ISTITUTO DI RICERCA SULLE ACQUE

del CONSIGLIO NAZIONALE DELLE RICERCHE

TOWARDS EUROPEAN INTER-CALIBRATION FOR THE WATER FRAMEWORK DIRECTIVE: PROCEDURES AND EXAMPLES FOR DIFFERENT RIVER TYPES FROM THE E.C. PROJECT STAR

CONTRACT No: EVK1-CT 2001-00089.

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

Buffagni A., Erba S., Birk S., Cazzola M., Feld C., Ofenböck T., Murray-Bligh J., Furse M. T., Clarke R., Hering D., Soszka H. & W. van de Bund

Rome, March 2005

Quaderni 123
(ISSN 0390-6329)

Quaderni, 123
2005, n. 1

Pubblicazione semestrale dell'Istituto di Ricerca sulle Acque del Consiglio Nazionale delle Ricerche (Aut. Trib. Roma n° 17228 del 14.4.1978)

Direttore responsabile: Roberto Passino

Direzione, Redazione e Amministrazione. Istituto di Ricerca sulle Acque, via Reno n° 1, 00198 Roma (tel. + 39 - 06 - 8841451; fax +39 - 06 - 8417861)
Redazione: Giulia Barbiero, Andrea Buffagni
Stampa: Centro stampa IRSA

This publication should be cited as follows:

Buffagni A., Erba S., Birk S., Cazzola M., Feld C., Ofenböck T., Murray-Bligh J., Furse M. T., Clarke R., Hering D., Soszka H. & W. van de Bund, 2005. **'Towards European Inter-calibration for the Water Framework Directive: Procedures and examples for different river types from the E.C. project STAR'**. 11th STAR deliverable. STAR Contract No: EVK1-CT 2001-00089. Rome (Italy), *Quad. Ist. Ric. Acque* 123, Rome (Italy), IRSA, 460 pp.

Contact for enquiries - Queries and further information on this Paper should be addressed to: Andrea Buffagni, Water Research Institute CNR-IRSA, Via della Mornera, 25 I-20047 Brugherio (MI), Italy - e-mail: buffagni@irsa.cnr.it

Copies of this publication may be obtained from the Consiglio Nazionale delle Ricerche – Dip. Attività Internazionali – Servizio V – Pubblicazioni e informazioni Scientifiche – Uff. Vendita, via S. Martino della Battaglia, 44 – 00185 Roma (Italy) (Tel. ++39 06 49932308; fax ++39 06 49933255), or may be downloaded (electronic version) from the STAR website www.eu-star.at or from the CNR-IRSA web site www.irsa.cnr.it

PREFACE

The implementation of the Water Framework Directive will be a long and crucial step towards achieving a common environmental quality in European water bodies.

At present, along with the development and testing of new methods and systems for the fulfilment of the WFD requirements, the Inter-calibration process represents one of the focal points for discussion, as it will have a notable effect on how our natural resources are managed.

In fact, a European agreement on the boundaries to be set between the Ecological Status classes will represent a major issue of the Inter-calibration process. The practical implications of the consistent setting of boundaries for assessment systems will it is hoped result in a desired and equally supported effort on the part of all European Member States to protect and restore our water bodies.

This paper presents the notable experience carried out within the EU co-funded STAR Project on Inter-calibration, achieved with the close collaboration of European Commission delegates and activities.

The main results of the Project on this important issue are being published concurrently, in order to render the huge amount of data generated on rivers available as a useful guide for everyone involved in future stages of the European Inter-calibration process.

.
Prof. Roberto Passino
Director of the Italian Water Research Institute
CNR-IRSA

Rome, March 2005

CONTENTS

TOWARDS EUROPEAN INTER-CALIBRATION FOR THE WATER FRAMEWORK DIRECTIVE: PROCEDURES AND EXAMPLES FOR DIFFERENT RIVER TYPES FROM THE E.C. PROJECT STAR[1]

Buffagni A.[1], Erba S. [1], Birk S. [2], Cazzola M. [1], Feld C. [2], Ofenböck T. [3], Murray-Bligh J. [4], Furse M. T. [5], Clarke R. [5], Hering D. [2], Soszka H. [6] & W. van de Bund[7]

[1] CNR – IRSA, Water Research Institute, Via della Mornera, 25, 20047 Brugherio (MI), Italy
[2] UDE, Department of Ecology, Faculty of Hydrobiology, University of Duisburg-Essen, D-45117 Essen, Germany
[3] BOKU, University of Natural Resources and Applied Life Sciences, Institute of Water Provision, Water Ecology and Waste Management; Department of Hydrobiology, Vienna, Austria
[4] EA, Environment Agency, Manley House Kestrel Way Exeter EX2 7LQ, UK
[5] CEH, Centre for Ecology and Hydrology - CEH Dorset, Winfrith Technology Centre Winfrith Newburgh, Dorchester Dorset DT2 8ZD, UK
[6] IOEP, Institute of Environmental Protection, Kolektorska, 401-692 Warszawa, Poland
[7] JRC, European Commission - Joint Research Centre, Institute for Environment and Sustainability, Inland and Marine Waters Unit T.P. 290, 21020 Ispra (VA), Italy

Extended Summary

In the present work, the topic of Inter-calibration of assessment methods is discussed in terms of the comparison and harmonization of their resulting classification (i.e. class boundaries). The Paper is the scientific contribution of the EU co-funded project STAR to outline the procedure for performing the European Inter-calibration process. The Inter-calibration (IC) process is one of the primary issues to be presently addressed in the context of the Water Framework Directive implementation. The European IC aims at consistency and comparability in the classification results of the monitoring systems utilized by each Member State for the WFD Biological Quality Elements. On the scientific side, the Inter-calibration exercise aims to compare the existing values for the

1 Buffagni, Erba, Birk, Cazzola, Feld, Ofenböck, Murray-Bligh, Furse, Clarke, Hering, Soszka & van de Bund (2005): 'Towards European Inter-calibration for the Water Framework Directive: Procedures and examples for different river types from the E.C. project STAR'. STAR Contract No: EVK1-CT 2001-00089. *Quad. Ist. Ric. Acque*, **123**, Rome, (I)

boundaries between the High/Good and Good/Moderate status across Europe and to provide tools and procedures for their harmonization. Out of the three IC options actually delineated, the procedures tested here mainly refer to Option 2 (and hybrids): 'Use of a common metric(s) method identified specifically for the purposes of the Inter-calibration exercise'. The main aim of the Paper is to evaluate the general applicability of the various approaches and procedures, by focusing on rivers and aquatic invertebrates.

The report is especially aimed at illustrating concrete possibilities for comparing and harmonizing the MSs' classification results by finally setting comparable boundaries to quality classes. The report is not focused on inter-calibrating biological methods or monitoring systems. The examples of harmonization of the National class boundaries presented here are intended to demonstrate the possibility of identifying and eliminating the possible differences between systems and the potential of using different approaches. The proposed procedure for Inter-calibration consists of two steps: a) comparison of existing National class boundaries; b) harmonization (adjustment) of boundaries. The different procedures were evaluated and tested through an application of datasets provided by STAR partners, other scientific institutions and environment agencies/Environmental Ministries from around Europe. A relevant part of the data presented and processed here was provided as part of the ongoing pilot IC exercises within the various Geographic Inter-calibration Groups'(GIGs). In particular, most test datasets refer to the Central GIG countries and activities, with notable additions from the South of Europe (Mediterranean GIG), where preliminary actions for the pilot started early during 2004 and led to important improvements in the delineation of the Inter-calibration Options.

Three main procedures were analyzed for comparison and, later on, harmonization: a) the direct comparison of the classification results of different assessment methods on the same dataset; b) the indirect comparison based on the selection of Inter-calibration Common Metrics (ICMs); c) the indirect comparison based on the selection of ICMs and the use of an external, benchmarking system. For the application of the last two procedures, a range of metrics providing information on tolerance to water pollution, abundance/habitat and richness/diversity of the community were calculated and combined into an ICM index (ICMi).

A significant part of the work was dedicated to the evaluation of the relationship between the ICMi and the National classification indexes by means of linear regression. In particular, the R-C1 river type was studied, with integrations for R-C2, RM1 and R-M5. The multimetric index (ICMi) developed for the purposes of Inter-calibration consists of six metrics covering all WFD requirements for macroinvertebrates: ASPT; Log_{10} (Sel_EPTD +1)(based on abundance of selected taxa belonging to the Ephemeroptera, Plecoptera, Trichoptera and Diptera Insect Orders); 1-GOLD (based on abundance of Gastropoda, Oligochaeta and Diptera); Total number of Families; Number of EPT Families (Families of Ephemeroptera, Plecoptera and Trichoptera); Shannon/Weaver diversity index.

In addition, some detailed analyses to identify biological metrics potentially suitable for the IC process were performed, clearly indicating the relationship between the metrics and the pressures acting on the studied rivers and sites. The results, while confirming the usefulness of species-level investigations for very precise assessment exercises, also demonstrated the suitability of selected metrics at a higher identification level (i.e. Family) for the purposes of Inter-calibration. Specifically, extended examples are provided for two river types in central Europe, by using large, trans-National datasets and for two Italian stream types (R-C1 and R-M1) with relation to the pressures acting on the river systems.

A description is provided of the results of comparison and harmonization on the basis of a direct comparison (a) of different assessment methods, which do not imply the use of the ICMi approach,.

One of the options based on the calculation of the ICMi (b) requires the indirect comparison *via* ICMi of National class boundaries. The linear regression models developed to analyze the relationship between ICMi and National methods were used to convert class boundary values of the National index to ICMi values for comparison with the boundaries of other MSs' National systems. The ICMi and the National classification method are converted to EQRs by normalization i.e. dividing them by the value of the reference state for that particular IC river type. This reference state was determined by means of a specified procedure (see below), which was applied in a fully comparable way for all methods, datasets and MSs.

The second evaluation procedure based on the calculation of the ICMi (c) involves the comparison of test datasets to a WFD-compliant, trans-national dataset (benchmark dataset), for which a Best Available Classification was provided (i.e. based on STAR/AQEM data). In this last example, the values of the ICMi calculated for the STAR/AQEM ecological quality classes were compared to those observed for the corresponding class of the National system (test dataset) by means of the Mann-Whitney U test. After the comparison phase, it was possible to start the harmonization phase.

Special focus was given to the two options for the harmonization following a comparison *via* ICMi. In one case, an example of harmonization based on the selection of median boundary values derived from (hypothetically) WFD-compliant National methods, is presented. No simple 'averaging' of existing class boundaries should be considered for European Inter-calibration, at least until all MSs' assessment systems are proved to be fully WFD-compliant.
Differently, in a further example, the harmonization was carried out on the basis of the statistical comparison executed between the ICMi values observed in the benchmark dataset and the same observed in the test dataset. For this approach, the Good status samples are compared first. If the ICM index values based on the two datasets (i.e. classification schemes) significantly differ and the National system appears too accommodating, the class boundary Good/Moderate for the National method is moved upwards in order to eliminate the differences. As a result of this procedure some samples classified beforehand as being in Good status are repositioned in Moderate status. After the adjustment of the Good/Moderate boundary - corresponding to no more significant differences between the ICMi values found in the Good quality class of the two datasets - the boundary High/Good is considered. The procedure of statistical comparison between values observed for the High status class, is thus performed. The new, harmonized boundaries for the National classification system are therefore set for the High/Good and Good/Moderate borders.

The general results reported in the Paper demonstrated that the use of common metrics for the IC process (ICMs) can be applied over a wide geographical range. The procedure for calculating the ICMi and comparing datasets was carefully described and it is now ready for application by European countries, GIGs or European Community delegates. The ICMi approach supports the use of existing datasets directly collected by MSs, which can guarantee a reliable availability of data for the IC process. The comparison exercise between European class boundaries and assessment systems led to different results for the different stream types and procedures used, but showed how systems and boundaries are actually comparable in the short term.

The direct 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 are available with collected samples that satisfy the requirements of the compared methods. Nonetheless, the results presented for the direct comparison indicate that the percentage of sites moving from High and Good status to Moderate status can be quite high (in most cases around 30%), when looking at different methods. In contrast, the used harmonization Options based on the ICMi approach did not highlight important differences among the methods tested, which always exhibited a low percentage of samples in need of re-allocation from the Good to the Moderate status. During the data analysis and test of the IC procedures some problems arose, which need further discussion e.g. Reference condition criteria and 'anchor value' setting.

In general, some of the Options used and especially the ICMi approach permit a large variety of Intra- and Inter-GIG comparisons. They also support the harmonization phase on a pan-European scale, which reliably helps to make the WFD European Inter-calibration process actually feasible.

VERSO L’INTERCALIBRAZIONE EUROPEA NELL’AMBITO DELLA DIRETTIVA QUADRO SULLE ACQUE: PROCEDURE ED ESEMPI PER VARI TIPI FLUVIALI DAL PROGETTO E.C. STAR

Buffagni A.[1], Erba S. [1], Birk S. [2], Cazzola M. [1], Feld C. [2], Ofenböck T. [3], Murray-Bligh J. [4], Furse M. T. [5], Clarke R. [5], Hering D. [2], Soszka H. [6] & W. van de Bund[7]

[1] CNR – IRSA, Water Research Institute, Via della Mornera, 25, 20047 Brugherio (MI), Italy

[2] UDE, Department of Ecology, Faculty of Hydrobiology, University of Duisburg-Essen, D-45117 Essen, Germany

[3] BOKU, University of Natural Resources and Applied Life Sciences, Institute of Water Provision, Water Ecology and Waste Management; Department of Hydrobiology, Vienna, Austria

[4] EA, Environment Agency, Manley House Kestrel Way Exeter EX2 7LQ, UK

[5] CEH, Centre for Ecology and Hydrology - CEH Dorset, Winfrith Technology Centre Winfrith Newburgh, Dorchester Dorset DT2 8ZD, UK

[6] IOEP, Institute of Environmental Protection, Kolektorska, 401-692 Warszawa, Poland

[7] JRC, European Commission - Joint Research Centre, Institute for Environment and Sustainability, Inland and Marine Waters Unit T.P. 290, 21020 Ispra (VA), Italy

Sommario

Viene affrontata la problematica dell’Intercalibrazione biologica dei sistemi di valutazione della qualità ecologica, in termini di comparazione ed armonizzazione dei risultati della classificazione di qualità (i.e. dei valori limite tra le diverse classi). Il lavoro rappresenta il contributo scientifico del Progetto STAR allo sviluppo di una procedura per la realizzazione del processo di Intercalibrazione su scala europea. L’Intercalibrazione è una delle principali attività connesse all’implementazione della Direttiva Quadro sulle Acque (EC 2000/60). Tale processo si prefigge di rendere comparabili i risultati della classificazione ottenuta con i differenti metodi di valutazione utilizzati dagli Stati europei e basati sugli elementi di qualità biologica richiesti dalla Direttiva. Dal punto di vista scientifico, l’esercizio di Intercalibrazione ha l’obiettivo di effettuare una comparazione su scala europea dei valori limite tra le classi di qualità Elevata/Buona e Buona/Moderata dei vari sistemi di valutazione attualmente in uso e di fornire strumenti per la loro armonizzazione. Le procedure

presentate fanno riferimento principalmente alla verifica dell'Opzione 2 di Intercalibrazione - e relativi ibridi: "Utilizzo di metriche comuni selezionate per l'esercizio di Intercalibrazione". Lo scopo principale del presente lavoro è quello di valutare l'applicabilità dei diversi approcci e procedure, con riferimento ai fiumi ed agli invertebrati acquatici.

Il Quaderno è focalizzato sulla descrizione delle concrete possibilità di comparazione ed armonizzazione tra i risultati della classificazione fornita dai metodi dei vari Stati Membri, attraverso la definizione di '*class boundaries*' effettivamente confrontabili tra loro, su scala europea. Il lavoro non è dedicato all'intercalibrazione tra diversi metodi biologici o sistemi di monitoraggio. Gli esempi di armonizzazione dei limiti tra le classi presentati intendono dimostrare la possibilità di identificare ed eliminare le possibili differenze riscontrate tra i diversi sistemi e le possibilità di utilizzo dei differenti approcci. Le procedure proposte per l'Intercalibrazione comprendono due distinte fasi: a) la comparazione tra i limiti attualmente esistenti tra le classi dei metodi nazionali; b) l'eventuale armonizzazione dei limiti tra le classi (*class boundaries*). Sono state valutate e testate diverse procedure, attraverso l'applicazione a set di dati forniti dal progetto STAR, da varie istituzioni scientifiche e da agenzie ambientali di varie aree europee. Una parte consistente dei dati presentati ed elaborati è stata fornita come contributo agli esercizi pilota di Intercalibrazione attualmente in corso di svolgimento tra i vari Gruppi Geografici di Intercalibrazione (GIGs). In particolare, la maggior parte dei set di dati valutati si riferiscono all'attività dei paesi coinvolti nel GIG Centrale, con alcuni consistenti contributi da parte dei paesi sud europei del GIG Mediterraneo. All'interno di tale GIG, l'esercizio pilota è stato precocemente avviato nel corso del 2004 ed ha condotto ad importanti sviluppi per la definizione delle Opzioni di Intercalibrazione.

Sono state considerate tre principali procedure per la comparazione e per la successiva armonizzazione: a) la comparazione diretta dei risultati della classificazione ottenuta con diversi metodi sul medesimo *dataset*; b) la comparazione indiretta basata sulla selezione di Metriche Comuni di Intercalibrazione (ICMs); c) la comparazione indiretta basata sulla selezione di Metriche Comuni di Intercalibrazione (ICMs) e sull'utilizzo di un sistema esterno di dati di riferimento (*benchmarking system*). Per l'applicazione delle ultime due procedure, è stato selezionato un pool di metriche in grado di fornire informazione sul grado di tolleranza all'inquinamento, sulle caratteristiche di abbondanza/habitat e sulla ricchezza/diversità della comunità bentonica. Tali metriche sono quindi state combinate in un semplice indice multimetrico sintetico (ICMi).

