Shredders (Moog, 1995)	-0,41	< 0.001	0,47	< 0.001	-0,31	-0,63	-0,13
42	BIOREG_A	Index of Biocoenotic Region (Austria)	0,41	< 0.001	-0,59	< 0.001	0,41	0,61	0,20

			Metric with PC1		Metric with pre-/post-classification		Metric with PC1: stream type-specific		
Order	Metric short	Metric Name	R	p	R	p	R mean	R max	R min
43	H_POM	[%] Particulate Organic Matter preferences (Moog, 1995)	0,58	< 0.001	-0,51	< 0.001	0,49	0,65	0,12
44	N_OD_TOT	[%] Oligochaeta + Diptera	0,64	< 0.001	-0,62	< 0.001	0,60	0,74	0,27
45	SI_ZM	Saprobic Index (Zelinka & Marvan, 1961)	0,62	< 0.001	-0,68	< 0.001	0,61	0,70	0,39
46	**P_OLIGOC**	[%] Oligochaeta + Diptera	0,67	< 0.001	-0,57	< 0.001	0,62	0,82	0,17
47	C_IN	[%] Indifferent current preferences (Moog, 1995)	0,44	< 0.001	-0,48	< 0.001	0,62	0,79	0,46
48	SI_CZ	Czech Saprobic Index	0,65	< 0.001	-0,68	< 0.001	0,64	0,76	0,45
49	H_PEL	[%] Pelal preferences (Moog, 1995)	0,52	< 0.001	-0,56	< 0.001	0,64	0,84	0,47
50	F_GATHCO	[%] Gatherers/collectors (Moog, 1995)	0,69	< 0.001	-0,64	< 0.001	0,67	0,83	0,47

7. Comparison

7.1 Direct comparison: Same sample, different calculation method

Direct comparison of class boundary values of national bioassessment methods based on AQEM/STAR data using bilateral correlation and regression

Introduction

For large geographic regions comprising several countries, which assessment systems are different in terms of taxonomic resolution and general approach, the intercalibration using "Intercalibration Common Metrics" (ICM) (Buffagni & Erba, 2004) is a suited procedure. However, it might result difficult to explain to water managers and the general public in the short period of the WFD IC process.
Thus, we outline an alternative, which is based on a simple comparison of assessments results from national assessment systems, without using the ICM-tool. The "direct comparison approach" can be used within transboundary river catchments and could also serve as an alternative or validation of the pan-European ICM approach.

In this chapter the "direct comparison approach" is exemplified on the basis of four case studies comprising assessment methods of altogether nine countries, using benthic invertebrates and macrophytes, and covering the common intercalibration stream types R-C3 and R-C4.

Methods

The procedure outlined in the following is the classical approach of methods' comparison conducted by various authors (e.g. Tittizer 1976, Rico et al. 1992, Friedrich et al. 1995). The consideration of a common stream typology and stream type-specific reference values to compare on the basis of Ecological Quality Ratios (EQRs) represent innovations to this approach.
In general, the "direct comparison approach" is very simple: Two different assessment systems (System A and System B) are calculated with a number of samples. The results are compared by a regression which leads to a "conversion formula" from System A into System B and vice versa.

In particular, the "direct comparison approach" comprises the following steps:

1. ***Compilation of a single test dataset including samples taken at a common stream type in various countries.***
   Benthic macroinvertebrate and macrophyte samples of the stream type groups "lowland" and "mountain" taken in the AQEM and STAR project are used (Table 1, see chapter 6.3 for details). These stream type groups correspond to the common intercalibration types R-C3 (small-sized, mid-altitude streams of siliceous geology) and R-C4 (medium-sized, lowland streams of mixed geology) according to CIS WG 2.A Ecological Status (ECOSTAT) 2004.
2. ***Calculation of index values of all methods included in the comparison for each sample in the dataset.***
   Benthic Invertebrates: For stream type R-C3 six and for stream type R-C4 five assessment indices are compared, respectively. Table 1 specifies the number of samples per country and the assessment indices. All samples taken at the same common intercalibration type are used for calculation of each index disregarding the sample's country-specific affiliation. In addition, assessment indices from Poland and

the United Kingdom are included in the analysis of R-C3, although the dataset does not comprise samples from these countries.

Both absolute index values and EQR values are calculated. The 95th percentile of all AQEM/STAR samples taken at sites of a common stream type which have been pre-classified as high status are chosen as reference values (see chapter 6.3 for details).

Macrophytes: For both stream types R-C3 and R-C4 three assessment indices are calculated (Table 1). EQR values are derived by using the 95th percentile value of each index based on all STAR samples.

3. ***Correlation and regression of index values of two assessment indices at a time.***
   Since the values of all indices are non-normally distributed Spearman rank correlation is applied.
4. ***Calculating regression formulae for correlations of all indices included in the comparison.***
5. ***Comparison of nationally defined class boundary values through conversion into respective national method-scale using regression formulae.***
   For the comparison of quality classes the high|good and good|moderate boundary values are expressed as EQR values, following the WFD requirements. Furthermore, this allows for integration of assessment methods specifying their quality class boundaries in EQR values (e.g. British ASPT, ASPT and DSFI applied in Sweden).
   To illustrate the discrepancies of the nationally defined quality classes all index values are correlated against the British ASPT (Benthic Invertebrates) or French IBMR (Macrophytes) as benchmark systems and boundary values are converted into the corresponding values of the benchmark system.

Table 1: Overview of assessment methods included in the class boundary comparison

biological quality element	common IC type	country	number of samples	assessment method	reference
Benthic Macroinvertebrates	R-C3 – small-sized, mid-altitude, siliceous geology	Austria	36	SI (AT) – Austrian Saprobic Index	Moog et al. 1999
		Czech Republic	100	SI (CZ) – Czech Saprobic Index	CSN 757716 1998
		Germany	110	SI (DE) – German Saprobic Index	Friedrich & Herbst 2004
		Poland	-	BMWP (PL) – Polish BMWP	*unpublished*
		Slovak Republic	48	SI (SK) – Slovak Saprobic Index	STN (Slovenská Technická Norma) 83 0532-1 to 8 1978/79
		United Kingdom	-	ASPT (UK) - Average Score Per Taxon	Armitage et al. 1983
	R-C4 – medium-sized, lowland, mixed geology	Denmark	46	DSFI (DK) – Danish Stream Fauna Index	Skriver et al. 2000
		Germany	86	SI (DE) – German Saprobic Index	Friedrich & Herbst 2004
		Sweden	79	ASPT (SE)- Average Score Per Taxon applied in Sweden DSFI (SE) – Danish Stream Fauna Index applied in Sweden	Swedish Environmental Protection Agency 2000
		United Kingdom	36	ASPT (UK) - Average Score Per Taxon	Armitage et al. 1983
Macrophytes	R-C3 – small-sized, mid-altitude, siliceous geology	France	in total 47 samples from sites in Austria, Czech Republic, Germany, Slovak Republic	IBMR (FR) – Indice Biologique Macrophytique en Rivière	AFNOR (Association Française de Normalisation) 2002
		Germany		RI-Moose (DE) – Reference Index (only mosses)	Schaumburg et al. 2004
		United Kingdom		MTR (UK) – Mean Trophic Ranking	Holmes et al. 1999

	R-C4 – medium-sized, lowland, mixed geology	France	in total 126 samples from sites in Denmark, Germany, Latvia, Poland, Sweden, United Kingdom	IBMR (FR) – Indice Biologique Macrophytique en Rivière	AFNOR (Association Française de Normalisation) 2002
		Germany		RI (DE) – Reference Index	Schaumburg et al. 2004
		United Kingdom		MTR (UK) – Mean Trophic Ranking	Holmes et al. 1999

7.1.1 Examples of the "direct comparison approach" based on AQEM/STAR data *- Benthic Invertebrates*

Benthic Invertebrates

R-C3 - small-sized, mid-altitude streams of siliceous geology

Correlation and regression
The correlation of the six assessment indices shows Spearman coefficients ranging from r = 0.86 (Austrian SI and German SI) to r = - 0.34 (Slovak SI and British ASPT). Correlation coefficients and diagrams as well as a matrix of regression formulae based on both absolute and EQR values are listed in Annex 1.1.

Reference values
For the Austrian and Czech indices reference values derived from the AQEM/STAR high status sites are lower (=representing higher quality) than the nationally defined references. The German SI shows a lower saprobic basic condition in the national definition (Table 2). Official reference values for the Slovak, British and Polish methods are not available.

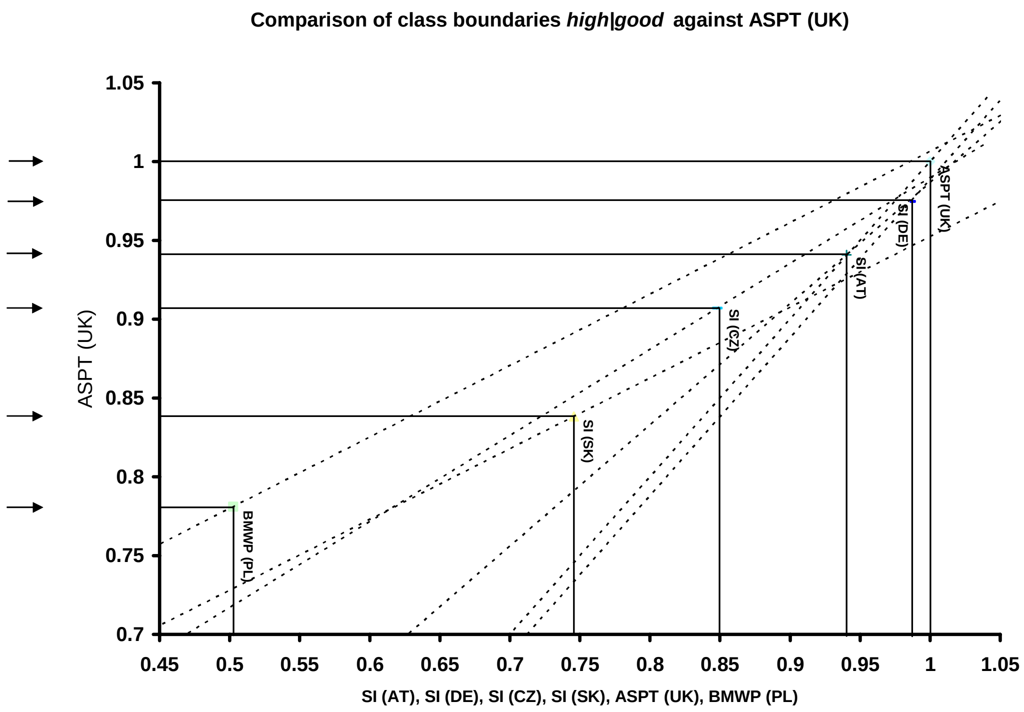
Table 2: Nationally defined and 95th percentile reference values (n.a. – not available)

	SI (AT)	**SI (DE)**	**SI (CZ)**	**SI (SK)**	**ASPT (UK)**	**BMWP (PL)**
nationally defined	1.5	1.25	1.2	n.a.	n.a.	n.a.
95^{th} percentile	1.34	1.36	0.70	1.04	7.49	199

Comparison of class boundary values
The direct comparison of EQR class boundary values reveals major discrepancies between the nationally defined values for both the high|good and good|moderate boundaries (Table 3). To compare the quality classes the boundary values of all indices are converted into values of the ASPT-scale (Figure 1).
For the high|good status boundary the largest deviation amounts to >0.2 ASPT-EQR units between ASPT (UK) and BMWP (PL). The smallest difference is between ASPT (UK) and SI (DE) (0.025 ASPT-EQR units).
The largest good|moderate class boundary value deviation of 0.177 units is observed between ASPT (UK) and BMWP (PL). For this boundary SI (AT) and SI (DE) show nearly similar values (difference of 0.002 units).

Table 3: Predicted values of high|good and good|moderate boundary values

R-C3			SI (AT)	SI (DE)	SI (CZ)	SI (SK)	ASPT (UK)	BMWP (PL)
predicted	high|good	SI (AT)	**0.940**	0.944	0.891	0.798	0.931	0.799
		SI (DE)	0.955	**0.986**	0.920	0.843	0.974	0.820
		SI (CZ)	0.879	0.900	**0.848**	0.728	0.881	0.725
		SI (SK)	0.924	0.925	0.875	**0.746**	0.903	0.781
		ASPT (UK)	0.941	0.975	0.907	0.838	**1.000**	0.781
		BMWP (PL)	0.821	0.872	0.766	0.652	0.903	**0.503**
	good|moderate	SI (AT)	**0.715**	0.761	0.836	0.750	0.867	0.756
		SI (DE)	0.775	**0.777**	0.868	0.805	0.899	0.770
		SI (CZ)	0.650	0.677	**0.757**	0.674	0.806	0.673
		SI (SK)	0.681	0.734	0.809	**0.675**	0.845	0.742
		ASPT (UK)	0.767	0.765	0.857	0.807	**0.890**	0.713
		BMWP (PL)	0.536	0.537	0.683	0.599	0.736	**0.352**



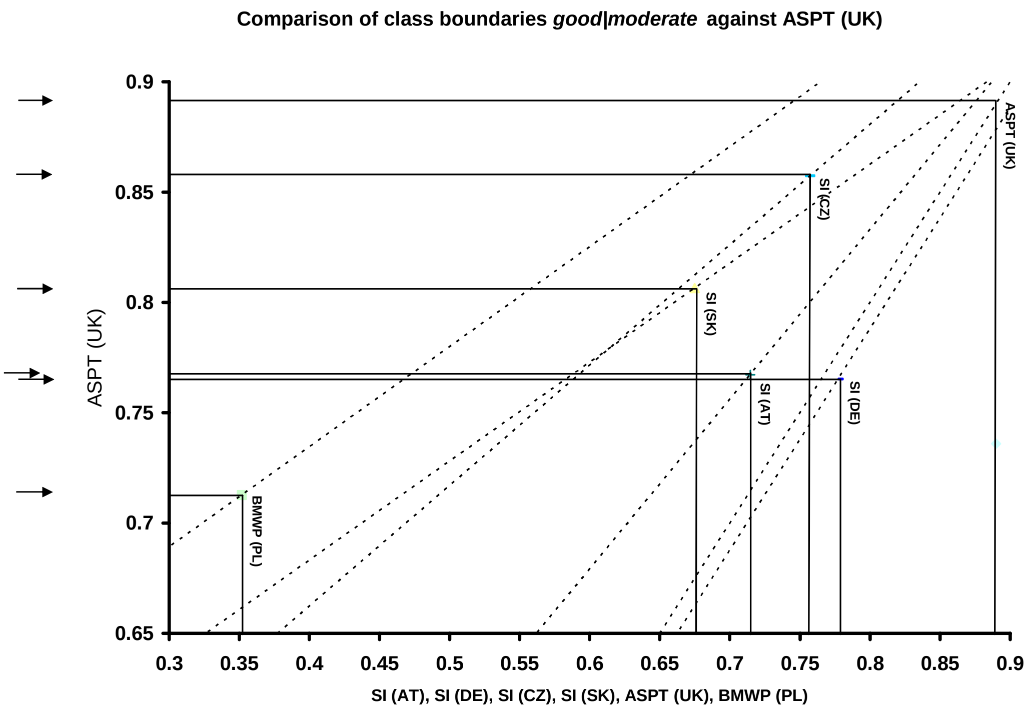


Figure 1: Class boundary comparisons through conversion of national boundaries into ASPT-EQR units using regression lines

R-C4 – medium-sized, lowland, mixed geology

Correlation and regression
Spearman correlation coefficients range from 0.79 (ASPT and DSFI) to -0.75 (German SI and DSFI). Annex 1.2 displays the correlation table and diagram, and lists regression formulae.

Reference values
The derivation of reference values using the 95th percentile of AQEM/STAR high status sites results in references of higher quality for all compared indices except DSFI (DK) (Table 4).

Table 4: Nationally defined and 95th percentile reference values (n.a. – not available)

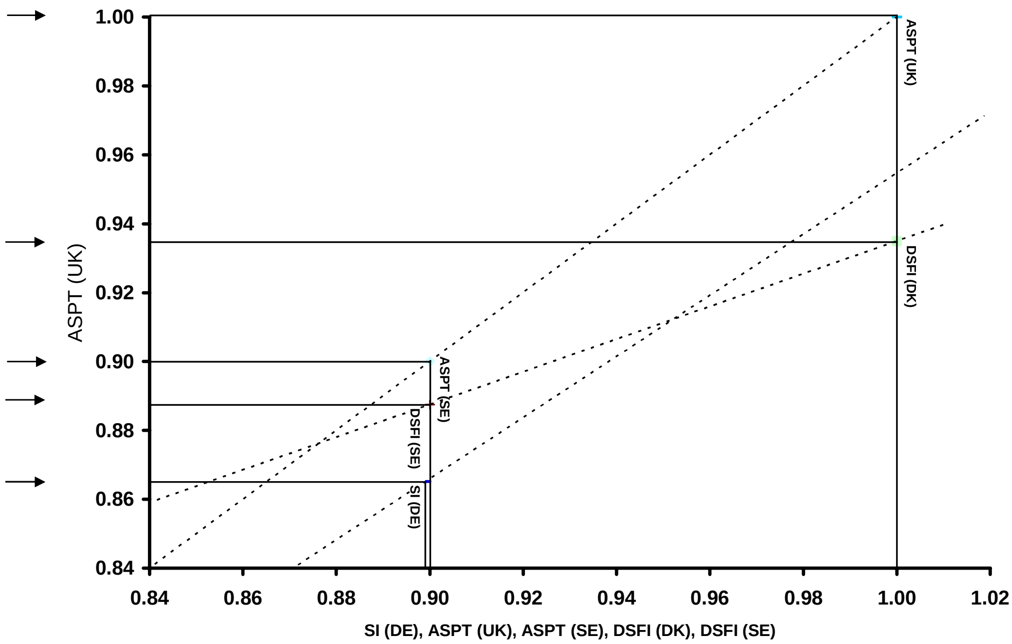
	SI (DE)	ASPT (UK)	ASPT (SE)	DSFI (DK)	DSFI (SE)
nationally defined	1.75	6.38	4.7	7	5
95th percentile	1.66	6.98	6.98	7	7

Comparison of class boundary values
None of the compared high|good class boundary values correspond. The highest difference amounts to 0.135 ASPT-EQR units between SI (DE) and ASPT (UK) (Table 5, Figure 2).
Comparing good|moderate class boundary values reveals almost no differences in boundary setting between ASPT (SE) and DSFI (DK) (difference of 0.001 units). As maximal difference 0.177 ASPT-EQR units exists between SI (DE) and ASPT (UK).

Table 5: Predicted values of high|good and good|moderate boundary values

R-C4			SI (DE)	ASPT (UK)	ASPT (SE)	DSFI (DK)	DSFI (SE)
predicted	high|good	SI (DE)	**0.899**	0.981	0.911	0.948	0.909
		ASPT (UK)	0.865	**1.000**	0.900	0.935	0.887
		ASPT (SE)	0.865	1.000	**0.900**	0.935	0.887
		DSFI (DK)	0.839	1.021	0.880	**1.000**	0.900
		DSFI (SE)	0.839	1.021	0.880	1.000	**0.900**
	good|moderate	SI (DE)	**0.728**	0.904	0.840	0.835	0.869
		ASPT (UK)	0.713	**0.890**	0.800	0.799	0.840
		ASPT (SE)	0.713	0.890	**0.800**	0.799	0.840
		DSFI (DK)	0.587	0.866	0.740	**0.714**	0.800
		DSFI (SE)	0.587	0.866	0.740	0.714	**0.800**

Comparison of class boundaries *high|good* against ASPT (UK)



Comparison of class boundaries *good|moderate* against ASPT (UK)

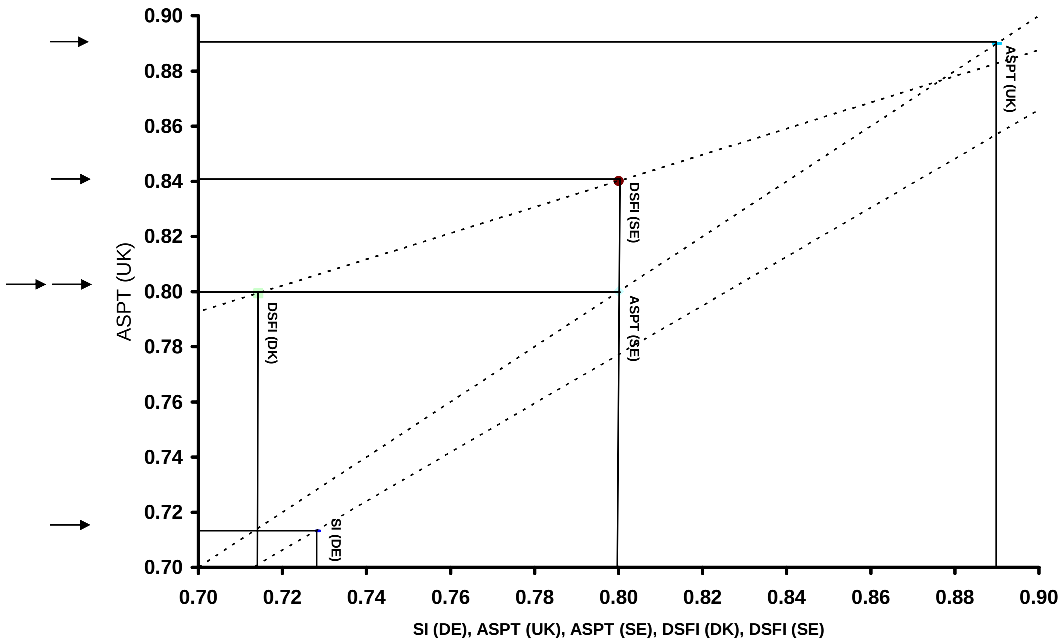


Figure 2: Class boundary comparisons through conversion of national boundaries into ASPT-EQR units using regression lines

7.1.2 Examples of the "direct comparison approach" based on AQEM/STAR data - Macrophytes

R-C3 - small-sized, mid-altitude streams of siliceous geology

Correlation and regression
Spearman correlation coefficients of the three macrophyte indices vary between 0.93 (British MTR and French IBMR) and 0.78 (German RI and British MTR). For the French and British indices 47 samples are included in the analysis. The German index only delivers validated results for 21 samples of the module "mosses" which is used in the analysis. A correlation overview and a table of regression formulae are provided in Annex 2.1.

Reference values
For both the French and British macrophyte indices nationally defined type specific reference values are not available. These values have been derived by using the 95th percentile index value of all STAR samples of R-C3. For the German RI this reference corresponds to the nationally defined reference value (Table 6).

Table 6: Nationally defined and 95th percentile reference values (n.a. – not available)

	MTR (UK)	**IBMR (FR)**	**RI-Moose (DE)**
nationally defined	n.a.	n.a.	100
95th percentile	80	15	100

Comparison of class boundary values
Currently no banding scheme of ecological status exists for the British MTR. Recommendations for the interpretation of MTR scores to evaluate the trophic state (Holmes et al. 1999) are used in the comparison as good ecological status boundaries.
The module "mosses" of the German Reference Index represents one out of two assessment compartments of the entire system. The overall quality class is derived by worst case. Since the other module "phanerogams" produced invalid index results for lack of sufficient plant quantities found at the sampling site, comparison is exclusively based on the classification of the module "mosses".
Expressed as IBMR-EQR units (Table 7, Figure 3) good ecological status boundary settings of the French and German indices are very similar (difference of 0.015 and 0.014 units, respectively). The largest deviation is between the good|moderate boundaries of MTR (UK) and the IBMR (FR) (0.377 units).

Table 7: Predicted values of high|good and good|moderate boundary values

R-C3			MTR (UK)	IBMR (FR)	RI-Moose (DE)
predicted	high|good	MTR (UK)	**0,825**	0,877	0,990
		IBMR (FR)	0,917	**1,000**	1,015
		RI-Moose (DE)	0,403	0,622	**0,810**
	good|moderate	MTR (UK)	**0,313**	0,690	0,670
		IBMR (FR)	0,423	**0,800**	0,786
		RI-Moose (DE)	-0,352	0,202	**0,120**

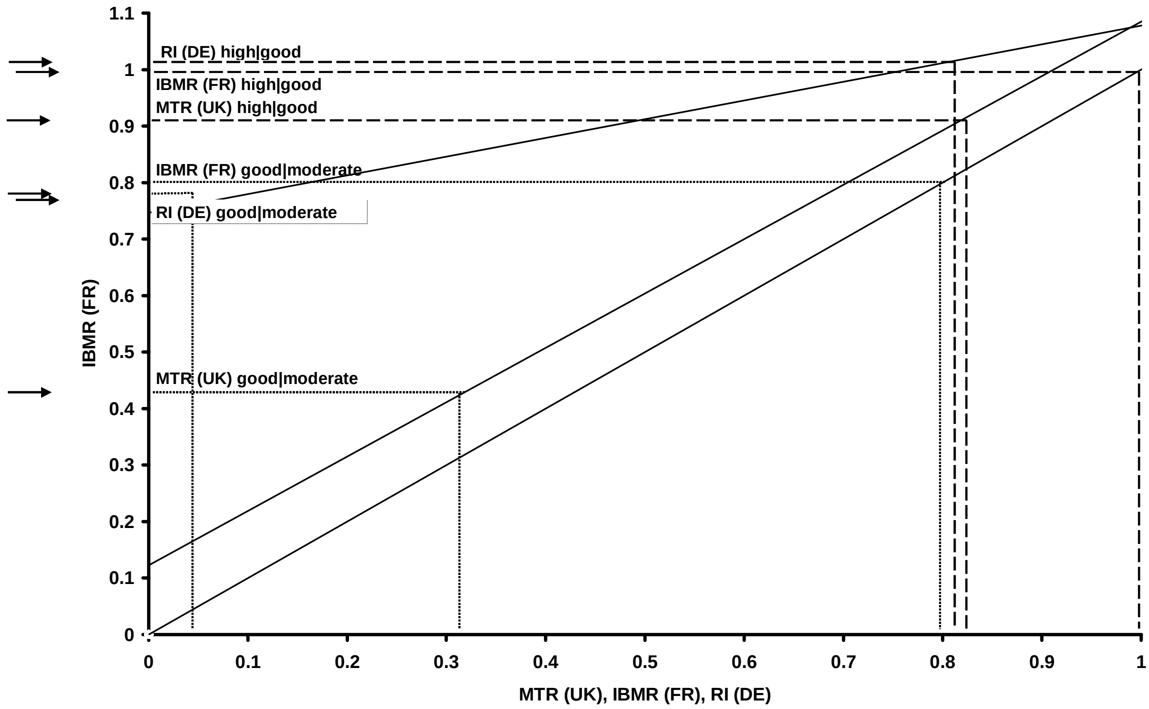


Figure 3: Class boundary comparisons through conversion of national boundaries into IBMR-EQR units using regression lines

R-C4 – medium-sized, lowland, mixed geology

Correlation and regression
Besides high Spearman coefficients of 0.83 between MTR (UK) and IBMR (FR) the correlation shows low coefficients (0.33) between IBMR (FR) and RI (DE). In Annex 2.2 results of the correlation and regression analysis are displayed.

Reference values
For both the French and British macrophyte indices nationally defined type specific reference values are not available. These values have been derived by using the 95th percentile index value of all STAR samples of R-C4. For the German Reference Index this reference is lower than the nationally defined reference value (Table 8).

Table 8: Nationally defined and 95th percentile reference values (n.a. – not available)

	MTR (UK)	IBMR (FR)	RI (DE)
nationally defined	n.a.	n.a.	100
95th percentile	60.35	13.20	66.73

Comparison of class boundary values
As in the comparison exercise for R-C3 class boundary values for MTR (UK) have been set based on the recommendations of Holmes et al. 1999.
The classification of ecological quality of the German RI for type R-C4 includes additional criteria which can individually modify the resulting quality class as obtained by the RI. These criteria have not been considered in the comparison.
Expressed in IBMR-EQR units all class boundary values are different (Table 9, Figure 4). The most similar values are those of the high|good boundaries of IBMR (FR) and MTR (UK) (difference of 0.073 units). The largest deviation of 0.365 units is between high|good boundaries of RI (DE) and IBMR (FR).

Table 9: Predicted values of high|good and good|moderate boundary values

R-C4			MTR (UK)	IBMR (FR)	RI (DE)
predicted	high|good	MTR (UK)	**1,094**	1,071	0,678
		IBMR (FR)	1,063	**1,136**	0,771
		RI (DE)	0,904	0,856	**0,600**
	good|moderate	MTR (UK)	**0,414**	0,823	0,550
		IBMR (FR)	0,596	**0,909**	0,694
		RI (DE)	0,346	0,672	**0,300**

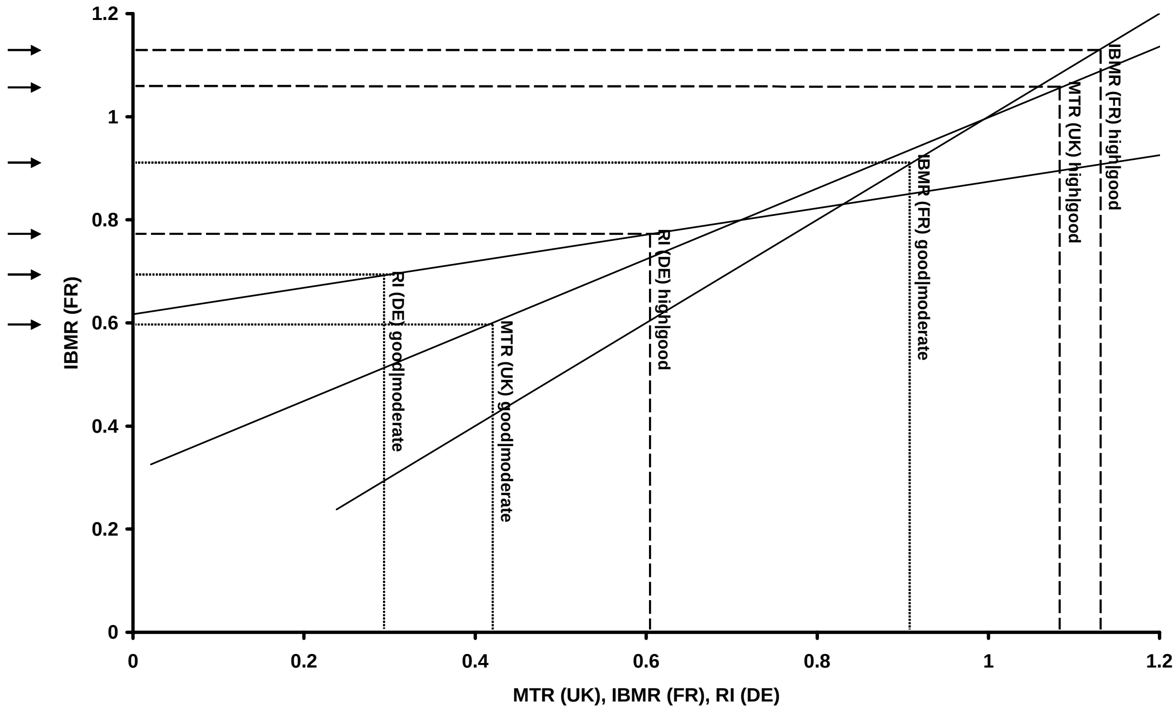


Figure 4: Class boundary comparisons through conversion of national boundaries into IBMR-EQR units using regression lines

7.1.3 Discussion

Reference values

In the intercalibration exercise class boundaries expressed as EQR values are compared. Prerequisite for intercalibration is therefore the availability of stream type-specific reference conditions to derive method-specific reference values. For some of the methods included in the comparison reference values are not available (Slovak SI, Polish BMWP, British MTR, French IBMR). Other methods use reference values derived by different approaches. The British assessment system is based on site-specific instead of type-specific reference conditions. Thus, the reference ASPT for the stream type is a range of values rather than a single number. In this exercise the value, which best corresponds to the abiotic data of the common type has been chosen as the reference for the national system (see Table 4). Austria defines the median of the Saprobic Index of all available reference sites of a certain type as the saprobic basic condition (=reference value). In Germany saprobic reference values have been derived by taking the 95th percentile of all available sites (minus 2 * standard deviation).

Calculation of EQR values in these examples of direct comparison is based on reference values that are defined by the 95th percentile index values of the AQEM/STAR sites that are pre-classified as high status (Benthic Invertebrates) or the 95th value of all STAR samples (Macrophytes), respectively. These values partly deviate from the values defined by the individual countries for the common stream type. But the approach enables to compare EQR class boundary values, even if no nationally defined references for a method are available. Furthermore, the comparison is based on homogeneously derived reference values. Nevertheless, the calculated boundaries have to be considered tentative, as they were in most cases calculated based on the pre-classification of sites. Actually, they still have to be checked to derive a fully WFD-compliant post-classification of sites. This will support an effective selection of reference sites, to be used for setting the reference value for each type.

Comparison of class boundary values

The country-specific assessment methods have either specified their ecological quality class boundaries as absolute numbers (e.g. Saprobic Index values) or EQR values (e.g. ASPT). For the latter the definition of reference values has no influence on the position of the respective class boundary in the EQR-scale. Contrary to that, the transformation of absolute class boundaries into EQR values is dependant on the defined reference, since lower reference quality results in EQR class boundary values closer to “1”. Therefore, the choice of reference values has an effect on the position of the quality boundary in the comparison.

The example of the “direct comparison approach” reveals major differences between class boundary settings of the methods included. Nevertheless, the significance of discrepancies between the individual methods needs to be specified. Additional analyses have to consider e.g. the level of confidence resulting from the degree of bilateral correlation between indices, and the influence of the benchmark index against which the comparisons are made (here: ASPT or IBMR).

General conclusion

The direct comparison of assessment methods has proven useful since more than 20 years. The applicability for WFD intercalibration purposes is shown in this study. Particularly, if EQR values based on reference conditions are used, the national methods can easily be compared between each other or to a benchmark system. The approach identifies inconsistencies in class boundary setting. Based on the defined reference conditions it would also be possible to suggest harmonised class boundaries (see chapter 8.1). Thus, the “direct comparison approach" is suited to validate the results of the ICM approach, or as an alternative.

Cited literature

AFNOR (Association Française de Normalisation) (2002): Norme Indice Biologique Macrophytique En Rivière (I.B.M.R.). AFNOR/T 95 F N 181. 15.02.2002.

Armitage, P.D., D. Moss, J.F. Wright & M.T. Furse (1983): The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Res. 17(3): 333-347.

Buffagni A. & S. Erba, 2004. A simple procedure to harmonize class boundaries of European assessment systems. Discussion paper for the Intercalibration process – WFD CIS WG 2.A ECOSTAT, 6 February 2004, 21pp.

CIS WG 2.A Ecological Status (ECOSTAT) (2004): Overview of common intercalibration types. First draft outline of guidance document for discussion at WG2A meeting, 4-5 March 2004. Version 4.0 - 26 February 2004.

CSN 757716 (1998): Water quality, biological analysis, determination of saprobic index. Prague, Czech Technical State Standard, Czech Standards Institute.

Friedrich, G., E. Coring & B. Küchenhoff (1995): Vergleich verschiedener europäischer Untersuchungs- und Bewertungsmethoden für Fließgewässer. Essen. Landesumweltamt Nordrhein-Westfalen.

Friedrich, G. & V. Herbst (2004): Eine erneute Revision des Saprobiensystems - weshalb und wozu? Acta hydrochim. hydrobiol. 32: 61-74.

Holmes, N.T.H., J.R. Newman, S. Chadd, K.J. Rouen, L. Saint & F.H. Dawson (1999): Mean Trophic Rank: A User's Manual. R & D Technical Report E38. Bristol. Environment Agency.

Moog, O., A. Chovanec, J. Hinteregger & A. Römer (1999): Richtlinie zur Bestimmung der saprobiologischen Gewässergüte von Fliessgewässern. Wien, Bundesministerium für Land- und Forstwirtschaft.

Rico, E., A. Rallo, M.A. Sevillano & M.L. Arretxe (1992): Comparison of several biological indices based on river invertebrate benthic community for assessment of running water quality. Annls Limnol. 28(2): 147-156.

Schaumburg, J., U. Schmedtje, C. Schranz, B. Köpf, S. Schneider, P. Meilinger, D. Stelzer, G. Hofmann, A. Gutowski & J. Foerster (2004): Erarbeitung eines ökologischen Bewertungsverfahrens für Fließgewässer und Seen im Teilbereich Makrophyten und Phytobenthos zur Umsetzung der EG-Wasserrahmenrichtlinie. Abschlussbericht. Januar 2004. München. Bayerisches Landesamt für Wasserwirtschaft.

Siligardi, M., G. Flaim, G. Ziglio, F. Ciutti, C. Monauni & C. Cappelletti (2000): L'esperienza di un corso sul confronto fra indici biologici europei (IBE, BBI, BMWP', RIVPACS). Biologia Ambientale 14(2): 39-41.

Skriver, J., N. Friberg & J. Kirkegaard (2000): Biological assessment of running waters in Denmark: Introduction of the Danish Stream Fauna Index (DSFI). Verh. Internat. Verein. Limnol. 27: 1822-1830.

STN (Slovenská Technická Norma) 83 0532-1 to 8 (1978/79): Biologický rozbor povrchovej vody. (Biological analysis of surface water quality.). Úrad pro Normalizaci a Mereni, Praha.

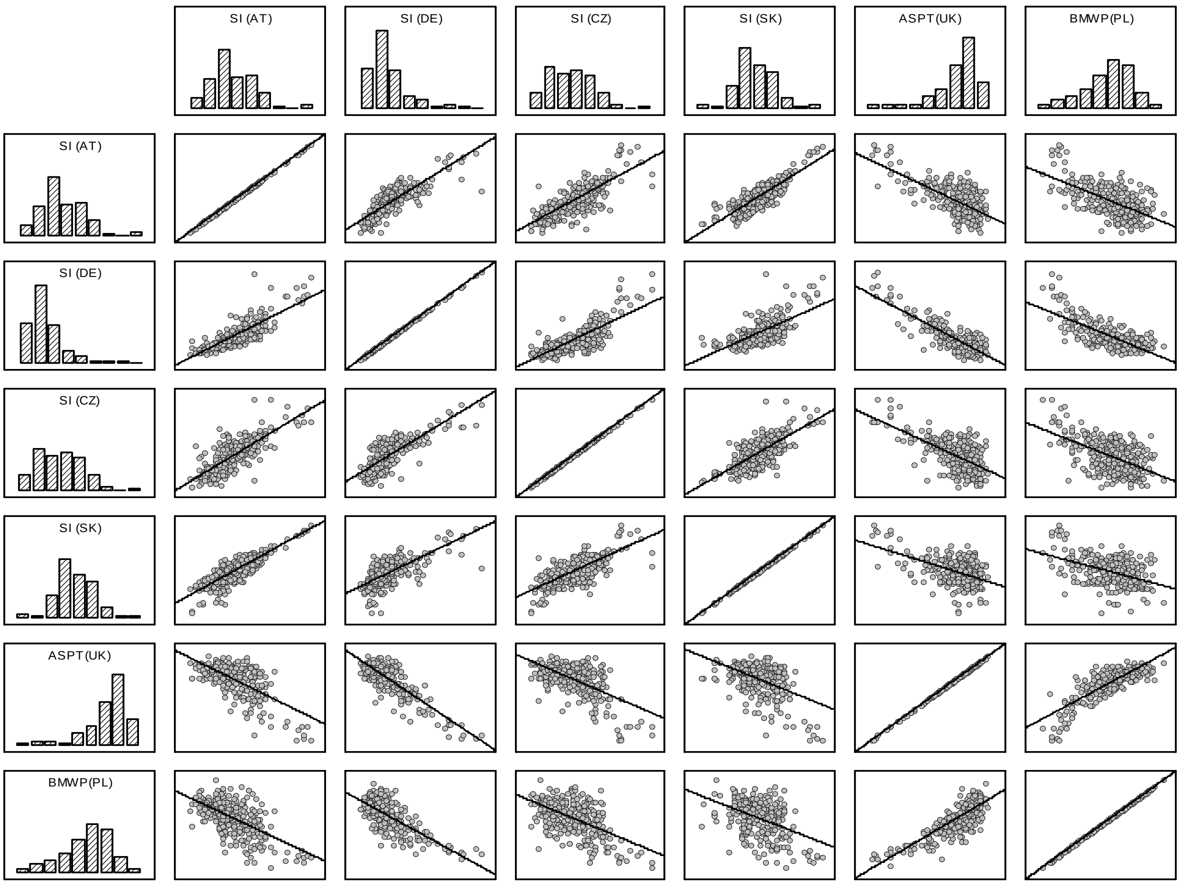
Swedish Environmental Protection Agency (2000): Environmental quality criteria: Lakes and watercourses. Stockholm. Swedish Environmental Protection Agency.

Tittizer, T. (1976): Comparative study of biological-ecological water assessment methods. Practical demonstration on the river Main. 2-6 June, 1975 (summary report). R.-J. Amavis (ed.^eds.): Principles and methods for determining ecological criteria on hydrobiocoenosis: Proceedings of the European Scientific Colloquium Luxembourg, Nov. 1975: 403-463.

Annex 1.1: R-C3 - small-sized, mid-altitude streams of siliceous geology (Benthic Invertebrates)

Correlation Coefficients (Spearman, $p < 0.01$; n=294; spring and summer)

	SI (AT)	SI (DE)	SI (CZ)	SI (SK)	ASPT (UK)	BMWP (PL)
SI (AT)	1.00	0.86	0.77	0.82	-0.51	-0.53
SI (DE)	0.86	1.00	0.76	0.70	-0.60	-0.63
SI (CZ)	0.77	0.76	1.00	0.71	-0.45	-0.46
SI (SK)	0.82	0.70	0.71	1.00	-0.34	-0.37
ASPT (UK)	-0.51	-0.60	-0.45	-0.34	1.00	0.77
BMWP (PL)	-0.53	-0.63	-0.46	-0.37	0.77	1.00



Correlations of SI (AT), SI (DE), SI (CZ), SI (SK), ASPT (UK), BMWP (PL)

R-C3: Regression formulae (absolute values)

	SI (AT)		SI (DE)		SI (CZ)		SI (SK)		ASPT (UK)		BMWP (PL)	
	a (intercept)	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*
SI (AT)	0	1	0.255	0.883	1.039	0.491	0.783	0.610	3.066	-0.206	2.255	-0.004
SI (DE)	0.295	0.790	0	1	1.010	0.466	0.935	0.469	3.242	-0.242	2.272	-0.004
SI (CZ)	-0.792	1.262	-0.837	1.337	0	1	0.081	0.846	3.357	-0.302	2.174	-0.006
SI (SK)	-0.528	1.196	-0.174	1.027	0.636	0.645	0	1	2.910	-0.212	2.081	-0.004
ASPT (UK)	10.294	-2.166	11.278	-2.847	8.281	-1.239	8.314	-1.135	0	1	4.144	0.017
BMWP (PL)	304.944	-94.477	342.729	-121.070	217.938	-54.594	219.108	-49.860	-121.802	40.244	0	1

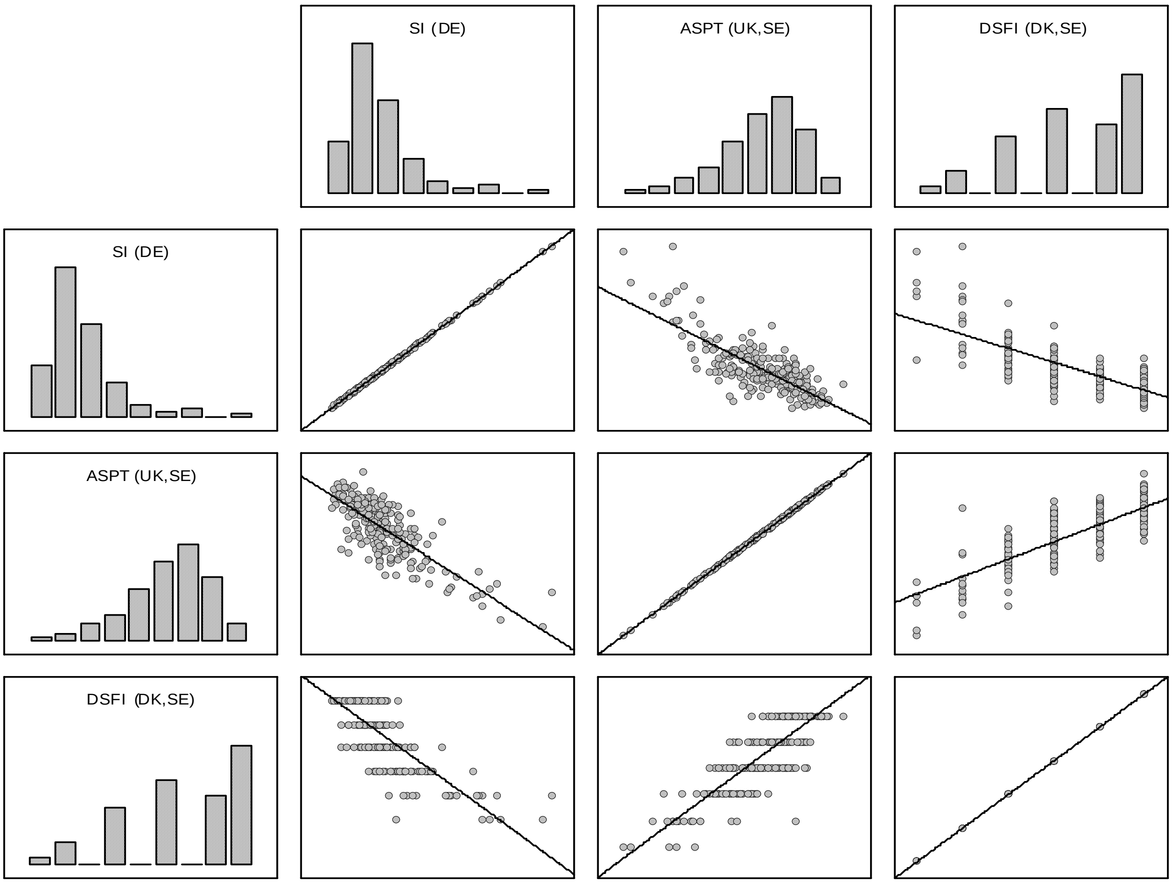
R-C3: Regression formulae (EQR values; based on 95th percentile of high status sites)

	SI (AT)		SI (DE)		SI (CZ)		SI (SK)		ASPT (UK)		BMWP (PL)	
	a (intercept)	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*
SI (AT)	0	1	0.0795	0.8768	0.3746	0.6093	0.2914	0.6792	0.3511	0.5798	0.656	0.2849
SI (DE)	0.2058	0.7964	0	1	0.4266	0.5823	0.4506	0.5258	0.2871	0.6871	0.6546	0.3291
SI (CZ)	-0.0768	1.0168	-0.1551	1.0698	0	1	0.1625	0.7586	0.1949	0.6865	0.5534	0.3408
SI (SK)	-0.0865	1.0747	0.0223	0.916	0.2645	0.7193	0	1	0.3682	0.5352	0.6484	0.2647
ASPT (UK)	0.2176	0.7692	-0.0148	1.0035	0.4441	0.5458	0.5037	0.4487	0	1	0.5533	0.4533
BMWP (PL)	-0.3667	1.2629	-0.7113	1.6062	-0.0022	0.9053	0.0988	0.7416	-0.6121	1.5147	0	1

Annex 1.2: R-C4 – medium-sized, lowland, mixed geology (Benthic Invertebrates)

Correlation Coefficients (Spearman, p < 0.01; n=247; spring, summer, autumn)

	SI (DE)	ASPT (UK)	ASPT (SE)	DSFI (DK)	DSFI (SE)
SI (DE)	1.00	-0.71	-0.71	-0.75	-0.75
ASPT (UK)	-0.71	1.00	1.00	0.79	0.79
ASPT (SE)	-0.71	1.00	1.00	0.79	0.79
DSFI (DK)	-0.75	0.79	0.79	0.79	0.79
DSFI (SE)	-0.75	0.79	0.79	0.79	0.79



Correlations of SI (DE), ASPT (UK, SE), DSFI (DK, SE)

R-C4: Regression formulae (absolute values)

	SI (DE)		ASPT (UK)		ASPT (SE)		DSFI (DK)		DSFI (SE)	
	a (intercept)	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*
SI (DE)	0	1	3.3474	-0.2353	3.3474	-0.2353	2.7041	-0.1318	2.7041	-0.1318
ASPT (UK)	11.057	-2.6466	0	1	0	1	3.2146	0.4731	3.2146	0.4731
ASPT (SE)	11.057	-2.6466	0	1	0	1	3.2146	0.4731	3.2146	0.4731
DSFI (DK)	14.2404	-4.4121	-2.6833	1.408	-2.6833	1.408	0	1	0	1
DSFI (SE)	14.2404	-4.4121	-2.6833	1.408	-2.6833	1.408	0	1	0	1

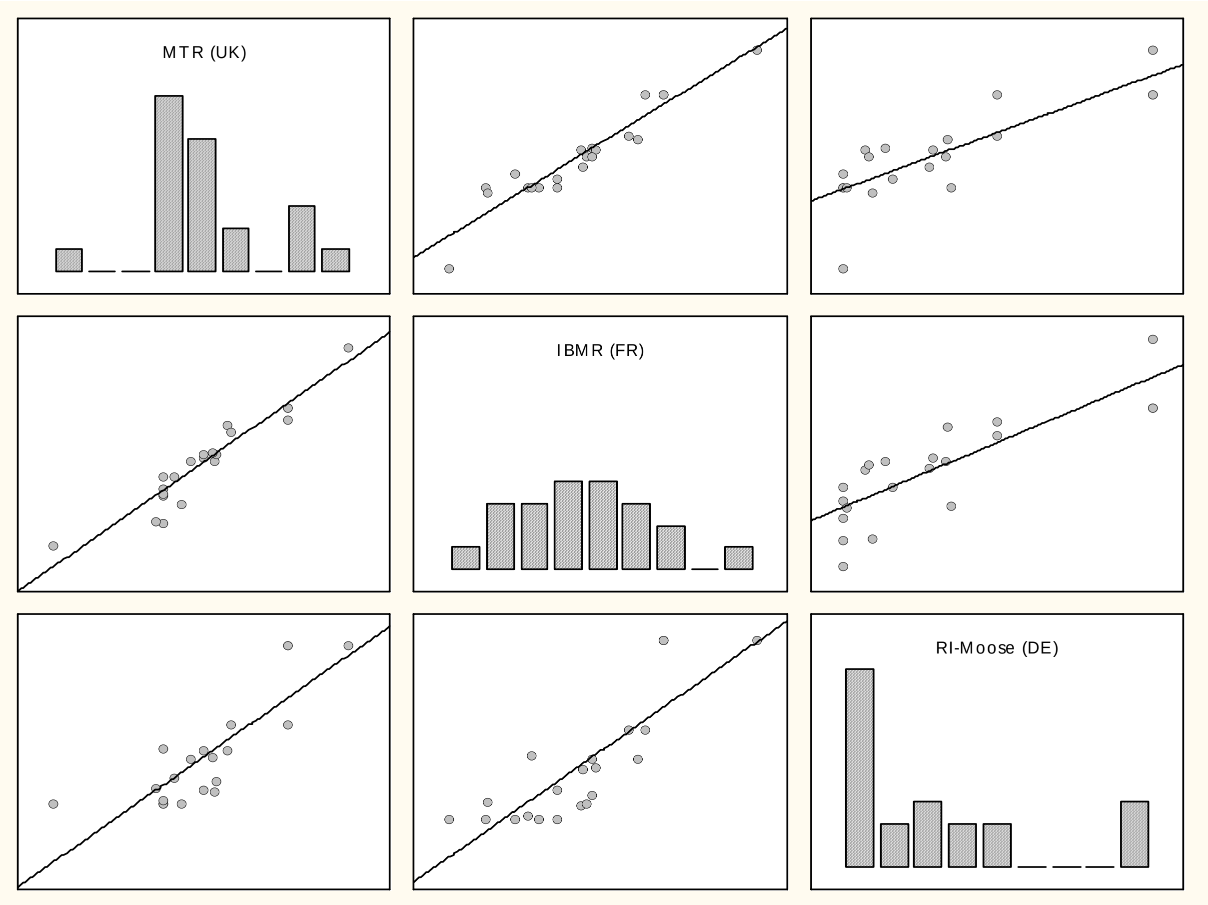
R-C4: Regression formulae (EQR values; based on 95th percentile of high status sites)

	SI (DE)		ASPT (UK)		ASPT (SE)		DSFI (DK)		DSFI (SE)	
	a (intercept)	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*	*a (intercept)*	*b (slope)*
SI (DE)	0	1	0.2789	0.7018	0.2789	0.7018	0.5538	0.3942	0.5538	0.3942
ASPT (UK)	0.0674	0.8872	0	1	0	1	0.4605	0.4744	0.4605	0.4744
ASPT (SE)	0.0674	0.8872	0	1	0	1	0.4605	0.4744	0.4605	0.4744
DSFI (DK)	-0.4869	1.4749	-0.3833	1.404	-0.3833	1.404	0	1	0	1
DSFI (SE)	-0.4869	1.4749	-0.3833	1.404	-0.3833	1.404	0	1	0	1