Una parte consistente del lavoro è stata dedicata alla valutazione della relazione tra l'ICMi e i diversi indici nazionali, attraverso l'analisi delle regressioni lineari tra i due indici. In particolare, è stato esaminato il tipo fluviale R-C1, con integrazioni dai tipi R-C2, R-M1 e R-M5. L'indice multimetrico qui utilizzato, sviluppato per gli scopi dell'Intercalibrazione (ICMi), è costituito dalla combinazione di sei metriche che soddisfano i requisiti della Direttiva Quadro per i macroinvertebrati: ASPT; Log_{10} (Sel_EPTD +1) (basata sull'abbondanza di taxa selezionati appartenenti agli Ordini degli Insetti Ephemeroptera, Plecoptera, Trichoptera e Diptera); 1-GOLD (basata sull'abbondanza di Gastropoda, Oligochaeta e Diptera); Numero totale di Famiglie; Numero di Famiglie di EPT (Famiglie di Ephemeroptera, Plecoptera e Trichoptera); indice di diversità di Shannon/Weaver.

Al fine di identificare metriche idonee per il processo di Intercalibrazione, sono state effettuate alcune analisi di dettaglio che hanno dimostrato l'esisitenza di una buona relazione tra le metriche selezionate e i fattori di alterazione presenti nei siti investigati. I risultati confermano, da un lato, la validità di indagini a livello di specie per studi di dettaglio e, dall'altro, dimostrano l'idoneità di metriche ad un livello di identificazione più superficiale (i.e. Famiglia) per gli scopi dell'intercalibrazione. In particolare, sono presentati alcuni esempi dettagliati per due tipi fluviali centro europei che dispongono di ampi *dataset* trans-nazionali e per due tipi fluviali italiani (R-C1 e R-M1), con riferimento ai fattori di alterazione presenti.

Viene presentata una descrizione dei risultati di comparazione ed armonizzazione sulla base del confronto diretto di diversi sistemi di valutazione (procedura a), che non implica l'utilizzo dell'approccio ICMi.

Una delle opzioni basate sul calcolo dell'ICMi (procedura b) richiede la comparazione indiretta *via* ICMi dei limiti tra le classi di qualità definiti secondo i metodi nazionali. I modelli di regressione lineare sviluppati per analizzare la relazione tra ICMi e i vari metodi nazionali, sono stati utilizzati per convertire i valori limite tra le classi dell'indice nazionale in valori di ICMi, al fine di consentire un confronto tra i diversi valori di *boundary*. L'ICMi e i valori dei metodi nazionali sono stati convertiti in EQR (*Ecological Quality Ratio*) attraverso una normalizzazione, cioé dividendo per il valore dello stato di riferimento definito per i singoli tipi fluviali. Tale stato di riferimento è stato determinato attraverso una specifica procedura (si veda oltre), applicata in maniera pienamente comparabile tra i vari set di dati.

La seconda procedura basata sul calcolo dell'ICMi (procedura c), implica il confronto dei *test dataset* con un *dataset* trans-nazionale conforme ai requisiti della Direttiva (*benchmark dataset*), i cui campioni biologici sono classificati in accordo con una *Best Available Classification* (i.e basata su dati STAR/AQEM). In quest'ultimo esempio, i valori dell'ICMi calcolato per le classi del database STAR/AQEM sono stati confrontati con quelli osservati per la corrispondente classe del sistema nazionale (*test dataset*) attraverso il test non parametrico Mann-Whitney U test. Dopo la fase di comparazione è stata effettuata la fase di armonizzazione.

Particolare rilievo è stato dato alle due opzioni di armonizzazione che prevedono un confronto *via* ICMi. É presentato un esempio di armonizzazione basato sulla selezione dei valori mediani dei *boundaries* derivati dai metodi nazionali ipoteticamente conformi alla Direttiva. Per gli scopi dell'Intercalibrazione, l'utilizzo di un semplice valore medio (o mediano) ottenuto dai metodi nazionali non dovrebbe essere consentito, almeno fino a che i vari metodi non saranno stati resi interamente conformi alle richieste della Direttiva Quadro.

In un esempio successivo, l'armonizzazione è stata effettuata sulla base del confronto statistico tra i valori di ICMi osservati nel *benchmark dataset* e i valori dello stesso indice osservati nei *test dataset*. Sono stati inizialmente considerati i campioni classificati in Buono stato di qualità. Se i valori dell'indice ICM dei due *dataset* (cioé dei due metodi di classificazione) mostrano differenze significative e il metodo nazionale appare troppo permissivo, il limite tra le classi Buona e Moderata per il metodo nazionale viene quindi riposizionato in corrispondenza di un valore più elevato, al fine di eliminare le differenze osservate. Come risultato di questa procedura, alcuni campioni precedentemente classificati in Buono stato di qualità verranno invece attribuiti allo stato Moderato. Dopo l'aggiustamento del *boundary* Buono/Moderato, viene considerato il *boundary* Elevato/Buono procedendo al confronto statistico per i campioni classificati in stato Elevato. Il tal modo vengono infine stabiliti i nuovi limiti armonizzati tra le classi Elevata/Buona e Buona/Moderata.

I risultati generali riportati nel presente Quaderno hanno dimostrato come l'utilizzo di metriche comuni per l'intercalibrazione (ICMs) possa essere proficuamente applicato su scala europea. La procedura per il calcolo dell'ICMi e per il confronto dei *dataset* è stata descritta in dettaglio e potrà essere applicata dagli Stati Membri, all'interno dei GIG, per la procedura formale di

Intercalibrazione. L’approccio ICMi favorisce l’utilizzo di *dataset* esistenti raccolti direttamente dagli Stati Membri, in grado di garantire una consistente disponibilità di dati per il processo di Intercalibrazione. L’esercizio di comparazione tra i *boundaries* europei ha consentito di evidenziare risultati differenti per i diversi tipi fluviali e tra le diverse procedure, ma ha dimostrato come i sistemi di valutazione e i relativi *boundaries* possano essere effettivamente comparati in tempi brevi.

L’approccio che prevede il confronto diretto risulta applicabile soprattutto quando i sistemi di valutazione confrontati sono piuttosto simili (ad es. per una regolazione fine dei limiti tra le classi) e nel caso di disponibilità di ampi set di dati raccolti in accordo con le procedure standard dei metodi applicati. Ciononostante, i risultati ottenuti mediante il confronto diretto mostrano come la percentuale dei siti la cui classificazione potrebbe variare dalle classi Elevata e Buona alla classe Moderata può essere decisamente elevata (nella maggior parte dei casi intorno al 30%). Al contrario, le opzioni di armonizzazione basate sull’approccio ICMi non evidenziano grandi differenze tra i metodi confrontati, i quali mostrano sempre modeste percentuali di siti che potenzialmente necessitano di una collocazione diversa rispetto alla classe originale (da Elevata/Buona a Moderata). Durante lo svolgimento del lavoro, sono emersi alcuni problemi che necessitano di discussioni più approfondite, ad esempiio per quanto riguarda i criteri per la definizione delle condizioni di riferimento.

In generale, le opzioni utilizzate, e soprattutto l’uso dell’indice ICMi, permettono un’ampia varietà di confronti, sia all’interno dei GIG, sia tra GIG differenti. Tali approcci potrebbero consentire un’armonizzazione su scala europea che renda concretamente realizzabile il processo di Intercalibrazione richiesto dalla Direttiva Quadro sulle Acque.

1 - INTRODUCTION

The Water Framework Directive (WFD, European Commission, 2000) creates a new legislative framework to manage, use, protect, and restore surface water and groundwater resources within river basins (river catchments), transitional (lagoons and estuaries) and coastal waters of the European Union. The WFD aims to achieve sustainable management of water resources, to achieve good ecological quality and prevent further deterioration of surface waters and groundwater and to ensure sustainable functioning of aquatic ecosystems (and dependent wetlands and terrestrial systems). The environmental objectives of the WFD (i.e. good ecological quality of natural water bodies and good ecological potential of heavily modified and artificial water bodies) should be reached by 2015.

The overall complexity of the Water Framework Directive and a very tight implementation timetable creates challenges for the fulfilment of requirements. Therefore the European Commission and the Member States started a Common Implementation Strategy (CIS) in 2001 (European Commission, 2003a). This has resulted in a number of guidance documents, where the various technical issues related to the WFD implementation requirements are interpreted according to the common understanding of Member States. They are not legally binding but present examples of best practises and common understanding of the legal requirements.

The Inter-calibration (IC) process (European Commission, 2003b) is the primary issue to be addressed. The Inter-calibration process aims at consistency and comparability in the classification results of the monitoring systems operated by each Member State for the biological quality elements. The Inter-calibration exercise must establish values for the boundary between 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, 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 address 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, through its research framework programmes, co-funded Europe-wide research projects, such as AQEM (Hering et al., 2004), FAME (Schmutz et al., 2001), STAR (Furse et al., 2001), REBECCA (Rekolainen et al., 2003), etc. The present Paper represents a contribution from the STAR and AQEM projects to outline the procedure for performing the Inter-calibration process for European rivers.

As a general tendency in the U.S.A. (e.g. Karr et al., 1986; Barbour et 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 ease 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 Inter-calibration Common Metrics to be calculated for river sites within or among Geographical Inter-calibration Groups (GIGs), Member States and stream types (European Commission, 2004).

1.1 - Objectives of the Paper

The present paper deals with the Inter-calibration of assessment methods in terms of harmonization of their resulting classification (i.e. class boundaries). According to Köhl et al. (2000), 'harmonization is based on existing concepts which should be brought together in such a way as 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: Harmonization is intended for results of biological assessment methods only.

Class boundary

The EQR value representing the threshold between two quality classes.

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

Among other important aims, the STAR Project worked to make stream assessment methods and results comparable in all of Europe, in order to achieve equivalent river quality in future. The STAR and AQEM projects results can support a number of different analytical approaches dealing with the EU 'Inter-calibration' process. Possible procedures to harmonize European class boundaries based on STAR/AQEM data, for aquatic macro-invertebrates are provided here. The present Paper deals with the Inter-calibration of assessment methods in terms of the harmonization of their resulting classification (i.e. class boundaries).

In considering the future standardization and harmonization of methods, the very different situations and traditions of European countries must be taken into consideration:

- It is unlikely that countries will change proven assessment methods, (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 without good reason.. Hence, comparability of results can only be achieved through an Inter-calibration, as required by the WFD. In addition, it is worth mentioning that the data being used for the IC process will largely be already existing data, thus requiring respective collection and calculation methods to be considered.

- Many of the existing assessment methods, which continue to be used in certain countries, do not entirely fulfil the demands of the EU Water Framework Directive. These methods need some adaptation and in particular, the 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 Inter-calibration, at least until all MSs' assessment systems are proved to be fully WFD compliant.
- Compared to invertebrates, stream assessment methods and systems using fish, macrophytes, phytoplankton and phytobenthos are less developed and far less data is available. 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 draft standards are already in existence. However, at present these groups are less commonly used than macroinvertebrates in water management and few widespread methods exist for calculating valid indices and converting the results into degradation classes. It is unlikely that initially 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 ecological data sources, defined and standardized methods are needed to integrate and Inter-calibrate the results obtained from different organism groups. The present paper will mainly deal with invertebrate data, which provides the most abundant 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, though simplified, to make it compatible with the timing and scopes of the IC process, will be considered in this paper.

Indirectly, we will deal with the problem of defining reference conditions and we will 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. criteria to accept/refuse sites as reference sites). After this step is 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 crucial step for implementing the Water Framework Directive. 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 advise the European Commission and the countries involved in the WFD inter-calibration process on how this information could be used, in conjunction with political considerations, in assisting the process of defining and delimiting the five grades of ecological status.

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

The main aims of the present paper 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 contest the results, with the idea of providing a general framework for discussion for people involved in the formal Inter-calibration process, that will be performed during the next years.
- To provide general information on potentially suitable large scale metrics, for a sample area in Europe (Central Europe).
- To outline the overall differences among test datasets from different countries in terms of their distance from one another, the average conditions or the tentative benchmarking system, which is supposed to fully satisfy WFD requirements.
- To provide a few full application examples of some of the considered harmonization procedures.

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

- To define new assessment systems (i.e. the proposed common approaches are explicitly dedicated 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 helpful examples for covering the whole gamut of Water Body Types, Stream Types and Biologcal Quality Elements to be inter-calibrated 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.

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, 2004). Three different Options are present in the Guidance, with their respective advantages and disadvantages listed briefly and 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

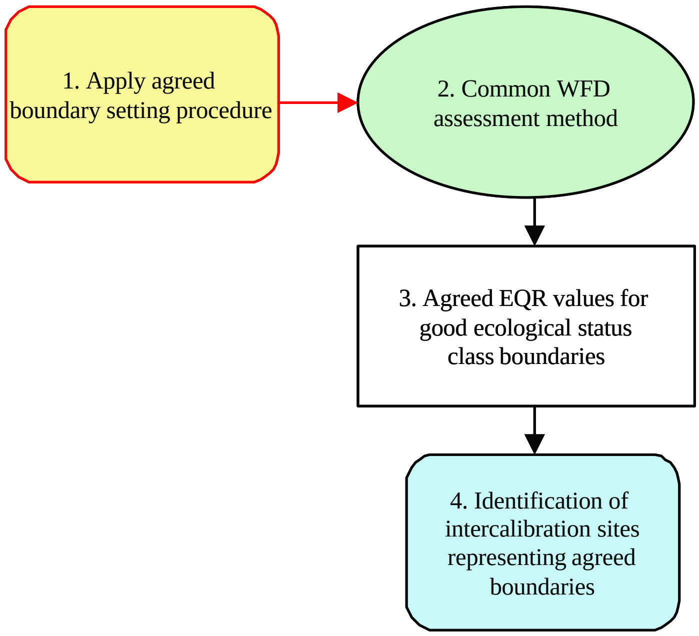


Figure 1.1 Example of how the application of Option 1 might take place (from the IC Guidance, European Commission, 2004)

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

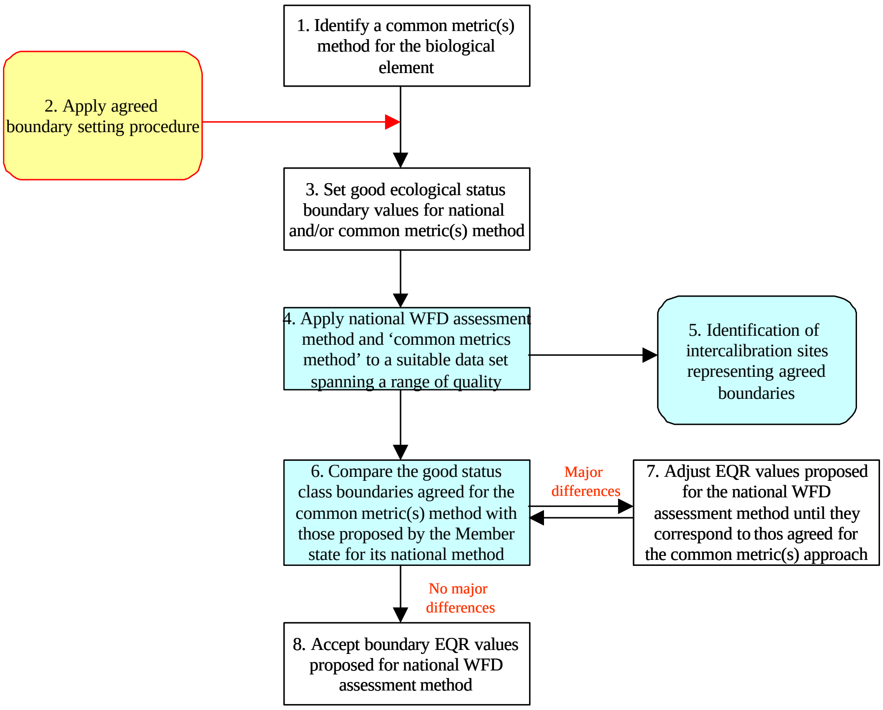


Figure 1.2 Example of how the application of Option 2 might take place (from the IC Guidance, European Commission, 2004)

1.2.3 - Option 3: Direct comparison of national methods at Inter-calibration sites

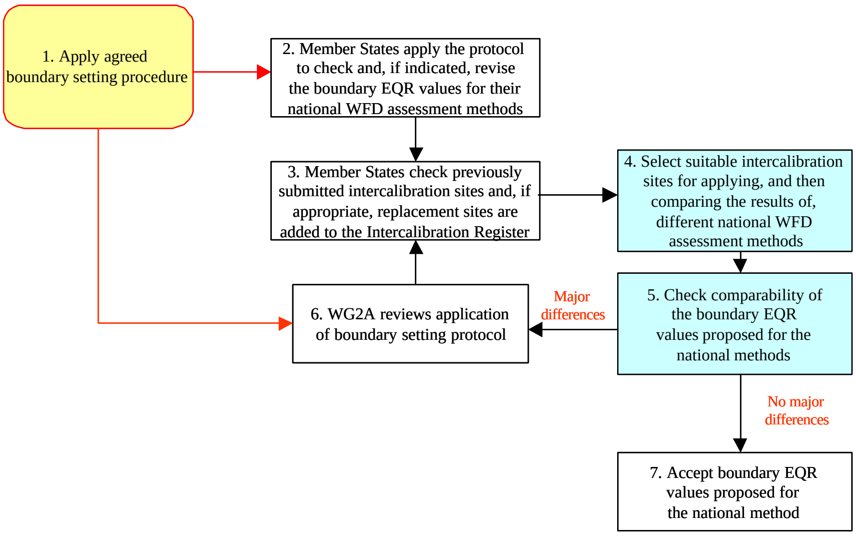


Figure 1.3 Example of how the application of Option 3 might take place (from the IC Guidance, European Commission, 2004)

A full use of Option 1, while ideal, will be possible only on a 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, because it is also the basis for the acceptance/refusal of sites as ‘reference’ sites. The possible steps of a ‘boundary setting protocol’ have been outlined in the IC Guidance and individual GIGs are expected to develop tight protocols adapted to their geographic areas and the main acting degradation factors.