Annex 2.1: R-C3 - small-sized, mid-altitude streams of siliceous geology (Macrophytes)

Correlation Coefficients (Spearman, $p < 0.05$; MTR (UK), IBMR (FR): n=47; RI (DE): n=21)

	MTR (UK)	IBMR (FR)	RI (DE)
MTR (UK)	1.00	0.93	0.78
IBMR (FR)	0.93	1.00	0.84
RI-Moose (DE)	0.78	0.84	1.00



Correlations of MTR (UK), IBMR (FR), RI (DE)

R-C3: Regression formulae (absolute values)

	MTR (UK)		IBMR (FR)		RI-Moose (DE)	
	a (intercept)	b (slope)	a (intercept)	b (slope)	a (intercept)	b (slope)
MTR (UK)	0	1	-30.35	7.1419	67.707	0.18562
IBMR (FR)	5.1194	0.12552	0	1	13.682	0.02484
RI-Moose (DE)	-262.3	3.6806	-396.1	28.029	0	1

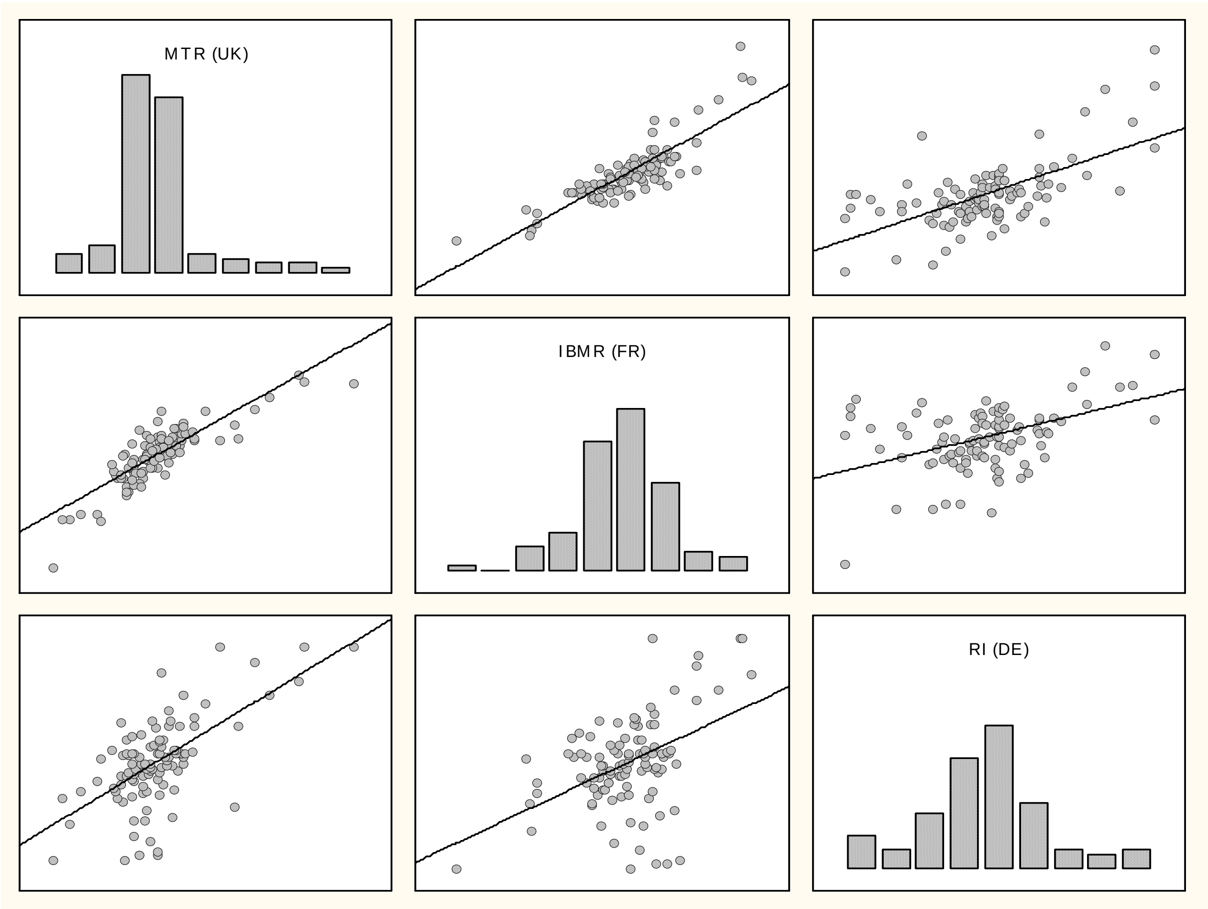
R-C3: Regression formulae (EQR values; based on 95th percentile of all AQEM/STAR samples)

	EQR:MTR (UK)		EQR:IBMR (FR)		EQR:RI-Moose (DE)	
	a (intercept)	b (slope)	a (intercept)	b (slope)	a (intercept)	b (slope)
EQR:MTR (UK)	0	1	-0.0609	0.9382	0.61431	0.46405
EQR:IBMR (FR)	0.1226	0.96236	0	1	0.74652	0.33126
EQR:RI-Moose (DE)	-0.8117	1.4722	-1.48	2.1022	0	1

Annex 2.2: R-C4 – medium-sized, lowland, mixed geology (Macrophytes)

Correlation Coefficients (Spearman, $p < 0.05$; MTR (UK), IBMR (FR): n=126; RI (DE): n=104)

	MTR (UK)	IBMR (FR)	RI (DE)
MTR (UK)	1.00	0.83	0.51
IBMR (FR)	0.83	1.00	0.33
RI (DE)	0.51	0.33	1.00



Correlations of MTR (UK), IBMR (FR), RI (DE)

R-C4: Regression formulae (absolute values)

	MTR (UK)		IBMR (FR)		RI-Moose (DE)	
	a (intercept)	b (slope)	a (intercept)	b (slope)	a (intercept)	b (slope)
MTR (UK)	0	1	-10.26	4.9918	40.949	0.15551
IBMR (FR)	4.1636	0.14691	0	1	10.141	0.01878
RI-Moose (DE)	-99.12	2.2709	-102.3	9.3201	0	1

R-C4: Regression formulae (EQR values; based on 95th percentile of all AQEM/STAR sites)

	EQR:MTR (UK)		EQR:IBMR (FR)		EQR:RI-Moose (DE)	
	a (intercept)	b (slope)	a (intercept)	b (slope)	a (intercept)	b (slope)
EQR:MTR (UK)	0	1	-0,17	1,0918	0,4208	0,4296
EQR:IBMR (FR)	0,3113	0,6873	0	1	0,6166	0,2574
EQR:RI-Moose (DE)	0,0053	0,822	-0,0638	0,8093	0	1

7.2 Indirect comparison: Different sample, same calculation method (ICMindex)

Comparison - and harmonization –of national methods' class boundaries through conversion in ICMi value: intra-GIG

The results of the conversion of the class boundaries of the national assessment methods in the ICMi are here discussed (see the Test datasets description, Chapter 4).
Within each IC type, a very simple harmonization option can be to set the class boundary value of each MS method at the median value of WFD-compliant methods. Following this approach, countries should increase or decrease the original boundary of their methods if existing values are, respectively, below or above the obtained median.
To be acceptable for the European IC exercise, this harmonization option is only suitable if WFD-compliant methods are considered. However, as already emphasized, national assessment systems can fulfil or not the WFD requests. The methods involved have to be fully WFD-compliant. Compliance verification must include reference conditions definition.

<u>**IMPORTANT WARNING**</u>

The option of averaging class boundary values of assessment methods is only applicable when all the considered biological methods are demonstrated as fully WFD-compliant.

In addition, the use of this option is acceptable if all MSs contribute in the calculation of the boundary values (i.e. they all have WFD-compliant methods for that stream type at the time of the IC process).

This option, while requiring consistency to normative definition, does not support a real comparability across stream types and GIGs, thus possibly limiting the aptitude of the European IC process.

The boundaries shown here represent merely an example, because the quality classification of the samples is based on the National assessment systems (i.e. not fully WFD-compliant for all countries).

The first part of the paragraph presents the results for all the below specified datasets (Table 7.2.1), for the comparison phase as well as for the harmonization phase (7.2.1 and 7.2.2). In the last part of the Chapter (6.2.1c) the results for WFD-compliant methods are shown.

MS	IC type	Classification method	type specific adaptations	tolerance	abundance	richness	available criteria for reference condition
ITALY	C1, M1	IBE	N	Y	Y/N	Y	N
FRANCE	C1, C2, M1	IBGN	Y	Y	Y/N	Y	Y
UK	C1	EQI-ASPT & EQI-NFAM	Y?	Y	Y/N	Y	Y
POLAND	C1	BMWP & Margalef div. ind.	N	Y	Y/N	Y	nk
DENMARK	C1	DSFI	Y	Y	Y	Y	N
ESTONIA	C1	ASPT	N	Y	Y/N	Y	N
GERMANY	C1	SI(DE) & GD(DE)	Y	Y	Y	Y	Y
SPAIN	C2	MMI-Spain	Y	Y	Y	Y	Y

Table 7.2.1 Characteristics of the tested method and their compliance with the WFD requirements.

7.2.1 Intra-GIG comparison

The compared Test datasets belong to IC stream type R-C1. The description of the datasets and the regression between ICMi and national methods are presented in a previous Chapter (4).

The ICMi is obtained by the sum of the ICMs weighted and normalized according to 75th percentile of the high status samples, according to the test method. The ICMi is re-normalized according to 75th percentile. In the same way, values of the national method are normalized according to the 75th percentile of the high status samples. For German data, the normalization was undertaken following a different approach, i.e. the reference value was obtained by regression with the GD(DE) index (see German dataset description in chapter 4 and Birk, 2004). In all figures and comments country names are secreted.

As some countries consider two indices for the classification, for such countries the mean value have been considered, except where indicated. The conversion of the boundaries of the national method quality classes into ICMi values is undertaken by means of linear regression (see each Test dataset description). At this purpose, the national method is positioned on the x axis, ICMi on the y. Results are shown in Figure 7.1.

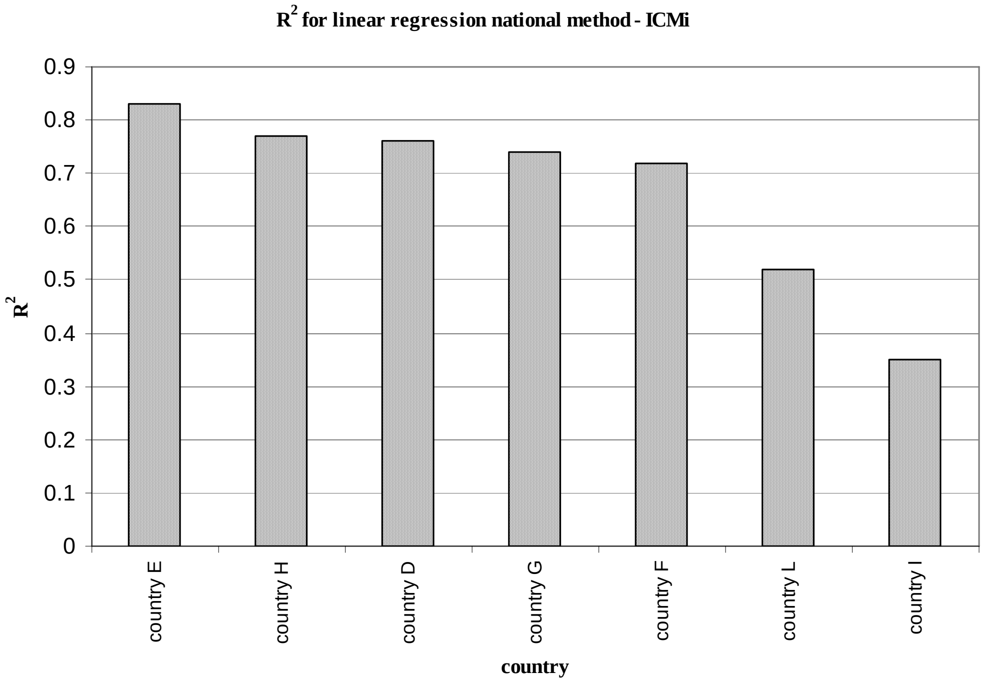


Figure 7.1 R^2 for linear regression, C1 test datasets. p<0.001 for all data

For most of the methods, the selected ICMi shows a good fit with the National test methods and well approximates the quality gradient in most datasets of the C1 type.

The lowest correlation is shown for Country I with a R^2 value of 0.35. The mediocre result for Country L (R^2 = 0.52) can be explained with the low values that some ICMs (i.e.: Shannon index and 1-GOLD) may have at some presumably minor impacted sites. Also, some taxa selected for the metric Log_EPTD may occur rarely in the considered streams.

Possible hypothesis that can be considered when low correlations of ICMi vs National method arise are presented here below.

As a first point, the structure of the data should be checked, in order to avoid the following conditions:

- o Differences in the sampling method within the same dataset may occur. In this case, the subsets of data should be normalized separately.
- o Datasets or stream types with different reference conditions were artificially merged into the dataset. This is the case when data from different areas or stream types are simultaneously considered. Also in this case the normalization should be separate.

Other possible situations:

- o For particular stream types or for not yet validate methods, attention should be paid to the capacity of the method in describing in a proper way the quality gradient.

- If the identification level for the national method is undertaken e.g. to species level, problems can arise if only few macroinvertebrates identified to species level are present in the stream type.
- Datasets can cover a short quality gradient, in particular if mainly poor quality samples are present.

Other possible hypothesis to be considered for low correlations between ICMi and National method in countries with stressor specific assessment modules (e.g. country I):

- If strong attention is paid to a single degradation factor (e.g. degradation in stream morphology), it might occur that the range of the gradient covered is not as long as that defined by a stressor acting stronger on the invertebtrate community (e.g. organic pollution). This would lead to overall higher values for most biological metrics.
- When invertebrates are identifiied to species level and data refer to large geographic areas, it might happen that natural variability of the communities is high (at least comparatively higher than that observed for family level data).
- If none, only one or a few sites belong to High/ Good status classes and a few as well are classified as Bad status, the dataset shows a short gradient, with most of the sites in the ‘central’ quality classes.
- If the National system is based on the ‘one-out, all-out’ principle at the level of different sub-indices, which are supposed to detect different alteration factors. This can determine a lower class for the considered sample (e.g. even if only the morphological quality is low) compared to the judgement provided by most other methods, which consider the average of the metrics. Again, it is important to verify if the quality gradients covered by the different stressors are similarly ample.

The Box&Whiskers representation in Figure 7.2, with the related results in tables 1, show the distribution of the values of the class boundaries converted in ICMi.

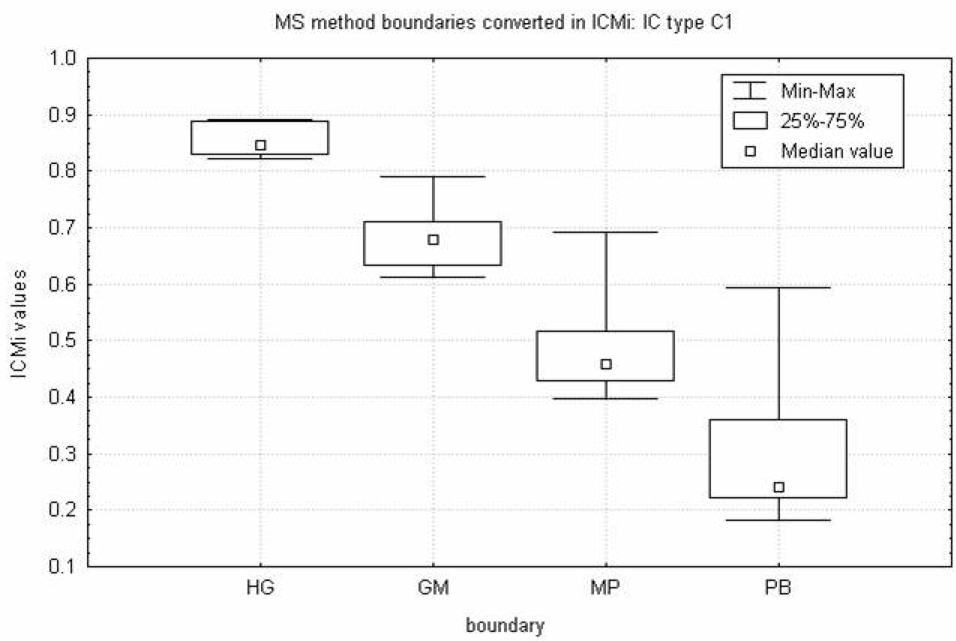


Figure 7.2 Box and whiskers for class boundaries converted in ICMi.

	min	25° perc	median	75° perc	max	mean	stdev
HG	0.822	0.827	0.845	0.888	0.892	0.854	0.028
GM	0.612	0.631	0.678	0.709	0.790	0.682	0.058
MP	0.398	0.426	0.458	0.516	0.692	0.491	0.106
PB	0.183	0.220	0.240	0.360	0.595	0.306	0.153

Table 7.1 Values of minimum, 25^{th} percentile, median, 75^{th} percentile, maximum, mean and standard deviation of class boundaries in ICMi.

Figure 7.2 shows a trend for ICMi values and median for each quality class as defined by MSs' methods, decreasing when quality class is decreasing. The interquartile range show no overlap among classes, while maximum values tend to show a relatively large variability for lower boundaries. Mann Whitney-U test shows significant differences among all the four boundaries (seven countries). This shows how the conversion in ICMi, with related normalization, maintains the boundary values among classes clearly separated.

	HG	GM	MP	PB
HG		0.002	0.003	0.003
GM	0.002		0.02	0.003
MP	0.003	0.02		0.04
PB	0.003	0.003	0.04	

Table 7.2 p-level for Mann-Whitney U test: C1 datasets, MSs existing boundaries.

Figures 7.3 and 7.4 show the deviation from the overall median of each MS' method, for the boundaries 'high-good' and 'good-moderate'.

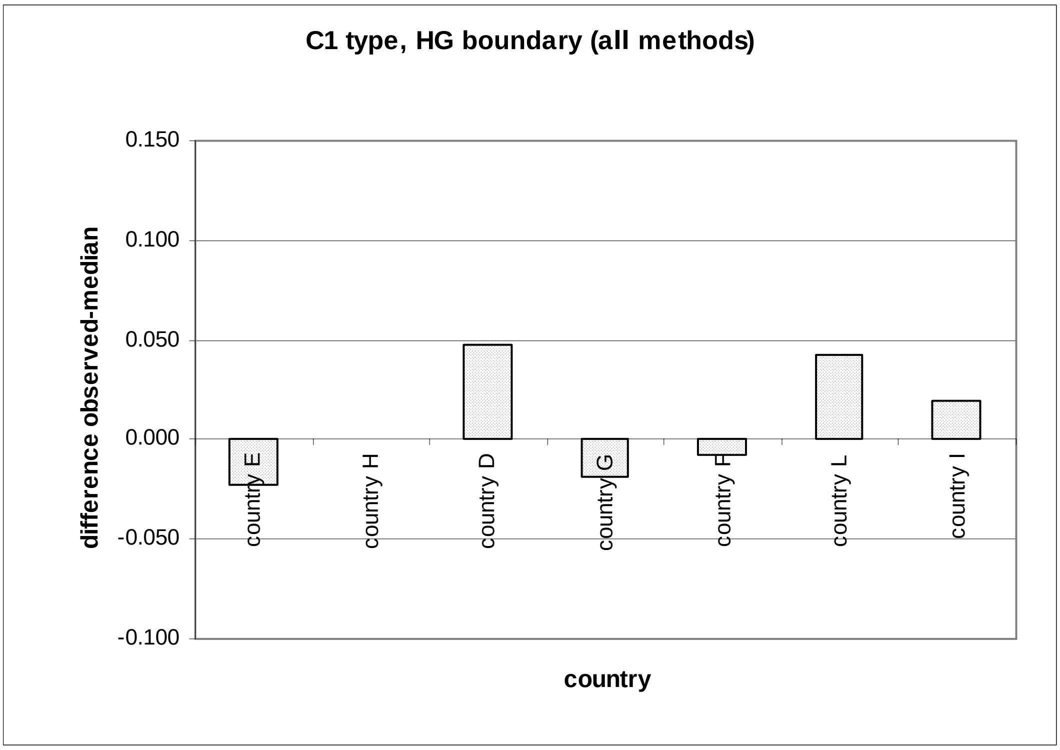


Figure 3: Boundary high-good setting according to median value in C1 streams

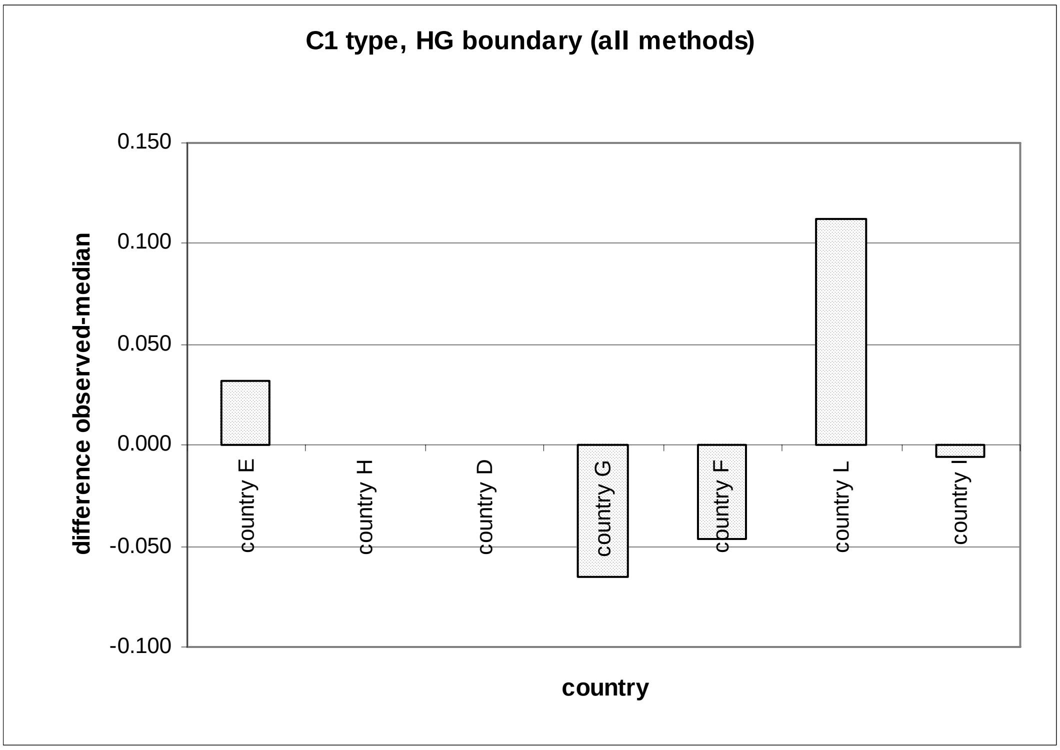


Figure 4: Boundary good-moderate setting according to median value C1 streams

For harmonization purposes, MSs showing a value lower than the median should increase their boundary. In the presented result, countries that should increase their boundaries are CountryE, CountryG and CountryF for boundary High-Good and CountryG, CountryF and CountryI for boundary Good-Moderate. The entity of the modification seems independent from the fit national method – ICMi.

On the contrary, if boundary are higher or equal to the median value, no change in boundary value should be expected. If boundary results much higher than the median value for a certain method, e.g. Country E and G in this example, it is possible that such MSs have already adopted more strict criteria, in order to adapt to the WFD requirements.

7.2.2 Inter-GIG comparison

The same approach described above can be applied to datasests and stream types belonging to different GIGs. The discussion of the results of the conversion of the class boundaries of the national assessment methods in the ICMi thus refers in this paragraph to different GIGs and types. Considered datasets are from types M1, M5, C2,. For type C1, three datasets were selected randomly. In Chapter 4, the description of such datasets and the regression between ICMi and national methods is reported.

For all the considered types, the national method of each MS is presented.

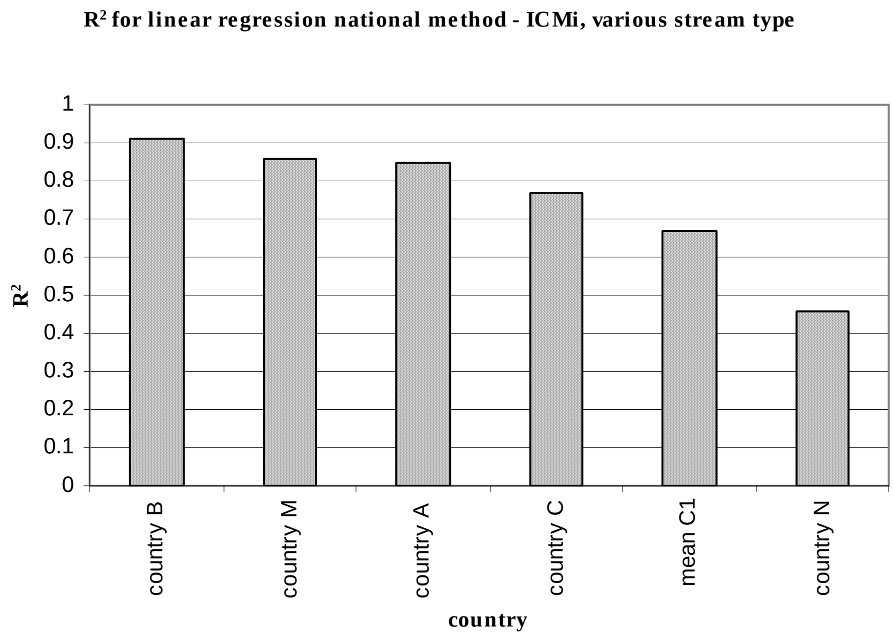


Figure 7.5 R^2 for linear regression, various IC types test datasets. $p<0.001$ for all data

In Figure 7.5, the mean value for all C1 datasets is also reported.
All the methods, except for one, show a R^2 higher than 0.60, in four cases higher than 0.70. The ICMi well fit with national methods in very different stream types.
The lowest correlation is observed for countryN (R^2= 0.46). In this instance, due to the characteristics of the stream type, i.e. intermittent streams, it is possible that the

national method can not properly describe the quality gradient in such river type (Buffagni *et al.*, 2004). Moreover, specific approaches, methods and metrics are needed for such particular stream types (e.g. for temporary streams). In this context, a discussion on the more appropriate ICMs to be used in type M5 is at the moment in progress among Mediterranean GIG partners.

The distribution of the values of the class boundaries converted in ICMi is showed in the Box and Whiskers representation of figure 7.6, with the related results in table 7.2. Country N data were excluded from the calculation, due to the particular features of the stream type.

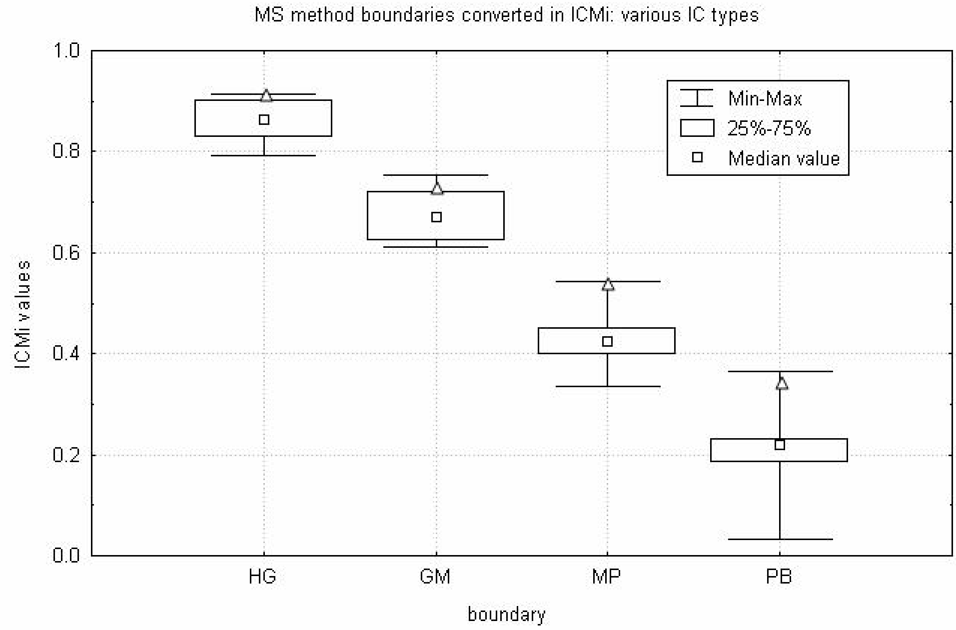


Figure 7.6 Box and whiskers for class boundaries converted in ICMi, various stream types. In triangles: Country N data.

	min	25° perc	median	75° perc	max	mean	stdev
HG	0.794	0.827	0.865	0.901	0.915	0.858	0.034
GM	0.612	0.624	0.672	0.722	0.754	0.671	0.046
MP	0.334	0.398	0.426	0.452	0.543	0.430	0.070
PB	0.032	0.183	0.220	0.232	0.364	0.206	0.115

Table 7.3 values of minimum, 25° percentile, median, 75° percentile, maximum, mean and standard deviation of class boundaries for ICMi, various stream types.

Even considering data from different stream types and GIGs, the interquartile ranges show no overlapping between boundaries. Mann Whitney U test shows significant differences among all the four boundaries.

	HG	GM	MP	PB
HG		0.002	0.004	0.004
GM	0.002		0.004	0.004
MP	0.004	0.004		0.02
PB	0.004	0.004	0.02	

Table 7.4: p-level for Mann-Whitney U test: various IC type

Figures 7.7 and 7.8 shows the deviation from the median of the WFD-compliant methods for the boundaries HG and GM for the single countries.

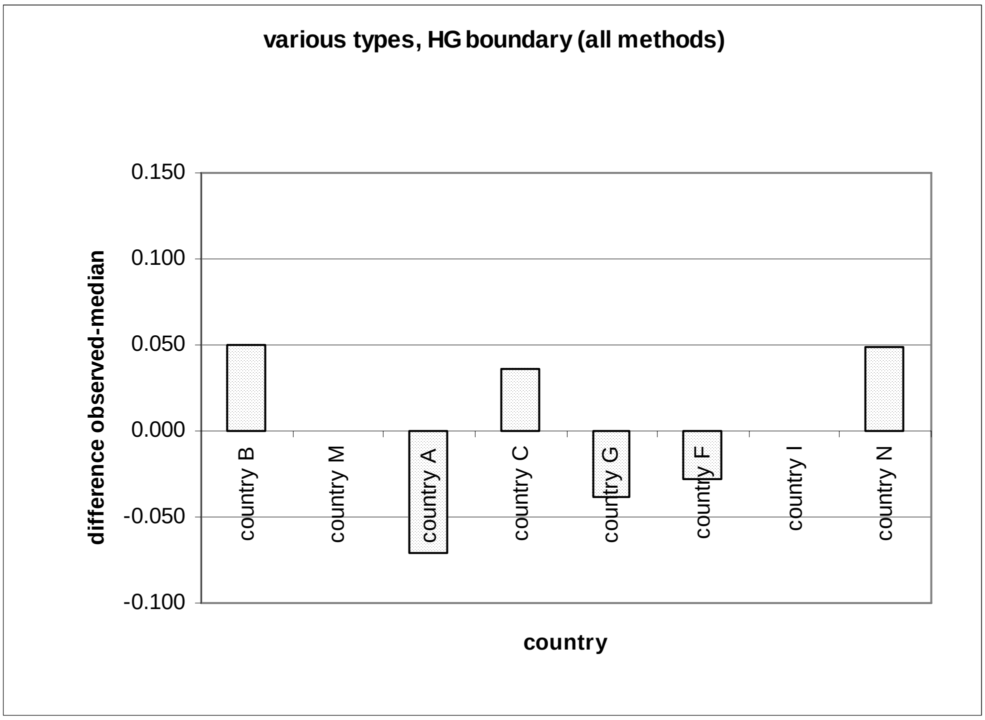


Figure 7.7 Boundary high-good setting according to the WFD-compliant median value, various IC types

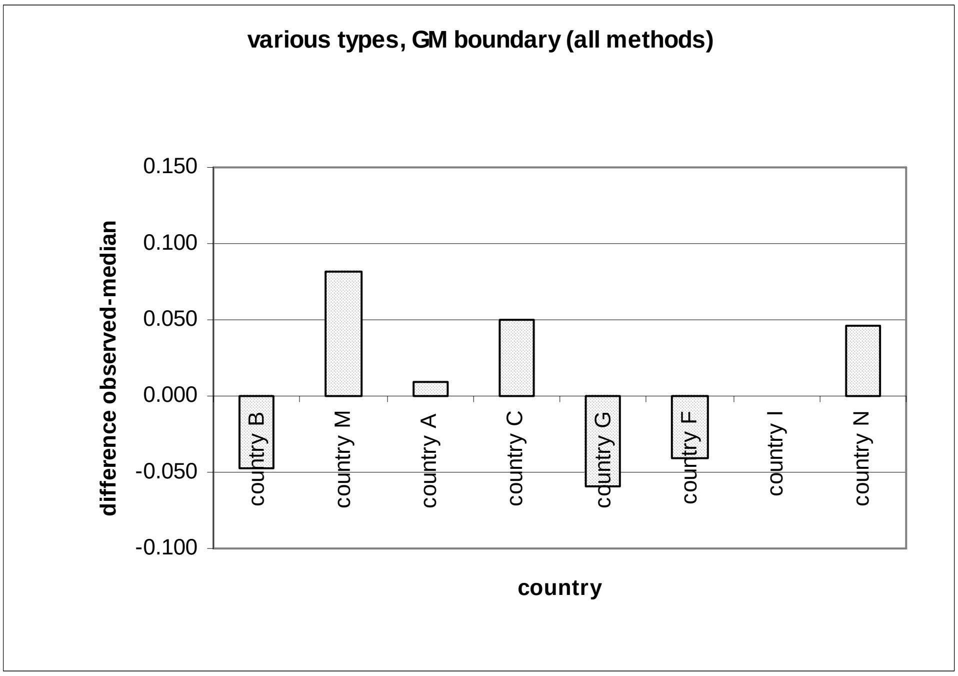


Figure 7.8 Boundary good-moderate setting according to the WFD-compliant median value, across GIGs

Country that might have to move up the boundary are country G and F for both High-Good and Good-Moderate boundary. Some countries have different results for the two boundaries: country A might have to move up the High-Good boundary, country B the Good-Moderate.

7.2.3 Inter-GIG comparison and harmonization for WFD compliant methods.

The compliance to the WFD for the considered methods is not going to be assessed here in any conclusive way. A tentative attribution has been made, accordingly to the availability of stream type-specific reference conditions and to the inclusion of tolerance and richness metrics. Abundance, anyway needed for the aim of the present exercise and for fully WFD-compliant methods, has not been stringently considered here, because none of the methods take it into acoount carefully.

The distribution of the values of the class boundaries converted in ICMi for WFD-compliant methods is showed in the Box and Whiskers representation of figure 7.9, with the related results in table 7.3.

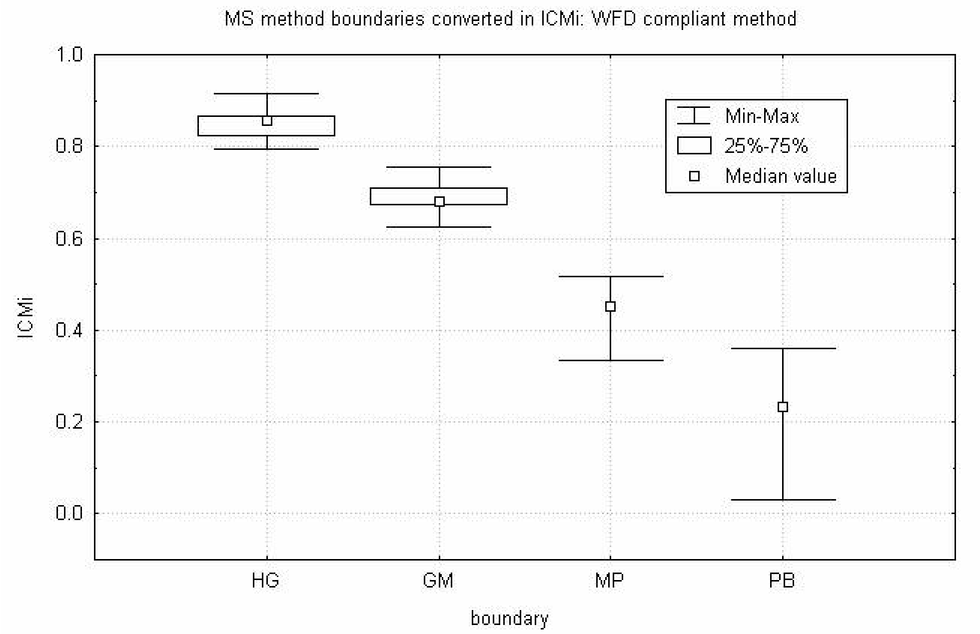


Figure 7.9 Box and whiskers for class boundaries converted in ICMi, WFD-compliant methods.

	min	25° perc	median	75° perc	max	mean	stdev
HG	0.794	0.822	0.855	0.865	0.915	0.851	0.042
GM	0.624	0.672	0.680	0.709	0.754	0.686	0.043
MP	0.334	-	0.452	-	0.516	0.434	0.092
PB	0.032	-	0.232	-	0.360	0.208	0.166

Table 7.5 values of minimum, 25° percentile, median, 75° percentile, maximum, mean and standard deviation of class boundaries in ICMi, various stream types.

The groups of boundaries HG and GM are well separated (significant difference for Mann Whitney U test). Non significant differences, p=0.13 are observed for boundaries MP and PB.

	HG	GM	MP	PB
HG		0.004	0.02	0.02
GM	0.004		0.02	0.02
MP	0.02	0.02		0.13
PB	0.02	0.02	0.13	

Table 7.6: p-level for Mann-Whitney U test: WFD compliant countries

Figures 7.10 and 7.11 show the deviation from the median of the boundaries HG and GM for the single countries. For countries that have two indices for the classification (i.e.: country I and country F) the histogram is split in part a and b.

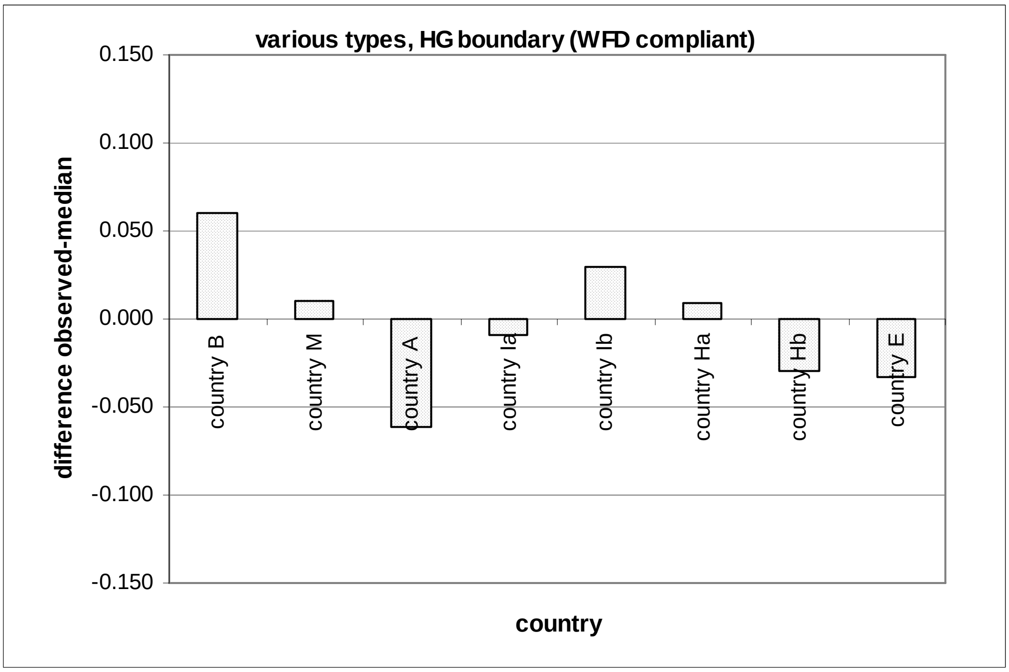


Figure 7.10: Boundary high-good setting according to median value WFD-compliant methods

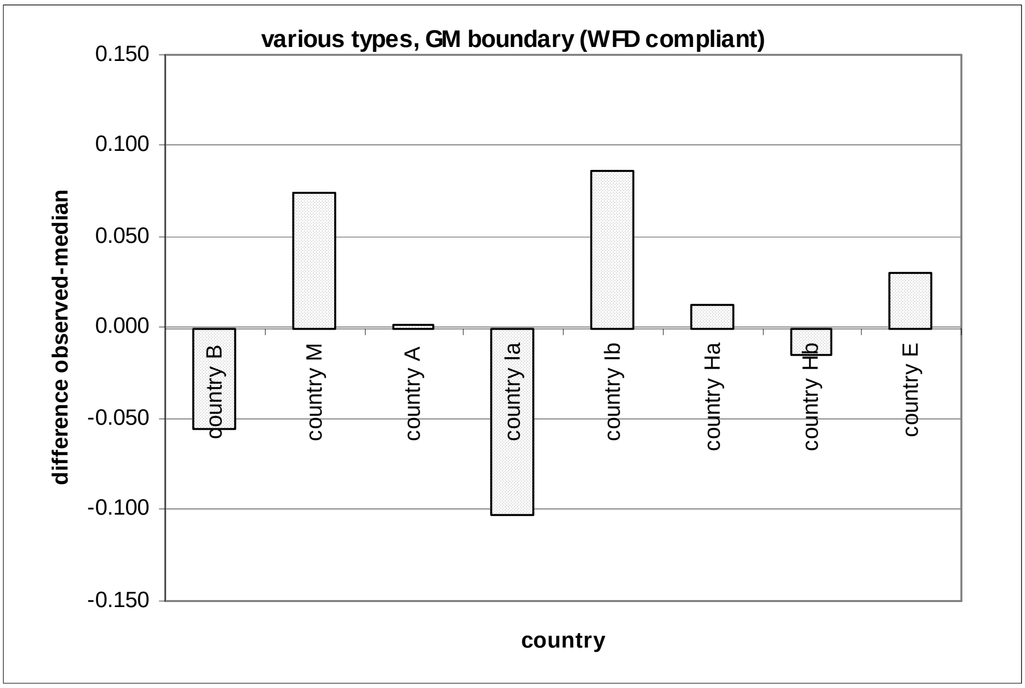


Figure 7.11: Boundary good-moderate setting according to median value WFD-compliant methods

Countries with two indices (i.e., country I and H) show a different response.

- For country I, boundaries may have to be refined for method 'a', while for method 'b' boundaries hare higher than the median. The method 'b' is a newly developed multimetric index.
- The same for country H where only method b should increase the boundary.
- The country that seems more restrictive from this results is country M.

The percentage of sites shifting their quality class after the harmonization process, is shown below. The classification before harmonization is undertaken according to the original national boundaries. With the harmonization step, the MS' boundaries have been ricalculated from the boundaries of the ICMi (median values) according to the linear regression formulae when necessary. According to the new boundaries, samples were re-classified. Examples from country A, B and I are shown

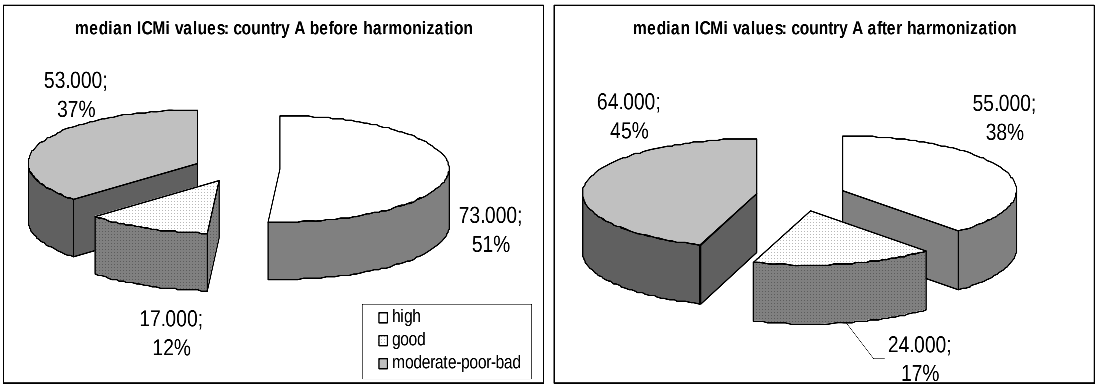


Figure 7.12 Percentage of samples belonging to High, Good and Moderate/Poor/Bad classes before (according to national method) and after harmonization in Country A.

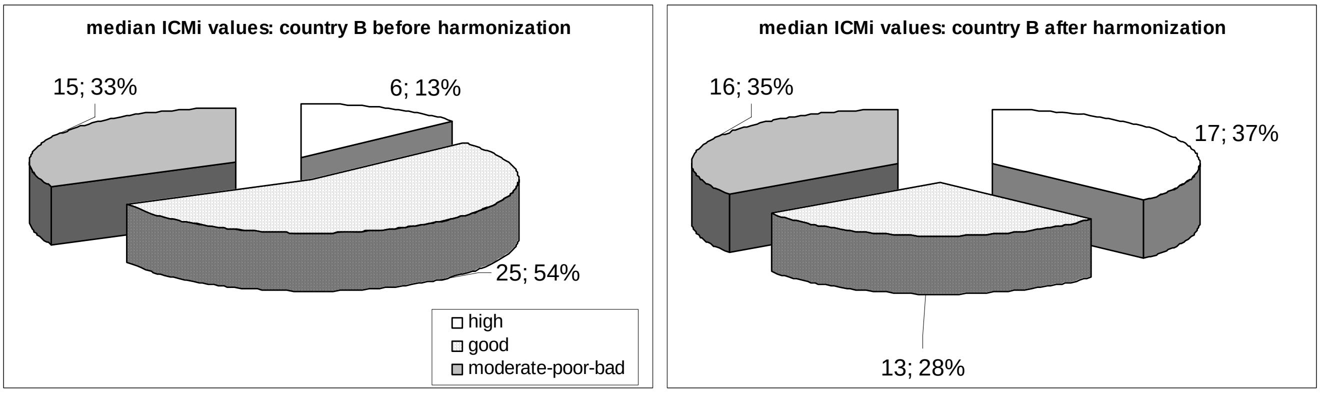


Figure 7.13 Percentage of samples belonging to High, Good and Moderate/Poor/Bad classes before (according to national method) and after harmonization in Country B.

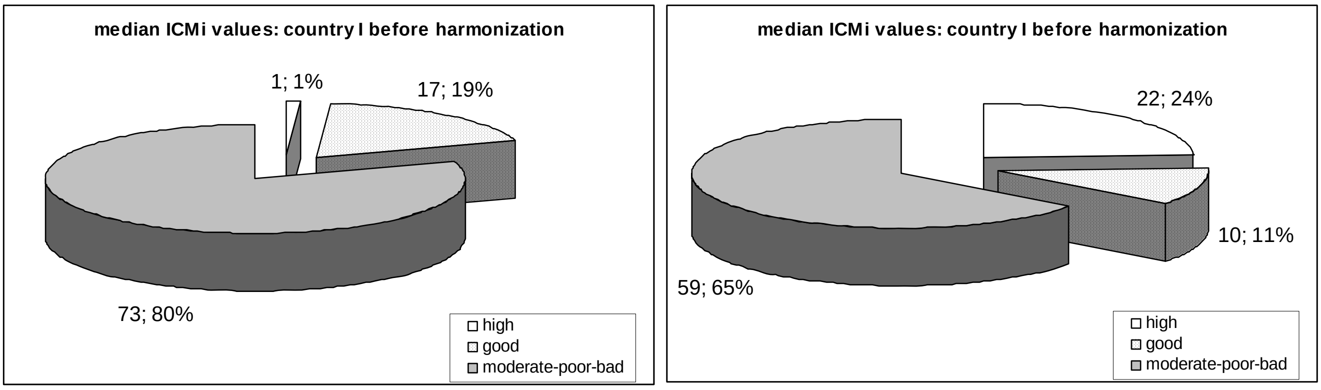


Figure 7.14 Percentage of samples belonging to High, Good and Moderate/Poor/Bad classes before (according to national method) and after harmonization in Country I.

Country A. Decreasing of the High status sites, increasing for good and moderate. Permissive boundary HG

Country B. Restrictive boundary HG: the number of samples in High status increases. It does not correspond to an increase of the Good status samples and the sum of High + Good status shows almost no variation.

Country I. The boundary for High status seems quite very restrictive. For this country the classification is done according to ‘one out all out’ between two indices.

7.2.4 Inter-GIG comparison and harmonization for non-WFD compliant methods.

For non WFD compliant methods, a comparison can be undertaken with the median values of the boundaries obtained for WFD-compliant methods. Then, an harmonization option can consist on the redefinition of the original boundary on the WFD-compliant median value.

Only countries G and F show an ICMi value below the WFD-compliant methods median. Thus, only those two countries/methods – according to this harmization approach – should adjust their HG and, especially GM, boundaries.

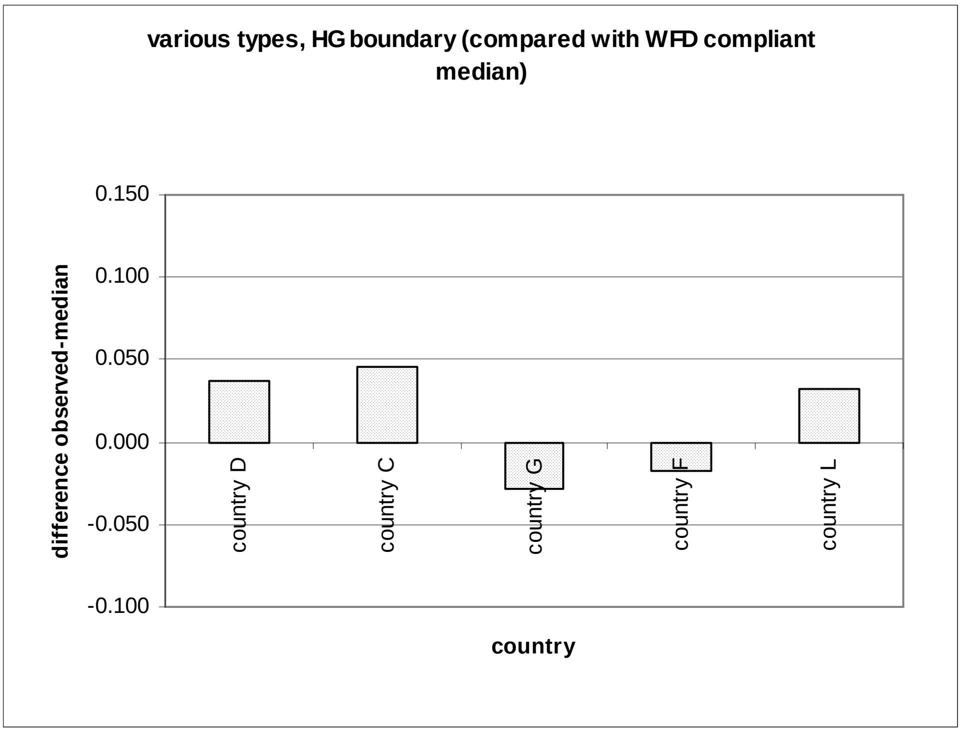


Figure 7.15 Boundary high-good set accordingly to the median value of WFD-compliant methods.

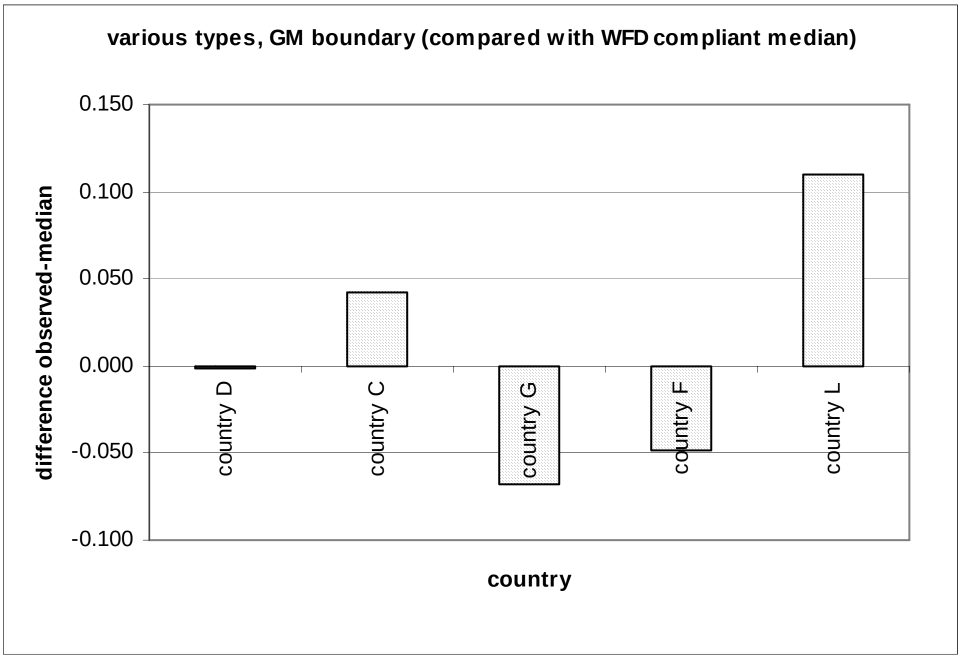


Figure 7.16 Boundary good-moderate setting according to median value WFD compliant methods.