Given the central aim of consistency with normative definitions, Option 2 put the emphasis on looking for a clear comparability among European class boundaries and assessment systems. To delineate this Option, an Inter-calibration Common Metrics (ICMs) approach was proposed for rivers (Buffagni & Erba, 2004). In addition, a full application of Option 2 assumes that a trans-National, benchmarking system can be agreed upon. In general terms, it means that the data provided by single MSs should be matched with 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 data is compared between countries directly, in the format that it is collected by each MS. This Option shows clear scientific and practical 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 3 would require very detailed data to be provided and jointly examined and/or large field activities, which are not planned at present. If not, there is a real danger that the application of this Option will result in an empty 'political agreement', which is in conflict with 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 Inter-calibration 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 made with the help of a ICMi approach (as in Option 2).

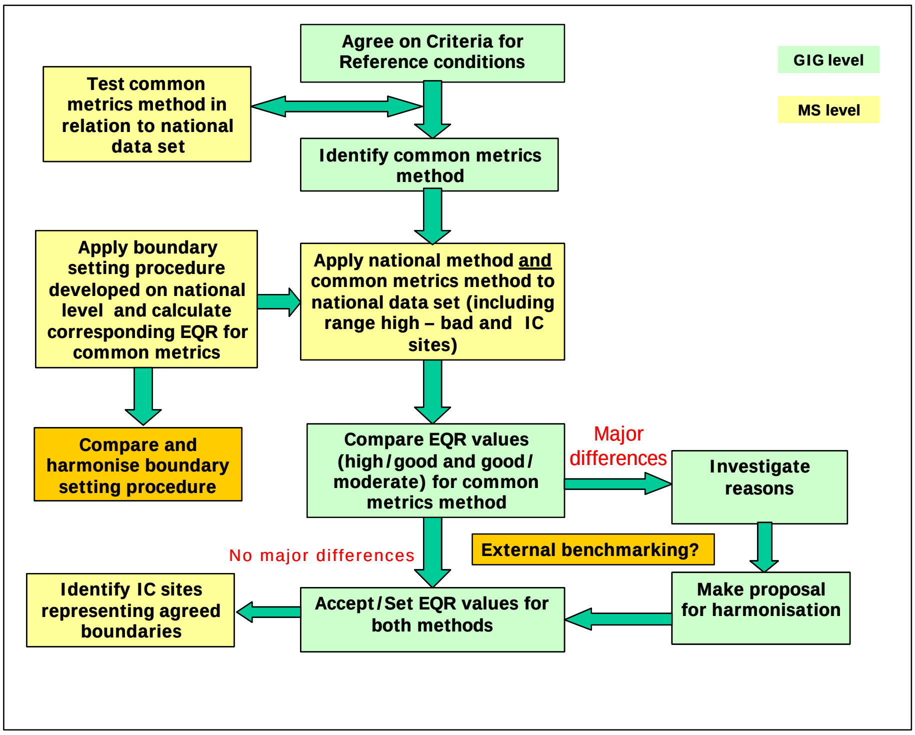


Figure 1.4 Example of a hybrid Inter-calibration approach, combining elements of Options 2 and 3 (from the IC Guidance, European Commission, 2004)

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

In the paper, we discuss some possible procedures for the determination of European class boundaries of Ecological Status. In doing so, we are aware that class boundaries should be agreed upon in the official Inter-calibration process. We therefore envisage that our results might serve as a proposal for WFD Inter-calibration 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 in performing such an important and potentially difficult task such as the IC process of European water bodies. It is not the intention of this paper and of the whole STAR Consortium to push one option or another. The general applicability of approaches – with the focus on rivers and aquatic invertebrates – is being evaluated, through a random application of datasets provided by STAR partners, other scientific institutions and environment agencies or Environmental Ministries from around Europe.

Hybrid Inter-calibration option applied for rivers
(from the IC Guidance, European Commission, 2004

An example of a hybrid Inter-calibration approach is given in Figure 1.4. In this approach boundaries are initially set by the Member State (as in Option 3), then compared to a common metric (as in Option 2), and harmonized where necessary). Common metrics enable a GIG-wide comparison of classification results. For this approach to be successful it is essential that there is agreement within the GIG on criteria to derive reference conditions.
With this approach it is not necessary to compile a single data set at the GIG level, thus avoiding the problem of collating data from different countries applying different methods. Instead, Member States apply a common metric to their own data sets, and compare this to their national assessment results. This approach is especially suitable in cases where Member States have relatively well-developed assessment methods in place at the start of the Inter-calibration exercise (e.g. macroinvertebrate assessment methods for rivers), and where a robust common metric is available. This procedure is undergoing testing in the Alpine, Mediterranean, and Central/Baltic river GIGs, with very promising results.
Because initially the class boundary setting procedure is only applied by Member States using their own data and methods, it will be necessary to compare and harmonize the different steps of the class boundary setting procedure within the GIG. If the comparison of Member State's classification results using the common metric show that there are no major differences between countries this should be a relatively trivial task; if there are major differences that cannot be resolved within the GIG it may be necessary to directly apply the class boundary setting procedure to a benchmarking data set (best available classification)

A relevant part of the data presented and processed here (see Chapter 4) was provided as part of the ongoing pilot IC exercises in the GIGs. In particular, most test datasets refer 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 led 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 Paper is reported.

STAR partners who participated to the compilation of the present Paper

CNR – IRSA, Water Research Institute	Italy (coordination)
University of Duisburg-Essen	Germany
Environment Agency	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

CNR – IRSA, Water Research Institute	Italy
Environment Agency	United Kingdom
Institute of Environmental Protection, Warsaw	Poland
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
CEMAGREF, Lyon	France
Royal Haskoning.	The Netherlands
Umweltbundesamt and LAWA	Germany

Institutions that provided benchmark datasets

STAR partners

CNR – IRSA, Water Research Institute	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

2 - PROCEDURE AND GENERAL TOPICS

2.1 - Summary of the STAR ICMi Inter-calibration procedure for macro-invertebrates – Comparison and Harmonization

2.1.1 - General statement

- The summary procedure for Inter-calibration presented here is a technical supplement to the ECOSTAT WG 2.A Discussion paper distributed in February 2004 (Buffagni & Erba, 2004 → Annex III).
- 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 Inter-calibration exercise* (ECOSTAT WG 2.A, European Commission, 2004).

For Inter-calibration 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 bio-geographical reasons even in similar physical contexts. In addition, the existing methods have different sampling strategies and laboratory procedures, and are to some extent based on different concepts. This is why an intermediate step such as the Inter-calibration Common Metrics index (ICMi) is needed (see Buffagni & Erba, 2004; ECOSTAT WG 2.A, European Commission, 2004).
The examples provided in the present report refer to river invertebrates but the general 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 Inter-calibration 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 the differences in sampling, analytical and computational methods used by different national monitoring and classification schemes. Such procedure is likely to be supplemented by more precise bilateral and multilateral Inter-calibration between similar national methods which may be based on more detailed taxonomic resolutions.

2.1.3 - Overview of Inter-calibration via ICMi

Inter-calibration involves two main steps:

1 Comparison of existing national class boundaries

2 Harmonization (adjustment) of boundaries

Provided there is consistency in each of the assessment methods to the normative definitions, harmonization 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 could be performed through a 'bilateral' comparison of two national methods (but this will hardly guarantee a complete European comparability). For both, a method of making the MSs' quality classifications comparable, would be to calculate a set of agreed Inter-calibration Common Metrics and to combine them into an ICM index.

2.1.4 - 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 Inter-calibration Common Metric index (ICMi)(Buffagni & Erba, 2004; Buffagni et al., 2005). The ICMi and the national classification method 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. A regression between EQRs of ICMi and national method values (after normalization) is derived and class boundary values expressed in terms of ICMi are then obtained for all methods/countries.

2.1.5 - Summary of the concept of harmonization

The concept of harmonization deals with the common understanding of ecological status, especially for what can be considered good and moderate. In the present report, the harmonization is carried out by ‘shifting’ boundaries - High/Good and Good/Moderate - in order to reduce/eliminate differences among different clusters of samples (i.e. grouped into quality classes) for datasets and methods, which are being tested. The basis for the harmonization is the calculation of a common Inter-calibration 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 boundary values derived from (hypothetically) WFD-compliant, national methods 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. based on STAR/AQEM data, 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).

2.1.6 - 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 data), which should be independent from 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 interpret 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 Good/Moderate 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 (see chapter 8.3 for examples).

2.1.7 - Criteria used for ICMs' selection

The Inter-calibration 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 inter-type datasets (e.g. M1, C1) and intra-type datasets (e.g. AQEM and STAR datasets), as well as considering recent metric selection experiences (e.g. AQEM Consortium, 2002; Buffagni et al., 2004a; Hering 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. The selected metrics are reported in chapter 3 and Annex 2.

The metrics used here should be considered tentative. Almost certainly, some changes might be necessary in the metric composition of the ICM indices to be used for the Inter-calibration exercise, e.g. in different GIGs. In particular, the metric 1-GOLD might result as not well suited to describe the quality gradients across the whole of European stream types. Abundance-based metrics are included here for two main reasons: a) to fit the normative (WFD) definitions; b) because they are often able to discriminate well between sites of different environmental quality (see chapter 3.2 for examples). The applicability and suitability of this category of metrics for different datasets and stream types across Europe will be checked within GIGs (see also chapters 4 and 5).

2.1.8 - 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 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 are normalized, i.e. the score of each ICM is divided by a reference value, typical for the dataset/stream type. Some possible Options were considered to define this value (Chapter 3.3).

2.1.9 - ICM index

After their normalization, the metrics are combined into an Inter-calibration Common Metric index (ICMi)(Chapter 3). Metrics are grouped into three groups, providing information on three major response areas: Tolerance, Abundance/Habitat and Richness/Diversity (see Chapter 3/Annex II). A different weight is attributed to the metrics within each group (see Chapter 3/Annex II), giving greater importance to the metrics based on the whole community (Buffagni et al., 2004a). To obtain the final multimetric score, the same weight is attributed to each of the three metric groups (0.333).

2.2 - Identification level

2.2.1 - Taxonomic resolution in freshwater research and biomonitoring

The issue of taxonomic resolution has been widely discussed (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 for monitoring purposes.

In any case, the level at which scientists identify freshwater macro-invertebrates varies due to the high number of species of different orders that compose the benthic community and to current available knowledge. Every so often, the capability to go to the species level is restricted to taxonomic experts and is not at all ready-to-use for end-users. The scientific community is used to arriving at different taxonomic resolution by depending on the final objective: e.g. 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 & McElravy 1993). The central issue is to define when it is necessary to identify at the lowest practicable level and if this option is obtainable in terms of human resources and time available. The lowest practicable level depends on the technical, taxonomical expertise on each different taxon constituting the benthic community and the availability of the time needed to reach a lower identification than e.g. family level.

The highest precision in bio-indication can be reached with data on species level. As illustrated in the niche concept, each species is expected to evolve special abilities to exploit resources and to cope with habitat heterogeneity. The occurrence of specialized species assemblages is therefore a result of the existing environmental conditions. In cases where the autoecological 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 bio-indicators for the ecological status of aquatic habitats. Because of autecological differences among related species, aggregating into the higher levels, the use of higher taxonomic levels may result in a loss of information potentially relevant for bio-indication purposes (e.g. for WFD investigative monitoring). Lack of species level information may reduce the ability to detect more subtle changes in ecological quality.

On the other hand, a species or genus level of identification may induce unjustified variability in quality classifications, due to the increased possibility of incorrectly identifying invertebrate specimens, a situation that can reduce the objectivity of biological assessment data and analysis.

For the STAR project (http://www.eu-star.at), the uncertainty of sorting and taxonomical identification were tested. Some preliminary results seem to indicate that, in some cases, the percentage of misclassification (or variability among different groups of surveyors) can be up to 20% even at family level. The aim of improving the amount of ecological information used by reaching a lower level of identification - while potentially very useful in assessing river quality status better - can be strongly effected by identification errors. Apart from the choice of taxonomic resolution, the call for an evaluation of accuracy and precision of taxonomic analysis has to be raised. Moulton et al. (2003) pointed out that often a key source of errors is the human factor. The employment of different taxonomists and the abilities of individual researchers can affect the quality and replication of results, usually due to problems related to lack of time, limited experience or the insufficient training of the taxonomist/surveyor. The Environment Canada (1993) provides some simple recommendations: the identification should be verified by an expert in the taxonomic group of interest; people who carry out the identifications should be named with details of their qualifications; literature and taxonomic keys used for benthos identification should be referenced.

Whereas bioassessment methods are rapidly growing and evolving, during the last 20 years a dramatic decline in taxonomic research has been obvious, despite the increased demand for taxonomic expertise (Moulton et al., 2003). Large funding resources have been diverted to other ecological fields while taxonomic, faunistic and autoecological investigations - essential as a basis for any reliable study of applied ecology - have been almost entirely abandoned. Research is now needed on benthic macroinvertebrate taxonomy and distribution to improve bioassessment as a water resource management tool (e.g. Buffagni et al., 2001). In some areas identification levels lower than family or genus are difficult for a lack of basic knowledge about the 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 mayflies in Italy, a number of studies have 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, 1998), 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 in correctly identifying individuals are major problems in the choice of a low taxonomic resolution. With the objective of preserving some of the species-level information, without an obligation to identify to species the collected organisms, an intermediate identification level between genera/family and species can be proposed (Buffagni, 1997). To approximate the specific composition of the community, a definition of benthic groups - either taxonomic or morphotaxonomic - with a fixed identification level could be employed. These groups, named Operational Units (OU), are created following the philosophy of grouping species by similar autoecological features and whenever possible, the most easily visible characteristics that can be singled out are used to discriminate among OUs, so as not to complicate identification. The OU permits the reduction of identification errors, as compared to species level and at the same time, offers more detailed ecological information. For instance, in Italy the following aggregation of mayfly species at different levels was proposed: genus level (24 OU), morpho-taxonomic group level (11 OU) and species level *(Baetis rhodani* and *B. buceratus* only). A linear regression analysis between species number and OU number on 150 samples collected in northern Italy was conducted and 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 (Buffagni, 1997).

As a general conclusion, it can be stated that there is a clear need for improving taxonomical expertise around Europe, especially considering the ‘balance’ between the effort of identification and the possible outcome in terms of biomonitoring and resource management (see the paragraph below). New and more detailed information being gained for data-poor areas (e.g. Southern Europe) should constitute the basis for the selection of metrics selectively sensitive to specific stressors and better assessment systems. In other areas, where the expertise available is greater (e.g. Central and Northern Europe), it is important to save the tradition of in-depth taxonomic investigations (e.g. for investigative monitoring), nevertheless considering the possibility of using cost-effective systems for wide scale, rapid bioassessment protocols (e.g. for basic operational monitoring), which could be based on higher taxonomic resolution in order to reduce the effort of identification.

2.2.2 - Identification level used for monitoring in Europe, the WFD requirements and the IC process

The Water Framework Directive (European Commission, 2000) establishes a new European standard for assessing river quality. Three main types of monitoring are indicated by the Water Framework Directive: surveillance, operational and investigative (European Commission, 2000). Each has different aims and frequency of application, focusing on the different information to be obtained for a river site (WFD, Annex V, 1.3.1/3). The WFD claims for 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, pressure specific assessment is required to evaluate the impact of different stressors, in order to guide future management. Better taxonomical resolution allows for more detailed ecological interpretation of monitoring data in the detection of the causes of degradation in investigative monitoring. To measure the exact biological response to a pressure in operational monitoring the use of species level data is, in some cases, inevitable.

Every European country has historically developed different bioassessment methods. The methods use different techniques for sampling and sorting, use a different taxonomic resolution and analysis of the data. All these features contribute to making the results obtained from the various national methods difficult to directly compare. 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 (AFNOR, 1992), etc. A more detailed identification level is in use 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 water, assessment systems based on family level are sometimes sufficient. The family level, easier to reach, seems to be appropriate whenever there is the need to determine large differences between sites, resulting in a coarse or primary value, that in cases of bioassessment means a preliminary rough classification. In taxonomically poor areas, where it is known that the number of species and genera is similar to the number of families, the family level is appropriate and enables a significant saving of resources, in terms of the time and money invested. Lenat & Resh (2001) advise aiming at a lower identification level where the small differences between sites or dates and conservation studies are to be detected,. These 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 or BMWP (family based), IBE, 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 pollution vanishes or the pollution gradients are less steep, the above mentioned systems should be replaced or supplemented by other assessment methods. In Central European countries (e.g. Germany, Austria), where organic pollution is now a lateral-problem in river management, assessment systems focussed on the detection of organic pollution give the same results almost everywhere. The dominant stressors, affecting Central European rivers today (e.g. hydromorphological degradation, catchments land use, eutrophication, pesticides) are acting in a much more subtle way. However, they still have detectable effects on the biocoenosis. There are indicator taxa for certain habitats, while others reflect catchments integrity or hydromorphological structures. In most cases, these are species level information (Feld & Hering, submitted).