Consistency of results for Countries F and G: in all the three comparison (intra-GIG, inter-GIG and against WFD-compliant) both HG and GM boundaries might have to be increased. The same for country L, always above the median.

Country C and D are both above the median.
Country F. Increase of the high and moderate status, decrease of the good status.

Discussion
The results of present exercise is strictly dependent on the included datasets. The inclusion or exclusion of one or more datasets has an influence on the value of the median and thus on the final boundary redefinition. It must be solely seen as an example of procedure.

The viewpoint of such exercise should be that all the involved countries have to include a set of data. Following the overall indication of the guidance (EC, 2004) this should be done for every stream type within a GIG. If suitable, the exercise can be extended to a more broad level, e.g. entire GIG, or even across GIGs.
As already stated, this exercise has to be undertaken by using to calculate the boundary values WFD-compliant methods only, i.e., in this case, among national methods that fulfil the requirements of the Directive.

8. Harmonization

In this Chapter, some examples of possible approaches to harmonization of class boundaries are provided, based on AQEM/STAR data only (7.1) and on both MSs and AQEM/STAR datasets (7.2). A few examples based on the sole use of MSs data have been presented in the previous Chapter (6.2, 6.3).
The first example shown, based on R-C4 data,

8.1 Bilateral harmonization

8.1.1 Averaging class boundaries of national methods - Same sample (no ICMi)

Harmonisation of ecological quality classification via averaging of class boundary values

Introduction
Based of the results of the “direct comparison approach” this section presents a simple procedure to obtain harmonised quality class boundaries via averaging of boundary values.

Methods
The “direct comparison approach” outlined in chapter 7.1 yields regression formulae for the bilateral relationships between assessment indices. Based on the national definition of the good ecological quality range, EQR boundary values are compared by converting each of them into the corresponding values of a benchmark system.
This example comprises the results of the comparison analysis of 247 AQEM/STAR samples from stream type R-C4. Table 7.1 lists converted boundary values for the good status of the indices SI (DE), ASPT (UK), ASPT (SE), DSFI (DK) and DSFI (SE). Harmonisation is done by averaging all EQR boundary values per national index.

Table 7.1: Individually predicted class boundaries and their average values per national method

R-C4			SI (DE)	ASPT (UK)	ASPT (SE)	DSFI (DK)	DSFI (SE)	average
predicted	high|good	**SI (DE)**	**0.899**	0.981	0.911	0.948	0.909	0.929
		ASPT (UK)	0.865	**1.000**	0.900	0.935	0.887	0.917
		ASPT (SE)	0.865	1.000	**0.900**	0.935	0.887	0.917
		DSFI (DK)	0.839	1.021	0.880	**1.000**	0.900	0.928
		DSFI (SE)	0.839	1.021	0.880	1.000	**0.900**	0.928
	good|moderate	**SI (DE)**	**0.728**	0.904	0.840	0.835	0.869	0.835
		ASPT (UK)	0.713	**0.890**	0.800	0.799	0.840	0.809
		ASPT (SE)	0.713	0.890	**0.800**	0.799	0.840	0.809
		DSFI (DK)	0.587	0.866	0.740	**0.714**	0.800	0.741
		DSFI (SE)	0.587	0.866	0.740	0.714	**0.800**	0.741

Results

The very right column of Table 1 displays the average EQR boundary values per assessment index. Figure 1 shows the effect of harmonisation on the distribution of quality classes in the test dataset. Before harmonisation of both reference conditions (95^{th} percentile of all AQEM/STAR samples pre-classified as high status, see chapter 7.1 for details) and class boundary values 18 percent of samples have been classified equally by all five indices. After harmonisation 44 percent of samples are of equal quality status (high, good, or moderate and worse).

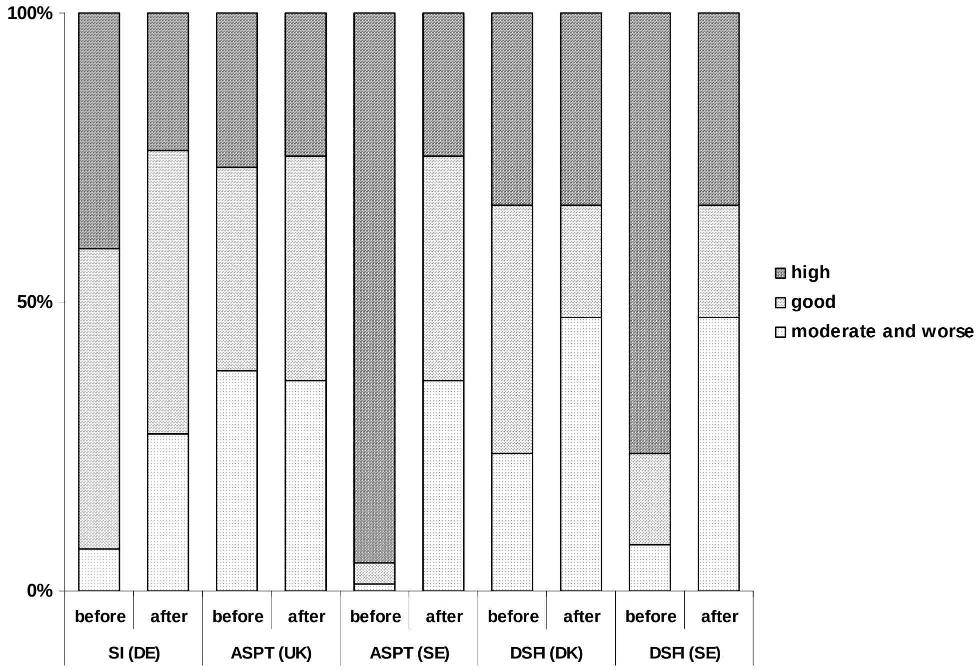


Figure 1: Distribution of quality classes before and after harmonisation via averaging of class boundary values (AQEM/STAR benthic invertebrate dataset; n=247)

Discussion

Prerequisite for harmonisation of class boundary values via averaging is including all assessment methods used to evaluate the quality of the respective common intercalibration stream type in the GIG. This chapter presents only an example of the option.

The performance of harmonisation is demonstrated by the increase of equally classified samples from 18 to 44 percent. For two reasons high percentages of conformity between individual classifications cannot be expected:

(1) The more assessment methods are involved in the process of harmonisation via averaging, the higher the difference can be between average class boundary and the optimal boundary value found in bilateral comparison (intersections located on regression line).

(2) The unexplained variance of the individual regression models included in the analysis causes different classifications of individual samples.

Conclusions

In general, harmonisation of quality classification has only to be executed if the direct or indirect comparison analysis reveals major discrepancies between national class

boundary settings. Since a large number of different methods are compared this is very likely to happen.
The option of averaging class boundary values of all assessment methods applied to a common stream type can represent an alternative to the harmonisation approach using benchmark datasets (cf. chapter 8.3). Both direct and indirect comparison approaches may serve as basis for this harmonisation option. Class boundary averaging is particularly recommended if compared class boundaries have been derived in full compliance with the WFD requirements, and appropriate benchmark datasets are not available. Averaging forms the least common denominator of the country-specific (WFD-compatible) concepts of ecological quality status.

IMPORTANT WARNING

The option of averaging class boundary values of assessment methods is only applicable when all the considered biological methods are demonstrated as fully WFD-compliant.

In addition, the use of this option is acceptable if all MSs contribute in the calculation of the boundary values (i.e. they all have WFD-compliant methods for that stream type at the time of the IC process).

This option, while requiring consistency to normative definition, does not support a real comparability across stream types and GIGs, thus possibly limiting the aptitude of the European IC process.

The boundaries shown here represent merely an example, because the quality classification of the samples is based on a pre-classification (i.e. not fully WFD-compliant).

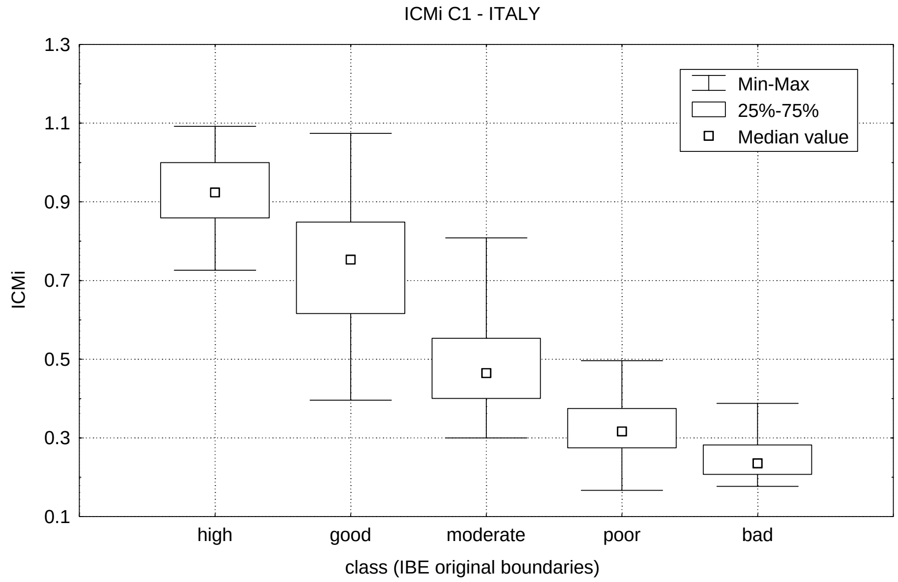
8.3 Harmonization of class boundaries: indirect comparison *via* ICMi

In this chapter some examples of harmonization of class boudaries based on the comparison of National datasets (test datasets) against a trans-National classification (benchmark datatset: STAR/AQEM data) are reported. These examples are taken from some test datasets described in Chapter 4. In particular, to illustrate the procedure, the results of the process of harmonization of class boundaries for C1 (6 examples), C2 (2 examples) and M1 (1 example) are described. A detailed example of the whole procedure is given for Italy and Poland.

8.3.1 Harmonization of class boundaries: indirect comparison, some specific examples from Italy and Poland R-C1

In this section are reported two specific examples of harmonization of class boudaries based on the comparison of National datasets (test dataset: R-C1 Italy and Poland) against a trans-National classification (based on samples form the whole of Europe) obtained following similar approaches and respecting the WFD requirements (benchmark datatset: STAR/AQEM data).
In figure 7.1 it is represented the variation of the ICM index in the five classes of Italian IBE for the sites belonging to intercalibration type C1.

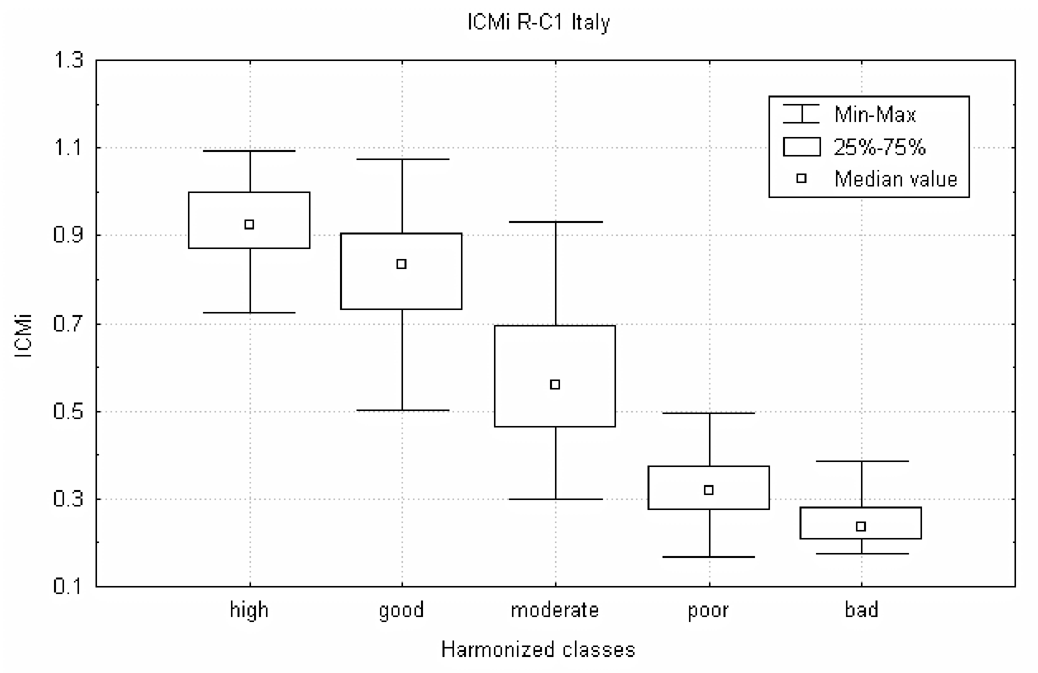
Figure 8.1 Variation of the ICM index for C1 – Italy within the IBE classes



The ICM index well reflects the quality classes derived from the IBE method, even if with some overlay between good and moderate classes if all the range is considered, but no overlap between interquartile range of high, good and moderate classes. The results of the Tukey test show as high good and moderate classes are statistically different ($p<0.000$).

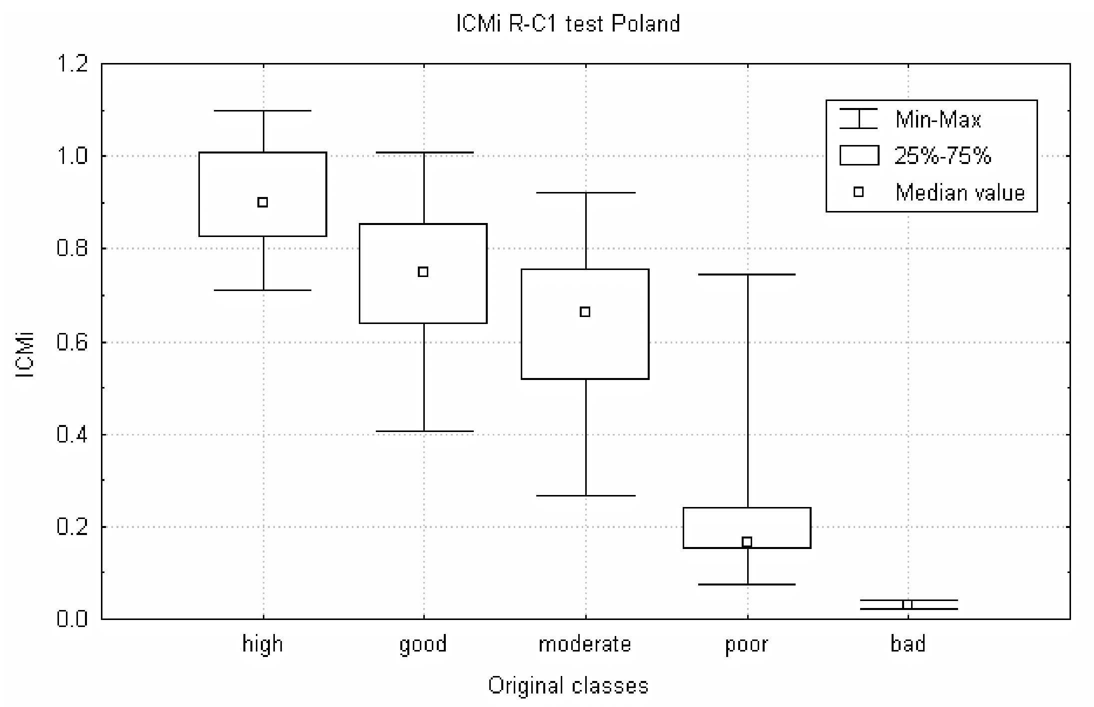
For the harmonization of class boundaries the values of the ICMi obtained for the test dataset are compared to the ICMi obtained for benchmark data. The median values of the ICMi in high and good classes are statistically compared with the median values of the ICMi calculated for the benchmark dataset in high and good status classes in order to see if differences exist. Firstly, the good status is tested: if differences exist (and test data lower than benchmark ones) the Good/Moderate boundary is shifted. The comparison of the median values of the ICM index from test and benchmark dataset reveals statistical differences for good status classes ($p<0.0000$). The new step of the process of harmonization involves the repositioning of the boundary until no more differences in the median value of the combined metrics are found by statistically comparing with the values observed in the STAR/AQEM samples. The boundary between Good and Moderate classes is shifted up, because median value is lower in test dataset with respect to benchmark. The threshold is repositioned step by step (e.g. from 7.6 to 8), until there are no more differences betweeen the values got by the ICMi according to the STAR/AQEM and IBE classification. What has been observed is that statistical differences are found until the boundary is shifted to 8.6, with a respective p level of 0.054. The new good/moderate boundary is so fixed at 8.6. After having compared and tested good status classes, the high status classes are compared and tested. The result of the Mann Whitney test shows a significant difference ($p=0.040$) for high status samples according to benchmark and test classification. Because of this difference, the boundary high/good is shifted up step by step as it was done for the previous boundary. To eliminate the differences it is enough to shift high/good boundary from 9.6 to 10, with a corresponding p value of 0.09. In figure 7.2 the variation of ICMi within IBE classes after harmonization is shown. The samples that move from high to good status are included in the good status box and the samples that move from good to moderate are incuded in the moderate status box. The interquartile range of good and moderate classes still remain separated. For IBE original classes, the separation between good and moderate classes is more evident because the values of ICMi included in this class are comparatively low (25^{th} percentile = 0.40). Moving the boundary the good status samples have a lower distance to the moderate status, but the two classes still remain statistically different ($p<0.000$). In general, after the harmonization the values of ICMi in good classes are higher with respect to the original classification and the same is observed for moderate status where the 25^{th} percentile is shifted up to aproximatively 0.5.

Figure 8.2. Variation of the ICM index for C1 – Italy within the IBE classes after harmonization.



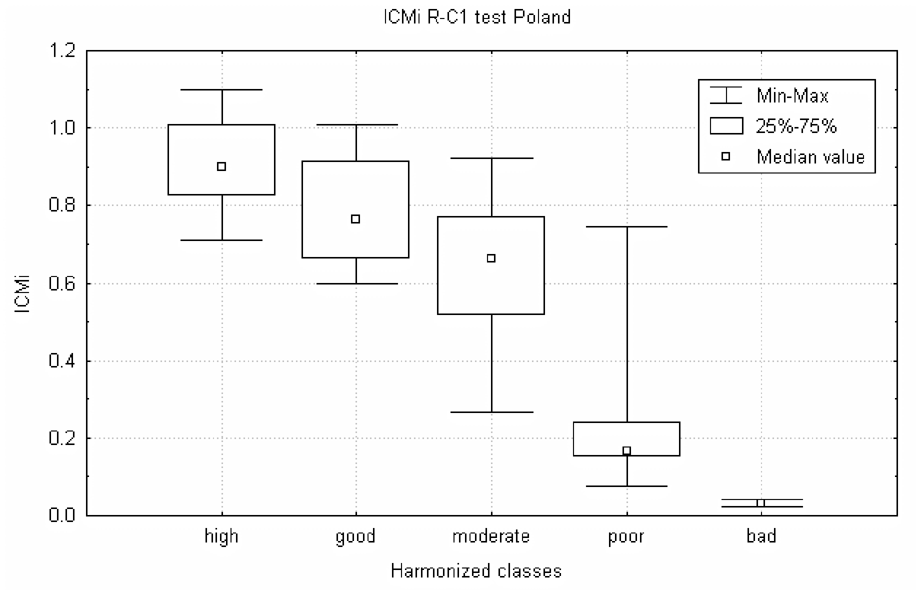
A second example for which it is necessary to shift boundaries because of the difference with benchmark data is R-C1 Poland. Figure 3 confirms the results of Chapter 4.3 indicating as the ICMi well follows the ecological gradient, even if the Tukey test does not indicate any significant difference between good and moderate class ($p = 0.46$). Neither significant differences were found between high and good status classes ($p = 0.15$)

Figure 8.3. Variation of the ICM index for C1 within Poland classes before harmonization.



In this example significant differences are found between good status test and benchmark samples (p=0.014). Due to the fact that for Poland sites'classification is derived from the combination of BMWP classification and Margalef classification, the repositioning of the boundaries was firstly done shifting up the threshold good/moderate for BMWP. In particular the boundary was moved up of 5 scores (from 70 to 75). The new classification produced for BMWP is than compared to the one of Margalef on the basis of the principle "one-out all-out". Is this last classification that was tested, in order to see if significant differences are still found. For the samples included in this example it is always the BMWP classification to determine the final classification, this mean that even with the harmonization of BMWP its classification it the worst one with respect to Margalef classification (as it was considering the original boundaries). The p level for the comparison of the new good test classification and benchmark one is 0.12, indicating that no more differences exist. For what regard the comparison of high status samples the test does not find any differences (p=0.22). With the harmonization of boundaries the variability of good samples is reduced. After the repositioning of the good/moderate boundary the lower samples of the good class are the ones shifting to the moderate status thus reducing the variability observed in the good class. It seems also that good and moderate classes are less overlapping. The Tuckey test still does not find statistical difference between good and moderate classes, but the p level is lower than the one with original clasification (p = 0.13).

Figure 8.4. Variation of the ICM index for C1 – Poland within the national classes after harmonization.



In table 8.1 the boundaries before and after harmonization for Italian IBE and Poland in R-C1 are reported.

Table 8.1. Class boundaries for the Italian and Polish standard assessment systems for R-C1 river type.

		Poland			Italy	
		BMWP score	BMWP harmonized	Margalef-DI score	IBE score	IBE harmonized
Limit	high-good	100	100	5.5	9.6	10
	good-moderate	70	75	4	7.6	8.6
	moderate-poor	40	40	2.5	5.6	5.6
	poor-bad	10	10	1	3.6	3.6

The figures 8.5 and 8.6 show the percentage of sites belonging to the different classes before and after harmonization in Italy and Poland. For Italy the number of samples moving from good to moderate status is 69, corresponding to 19%. At the beginning, 72% of the samples did not require a restoration action (including high and good samples) while after harmonization the samples that require a rehabilitation (50%) correspond to a more realistic picture for a river type located in an highly urbanized area. Anyway, some further considerations are necessary. The Italian legislation (D.L.vo 152/99) requires, to derive a final site classification, a comparison of the biological and chemical classification, by finally classifyibg according to the one-out all-out principle. The final classification will then be determined by the worst of the two, so it could happen that some of the sites classified in good status for biology, shift to moderate because of chemistry and *vice versa*. Further analysis should be addressed to the comparison of chemical and biological data. Furthermore, detailed investigations are required, because many samples from the same site get different classification depending on the year/season of investigation. Of aproximatively 16 sites belonging to high status (84 samples in total), 4 sites get a stable classification in high status, 4 are present both in high and good status and the left are more often classified in good status, with two sites equally present in high, good and moderate status. For the good classification, out of 29 sites belonging to this status, 10 sites get a stable classification (not dependent from season or year). For moderate status, out of 16 sites, 6 can be considered as always classified in moderate status.

For Poland the percentage interested to a shifting to a moderate status is smaller (9%) with respect to Italy.

Figure 8.5. Sample distribution according to national italian classification before and after harmonization.

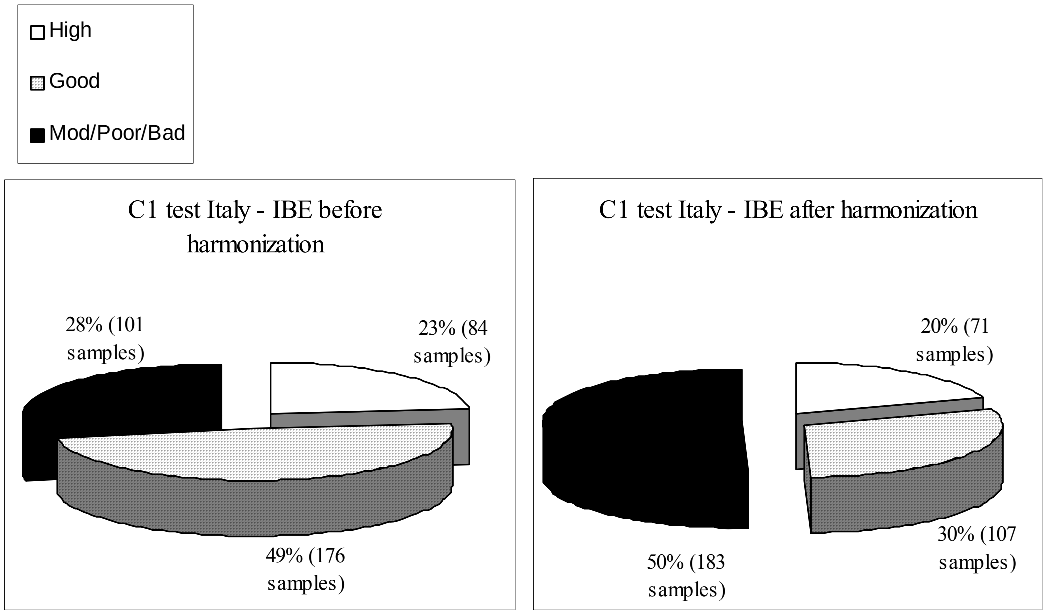
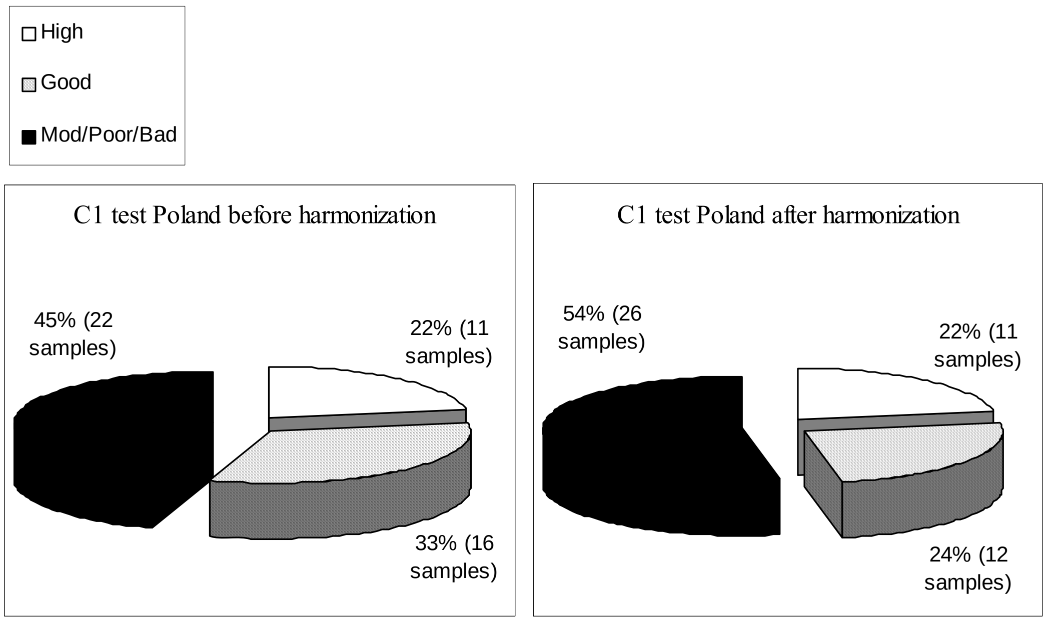


Figure 7.6. Sample distribution according to national polish classification before and after harmonization.



As a further test, a direct comparison (without comparing to a benchmark dataset) can be carried out between these two datasets. The result of the Mann Whitney U test shows that no differences are found, nor for high status samples ($p= 0.52$), neither for good status samples ($p=0.84$) for the two datasets. Making the comparison undirectly, *via* the benchmark dataset, shows that boundaries should be adjusted for both countries in order to eliminate the differences between benchmark and test datasets classification. Thus, the comparison with an external dataset is recommended, expecially when the compared methods are not WFD-compliant (e.g. IBE). In particular, the indirect comparison with a benchmark dataset, which has to be WFD-compliant, works out the problem of not having WFD fulfilling methods.

8.3.2 Overall comparison of C1 test data from different MSs

The ICM index was calculated for the benchmark dataset (see chapter 5 and 6) and for test datasets (see chapter 4 and 6). The procedure of indirect comparison *via* benchmark and ICMi, as seen in the previous paragraph, requires the contrast of a test dataset against a benchmark one, through the values of the ICM index. The values of the ICM index are compared among high status classes of benchmark data and high status classes of test data. In figure 8.8 the variation of the ICMi in high status classes for some tested datasets belonging to C1 type in comparison with benchmark values (left part of the graph) is presented. As it has been done for high status, the comparison of ICMi values was done also in good status (figure 8.9). The lower median values are observed for Country F and G; the value reported for Country I refers to the only one high status sample, for that stream type.
The second step is the statistical comparison among the two datasets in order to see if there are ant significant differences. If differences are found, the process requires the repositioning of the appropriate boudaries in the test dataset in order to eliminate these differences.

Figure 8.8. Variation of the ICM index for R-C1 for high status classes according to different datasets (benchmark vs national standard methods and datasets). National standard boundaries are considered for test data.

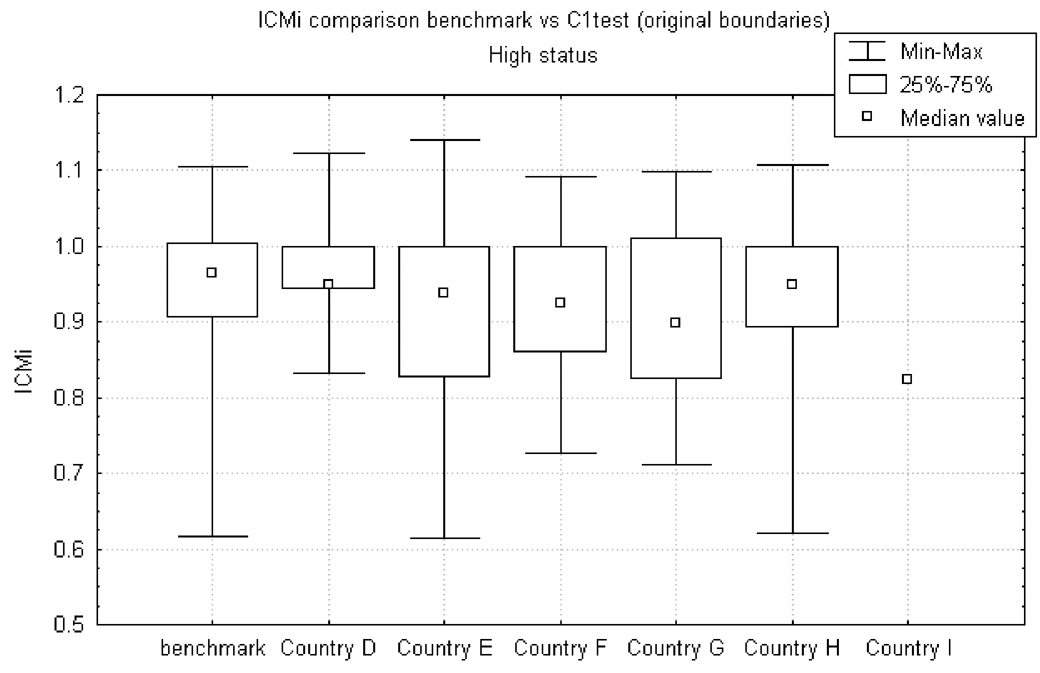
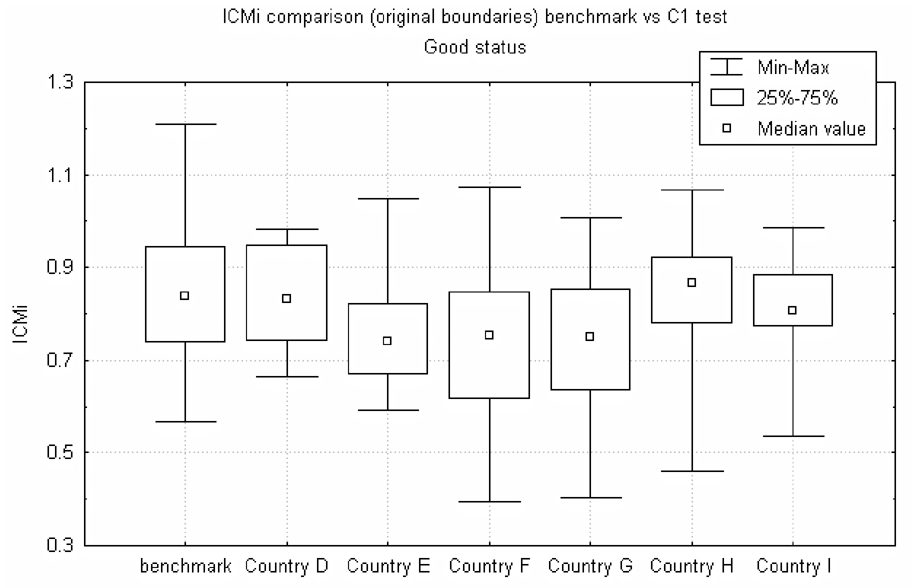


Figure 8.9. Variation of the ICM index for R-C1 for good status classes according to different datasets (benchmark vs national standard methods and datasets). National standard boundaries are considered for test data.



Firstly, the good status samples are tested. The application of the Mann Whitney test reveals differences for good status between benchmark and test data for countries E (p=0.0007), F (p<0.0000) and G (p=0.01). For high status, country E and F have significant differences compared to the benchmark with a respective p level of 0.035 and 0.04.

In general, the repositioning of the boundaries does not imply large adjustments. In figure 8.10, the variation of the ICM index after the harmonization for good status samples, including also the examples for which no difference was found with benchmark, is presented. In figure 7.11, the variation of ICMi in high status after harmonization is shown.

Figure 8.10. Variation of the ICM index for R-C1 for good status classes after harmonization.

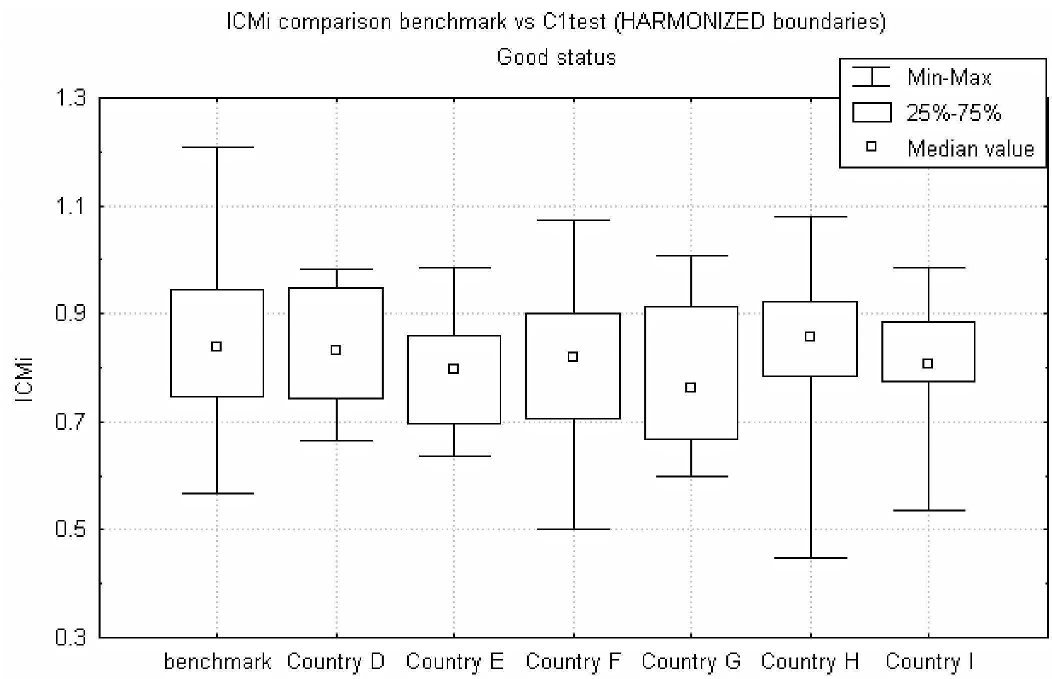
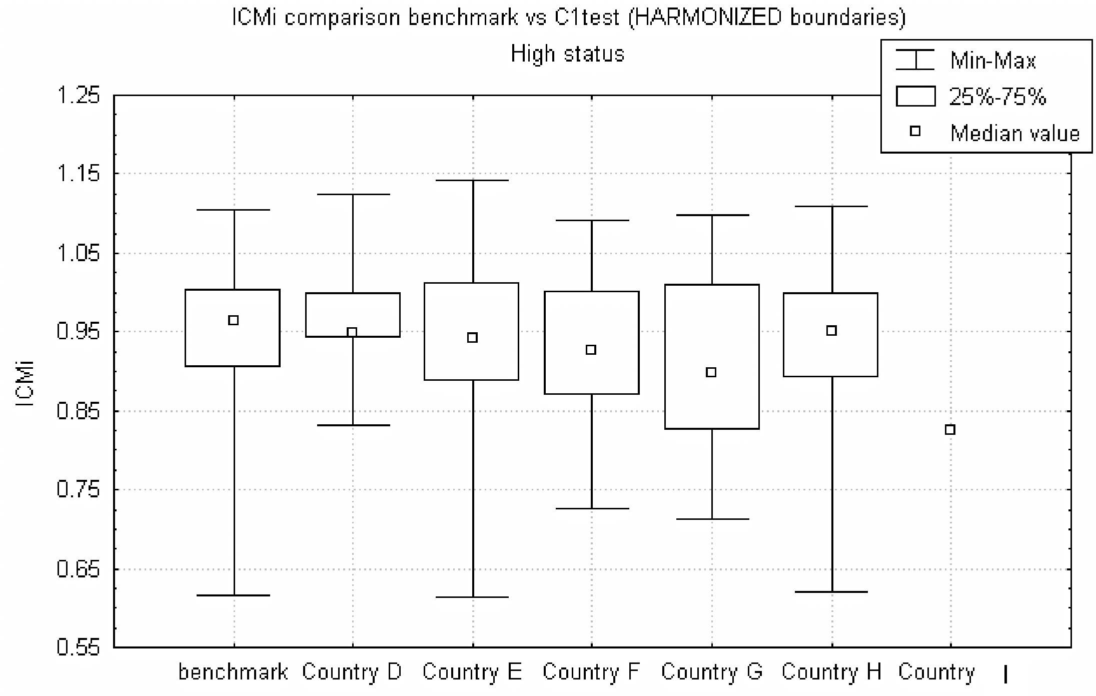


Figure 8.11. Variation of the ICM index for R-C1 for high status classes after harmonization.



8.3.3 Comparison of C2 and M1test data

In this section, the comparison and statistical test for some examples derived from C2 and M1 test datasets - in the same way as previously done for C1 - are reported. In these examples, differences were found both for high and good status samples, for one of the two countries belonging to R-C2 (country A). Figures 12 and 13 show respectively the variation of ICMi in high and good status before harmonization, accordingly to the original national boundaries. The lowest median values are observed for Country A (R-C2), both for high and good status classes with a significant difference with the benchmark (p level 0.0008 for high status and p<0.000 for good status).

Figure 8.12. Variation of the ICM index for R-C2 and R-M1 for high status classes according to different datasets (benchmark vs national standard methods and datasets). National standard boundaries are considered for test data.

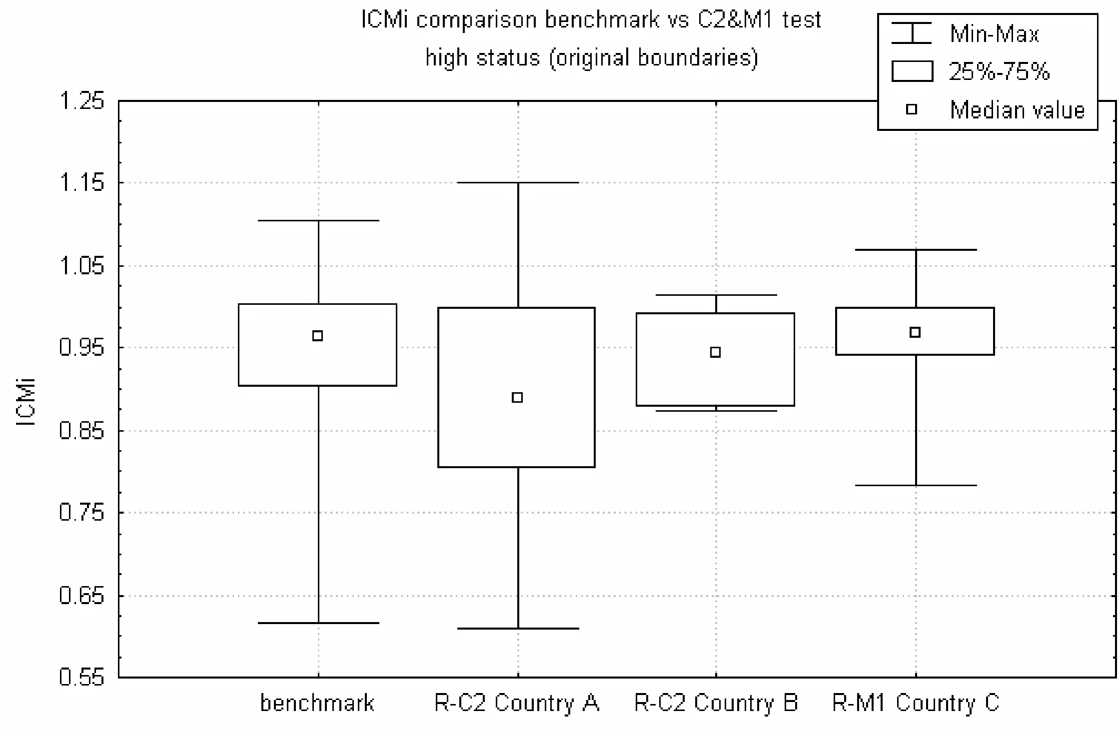
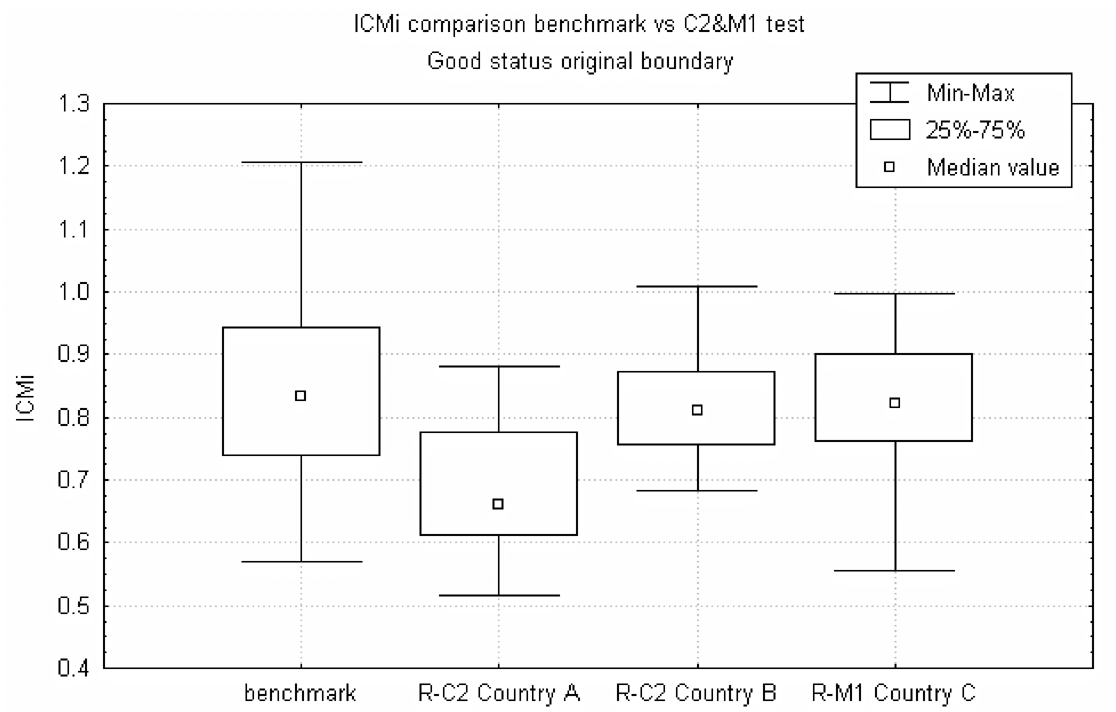


Figure 8.13. Variation of the ICM index for R-C2 and R-M1 for good status classes according to different datasets (benchmark vs national standard methods and datasets). National standard boundaries are considered for test data.



The process of harmonization involved the shifting up of the Good/Moderate and High/Good boundaries for country A in order to eliminate these differences. The results of the process of harmonization are represented in figures 8.13 and 8.14, in which it can be seen as in country A, for good status samples, the process of harmonization gives the additional result of diminishing variability.

Figure 8.14. Variation of the ICM index for R-C2 and R-M1 for high status classes after harmonization.

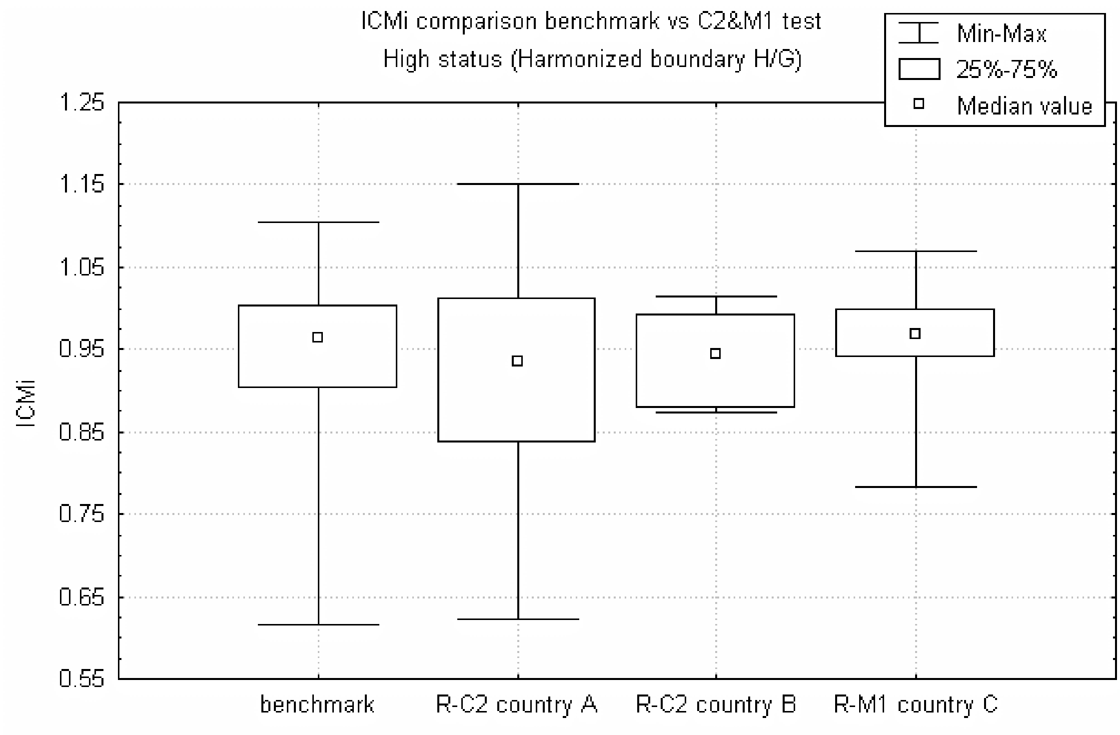
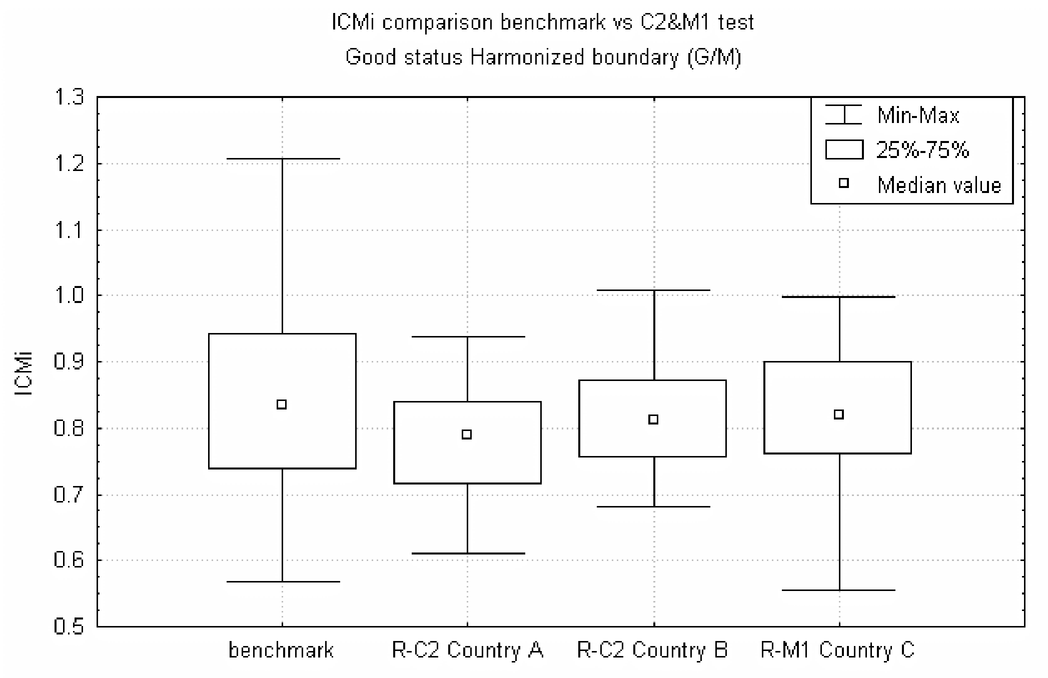


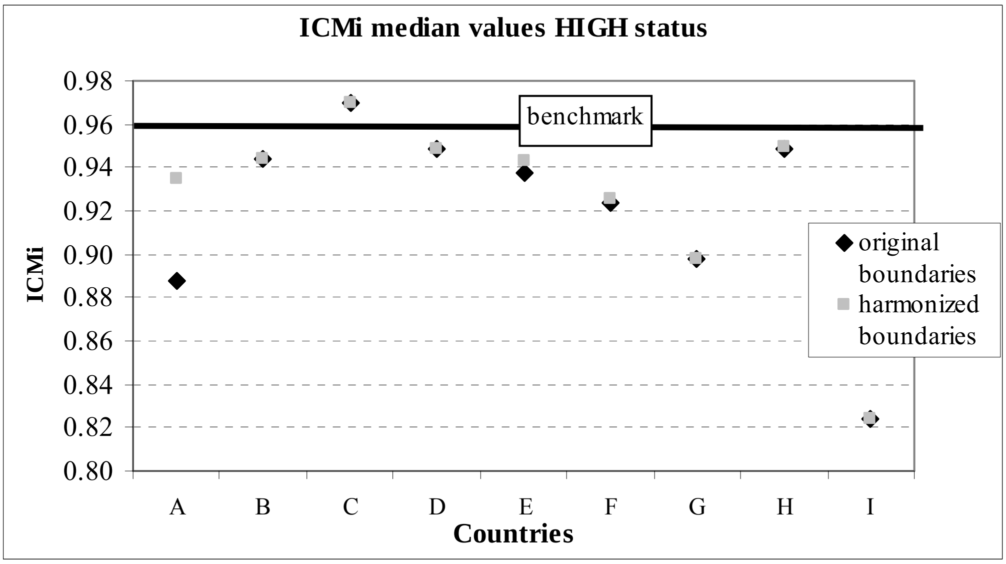
Figure 8.15. Variation of the ICM index for R-C2 and R-M1 for good status classes after harmonization.