It is important to state in this paper that most important factor in the selection of an identification level is the <u>search for comparability</u> across Europe. At least four situations should be borne in mind:

a) The data available across Europe for the Inter-calibration process must be suited to the application of any proposed procedure → i.e. the minimum common identification level should be selected;
b) For the WFD Inter-calibration activity, only major changes between assessment systems and classes of European countries are being analyzed → i.e. a relatively high identification level might be enough to detect major differences;

c) What is being inter-calibrated for European water bodies is the overall 'biological status' of a site, not the effect and quality derived for any single stressors acting → i.e. there is no need to select and inter-calibrate very detailed, species level biological metrics, which might be more suitable than those calculated at higher taxonomic level for stressor-specific assessment systems;
d) The IC process itself is not aimed at developing new assessment systems → i.e. it can be assumed that each MS is developing WFD-compliant methods, which are well adapted e.g. to bio-geographic, hydrological, environmental conditions and to the available taxonomic knowledge.

Such considerations lead to the selection of the <u>family level of identification</u> as the most suitable assessment for the pan-European Inter-calibration of rivers.

2.2.3 - Taxonomic requirements of the ICM index

The ICM index for benthic invertebrates (see chapter 3) is then calculated with taxonomic data on family level. This allows for integrating monitoring data of different countries to gain comparability within and across GIGs. The standard monitoring programmes of certain 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 Inter-calibration exercise outlined in this paper, 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 lays down the common ground for the Inter-calibration 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 the classification results of the national assessment system which were obtained by application of the taxonomic resolution used for the national monitoring programme. Even if the ICM approach does not imply any recommendations for the adequate taxonomic level needed in biomonitoring, a high correlation with the ecological quality gradients as defined by the national methods can be seen in most datasets (chapter 4) as the different metrics calculated at family level demonstrate. 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 the overall quality of sites when assessing their 'general degradation', i.e. to define their ecological status.

2.2.4 - Level of taxonomic identification needed for bilateral comparison

Bilateral comparison of national assessment results, as a direct method of inter-calibration e.g. in trans-boundary catchments, can be proficiency based on species level data. In relation to the ICMi approach, which aims at a coarse comparison of assessment systems on a Europe-wide scale, a direct Inter-calibration between neighbouring countries can focus better on subtle adjustments of class boundaries.

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

3 - INTER-CALIBRATION COMMON METRICS (ICMS) AND ICM INDEX APPROACH

Based on previous experience with similar sets of metrics (e.g. AQEM Consortium, 2002; Buffagni et al., 2004a; Hering et al., 2004; Pinto et al., 2004), six metrics were selected to test the procedure of Inter-calibration and give examples of a possible harmonization of European class boundaries. From the initially proposed metrics (Buffagni & Erba, 2004), the selected metrics resulted from a wide discussion that took place at 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/diversity 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' behavior, the following criteria were followed:

- 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;
- the variability of the metrics at reference sites should preferably be low;
- 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).

Some of the points listed above (i.e. ability in describing ecological gradients and ability in discriminating different quality classes) will be discussed further (chapter 6).

The process of metrics selection involved the analysis of more than 50 metrics (e.g. Eveness, Margalef, Pielou indices, single taxa' abundance). In particular, linear regression coefficients between metrics and quality classes were examined: in general, the metrics having the highest R^2 were thus selected. In

some cases, e.g. for diversity indices, the observed differences between the responses of metrics of a similar kind were very minor. The capability of metrics to separate well among quality classes was evaluated both for AQEM/STAR data (covering a wide range of different European stream types) and national datasets (see chapter 6).
The selected metrics, here termed Inter-calibration Common Metrics: ICMs (Buffagni & Erba, 2004; Buffagni et al., 2005), 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-CNR, 2004). Later on, after a joint discussion at the ECOSTAT and Inter-calibration meetings on data availability around Europe, they were set at the Family level (Erba et al., 2004; see also chapter 2.2). The identification level adopted here for the calculation of these metrics is family.

A few definitions

(Biological) Metric

A metric quantifies 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).

Inter-calibration 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 Paper are example metrics which, while showing an overall applicability, might be profitably substituted or integrated by others at the GIG scale.

Inter-calibration Common Metric Index (ICMi)

The combination of the values obtained by ICMs into a single multimetric index.

Notes: As multimetric systems are more suitable than single metrics to assess ecological quality and to describe biological communities, preferably more than one metric should 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.

Table 3.1 The Inter-calibration Common Metrics (ICMs) used in the analysis and comparisons shown in the present report

Intercalibration Common Metrics (ICMs) selected for STAR Inter-calibration 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. Ofenböch 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 - Weights of the ICMs in the calculation

The selected metrics have been weighted according to the conceptual group to which they belong (see Table 3.1), giving the same weight to each of the three groups. Even if some conceptual groups were identified, it has to be stated that it is difficult to associate unique information to each individual metric. ASPT was assigned to the group of tolerance metrics because of literature information; EPTD abundance and 1-GOLD were assigned to the group abundance/habitat, because these taxa often react to habitat alteration (Buffagni et al., 2004a; Pinto et al., 2004), even though they are also influenced by water pollution (i.e. tolerance). Into each group, more weight is given to more robust metrics (i.e. metrics taking into account the whole invertebrate community). Each ICM normalized value is multiplied 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 from different combinations with certain example test datasets (especially C1, M1, M2).

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 diversity: 0.083

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

3.2 - Why to use ICMs?

Certain simple and apparent concerns make the use of the ICMs approach advantageous. Some of them are summarized below.

The use of Inter-calibration Common Metrics for the IC process can be adopted because:

- They support the translation of MSs' assessment systems results into a single type 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 expressed by MSs' systems closer to an interpretation along the lines of WFD definitions, 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 us to return to the original information collected by each MS (i.e. the invertebrate samples) with its own method, thus supporting the use of large, existing datasets all over Europe;
- They can be selected as metrics for general application and can support a large scale comparison of European streams and rivers
- The use of ICMs is encouraged by the European Commission Inter-calibration Guidance (2004).

In more general terms, a major advantage of the ICMi approach is its suitability for making different assessment systems comparable and not in its greater power to discriminate among different quality classes (→ comparability as a starting point). It is therefore suggested that an attempt be made to retain the same ICMi for different types and GIGs in order to guarantee a full European comparability, even if for specific datasets and types it does not perform as well as others. Nevertheless, it is important to state that particular stream types, such as e.g. temporary streams or large lowland rivers, do need dedicated metrics and specific approaches for the IC process.

3.3 - Scaling and Normalizing EQR values: a focal point in the WFD Inter-calibration process

3.3.1 - Introduction

The Water Framework Directive demands that numerical values be used to describe ecological quality 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 the highest achievable condition (i.e. highest quality, reference conditions). An additional statement for the derivation of EQRs, declares 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 Inter-calibration process. The combination of the two factors 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. the 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’ is applied (see the ECOSTAT Inter-calibration Guidance, European Commission, 2004). <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 is established according to the (agreed) boundary setting protocol, the High/Good boundary being derived will offer a high degree of confidence (see Table 3.4). The median value can then be used as the most robust measure for setting the reference condition, to be used in the EQRs calculation.

Table 3.4 Options for calculation of anchor value for reference conditions

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 vary greatly according to the number of observations and e.g. natural variability. When a larger number of sites are included in the dataset, the observed maximum will 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 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 re-sampling technique (e.g. by bootstrapping) to extract 12 samples for each re-sampling. This will support a more robust estimation of the maximum value, calculated as the average of the re-sampled 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 75th %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 potential presence of samples classified as High status but not being acceptable reference sites can be partly overcome. In the meantime, an intrinsic 'correction' to exclude possible outliers and to reduce extreme values due to a potentially high natural variability will be provided.

From the statistical point of view, 90th %ile have often been used when defining boundaries for biological methods. In this paper 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), whereas 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 supports the statistical evaluation, but it strongly depends 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 the present time it seems unrealistic that all MSs will be able to 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, etc. - the supporting data may not be provided in due time for the IC process. This will result in an incomplete, fragmented scenario, where some MSs will be able to provide the required data and others not. To aid comparison and in respect of 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 also be suitably used when the WFD compliant boundary setting protocol is followed. For this reason, if not specified differently, the data presented and discussed in this Paper has been normalized on the basis of the 75th %ile of the High status sites/samples (or reference sites when specified). 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). The option of normalization on the basis of the value of the 75th %ile of the High status or reference sites/samples was also applied to the final ICMi.

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 e.g. physico-chemical and hydromorphological information (i.e. pressures data).

The use of the 75th %ile in the present Paper has been adopted because for most MSs WFD-compliant systems, the 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 assumed here that the biological metrics considered in the assessment systems of MSs for the purposes of ecological quality classification and the Inter-calibration Common Metrics (ICMs) used here to aid the illustration of possible Options for the IC process can obtain 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 obtain a value higher than one. To keep this calculation option – 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 Options discussed to normalize data for further comparison for the formal IC process do need the availability of at least a few sites/samples that have been classified as reference sites and fulfil the prerequisite of the agreed boundary setting protocol. For each of the ecological status classes and for each dataset, a minimum number of samples/sites (e.g. 12) should be available to make calculations reliable.

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 re-scale single metrics or multimetric indices to make them fit into the 0-1 scale, in spite of 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 inhomogeneous 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 serious 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 one would highly increase the direct comparability of classification results across Europe.

Different European countries are actually employing dissimilar options to scale the values used to describe the quality gradient for classification purposes. In some MSs, e.g. France, the median value of Reference site samples is equated to one and the 25^{th} %ile 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 general opinion is to attribute the value of 1 to the Good/Moderate status boundary, so that it will become immediately obvious 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 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 knowledge available in the period when the assessment method (IBE) was developed and a coherent 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) Test data is presumably going to be the basis of the IC process. (b) It 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 - Requisite characteristics for test data

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

In general terms, the characteristic for each test dataset is:

- 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
- the boundaries between classes according to such assessment method must be known
- preferably the sampling 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 other elements considered (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, semi-quantitative)
- calculation formulae
- final classification (BAC, MS's) for the presented data

4.3 - Test database presentation

The present paragraph contains the description of the considered test datasets. These are presented by groups of the same Inter-calibration type (according to European Commission, 2003c).

The information provided are:

General features

Very brief overall description of the area and characteristics fitting with the IC type requirements, such as catchments 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 that collected the data and/or made the data available, together with a contact person is included. The aim of the collection, the number of sites and samples, the period of the collection of the data are also declared.

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 to be reported. The boundaries between classes are to be 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 Inter-calibration Common Metrics and test methods have been recalculated according to 75th percentile observed in 'High Status' sites according to test methods, in order to uniform the criteria and to make comparison possible. 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 a 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. The ICMi is is located on the x axis in the graphs, and on the y axis in the tables.

In all the calculations performed in the present work (see results in chapters 7 and 8), only one of the two presented regression formulae was used i.e. that one with the ICMi on the y axis: ICMi = a(NatMet) + b.

General remarks/comments

Comments, problems encountered during the treatment of the data.

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

4.4.1 - 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 regional 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 et al., 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 et al., 2004).

Notes on classification

MIF: 10 samples on 208 are classified as 'high status' according to MIF assessment method.

No Best Available Classification nor pressures based classification available.

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

For the calculation of the ICMi, metrics were normalized according to 75th percentile of high status samples according to MIF method (see explanations in previous chapters). Final ICMi is re-normalized according to 75th 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 normalization according to the high status samples' 75th percentile.

Figures below represent the linear regression between ICMi, single metrics and the MIF. Figure 1A show the relationship between ICMi and MIF and the following figures the relationship between single ICMs and MIF (Figg. 1 B-G).

Regression coefficient found between ICMi and MIF is 0.74.

The conversion of the class boundary values for the MIF method from the original boundaries to ICMi values is done according to Table 1. Original boundaries are reported in Gabriels et al. (2004)

Table 1 MIF class boundaries conversion

	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 = MIF EQR*0.9698 - 0.0258			
R^2 =0.74; p<0.001			

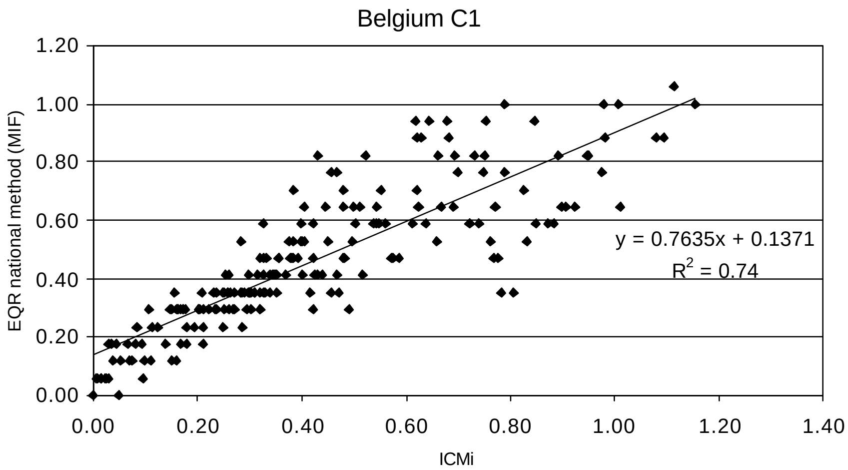


Figure 1A ICMi vs MIF - R^2 = 0,74; p<0.001

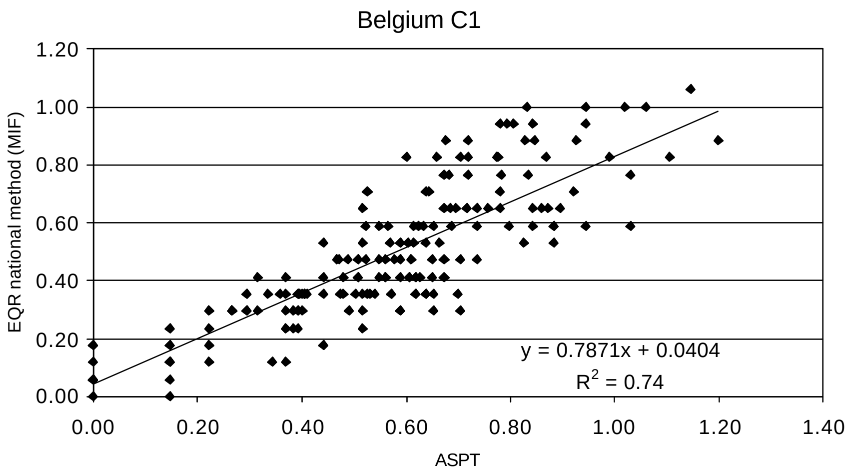


Figure 1B ASPT - $R^2 = 0.74$; $p<0.001$

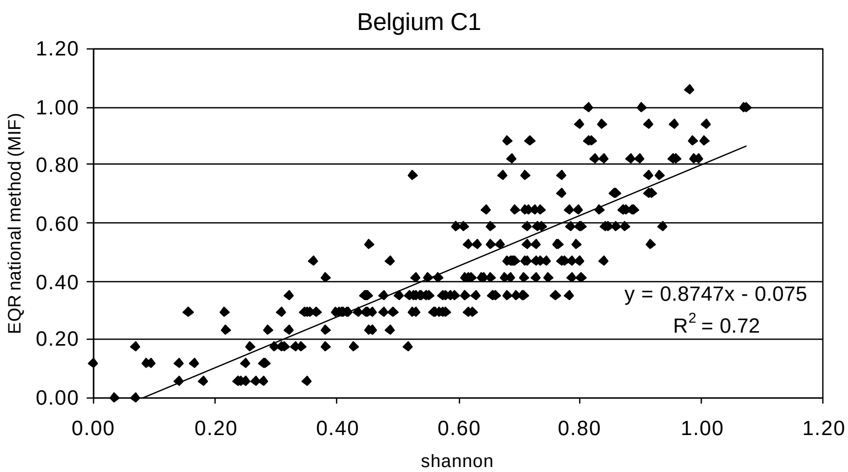


Figure 1C Shannon - $R^2 = 0.72$; $p<0.001$

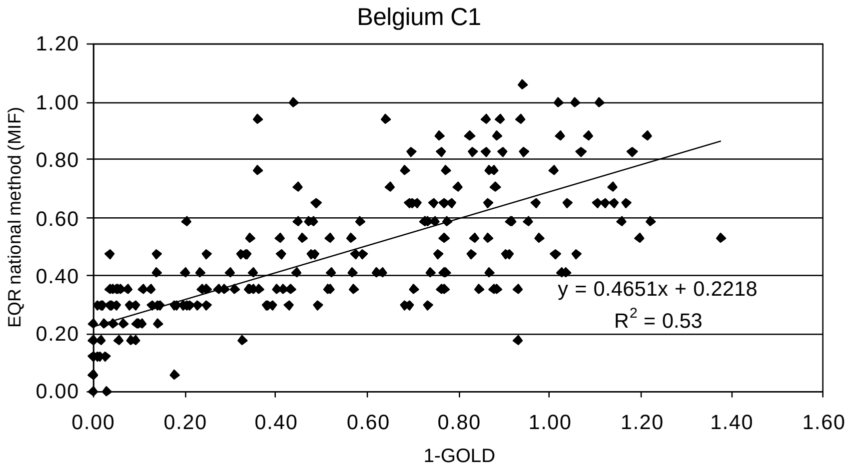


Figure 1D 1-GOLD - R^2 = 0.53; p<0.001

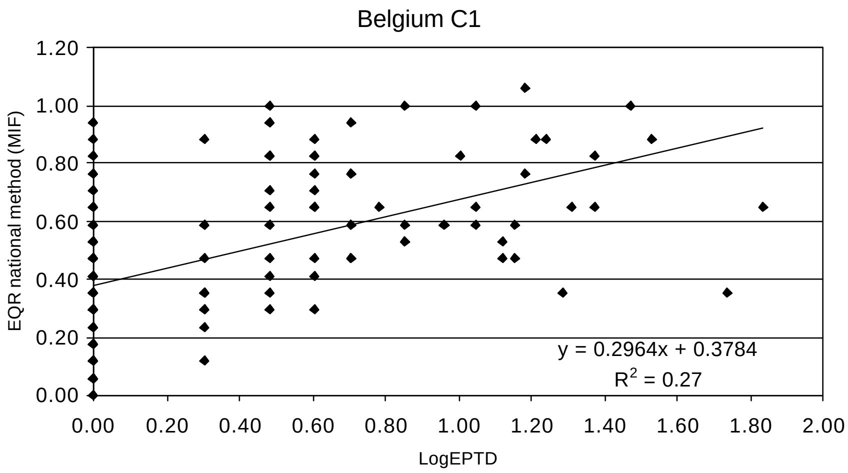


Figure 1E Log EPTD - R^2 = 0.27; p<0.001

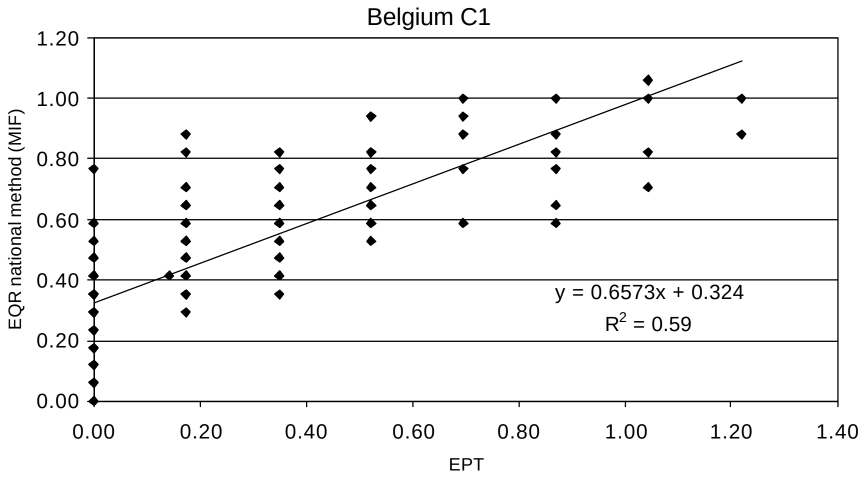


Figure 1F EPT - R^2 = 0.59; p<0.001

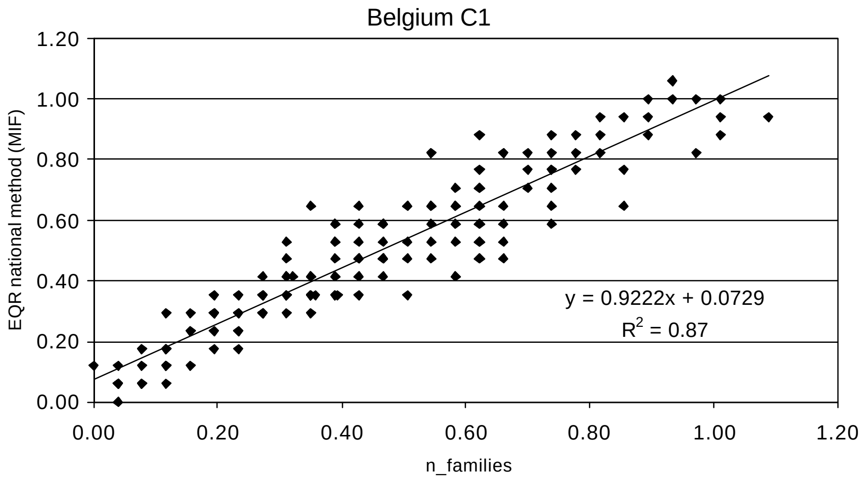


Figure 1G Number of families - R^2 = 0.87; p<0.001

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 presented 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.

4.4.2 - Denmark C1

General features

Streams with moderate alcalinity can be found only in western parts of Jutland. Here, the landscape is flat and sandy soils dominated. 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 intensive agriculture landuse (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 includes a physical description (substrate types, current velocity etc.). The main degradation factors being physical degradation (weed-cutting and regulation) and ochre pollution (because of drainage activities in the catchment). Information on water quality, based on abiotic parameters, only exists 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 general guidelines (not very detailed). 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 8% (29 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 has 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 were normalized according to 75th percentile observed in the ‘high status’ samples (see explanations in previous chapters). ICMi was re-normalized according to the 75th 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 0.52 was found (see Figure 1A).

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

The scores of DSFI in the graphs are expressed in EQR values, calculated dividing the DSFI score for each sample by the 75th 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. Original boundaries are provided by Skriver (pers. comm.).

Table 1 DSFI class boundaries conversion

	DSFI score	DSFI EQR	ICMi EQR
Limit high-good	7	1.000	0.963
Limit good-moderate	5	0.714	0.763
Limit moderate-poor	4	0.571	0.663
Limit poor-bad	3	0.429	0.564

$$\text{ICMi EQR} = \text{DSFI}*0.6984 - 0.2642$$

$R^2 = 0.51$; $p<0.001$

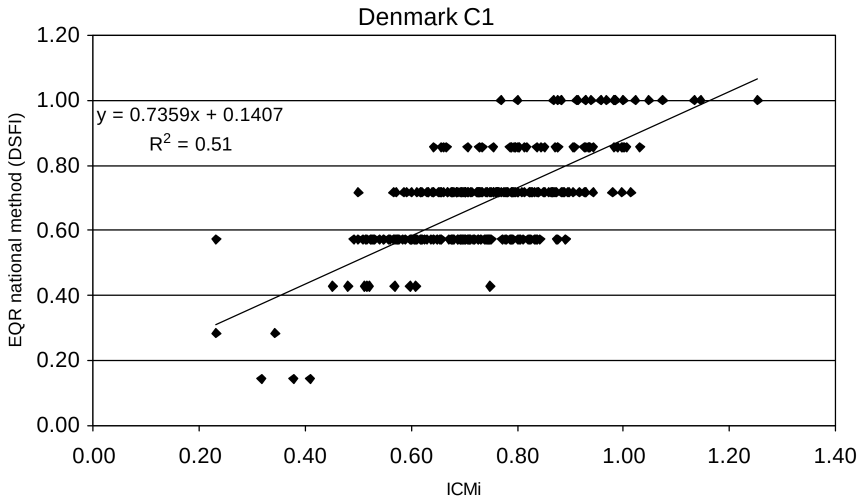


Figure 1A ICMi vs DSFI - R^2 = 0.51; p<0.001

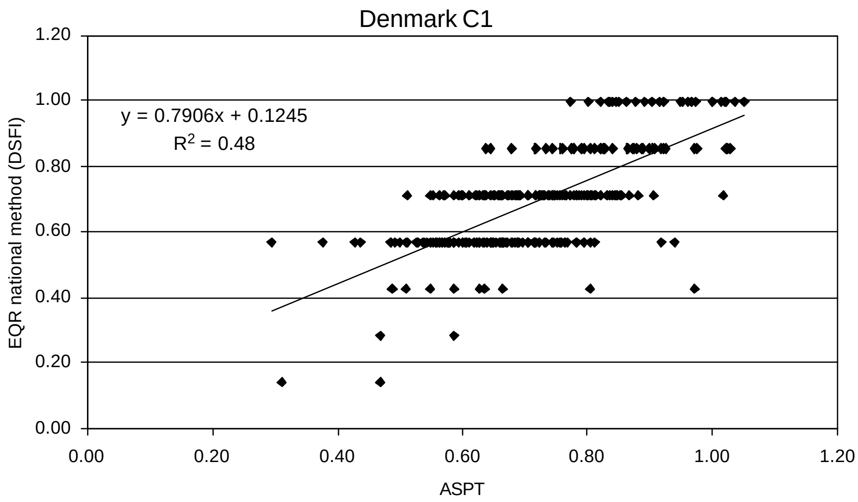


Figure 1B ASPT - R^2 = 0.48; p<0.001

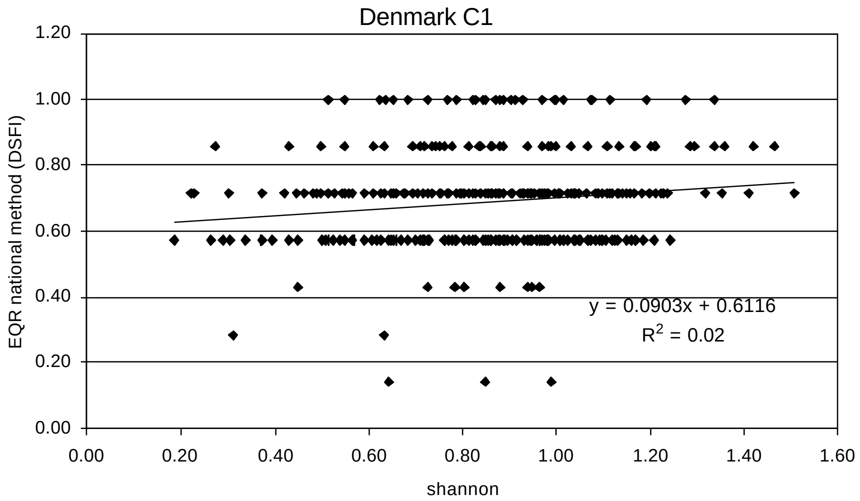


Figure 1C Shannon - R^2 = 0.02; p<0.001

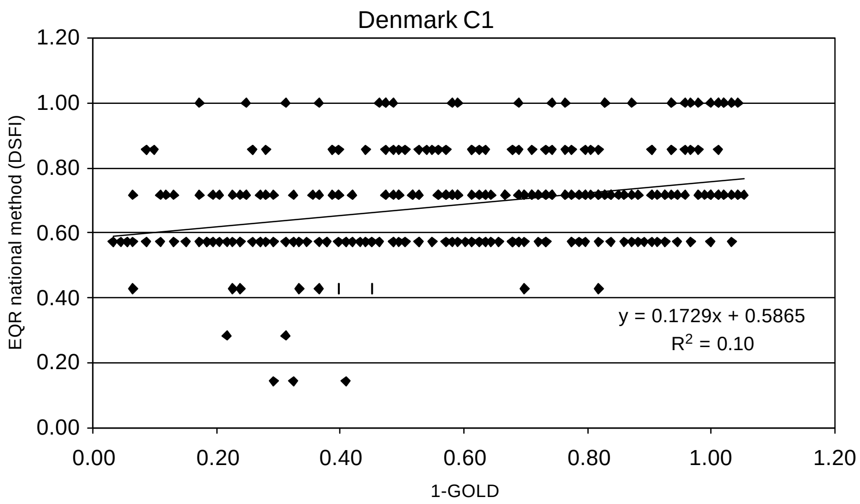


Figure 1D 1-GOLD - R^2 = 0.10; p<0.001

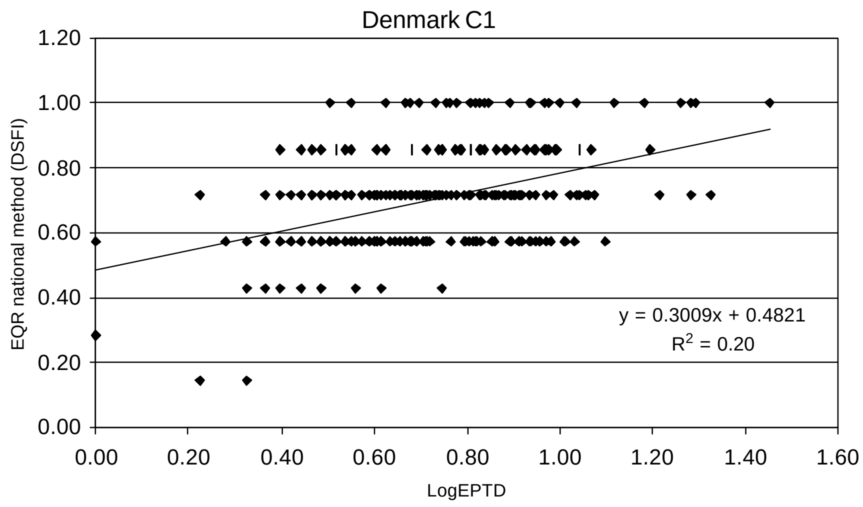


Figure 1E Log EPTD - $R^2 = 0.20$; $p<0.001$

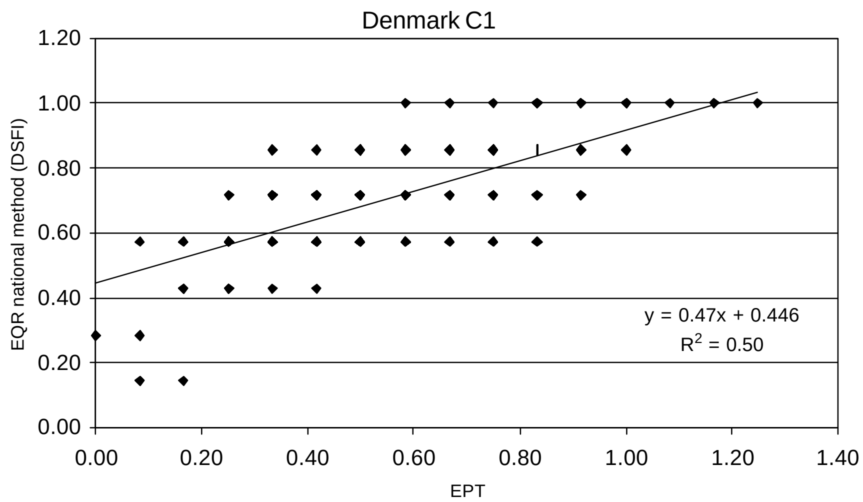


Figure 1F EPT - $R^2 = 0.50$; $p<0.001$

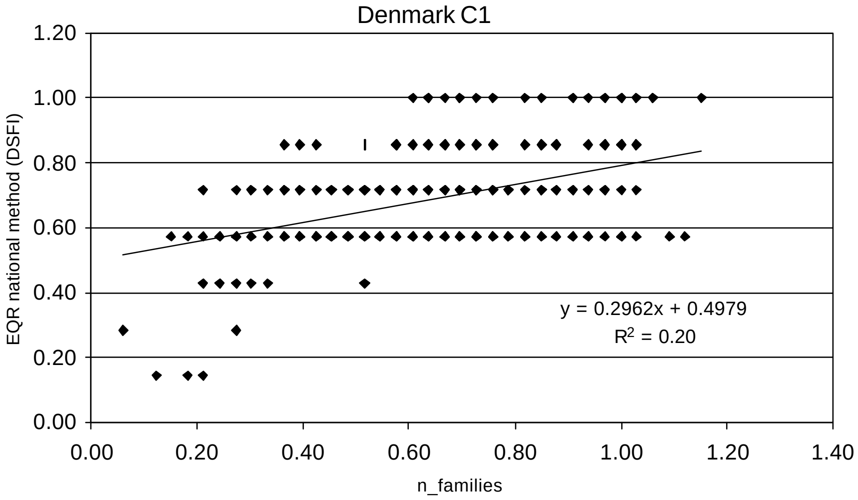


Figure 1G Number of families - R^2 = 0.20; p<0.001

General remarks, comments

The regression in this dataset may be influenced by different sorting and identification procedures. 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 occur 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.

4.4.3 - 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 characterized 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 high-water 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 channellization) 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 was 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 related measures, 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 (Timm, pers. comm.).

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.

No Best Available Classification nor pressures based classification available.

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

For the calculation of the ICMi, metrics were normalized according to 75^{th} 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 ICMs and ASPT are shown in Figures 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 75th 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. Original boundaries are provided by Timm (pers. comm.).