8.4 Summary of harmonization results

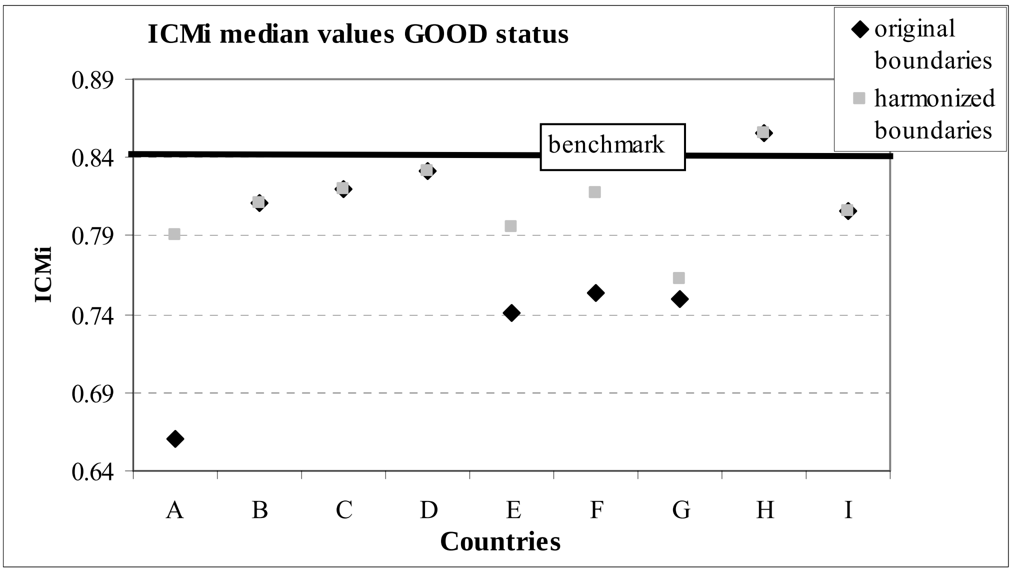
As a general result, it can be said that out of all the examples considered only 33% of countries had to adjust the High/Good boundary (3 countries out of 9 fig. 7.16). In general terms, the median values of the ICM index in high status samples according to the test datasets is lower than the median value of ICMi in benchmark and significant differences are found only for country A, E and F. The repositioning of the boundary involves very minor changes for country E and F. Country C is the only country which presents slightly higher median value compared to the benchmark. Country G presents quite low median value but no significant differences were found. This is probably due to the variability of the ICMi in the dataset, with quite a high maximum value (Tab. 7.1), but also because the number of tested samples is quite low (i.e. 11 high status samples). In general terms, it has to be stated that probably, when more samples will be included, the results of the statistical test can change, determining the need of a slight harmonization. In general, the percentage of samples moving from good to moderate status is around 10% for all the examples considered (with the exception of C1 Italy).

Figure 8.16 median values of ICMi for benchmark dataset (black line) compared to test datasets within high status (before and after harmonization)



In fig. 8.17 the median values of the ICM index for good status samples according to national classification before and after harmonization are presented. The black line represents the median value of ICMi for good samples within the benchmark dataset. It can be seen as all the samples have lower median values with respect to the benchmark with the exception of country H. For the Good/Moderate boundary, 4 out of 9 countries have to shift their boundaries, because only in 4 cases significant differences were found between test and benchmark data. The median values of almost all countries get close to the median value of benchmark after harmonization, with the exception of country G, whose median remains the lowest and correspond to 0.76.

Figure 8.17 median values of ICMi for benchmark dataset (black line) compared to test datasets within good status (before and after harmonization)



In table 8.2 (above), a summary of the basic statistic for the examples tested and harmonized and for the benchmark dataset is reported.

ICMi

	benchmark high	C2-A	C2-B	M1-C	C1-D	C1-E	C1-F	C1-G	C1-H	C1-I	
min	0.616	0.610	0.874	0.780	0.832	0.614	0.726	0.712	0.621		test high (Original Boundaries)
max	1.105	1.152	1.015	1.070	1.124	1.141	1.092	1.098	1.098		
25percentile	0.905	0.803	0.877	0.940	0.942	0.827	0.859	0.825	0.896		
75percentile	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.001		
median	0.965	0.888	0.944	0.970	0.949	0.938	0.924	0.898	0.949	0.824	
min	0.616	0.481	0.874	0.780	0.832	0.614	0.726	0.712	0.621		test high (harmonized Boundaries)
max	1.105	1.124	1.015	1.070	1.124	1.141	1.092	1.098	1.109		
25percentile	0.905	0.897	0.877	0.940	0.942	0.887	0.868	0.825	0.892		
75percentile	1.000	1.005	1.000	1.000	1.000	1.011	1.001	1.000	1.000		
median	0.965	0.957	0.944	0.970	0.949	0.943	0.925	0.898	0.950	0.824	

	benchmark good	C2-A	C2-B	M1-C	C1-D	C1-E	C1-F	C1-G	C1-H	C1-I	
min	0.569	0.516	0.682	0.560	0.664	0.591	0.396	0.405	0.447	0.536	test good (Original Boundaries)
max	1.207	0.880	1.008	1.000	0.982	1.049	1.074	1.010	1.081	0.986	
25percentile	0.743	0.611	0.755	0.760	0.741	0.669	0.616	0.630	0.782	0.771	
75percentile	0.944	0.776	0.872	0.900	0.948	0.823	0.849	0.850	0.922	0.885	
median	0.839	0.660	0.811	0.820	0.832	0.741	0.753	0.750	0.856	0.806	
min	0.569	0.610	0.682	0.560	0.664	0.635	0.502	0.600	0.447	0.536	test good (harmonized Boundaries)
max	1.207	0.938	1.008	1.000	0.982	0.692	1.074	1.007	1.081	0.986	
25percentile	0.736	0.713	0.755	0.760	0.741	0.859	0.702	0.663	0.782	0.771	
75percentile	0.944	0.839	0.872	0.900	0.948	0.859	0.900	0.913	0.922	0.885	
median	0.839	0.790	0.811	0.820	0.832	0.796	0.818	0.763	0.856	0.806	

8.5 Discussion

The comparison between the classification derived by the standard methods of some European countries and that based on the best available information (i.e. BAC, based on AQEM, STAR and some additional data) was performed for a number of test datasets belonging to three stream types belonging to the Central and Mediterranean GIGs. This comparison was based on the values obtained at each site by calculating the selected ICMs. The general results for the studied stream types highlighted how a small refinement was usually sufficient to set new boundaries to the National methods quality classes to fit the benchmarking dataset (WFD compliant, BAC-based clustering of sites).

As expected, the situation found across Europe is not fully homogeneous, with some countries and boundaries being not statistically different from the established benchmarking dataset for all classes, some with discrepancies for one of them and others with both the relevant boundaries. The large differences in the conceptual basis between some of the methods compared, the different 'age' and the consistency with the WFD requirements can easily account for such disagreement. Nevertheless, it must be noted that the observed differences – whilst sometimes statistically significant – are not very high.

In general terms, a very low percentage (ca 0-6%) of samples moved from High to Good quality class, looking at all types and datasets. In such cases, it might mean that the initial setting of the boundary for a specified method for quality class I (High quality) is too generous for the studied stream type.

A slightly larger percentage (0-10%, in one circumstance 19%) of the samples initially classified into quality class II (Good) by the national classification schemes moved to class III (Moderate).

Thus, the harmonization by re-adjusting class boundaries *via* ICMi according to a trans-National, WFD-compliant classification, did not lead to the need for a weighty adaptation of the National classification schemes. The comparison was performed at the sample level, which means that in many cases the refinement will lead to restoration measures to be taken for a sub-set of the sites only.

The procedure applied here and the illustrated results can support the involved MSs – which will get access to their own results in a non-blind version - in the revision of the class boundaries to make their methods in better accordance with WFD requirements.

At a larger scale, like presented here for the stream type R-C1 or for the trans-GIG comparison – which, in principle, is what the European IC process is addressed to - we preliminarily tested the equivalence of boundaries among countries and classification systems. In fact, while in a previous example of application benchmark data were derived form a single ecoregion (Italy: Buffagni & Erba, 2004), here we used a trans-National, inter-GIG database, which is expected to include rivers even quite dissimilar in their general character. In IC applications the boundaries for a stream type are expected to be re-set not lower than at the GIG scale. Based on the examples provided here, we wonder if a trans-GIG Intercalibration might be tentatively adopted.

Even if exclusively comparing metrics based on a high taxonomic resolution (e.g. to the Family level), the variabilty among stream types belonging to different GIGs is calculated to be higher than within a GIG. The main test stream type considered here (small, lowland, sandy streams) has a counterpart in the benchmark dataset, which nevertheless contains more data form other stream types. After normalizing, stream types not differing too much in character can be satisfactorily compared for the purposes of the IC process. Among the tast datasets and stream types considered in the present Deliverable, one only of the types resulted harly comparable with others. Not surprisingly, it is R-M5, which corresponds to South European temporary rivers. Apart from this lone illustration, the other types and datasets – covering a very wide geographic range, from UK to Poland, from Germany to Southern Italy – provided highly comparable results.

9 General conclusion

An extended overview of test datasets obtainable around Europe for the IC process has been provided for selected stream types, together with examples of the AQEM/STAR datasets, which might be used as benchmarking systems.

Many approaches to the European Intercalibration of class boundaries of biological assessment methods have been outlined and preliminarily tested. Each of them can have potential application depending on the kind and amount of data available, proximity of methods to be harmonized, availability of reference sites, etc.

The main features and use of common metrics for the IC process (ICMs) have been described and applied to a number of problems and situations in Europe, covering a wide geographical range.

The comparison exercise between European class boundaries and assessment systems lead to different results for different stream types and options used, but showed how systems and boundaries are actually comparable in short time.

Examples of harmonization have been presented following three different approaches, which might be used individually or combined in different GIGs and European ares.

In general terms, some conclusions can be drawn:

- The comparison of a relatively high number of European MSs' datasets have been performed by using a simple ICM index for making them fully comparable. The general outcomes indicate that the ICM approach is suitable for comparing rivers and invertebrate communities along a wide range of situations.
- ICMs and ICMi were stressed against a high number of European biological assessment methods and resulted in very high correlations with most of them (i.e. they are able to describe the quality gradient actually detected by the methods presently in use).
- The response of many metrics, including ICMs, was analized for groups of test stream types along observed pressure gradients. While a few metrics performed sometimes slightly better (i.e. species-level metrics), ICMs demonstrated a very good general attitude (especially if we consider they are based on a Family level identification).
- The ICM approach supports the use of existing datasets directly collected by MSs, which can guarantee a good availability of data for the IC process.
- The procedure to calculate the ICMi and compare datasets is now well described and readily applicable by European countries, GIGs or European Community delegates.

- A common restriction to all the possible procedures for the IC process is linked to scarce availability of data from reference sites.
- The direct comparison approach (i.e. not using ICMs) has been used to demonstrate apparent discrepancies between MSs' assessment systems boundaries (up to 50% for the High/Good boundary).

- The same approach has potential for IC harmonization purposes especially when the compared systems are quite similar (e.g. for bilateral, fine tuning of class boundaries) and when large datasets with collected samples, which satisfy the requirements of the compared methods, are available.
- The ICMi approach, by using Reference conditions and data normalization set within each of the datasets being compared, allows a large variety of Intra- and Inter-GIG comparisons (i.e. it supports a large pan-European comparability).
- As well, by using an entirely external benchmarking system (in the present Deliverable, the AQEM/STAR WFD-compliant dataset), the ICMi can be used to harmonize class boundaries within and between GIGs, getting a full comparability and unambiguousness of results.
- The examples presented show how direct comparison - applying different assessment methods to the same sample – found up to ca 50% of difference between existing boundaries. The differences among boundaries observed with comparison *via* ICMi were much lower – even comparing datasets from different GIGs - ranging from 0 to around 10% (usually lower than 5%).
- The harmonization approach via ICMs and external benchmarking do not require fixed class boundaries to be defined. Moreover, it allows a step by step adaptation of National methods boundaries until no more differences are observed between National samples and benchmarck samples, for Good and High quality classes, in sequence.
- The complete harmonization exercise provided here via ICM approach (full IC Option 2 application) lead to quite interesting results.
- More than half of the considered assessment systems resulted already aligned to the benchmarking system (i.e. no statistical differences observed, which means that no boundaries should be refined).
- Comparison and harmonization using benchmark datasets handles the problem of not having fully WFD-compliant systems presently available. If the comparison of the tested datasets with benchmark ones does not show significative differences, it means that the tested method can be considered provisionally fulfilling WFD.

A few very important, general warnings can be highlithed:

- The option of averaging the values of class boundaries of MSs' assessment methods is only applicable when all the considered biological methods are demonstrated as fully WFD-compliant.
- If calculated on the basis of MSs biological protocols only, the simple agreement on the use of any statistical values (e.g. median, 75^{th} %ile) as an anchor value to set Reference conditions for EQRs calculation, is not acceptable for the formal IC process, because it would not guarantee conformity to the WFD.
- An important requirement for a successful application of most of the described procedures is the availability of datasets covering the whole degradation gradient.

Acknowledgments

We are especially grateful to Jean-Gabriel Wasson (CEMAGREF, Lyon, France), who participated in all the major discussions and meetings that lead to the preparation of the present Deliverable, nevertheless not being part of the STAR Consortium. For the same reason, we thank Wouter van de Bund (JRC, Ispra, Italy), who effectively managed the continuous relationship between STAR partners and JRC CIS activities.

Also, we are grateful to Dr. Pietro Genoni (ARPA Lombardia, Italy), who provided the test data systematically used in the preparatory phases (i.e. ECOSTAT Discussion paper) and in the Deliverable itself to describe the ICMi harmonization procedure.

We want to thank all Central and Mediterranean GIG delegates who provided data presented in the Deliverable, as well as the STAR partners who contributed to the discussion on Intercalibration.

10 Main references

Alba-Tercedor, J. & S. Ortega, 1988. Un método ràpido y simple para evaluar la calidad biològica de las aguas corrientes basato en el de Hellawell (1937) de Limnetica 4: 51-57.

Alba-Tercedor J., Jáimez-Cuellar P., M, Álvarez, Avilés, J., Bonada, N., Casas, J., Mellado, A., Ortega M., Pardo, I., Prat, N., Rieradevall, M., Robles S., Sainz-Cantero C.E., Sanchez-Ortega A., Suárez, M.L., Toro, M., Vidal-Abarca, M.R., Vivas, S & C. Zamora-Munoz, 2002. Caracterizacion del estado ecologico de rios mediterraneos ibericos mediante el indice IBMWP (antes BMWP'). Limnetica, 21(3-4): 175-185.

APAT & IRSA-CNR, 2004. Metodi Analitici per le Acque. Indicatori biologici. 9010. Indice biotico esteso (I.B.E.). APAT Manuali e Linee guida 29/2003 (vol.3): 1115-1136.

AQEM Consortium, 2002. Manual for the application of the AQEM system. A comprehensive method to assess European streams using benthic macroinvertebrates, developed for the purpose of the Water Framework Directive. Version 1.0, February 2002, 202 pp.

Armitage, P. D., D. Moss, J. F. Wright & M. T. Furse, 1983. The performance of a new biological water quality scores system based on macroinvertebrates over a wide range of unpolluted running-water sites. Wat. Res. 17: 333-347.

Arnwine, D. H. & G. M. Denton. 2001. Development of Regionally-Based Numeric Interpretations of Tennessee's Narrative Biological Integrity Criterion. Tennessee Department of Environment and Conservation. Division of Water Pollution Control, Nashville, 136 pp.

Barbour, M. T., J.L. Plafkin, B.P. Bradley, C.G. Graves & R. Wisseman, 1992. Evaluation of EPA's rapid bioassessment benthic metrics: metri redudancy and variability among reference stream sites. Environmentsl Toxicology and Chemistry, 11: 437-449.

Barbour, M. T., J. Gerritsen, G. E. Griffith, R. Frydenborg, E. McCarron, J. S. White & M. L. Bastian, 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. J. N. Am. Benthol. Soc. 15 (2): 185-211.

Barbour, M. T., J. Gerritsen, B. D. Snyder, & J. B. Stribling, 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C.

Birk, S., 2004. Description of how stream type-specific reference conditions using macrozoobenthos have been derived in Germany. 2pp. Essen, 15 November 2004.

Böhmer, J., A. Zenker, B. Ackermann & B. Kappus, 2001. Macrozoobenthos communities and biocoenotic assessment of ecological status in relation to degree of human impact in small streams in southwest Germany Journal of Aquatic Ecosystem Stress and Recovery 8: 407–419.

Bonada, N., N. Prat, T. Munné, M. Plans, C. Solà, M. Álvarez, I. Pardo, G. Moyà, G. Ramon 3 , M. Toro, S. Robles, J. Avilés, M.L. Suárez, M.R. Vidal-Abarca, A. Mellado, J.L. Moreno, C. Guerrero,S. Vivas, M. Ortega, J. Casas, A. Sánchez-Ortega, P. Jáimez-Cuéllar, J. Alba-Tercedor, in press. INTERCALIBRACIÓN DE LA METODOLOGÍA GUADALMED. SELECCIÓN DE UN

PROTOCOLO DE MUESTREO PARA LA DETERMINACIÓN DEL ESTADO ECOLÓGICO DE LOS RÍOS MEDITERRÁNEOS. Limnética.

Buffagni, A., J. L. Kemp, S. Erba, C. Belfiore, D. Hering & O. Moog, 2001. A Europe wide system for assessing the quality of rivers using macroinvertebrates: the AQEM project and its importance for southern Europe (with special emphasis on Italy). J. Limnol. 60 (suppl.1): 39-48.

Buffagni, A., R. Alber, E. Bielli, S. Erba, J. L. Kemp, M. Cazzola, V. Albrigo, A. Mutschlechner & R. Pagnotta, 2002. Il Progetto AQEM in Italia: aree studiate, metodologia di campionamento e potenziali interazioni con altri settori di indagine. Studi Trentini di Scienze Naturali. Acta Biologica 78 (1): 181-193.

Buffagni A. & S. Erba, 2004. A simple procedure to harmonize class boundaries of European assessment systems. Discussion paper for the Intercalibration process – WFD CIS WG 2.A ECOSTAT, 6 February 2004, 21pp.

Buffagni A., S. Erba, M. Cazzola & J. L. Kemp, 2004. The AQEM multimetric system for the southern Italian Apennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe. Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 313-329.

Buffagni, A., S. Erba & R. Pagnotta, 2004. Carattere lentico-lotico dei fiumi mediterranei e classificazione biologica di qualità. Quad. Ist. Ric. Acque, 122: 165-189.

Bundi, U., A. Peter, A. Frutiger, M.Hütte, P. Liechti & U. Sieber, 2000. Scientific base and modular concept for comprehensive assessment of streams in Switzerland. Hydrobiologia 422/423: 477–487.

Crosa, G., P. Genoni & F. Stefani. 2001. Progetto di ricerca – Taratura e standardizzazione di metodi per la valutazione qualitativa dei corsi d'acqua minori della Provincia di Milano. ARPA Agenzia Regionale per le Protezione dell'Ambiente e Università degli Studi di Milano Bicocca, 111pp.

De Wilde A.J. Setting Classboundaries for the classification of rivers and lakes in Europe. Final Report. 39534/R0331/AdW/DenB.

D. L. 152/99 - Decreto Legislativo 11 maggio 1999, n. 152 - Disposizioni sulla tutela delle acque dall'inquinamento e recepimento della direttiva 91/271/CEE concernente il trattamento delle acque reflue urbane e della direttiva 91/676/CEE relativa alla protezione delle acque dall'inquinamento provocato dai nitrati provenienti da fonti agricole. Suppl. Ord. n. 101/L alla Gazzetta Ufficiale 29 maggio 1999, n. 124.

European Commission. 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327, 22.12.2000, 1-72.

European Commission. 2003 (a). Technical support in relation to the implementation of the Water Framework Directive (2000/60/EC) - TECHNICAL ANNEX, Contract for technical and administrative assistance (ATA model), DIRECTORATE-GENERAL ENVIRONMENT 17pp.

European Commission. 2003 (b). Common Implementation Strategy for the Water Framework Directive (2000/60/EC). *Guidance document n.o 6.* Towards a guidance on establishment of the intercalibration network and the process on the intercalibration exercise. Produced by Working Group 2.5 – Intercalibration, 54pp.

European Commission. 2003 (c). Overall Approach to the Classification of Ecological Status and Ecological Potential. Water Framework Directive, Common Implementation Strategy Working Group 2 A Ecological Status (ECOSTAT), 52pp.

European Commission. 2003 (d). Overview of common Intercalibration types and Guidelines for the Selection of Intercalibration sites. Water Framework Directive Common Implementation Strategy Working Group 2A 'Ecological Status' Version 3.1 – 9 October, 2003, 70pp.

European Commission. 2004 (e). ECOSTAT WG 2.A, 2004. Guidance on the Intercalibration Process. Water Framework Directive Common Implementation Strategy Working Group 2.A 'Ecological Status' Final Draft 4.1. October, 14th 2004.

Furse, M.T. 2001. "Standardisation of River Classifications: Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive" STAR Project, Part B - EU Contract number: EVK1-CT 2001-00089.

Genoni P., P. Beati, F. Buzzi, P. Casarini, M. Girami E. Gozio, V. Mafessoni, P. Roella & A. Sarzilla 1997. Intercalibrazione del metodo Indice Biotico Esteso I.B.E. (IRSA-CNR, 1995) per la valutazione della qualità dei corsi d'acqua. REGIONE LOMBARDIA DIREZIONE GENERALE SANITÀ SERVIZIO PREVENZIONE SANITARIA. 62 PP.

Genoni P., P. Beati, F. Buzzi, P. Casarini, M. Girami E. Gozio, V. Mafessoni, P. Roella & A. Sarzilla 1998. Intercalibrazione del metodo Indice Biotico Esteso I.B.E. (IRSA-CNR, 1995) per la valutazione della qualità dei corsi d'acqua. REGIONE LOMBARDIA DIREZIONE GENERALE SANITÀ SERVIZIO PREVENZIONE SANITARIA. 34 PP.

Genoni P., 2003. Influenza di alcuni fattori ambientali sulla composizione delle cenosi macrobentoniche dei corsi d'acqua planiziali minori. Biologia Ambientale 17 (1): 9-16,

Ghetti, P. F., 1997. Indice Biotico Esteso (I.B.E.). I macroinvertebrati nel controllo della qualità degli ambienti di acque correnti. Provincia Autonoma di Trento: 222 pp.

Helsel, D. R. & R. M. Hirsch, 1992. Statistical methods in water resources. Ed. Elsevier: 522 pp.

Hering, D., A. Buffagni, O. Moog, L. Sandin, M. Sommerhäuser, I. Stubauer, C. Feld, R. K. Johnson, P. Pinto, N. Skoulikidis, P. F. M. Verdonschot & S. Zahrádková, 2003. The development of a system to assess the ecological quality of streams based on macroinvertebrates – design of the sampling programme within the AQEM project. Internat. Rev. Hydrobiol. 88: 345-361.

Hering, D. & J. Strackbein, 2002. STAR stream types and sampling sites. STAR 1st deliverable, due 30/06/02. Standardisation of river classifications: Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive (Contract No: EVK1-CT 2001-00089), 28pp.

Hering D., O. Moog, L. Sandin & P. F. M. Verdonschot. 2004. Overview and application of the AQEM assessment system. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 1–20.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant & I. J. Schlosser, 1986. Assessing biological integrity in running waters: A method and its rationale. Special publication 5. Illinois Natural History Survey.

Kerans, B. L. & J. B. Karr, 1994. A Benthic index of biotic integrity (B-IBI) for rivers of the Tennessee valley. Ecological Applications 4 (4): 768-785.

Knobel, R. & M. van den Berg, 2004. Report on Intercalibration pilot exercise Central GIG Rivers from The Netherlands. Pilot exercise report. 7 pp. November 2004.

Köhl, M., B. Traub & R. Päivinen. 2000. Harmonisation and standardisation in multi-national environmental statistics – mission impossible? *Environmental Monitoring and Assessment* 63: 361–380.

Lafont, M. , J. C. Camus, A. Fournier & E. Sourp, 2001. A practical concept for the ecological assessment of aquatic ecosystems: application on the River Dore in France. Aquatic Ecology 35: 195–205.

Lazaridou-Dimitriadou, M. , V. Artemiadou , G. Yfantis, S. Mourelatos & Y. Mylopoulos, 2000. Contribution to the ecological quality of Aliakmon river (Macedonia, Greece): a multivariate approach. Hydrobiologia 410: 47–58.

Legendre P. & L. Legendre, 1998. Numerical ecology. Developments in Environmental Modelling 20. Elsevier, Amsterdam, 853pp.

Milner, A. M. & M. W. Oswood, 2000. Urbanization gradients in streams of Anchorage, Alaska: a comparison of multivariate and multimetric approaches to classification. Hydrobiologia 422/423: 209-223.

Morpurgo M., 1996. CONFRONTO FRA INDICE SAPROBICO (Friedrich e DIN, 1990) e INDICE BIOTICO ESTESO (Ghetti e IRSA, 1995). Biologia Ambientale 2-3: 30-36.

Morais, M., P. Pinto, P. Guilherme, J. Rosado & I. Antunes, 2004. Assessment of temporary streams: the robustness of metric and multimetric indices under different hydrological conditions. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 229-249.

Nijboer, R. C., R. K. Johnson, P. F. M. Verdonschot, M. Sommerhäuser & A. Buffagni, 2003. Establishing reference conditions of European streams. Hydrobiologia (in press).

Ofenböck, T., O. Moog, J. Gerritsen & M. Barbour, 2004. A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 251-268.

Parco V., A. Soana, G. Crosa, P. Genoni, F. Stefani, M. Bogani & R. Venturelli. 2003. Approcci metodologici per la valutazione dell'integrità ecologica dei corsi d'acqua minori della provincia di Milano. 1 Atti Seminario "Nuovi orizzonti dell'ecologia", Trento, 18-19 aprile 2002. G.N. Baldaccini & G. Sansoni (eds.), Ed. C.I.S.B.A., Reggio Emilia, 2003, 6pp (in press).

Pinto, P., J. Rosado, M. Morais & I. Antunes, 2004. Assessment methodology for southern siliceous basins in Portugal. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 191-214.

REYNOLDSON, T. B., K. E. DAY & T. PASCOE, 2000. THE DEVELOPMENT OF BEAST: A PREDICTIVE APPROACH FOR ASSESSING SEDIMENT

QUALITY IN THE NORTH AMERICAN GREAT LAKES. IN: WRIGHT, J. F., D. W. SUTCLIFFE & M. T. FURSE (EDS), ASSESSING THE BIOLOGICAL QUALITY FO FRESH WATERS – RIVPACS AND OTHER TECNIQUES. FBA: 165-180.

Rossaro, B. & A. Pietrangelo, 1993. Macroinvertebrate distribution in streams: a comparison of CA ordination with biotic indices. Hydrobiologia 263:109-118.