Table 1 Estonian ASPT class boundaries conversion.

	estASPT score	estASPT EQR	ICMi EQR
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; $p<0.001$

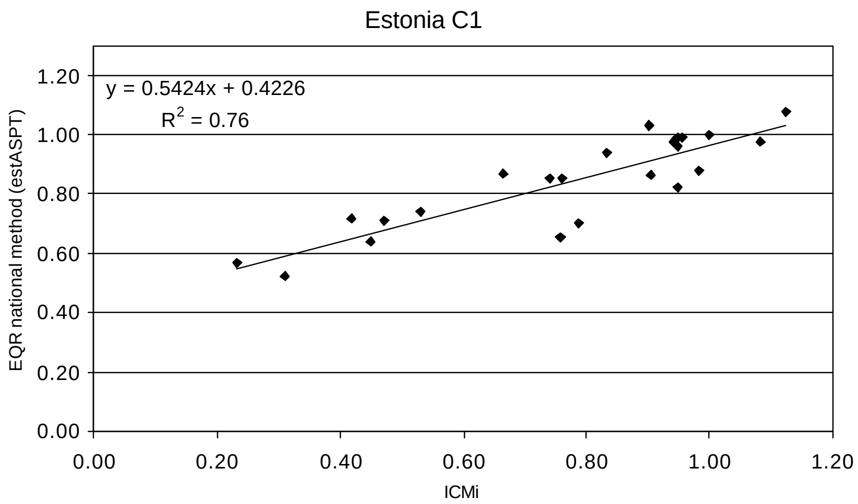


Figure 1A ICMi - R^2 = 0.76; $p<0.001$

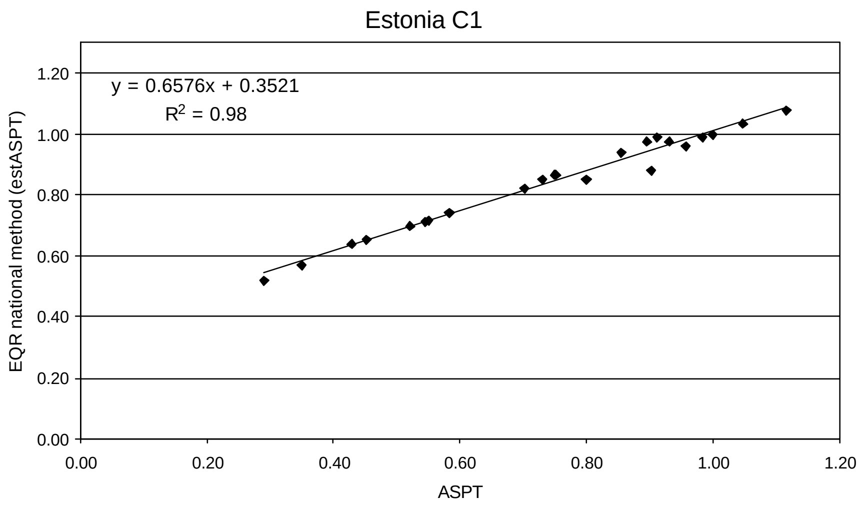


Figure 1B ASPT - R^2 = 0.98; p<0.001

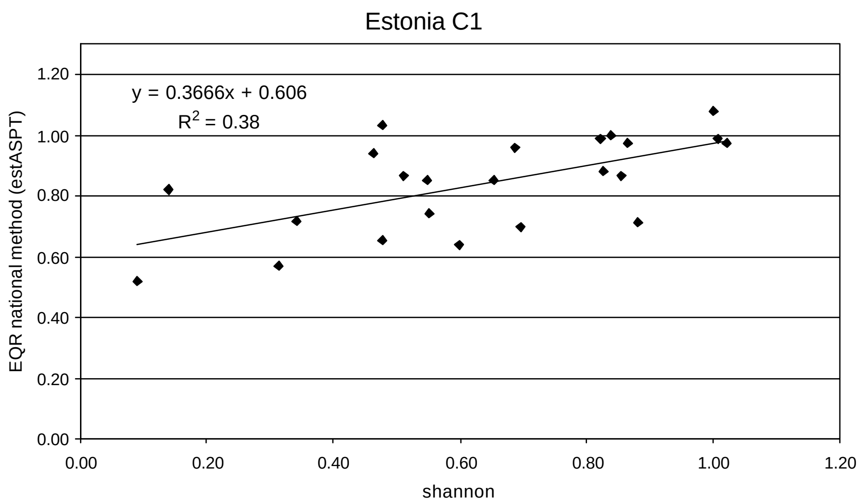


Figure 1C Shannon - R^2 = 0.38; p<0.001

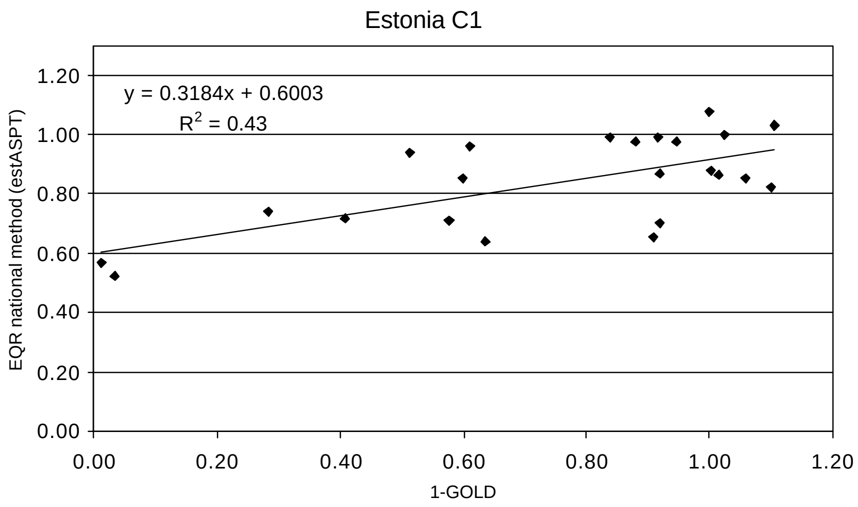


Figure 1D 1-GOLD - $R^2 = 0.43$; $p<0.001$

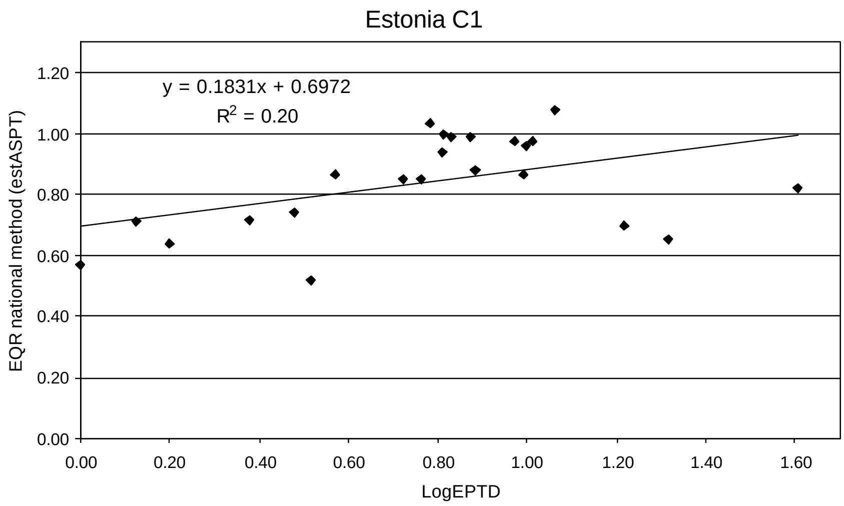


Figure 1E Log EPTD - $R^2 = 0.20$; $p<0.001$

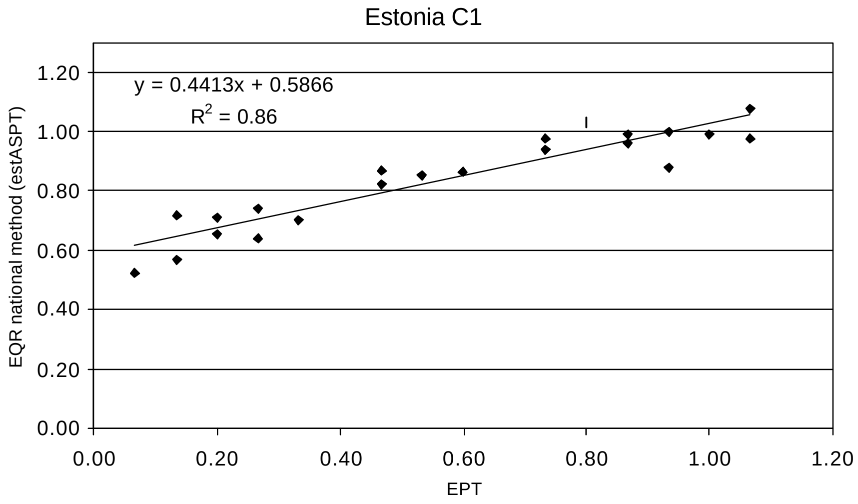


Figure 1F EPT - $R^2 = 0.86$; $p<0.001$

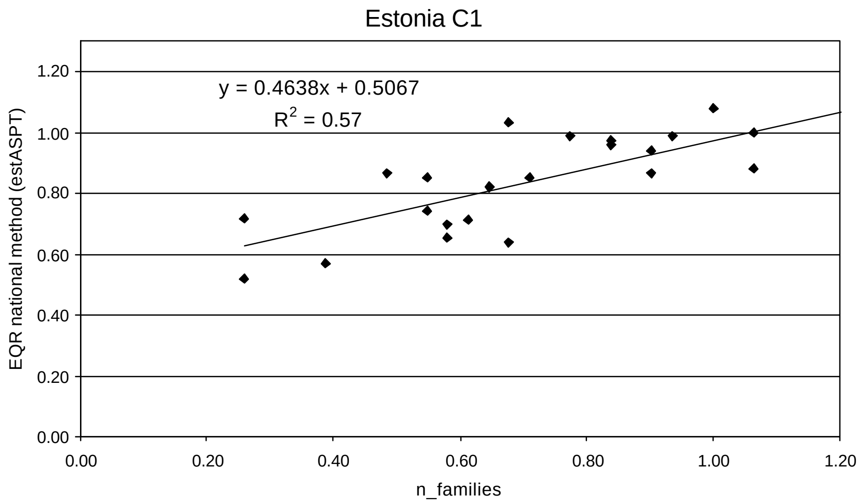


Figure 1G Number of families - $R^2 = 0.57$; $p<0.001$

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.

4.4.4 - 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, AFNOR, 1992). 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: high-good, 14; good-moderate, 12 (Wasson, pers. comm.).

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.

For each type, the classification is done following a WFD compliant procedure, according to the REFCOND Guidance recommendations, in relation to the type specific reference conditions. The Reference value (RV) for the normalization (EQR calculation) is the median of the IBGN values observed in reference sites. The H/G boundary is set at the 25^{th} percentile of the values observed in reference sites. The G/M boundary is first calculated separately for the two metrics (H/G boundary minus 1 for the GFI, and 1/4 of the range below H/G boundary for the number of taxa), and the combination of the two metrics gives the IBGN G/M boundary.

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^{th} percentiles of the reference samples. According to this procedure, the boundary High/Good for IBGN in C1 type changed from 17 to 14 and boundary Good/Moderate changed from 13 to 12.

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^{th} 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 75th 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 ICMs and IBGN are shown in Figures 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 75th 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. Original boundaries are provided by Wasson (pers. comm.).

In Table 2 is also reported the conversion of boundaries according to the original normalization, i.e. to the median value of the reference samples.

Table 1 IBGN class boundaries conversion for C1 dataset. Normalization 75th percentile high status samples

	IBGN score	IBGN EQR	ICMi EQR
Reference value	16	0.941	0.935
Limit high-good	14	0.824	0.822
Limit good-moderate	12	0.706	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 2 IBGN class boundaries conversion for C1 dataset. Normalization median reference samples

	IBGN score	IBGN EQR	ICMi EQR
Reference value	16	0.938	0.846
Limit high-good	14	0.813	0.743
Limit good-moderate	12	0.688	0.639
Limit moderate-poor	nd	nd	nd
Limit poor-bad	nd	nd	nd

ICM index = IBGN EQR * 0.8267 + 0.0709

R^2=0.80; $p<0.001$

Important notice

The reference values and class boundaries tested here are provisional, and may change due to ongoing work on reference sites selection and sampling.