Roux, D. J., C. J. Kleynhans & C. Thirion, 1999. Biological monitoring and assessment of rivers as a basis for identifying prioritising river management options. Water Science and Technology, 39 (10-11): 207-210.

Sandin, L., M.T. Furse, Hering D., O. Moog, P. Verdonschot, R. Johnson, K. Brabec, K. Gritzalis, A. Buffagni, P. Pinto, N. Friberg, J. Murray-Blight, J. Kokes, R. Alber, P. Usseglio-Polatera, P. Haase, R. Sweeting, B. Bis, K. Szoszkiewicz, H. Soszka, G. Springe, F. Sporka and J. Krno. 2003. Standardisation of river classifications, a EU project contributing to the implementation of the Water Framework Directive. In: How to assess and monitor ecological quality in freshwaters. Ruoppa, Heinonen, Pilke, Rekolainen, Toivonen and Vuoristo (eds.), TemaNord, 547: 53-56.

Shannon, C. E. & W. Weaver, 1949. The Mathematical Theory of Communication. The University of Illinois Press, Urbana, IL.

Skriver, J., 2004. European intercalibration: Stream type R-C1 in Denmark. Pilot exercise report. 4pp. November 2004.

Spaggiari, R. & S. Franceschini. 2000. Procedure di calcolo dello stato ecologico dei corsi d’acqua e di rappresentazione grafica delle informazioni. Biologia Ambientale, 14 (2): 1-6.

Thorne R.S. & W.P. Williams. 1997. The response of benthic macroinvertebrates to pollution in developing countries: a multimetric system of bioassessment. Freshwater Biology 37: 671-686.

USEPA, 1999. Rapid Bioassessment Protocols for Use in Streams and Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Draft Revision June 2, EPA 841-D-97-002.

Vlek H.E., P. F. M. Verdonschot & R. C. Nijboer, 2004. Towards a multimetric index for the assessment of Dutch streams using benthic macroinvertebrates. In: *D. Hering, P.F.M. Verdonschot, O. Moog & L. Sandin (eds), Integrated Assessment of Running Waters in Europe.Kluwer Academic Publishers. Printed in the Netherlands.* Hydrobiologia 516: 173–189.

Woodiwiss, F. S., 1964. The biological system of stream classification used by the Trent River Board. Chemistry and Industry 14: 443-447.

11 Short glossary

Harmonization. The process by which the class boundaries of MS National methods should be accomodated to correspond to a common understanding of ecological status trans-National benchmarking. It must be preformed for High/Good and Good/Moderate status borders.

Class boundary. The EQR value representing the threshold between two quality classes.

EQR Ecological Quality Ratio. Calculated from the ratio Observed value / Reference value. Each Member State shall divide the ecological quality ratio scale for their monitoring system for each surface water category into five classes ranging from high to bad ecological status, by assigning a numerical value to each of the boundaries between the classes (from WFD text)

EQR setting criteria. The calculation options used to define the range of variation of EQRs, i.e. how to set the highest (EQR=1) and lowest (EQR=0) benchmarking, and to derive class boundaries.

National Standard Classification. The biological classification obtained by applying the current MS quality classification scheme for each BQE.

Best Available Classification (BAC). The biological classification obtained by applying a WFD compliant procedure and all the available, relevant information on a site. Depending on the main kind of pressure acting, it may results from integrating biological, physico-chemical and hydromorphological information. It is based on detailed community analysis (e.g. by multivariate analysis on one or more BQEs) and not on the standard National methods of classification.

(Biological) Metric. A metric is a calculated value representing some aspect of the biological population's structure, function or other measurable characteristic that changes in a predictable way with increased human influence (Barbour et al., 1999).

Qualitative metric. A metric that can be calculated from field samples collected following a qualitative sampling protocol. Its calculation does not require any abundance estimation (e.g. BMWP, ASPT, number of EPT taxa, etc.).

Quantitative metric. A metric that can be calculated from field samples collected following a quantitative sampling protocol (i.e. area-based sampling). Its calculation requires abundance estimation (e.g. number of specimens of selected taxa, diversity indices, etc.).

Intercalibration Common Metric (ICM). A biological metric widely applicable within a GIG, which can be used to derive comparable information among different countries/stream types.

Test data Data. Derived by standard monitoring according to MS legislation and tradition. They refer to a stream type.

Benchmark data. Data collected with the explicit aim of satisfying the WFD demands (e.g. stream type specific data, reference conditions established, EQRs, five quality classes considered, etc.), including biological, chemical and general pressure data.

WFD Water Framework Directive European Commission. 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327, 22.12.2000, 1-72.

AQEM. "The Development and Testing of an Integrated Assessment System for the Ecological Quality of Streams and Rivers throughout Europe using Benthic Macroinvertebrates". EU funded project within 5th Framework Program, Energy, Environment and Sustainable Development, Key Action Water, AQEM Contract no. EVK1-CT1999-00027.

STAR. "Standardisation of River Classifications: Framework method for calibrating different biological survey results against ecological quality classifications to be developed for the Water Framework Directive project" EU funded project within 5th Framework Program, Energy, Environment and Sustainable Development, Key Action Water, STAR Contract no. EVK1-CT-2001-00089.

GIGs. Geographical Intercalibration Groups. Cluster of European countries whose water bodies are supposed to be directly comparable for the IC process. For rivers, five GIGs were agreed: Northern, Central European, Alpine, Mediterranean and Eastern Continental.

ASPT. Biotic index: Average score per taxon (Armitage et al. 1983). Used as a standard basis in the U.K. to classify rivers based on aquatic invertebrates.

BMWP Biotic index: Biological Monitoring Working Party score (Armitage et al. 1983)

BQE Biological Quality Element (Water Framework Directive)

CIS. European Common Implementation Strategy for the Water Framework Directive.

ECOSTAT. CIS Working Group 2 A dedicated to the Ecological Status of surface water bodies within the implementation of the Water Framework Directive.

MSs States members of the European Union.

EPT. Total number of taxa belonging to the Insect Orders of Ephemeroptera, Plecoptera and Trichoptera

GOLD. Total number of taxa belonging to the Orders of Gasteropoda, Oligocaeta and Diptera.

IBE. Biotic index: Indice Biotico Esteso (Ghetti, 1997; APAT-IRSA, 2004). Used as standard in Italy to classify rivers based on aquatic invertebrates.

IC. European Intercalibration Process for the WFD.

Reference conditions. For any surface water body type reference conditions or high ecological status is a state in the present or in the past where there are no, or only very minor, changes to the values of the hydromorphological, physico-chemical, and biological quality elements which would be found in the absence of anthropogenic disturbance (from REFCOND guidance 14/06/2002).

REFCOND. Working Group 2.3 on "Development of a protocol for identification of reference conditions, and boundaries between high, good and moderate status in lakes and watercourses".

Ecological status. It is an expression of the quality of the structure and functioning of aquatic ecosystems associated with surface waters, classified in accordance with annex V (from Article 2 (21) in WFD).

Intercalibration exercise. Exercise that should be carried out to establish the value for the boundary between the classes of high and good status, and the value for the boundary between good and moderate status. The Commission shall facilitate this intercalibration exercise in order to ensure that these class boundaries are established consistent with the normative definitions in Section 1.2 and are comparable between Member States (WFD 1.4.1 (iv)).

Intercalibration Network. As part of this exercise the Commission shall facilitate an exchange of information between Members States leading to the identification of a range of sites in each ecoregion in the Community; these sites will form an intercalibration network. The network shall consist of sites selected from a range of surface water body types present within each ecoregion. For each surface water body type selected, the network shall consist of at least two sites corresponding to the boundary between the normative definitions of high and good status, and at least two sites corresponding to the boundary between the normative definitions of good and moderate status. The sites shall be selected by expert judgement based on joint inspections and all other available information (WFD 1.4.1 (v)).

JRC. EC Joint Research Centre with the role of facilitating the intercalibration process.

CEN. European Committee for Standardization with the role of contributing to the objectives of the European Union with voluntary technical harmonization in Europe and standards which promote, among others, environmental protection, exploitation of research and development programmes, and public procurement.

Annex I - Operational summary of the procedure

This summary procedure is reported to help deriving a part of the steps and calculations shown in the Deliverable.

All data have to be included at family level. If the identification level in your dataset is more detailed than family, you have to merge all the data to family level. For the taxonomic list, you should refer to the species list used in AQEM Project. Please check on web site www.aqem.de 'list of key taxa values'[1].

1) Prepare spreadsheet of data
DATA FOR EACH SITE MUST BE ON A SEPARATE ROW. ALL DATA MUST BE FROM THE SAME IC TYPE. IF NUMERICAL ABUNDANCE DATA IS NOT AVAILABLE IT MUST BE ESTIMATED.

2) Calculation of ICMs

a. Use AQEM assessment software Version 2.3 (free download from http://www.aqem.de/start.htm, link from the STAR web site http://www.eu-star.at) to calculate EPT, N-taxa, ASPT and Shannon-Weiner diversity index.

   In Excel, transpose biological data, so that each column represents a separate site and each row a separate taxon.
   If present in your dataset, the following taxa have to be excluded: Hydracarina, Cladocera, Copepoda, Ostracoda, Nematoda, families of Lepidoptera.
   You must keep the Oligochaeta, even if identified to Class level.

   Add 'shortcode' taxon names (by hand or by using ID-Art software)
   The shortcodes are essential to input the data into AQEM assessment software. You should obtain the shortcodes that you need to enter data into AQEM assessment software Version 2.3 from the list that was distributed to you all (Translation-table_FAM-161004.xls). This has the shortcodes for all families.

   Split into sheets of 90 samples
   Each excel file to import data into AQEMsoft must contain an only sheet with the following columns in this order: shortcode, taxon name, site1, site2 etc..

   Enter into AQEM assessment software
   Input instruction:
   - execute the program
   - select a country, if your country is not present select whatever you want
   - click 'import'
   - choose your input file
   - select 'shortcode' from the import file settings
   - click OK on 'replace taxa names'
   - click OK if 'replace unknown taxa' appears
   - click OK to 'sample characterization'

- click 'calculate'
- wait…
- select the 'sheet' 'metrics'
- export to excel the metrics' results.

Copy data for ICM metrics and transpose into original data sheet.

b. Calculate remaining ICMs log(Sel_EPTD+1) and 1-GOLD
Sel_EPTD is the sum of abundance of Heptageniidae, Ephemeridae, Leptophlebiidae, Brachycentridae, Goeridae, Polycentropodidae, Limnephilidae, Odontoceridae, Dolichopodidae, Stratyomidae, Dixidae, Empididae, Athericidae & Nemouridae.
You can sum up these abundance from your original data file.
Use Excel function to calculate the ICM log (Sel_EPTD+1).
Log is 10 base.

GOLD is the relative abundance of all families of Gastropoda, Oligochaeta and Diptera. Sum up the abundances of the specimens of Gastropoda, Oligochaeta and Diptera and divide it by the total abundance of the site. 1-GOLD is the ICM.

3) Determine Reference state value for the normalization

a) Remove outliers (optional). The values higher than 1.5-3 times the interquartile range + the value of the 75 percentile are considered outliers. For this, you will need to calculate the inter-quartile distance. This is the range between 25%ile and the 75%ile. There is a function in Excel to calculate the quartiles (see help in Excel).

b) Determine reference state value

- If you have not defined reference state samples according to a pressure based classification, but only 'high status' sites biologically based (applicable to UK), take the 75th percentile of the reference state samples metric for the river type (excluding outliers)
- If you have defined reference state samples, take the median value of the reference state samples (this option, while being the best statistically, will not support the comparison with datasets in the previously described situation).

4) Normalise (convert to EQRs) both the ICMs and the national classification metrics by dividing them by the reference values.
NB for ICM ASPT: it is considered as ASPT does not reach 0; as a general rule it is always considered a value of 2 as minimum (even if in some cases a value lower than 2 can be reached. Values lower than 2 have anyway to be reported to 0). To normalise ASPT, subtract 2 to the observed value and to the reference value before dividing by the reference value.

5) Calculate ICMi values.
The ICMi value is calculated by sum of all the ICMs. Each ICM is previously multiplicated by its weight (see also table 1):

- ASPT*0.333
- Log_{10}(sel_EPTD+1)*0.266
- 1-GOLD*0.067
- N-taxa*0.167
- EPT*0.083
- Shannon-Weiner*0.083

6) Determine the reference value for ICMi by the same procedure that you used to determine the reference values of the individual ICMs and the national classification metrics. Re-normalise ICMi values by dividing them by the reference value.

7) Undertake regression in Excel, with ICMi on y-axis and national classification metric on x-axis. Show the Excel scatter plot, add the trendline and show the regression statistics. Add the national class boundaries and read across the boundary values in terms of ICMi.

For the comparison:
8) Determine the value of your national WFD class boundaries in units of ICMi using the regression formula.

9) Compare the position of your national class boundaries with those of other countries or other IC river types.

The basic requirements for a test dataset are the following :

- Data is needed for as many samples as possible, taken for example from the national monitoring network.
- It is indispensable that a sufficient number of reference sites (or samples) were included in the dataset, and that all the sites included in a dataset were of the same IC type. Criteria for the definition of the reference status sites have to be provided and should follow the general principles expressed by the Water Framework Directive and by REFCOND guidance. The reference sites correspond to the ‘high status’ sites according to national assessment method, if it is the only available classification method. As rough suggestion, a single dataset should include a minimum of 25 samples, with at least 3 high status/reference samples.
- Samples of different regions or districts can be included in the same dataset, nevertheless the IC type has to be the same. If so, should be indicated if samples belonging to different regions have different reference faunistic conditions or not.
- Each dataset must be homogeneous, i.e. all samples were taken with the same method (sampling and laboratory procedures).
- The dataset must contain the widest range of ecological quality from high to bad status (if possible), at the minimum from reference sites to the whole extent of the “moderate” class.

[1]: You can download the taxonomic codes (shortcodes) for all taxa (List of key taxa values) from the same web site as AQEM assessment software Version 2.3. You may find this list useful to understand the taxonomy used by AQEM assessment software Version 2.3, i.e. which genera are assigned to the families recognised by AQEM. This differs from current UK practice. However, we recommend you to use the table of shortcodes for families rather than this full list to assign shortcodes to families for this intercalibration pilot, to avoid confusion.

Annex II: Intercalibration Common Metrics (ICMs) selected for STAR Intercalibration procedure

Intercalibration Common Metrics (ICMs) selected for STAR Intercalibration procedure

Information type	Metric type	Metric name	Taxa considered in the metric	Literature reference	weight
Tolerance	Index	ASPT	Whole community (Family level)	e.g. Armitage *et al.*, 1983	0.333
Abundance/Habitat	Abundance	Log_{10} (Sel_EPTD +1)	Log(sum of Heptageniidae, Ephemeridae, Leptophlebiidae, Brachycentridae, Goeridae, Polycentropodidae, Limnephilidae, Odontoceridae, Dolichopodidae, Stratyomidae, Dixidae, Empididae, Athericidae & Nemouridae)	Buffagni *et al.*, 2004; Buffagni & Erba, 2004	0.266
	Abundance	1-GOLD	1 - (relative abundance of Gastropoda, Oligochaeta and Diptera)	Pinto *et al.*, 2004	0.067
Richness and Diversity	Taxa number	Total number of Families	Sum of all Families present at the site	e.g. Ofenboch et al., 2004	0.167
	Taxa number	number of EPT Families	Sum of Ephemeroptera, Plecoptera and Trichoptera taxa	e.g. Ofenboch et al., 2004; Böhmer et al., 2004.	0.083

Diversity index	Shannon-Wiener diversity index	$$D_{S-W} = -\sum_{i=1}^{s}\left(\frac{n_i}{A}\right)\cdot\ln\left(\frac{n_i}{A}\right)$$	e.g. Hering et al., 2004; Böhmer et al., 2004.	0.083

Annex III - STAR IC Internat. activities

STAR Consortium			
	date	products	Involved STAR institute
"Potential contribution of the STAR and AQEM projects to the Intercalibration process" A. Buffagni, approccio per una possibile procedura di armonizzazione delle *class boundaries* sulla base di un Indice Comune di Intercalibrazione (multimetrico). ECOSTAT and INTERCALIBRATION WG 2A, JRC- Ispra (I)	15-17 ottobre 2003	presentatione .ppt (su CIRCA e STAR web site)	CNR-IRSA (Italy)
caratteri generali dell' IC *guidance*			CNR-IRSA (Italy)
"A SIMPLE PROCEDURE TO HARMONIZE CLASS BOUNDARIES OF EUROPEAN ASSESSMENT SYSTEMS" Discussion paper for the Intercalibration process, A. Buffagni & S. Erba, Explanation of the STAR ICM_index approach - WFD CIS WG 2.A ECOSTAT.	6 Feb. 04	WFD CIS WG 2.A ECOSTAT - Discussion paper; distributed to all MS experts (on CIRCA and STAR web site)	CNR-IRSA (Italy)
Progress with the IC guidance	10 Feb. 04		CNR-IRSA (Italy)
Intercalibration of river classification results – a practical example from the STAR project. "A contribution from the STAR Project to the Intercalibration process: A simple procedure to Harmonize class boundaries of European assessment systems", A. Buffa	11 February 2004	ppt presentation Wed. (on CIRCA and STAR web site)	CNR-IRSA (Italy)
Procedure to harmonise class boundaries of European river assessment systems – an example from STAR project - revisited. "A contribution from the STAR Project to the Intercalibration process: A simple procedure to Harmonize class boundaries of European as	13 February 2004	ppt presentation Fri. (on CIRCA and STAR web site)	CNR-IRSA (Italy)
Lednice STAR meeting. Discussed topics: ICMi approach, comparisons of different datasets and methods, harmonization of class boudaries	March04	ppt (STAR web site)	CNR-IRSA (I), UDE (D), all partners

joined STAR/GIG activity - Support to JRC			
	date	products	Involved institute
First meeting drafting group Intercalibration, Vienna (A)	4 December 2003	General outline of the IC guidance	Germany, Joerg Janning; Spain, Manuel Toro; France, Pierre-Jean Martinez; Austria, Gisela Ofenböck; STAR/AQEM, Italy, Andrea Buffagni; COAST, Norway, Kari Nygaard; JRC, Wouter van de Bund
Data provided to CNR-IRSA for the IC procedure development	Jan/Feb. 04	Test dataset	ARPA Parabiago (I), P. Genoni
Second meeting drafting group Intercalibration, Ispra, JRC (I)	10-11 Feb. 2004	Progress with the IC guidance	Germany, Joerg Janning; Spain, Joze Ortiz-Casas; France, Pierre-Jean Martinez, Jean-Gabriel Wasson; UK, Peter Pollard; STAR/AQEM, Italy, Andrea Buffagni; COAST, Norway, Kari Nygaard; JRC, Wouter van de Bund
Nordic GIG workshop (Coordination: Frida Löfström/Anette Björlin, SWEDISH ENVIRONMENTAL PROTECTION AGENCY)	March04		

STAR IC Activities - international level 2

STAR Consortium			
	date	products	Involved STAR institute
"STAR Intercalibration", John Murray-Bligh, Central Europe GIG meeting, Brussels	20-apr-04	ppt (STAR partners)	EA (UK)
Mediterranean GIG meeting, Evora, Portugal. Pilot application of STAR ICMi approach, ICMs and comparison	19-21 May 2004	Relationships between National methods and ICMi	CNR-IRSA (I), INAG (P), Uni Evora (P)
Establishment of a common dataset for R-M1, based on AQEMdip software, at the Family level (ca 200 samples included from F, I, P, S)	may 2004	database (available to participant Institutes)	CNR-IRSA (I), INAG (P), Uni Evora (P)
Attempt to bilateral approach, based on averaging of class boundaries "Bilateral comparison of assessment and classification methods" Draft document - S. Birk	June04	Draft document (internally circulated)	UDE (D)
Data provided to CNR-IRSA for testing and application of the whole procedure, including harmonization of class boundaries. EA (test dataset); AQEM and STAR consortium (benchmark dataset)	June04	datasets	EA (UK), EA (PL) CEH (UK) AQEM consortium,
AQEM partners' data (elaborated within the AQEM project) were used for the construction of a benchmark dataset (family level)	June04-July04	benchmark dataset (restricted to AQEM/STAR consortium)	CNR-IRSA (Italy)
Examples of application of the STAR ICMi approach for: R-C1: Italy, UK, Poland; R-M1: Italy, France.	June04-July04	ppt presentation ECOSTAT meeting	CNR-IRSA (Italy)

joined STAR/GIG activity - Support to JRC			
	date	products	Involved institute
Central GIG meeting, Brussels, Belgium. Discussed topics: presentation on the status of the intercalibration planning process; presentation on tasks of the GIG coordinators; discussion on the options for the Central GIGs coordination; presentation on opti	20-apr-04	Central GIG minutes	VMM and CRNFB (Belgium), Nieders Landesamt.f. Okologie (Germany), Ministry of Environment (Denmark), Ministry of Agriculture, Forestry (Austria), RIZA (Netherlands), Institute of Environmental Protection (Poland), Ministère de l'écologie et du développeme
Mediterranean GIG meeting, Evora, Portugal. Pilot application of STAR ICMi approach (ICMs and comparison) on R-M1 for France, Italy, Portugal & Spain: Med GIG - Evora (P). The practical work comprised incorporating metrics calculated by AQEMrap and calcul	19-21 May 2004	Med GIG minutes (draft); JGW manuscript (restricted); Relationships between National methods and ICMi	Jean-Gabriel Wasson (Cemagref, F), Andrea Buffagni (CNR-IRSA, I), Manuel Toro (Min. Medio Ambiente, SP), Maria Helena Alves (INAG, P), João Manuel Bernardo (Uni Evora, P) and collaboration of Paulo Pinto (Uni Evora, P) and Hélio Figueiredo (Uni Evora, P).
Data provided	may 2004	dataset	Cemagref, F; Min. Medio Ambiente, SP; Uni Evora, P
Examples of application of the STAR ICMi approach for: R-M1: Italy, France.	June04-July04		CEMAGREF (F)
Alpine GIG meeting - Vienna (A). Discussion on the applicability of option 2 through STAR ICMi approach (input coming from Med-GIG by JG Wasson)	29 June04	Alpine GIG minutes	Jean-Gabriel Wasson (F), Paolo Negri (I), Manuel Toro (SP), Bernarda Rotar (SI), Gisela Ofenböck (AT) and Franz Wagner (AT)

STAR IC Activities - international level 3

STAR Consortium			
	date	products	Involved STAR institute
The procedure (based on STAR ICMi approach) was analysed, examining some of the crucial points of the process of intercalibration (data preparation, metrics calculation and testing, normalization, ICMi calculation, criteria for reference conditions)	8-11Sep04	Relationships between National methods and ICMi	CNR-IRSA (I)
Pre-pilot STAR meeting: datasets preparation and calculations (ICMi) for R-C1, R-C4, R-C2. Analysis of the correlation between ICMi and national standard methods.	13-14 Sep04	Relationships between National methods and ICMi	CNR-IRSA (I), UDE (D), EA (UK), CEH (UK)
Coordination & provision of technical examples on the application of the procedure within Central and Baltic GIG	15-16 Sep04	CentralGIG meeting	CNR-IRSA (Italy), EA (UK)
Examples of the averaging approach for the harmonization of class boundaries (bilateral):	16 Sep04	ppt presentation Central GIG meeting	UDE (Germany)
Preparation and circulation of a document on ICMs & standardization	18 Sep04	ICMs document (Central GIG delegates; STAR partners)	CNR-IRSA (I)
Preparation and distribution of the draft index of the STAR Intercalibration Deliverable	22 Sep04	Draft index restricted to STAR consortium	CNR-IRSA (I)
"Comparison of results of national bioassessment methods based on AQEM/STAR data using bilateral correlation and regression – description of general procedure", S. Birk	23 Sep04	discussion document for the 6th STAR meeting	UDE (Germany)
ECOSTAT meeting	7-8Oct04	ECOSTAT meeting partecipation to Rivers' presentation	CNR-IRSA (I)

joined STAR/GIG activity - Support to JRC			
	date	products	Involved institute
The procedure (based on STAR ICMi approach) was analysed, examining some of the crucial points of the process of intercalibration (data preparation, metrics calculation and testing, normalization, ICMi calculation, criteria for reference conditions)	8-11Sep04	Relationships between National methods and ICMi	CEMAGREF(F)
Pre-pilot STAR meeting: datasets preparation and calculations (ICMi) for R-C1, R-C4, R-C2. Analysis of the correlation between ICMi and national standard methods.	13-14 Sep04	Relationships between National methods and ICMi	CNR-IRSA (I), CEMAGREF(F), UDE (D), EA (UK), CEH (UK)
Central and Baltic GIG meeting. Pilot application of STAR ICMi approach (see R-M1 exercise) on R-C2 for Spain & France: Central GIG meeting- Milan (I)	15-16 Sep04	Central GIG minutes	APAT (Italy), STAR UniEssen (Germany), STAR CNR-IRSA (Italy), STAR, CEH (UK), CRNFB (Belgium), Nieders Landesamt.f. Okologie (Germany), Danish Forest and Nature Agency (Denmark), RIZA (Netherlands), Institute of Environmental Protection (Poland), Cemagref
Preparation and circulation of a document on ICMs & standardization	18 Sep04	ICMs document (Central GIG delegates; STAR partners)	CEMAGREF(F)

STAR IC Activities - international level 4

STAR Consortium	date	products	Involved STAR institute
Workshop on "Ecological quality assessment and Intercalibration in the EU Water Framework Directive" - INFRA 10899 - Technical Assistance Information Exchange Office and the Joint Research Centre	25-26 Nov04	ppt presentation "STAR: harmonization of river assessment systems"	CNR-IRSA (I)
Compilation of the 11th STAR deliverable for EC	22nd Dec04	"Matrix of possible class boundaries of grades of 'Ecological Status' associated with different methods and stressors" "Contribution of the STAR Project to the European CIS Intercalibration process"	CNR-IRSA (I), BOKU (A), UDE (D), EA (UK), CEH (UK)

joined STAR/GIG activity - Support to JRC	date	products	Involved institute
Mediterranean GIG meeting, Lion, France. ICMs discussion, selection of datasets, which BQEs using for IC process (in addition to invertebrates). Pilot exercise to test abundance' metrics	18-19 Nov04	Med GIG minutes (draft);	Jean-Gabriel Wasson (Cemagref, F), Nicolas Mengin (Cemagref, F), Marcello Cazzola (CNR-IRSA, I), Maria Belli (APAT, I), Manuel Toro (Min. Medio Ambiente, SP), Maria Helena Alves (INAG, P), João Manuel Bernardo (Uni Evora, P).
JRC Scientific support to River intercalibration process	since Dec04		CNR-IRSA (I)