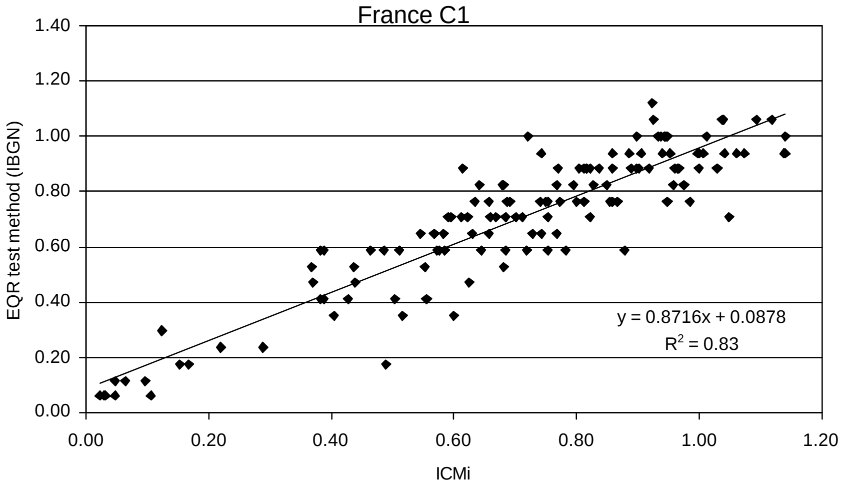


Figure 1A ICMi - R^2 = 0.83; p<0.001

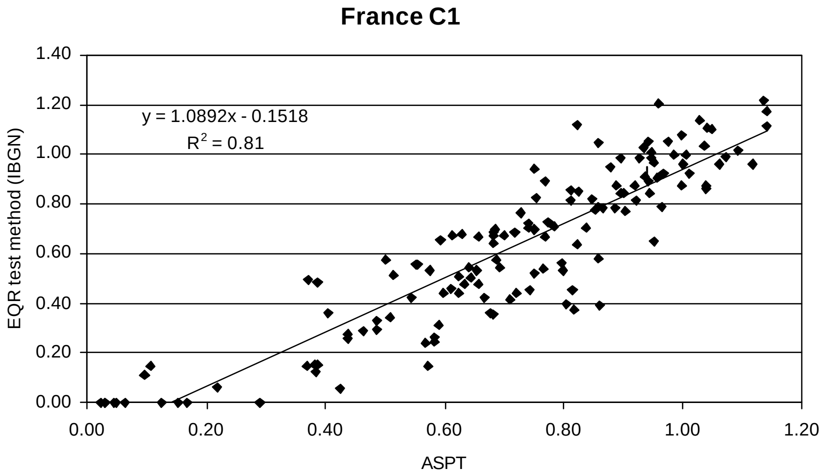


Figure 1B ASPT - R^2 = 0.81; p<0.001

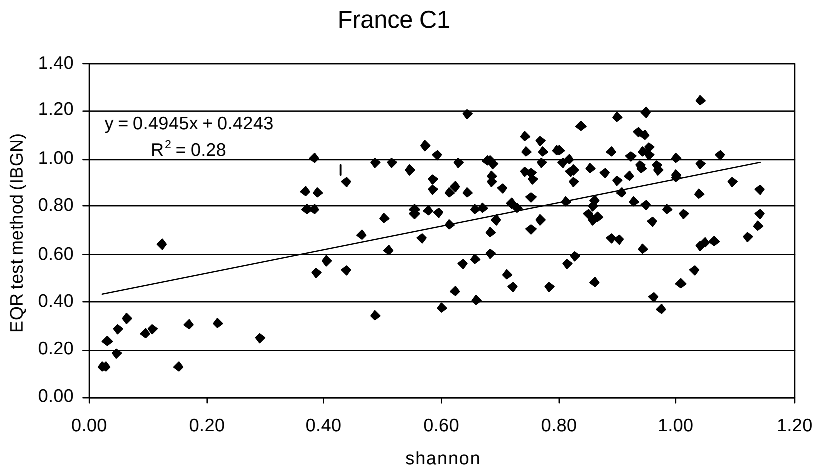


Figure 1C Shannon - R^2 = 0.28; p<0.001

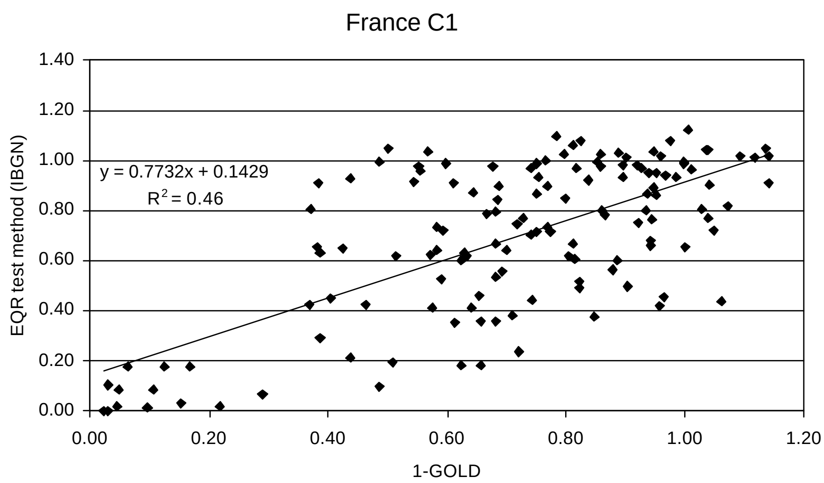


Figure 1D 1-GOLD - R^2 = 0.46; p<0.001

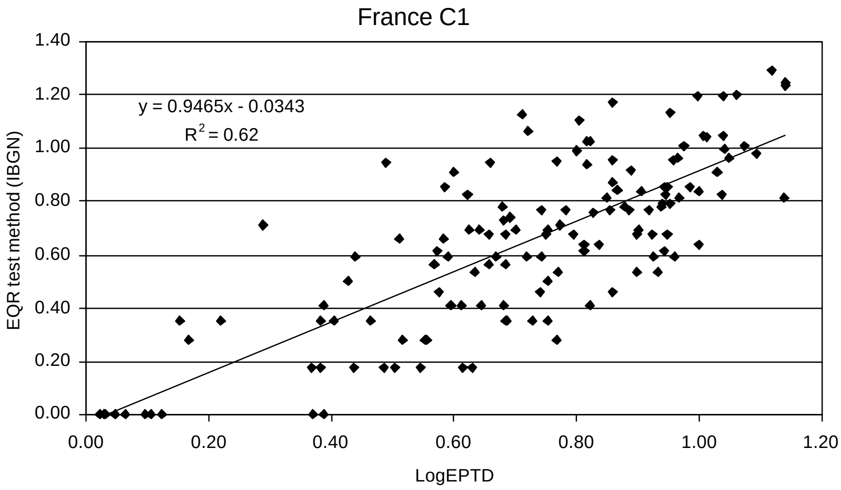


Figure 1E Log EPTD - R^2 = 0.62; p<0.001

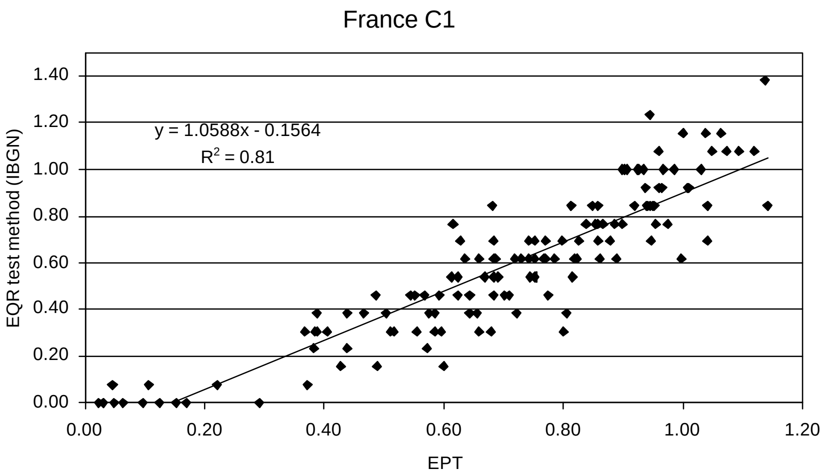


Figure 1F EPT - R^2 = 0.71; p<0.001

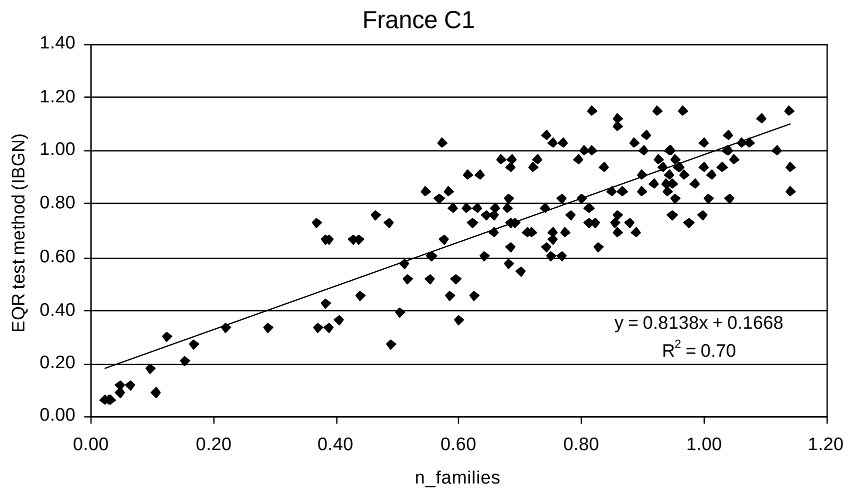


Figure 1G Number of families - R^2 = 0.70; p<0.001

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

AFNOR (Association Française de Normalisation), 1992. Détermination de la qualité biologique des eaux courantes : Indice Biologique Global Normalisé. Norme NF T 90-350.

4.4.5 - 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 75^{th} percentile of high status sites – for German R-C1 only one site has high status. And even if only the saprobic index is regarded, 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 is 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.

No Best Available Classification nor pressures based classification available.

The delineation of type-specific reference conditions in German river assessment using benthic invertebrates comprises two aspects:

(1) Compilation of reference taxa lists

- based on taxa lists derived from sampling existing reference sites (available for small and medium sized alpine, mountain and (partly) lowland streams)
- analysis of national database containing > 6000 samples through filtering for samples meeting the following criteria
    - hydromorphological quality according to Structure Index (Meier et al. 2004) at least "good"
    - low level of anthropogenic land-use in catchment area
    - type-specific Saprobic Index (Rolauffs et al. 2003) at least "good"

    → represents "best available" which includes reference conditions for alpine and mountain streams; lowland and large rivers deviate from reference state due to unavailability of existing sites in near-natural condition

(2) Definition of type-specific reference values for relevant assessment metrics

The assessment of ecological quality of running waters in Germany using macrozoobenthos is based on type-specific reference conditions expressed in reference values of relevant assessment metrics. These values have individually been derived by the following procedures:

type-specific "true" reference sites available

- calculation of reference values of relevant assessment metrics on the basis of a dataset including "true" type-specific reference sites. Reference value is the 95th percentile of all metric values.
- only sites slightly deviating from type-specific reference state available

- correlation of metric values against structure or land-use indices and extrapolation of best-fit-straight-line to reference values

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 75th 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 ICMs and SI(DE)mod are shown in Figures 1B to 1G.

The regression coefficient between ICMi and GD(DE) is 0.38 (see Figure 2a). For regression of the single ICMs and the GD(DE) see Figures 2B to 2G.

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. Original boundaries are provided by Birk (pers. comm.).

Table 1 SI(DE) class boundaries conversion

	SI(DE) score	SI(DE) EQR	ICMi EQR
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 2 GD(DE) class boundaries conversion

	GD(DE) score	GD(DE) EQR	ICMi EQR
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			

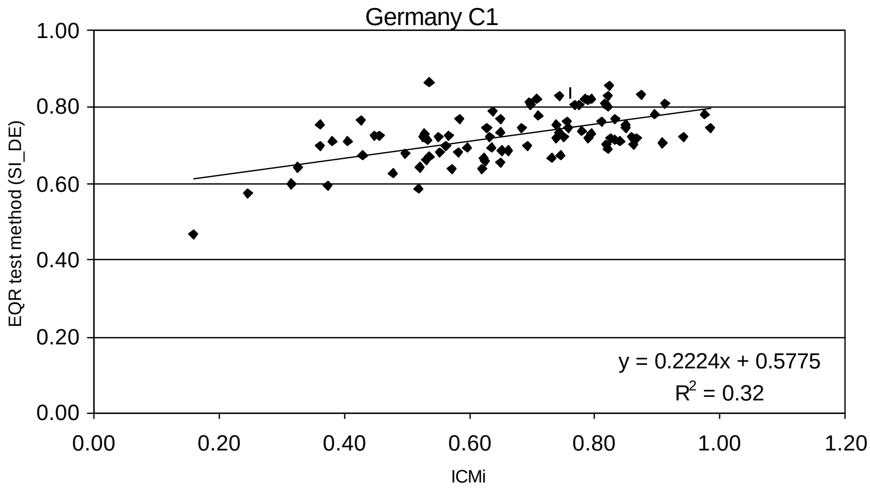


Figure 1A ICMi - R^2 = 0.32; p<0.001

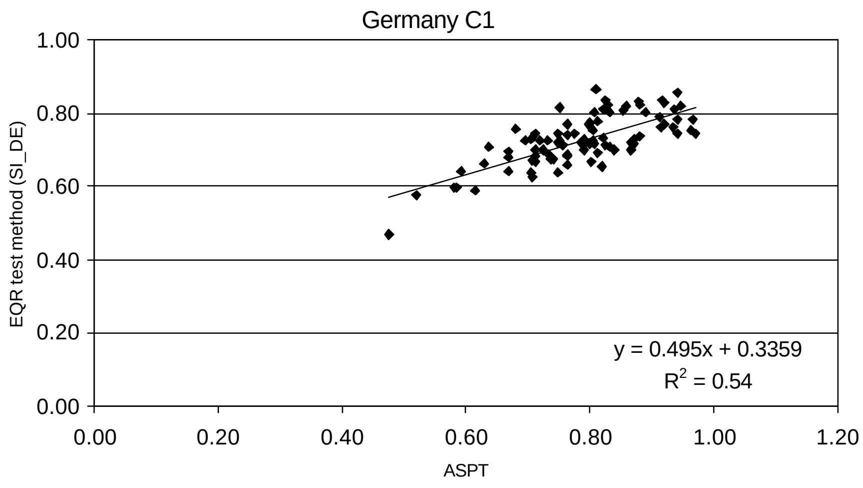


Figure 1B ASPT /SI(DE) - R^2 = 0.54; p<0.001

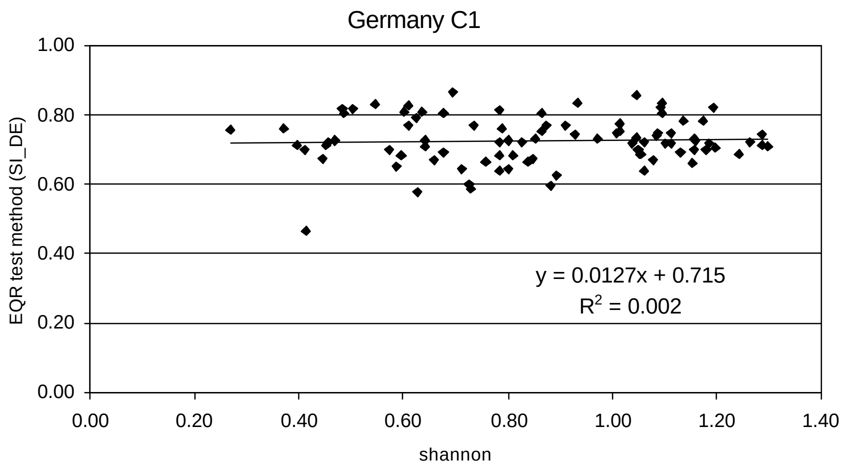


Figure 1C Shannon /SI(DE) - R^2 = 0.002; p=0.648

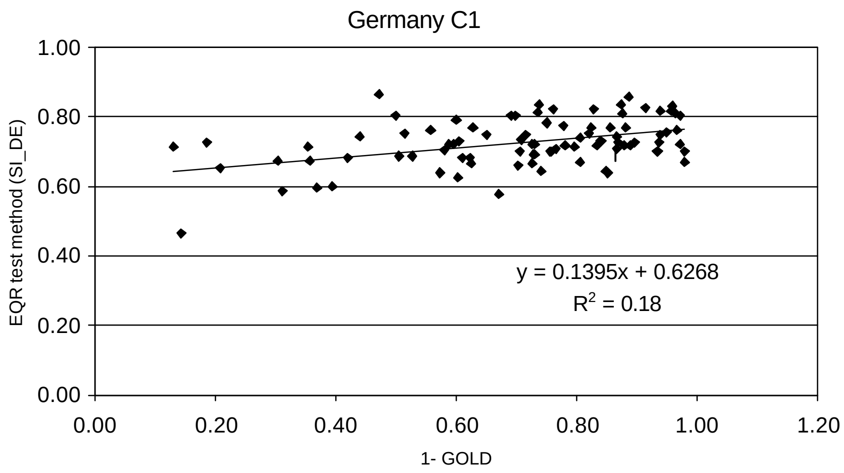


Figure 1D 1-GOLD /SI(DE) - R^2 = 0.18; p<0.001

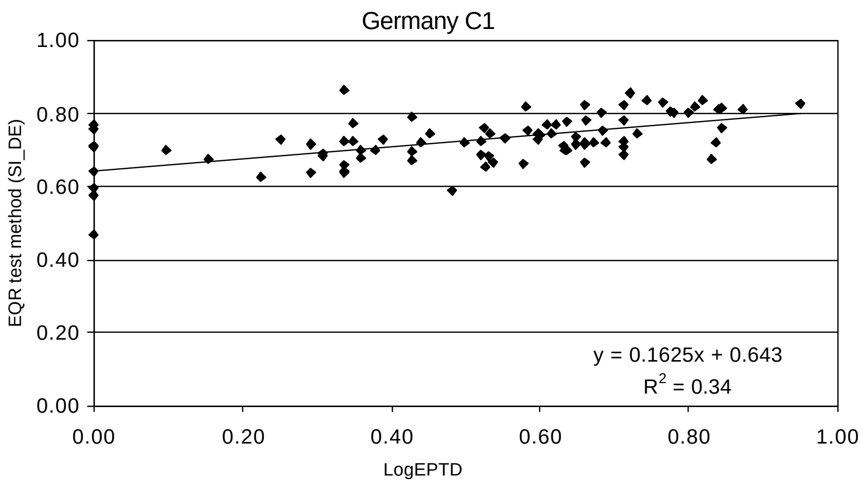


Figure 1E Log EPTD/SI(DE) - R^2 = 0.34; p<0.001

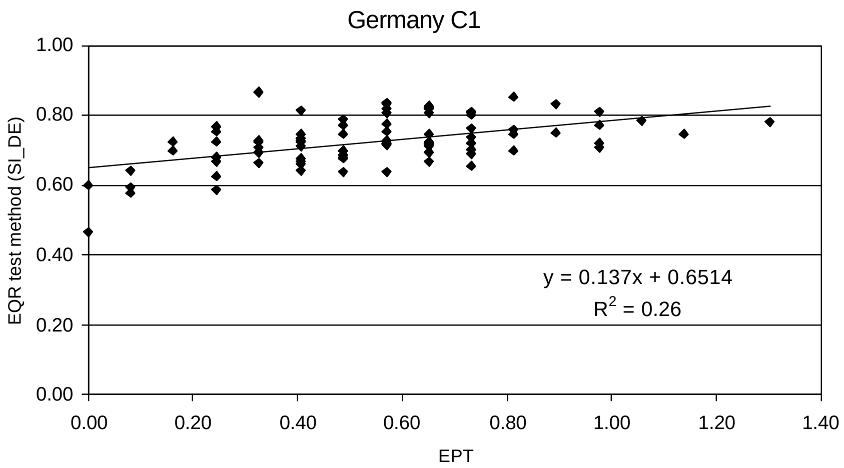


Figure 1F EPT /SI(DE) - R^2 = 0.26; p<0.001

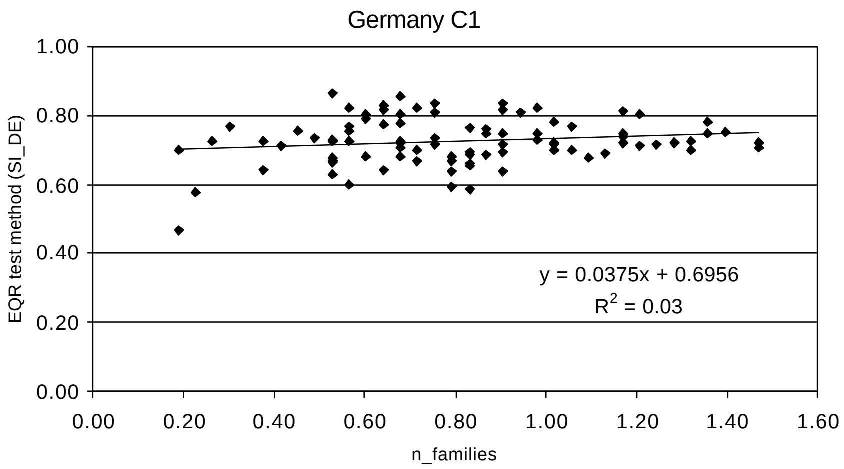


Figure 1G Number of families /SI(DE) - R^2 = 0.03; p=0.121

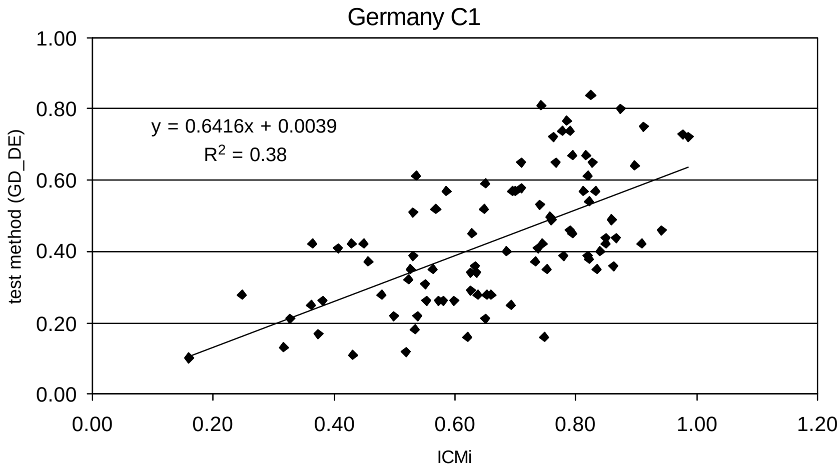


Figure 2A ICMi - R^2 = 0.32; p<0.001

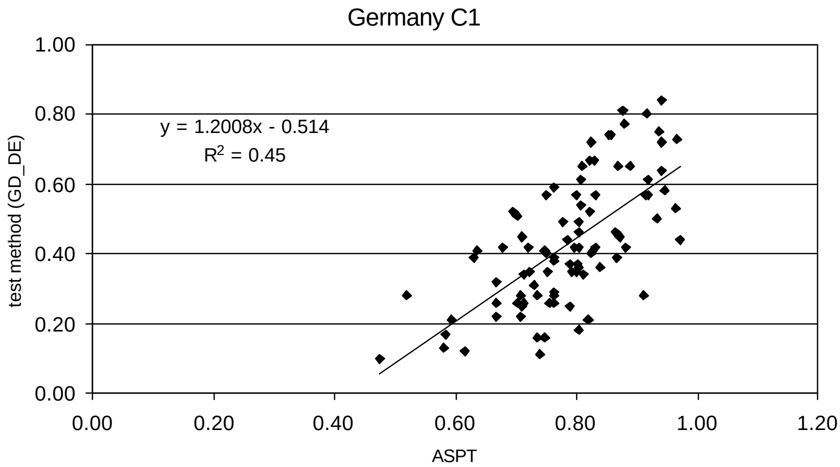


Figure 2B ASPT /GD(DE) - R^2 = 0.45; p<0.001

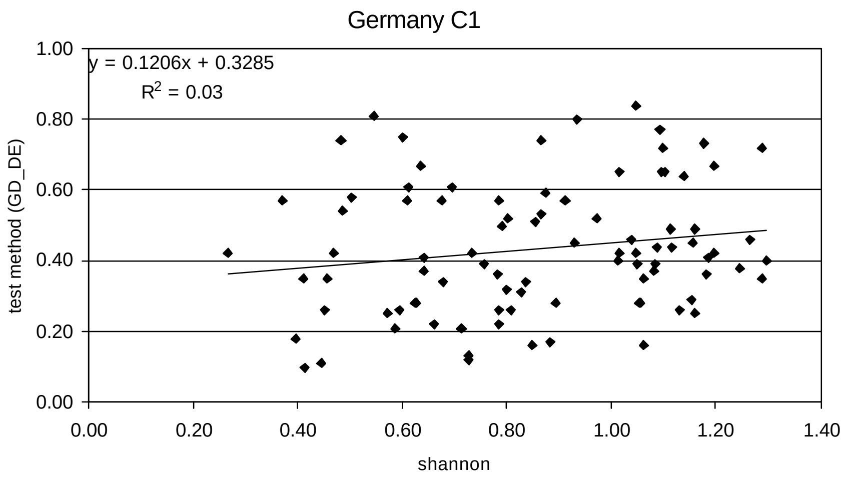


Figure 2C Shannon /GD(DE) - R^2 = 0.03; p=0.10

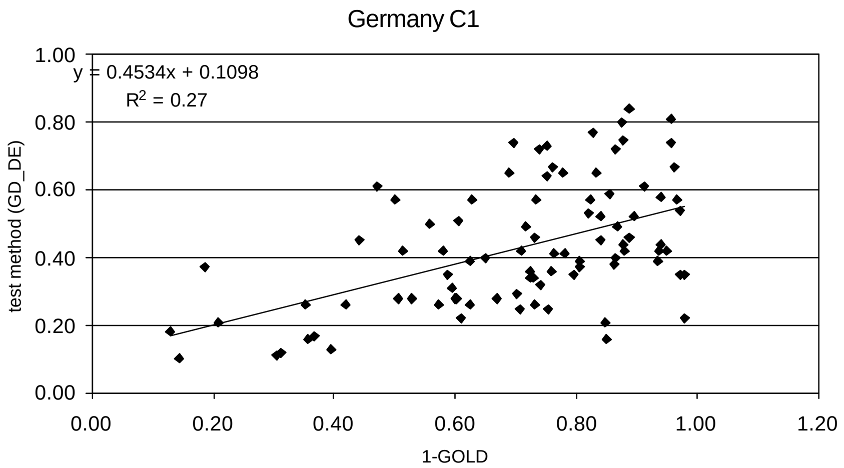


Figure 2D 1-GOLD/GD(DE) - R^2 = 0.27; p<0.001

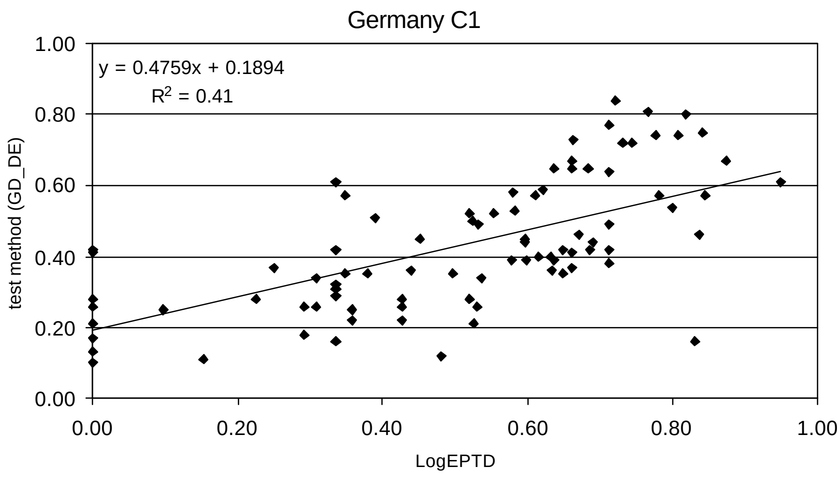


Figure 2E Log EPTD/GD(DE) - R^2 = 0.41; p<0.001

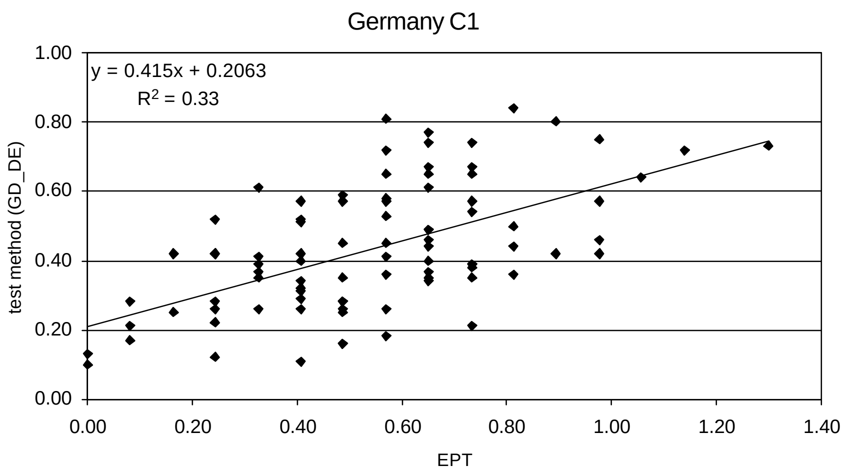


Figure 2F EPT /GD(DE) - R^2 = 0.33; p<0.001

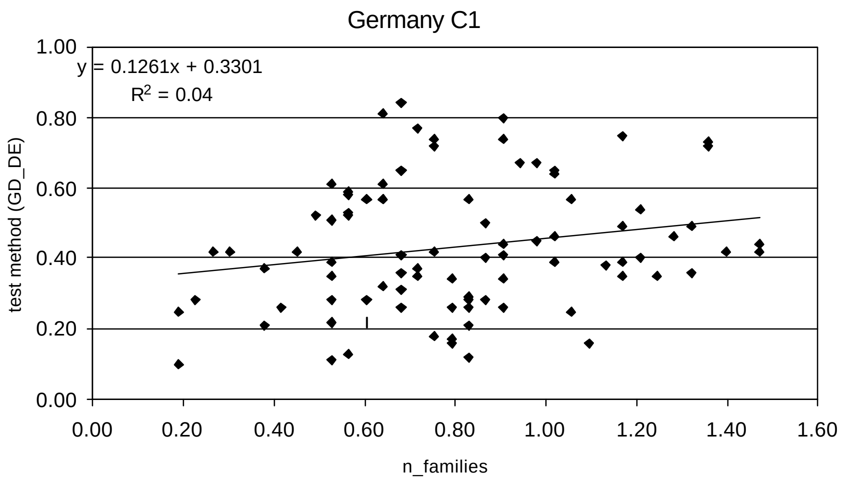


Figure 2G Number of families/GD(DE) - R^2 = 0.04; p=0.05

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:

- For German lowland rivers, organic pollution is not the dominating stressor degrading river quality but hydromorphological pressure and land use.
- Since the German assessment method is based on species level data, this can account for variability in the family-based results of the ICMi.
- 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 quality appraisal of the German multimetric assessment system is based on the principle of 'one out, all out' among different modules that consider different alteration factors. This can determine a lower quality 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).

Friedrich, 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).

Meier, C., D. Hering, R. Biss, J. Böhmer, C. Rawer-Jost, A. Zenker, P. Haase & F. Schöll, 2004. Weiterentwicklung und Anpassung des nationalen

Bewertungssystems für Makrozoobenthos an neue internationale Vorgaben. Berlin, Umweltbundesamt.

Rolauffs, P., D. Hering, M. Sommerhäuser, S. Rödiger & S. Jähnig, 2003. Entwicklung eines leitbildorientierten Saprobienindexes für die biologische Fließgewässerbewertung. Berlin, Umweltbundesamt.

4.4.6 - Italy C1

General features

This dataset contains samples from typical Northern Italian spring fed streams in the lowland of Po valley 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 $20km^2$).

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, 2004). 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. 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 23% (84 on 361) of the samples are classified as 'high status' according to national assessment method.

No Best Available Classification nor pressures based classification available.

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

For the calculation of the ICMi, metrics was normalized according to 75th 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 Figures 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 75th 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. Original boundaries are provided according to APAT/IRSA-CNR, 2004 and Spaggiari & Franceschini, 2000 for the conversion in values.

Table 1 IBE class boundaries conversion

	IBE score	IBE EQR	ICMi EQR
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

$R^2=0.72$; $p<0.001$

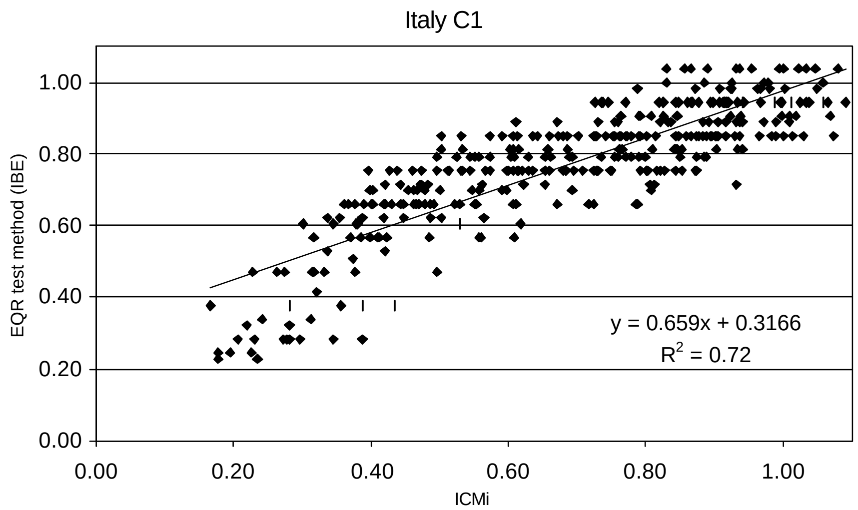


Figure 1 A ICMi - R^2 = 0.72; $p<0.001$

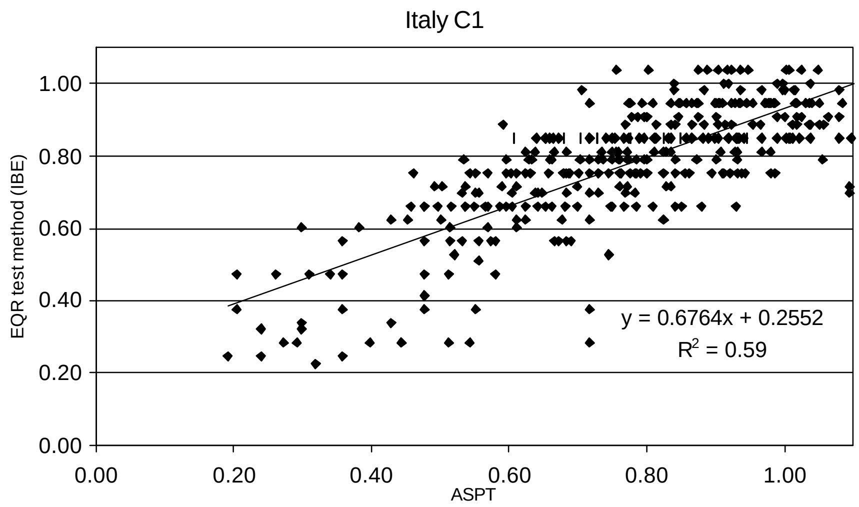


Figure 1B ASPT - R^2 = 0.59; p<0.001

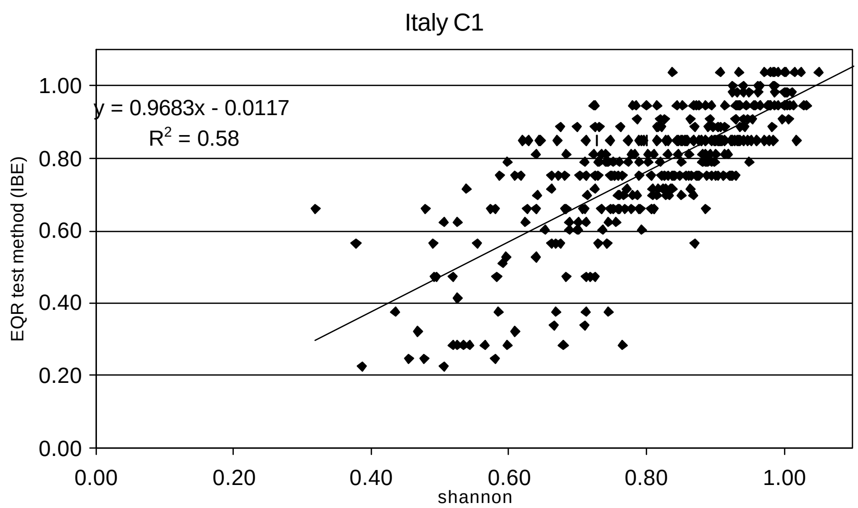


Figure 1C Shannon - R^2 = 0.58; p<0.001

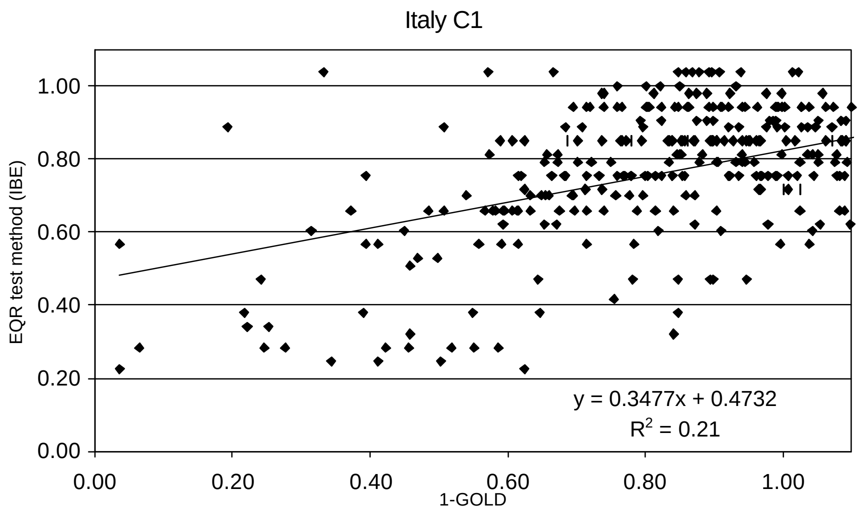


Figure 1D 1-GOLD - R^2 = 0.21; p<0.001

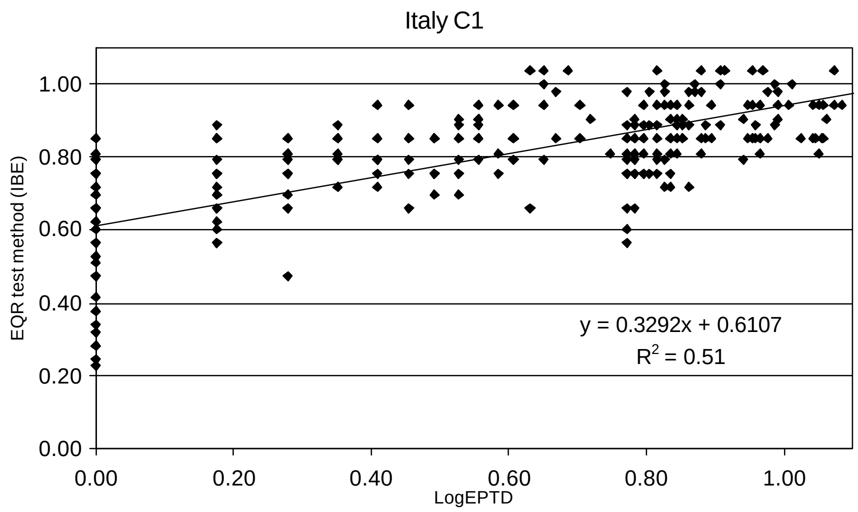


Figure 1E Log EPTD - R^2 = 0.51; p<0.001

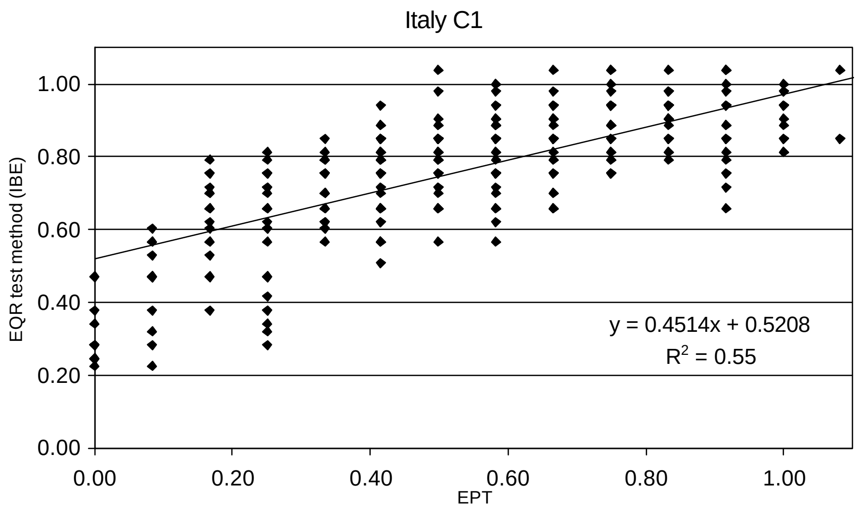


Figure 1F EPT - R^2 = 0.55; p<0.001

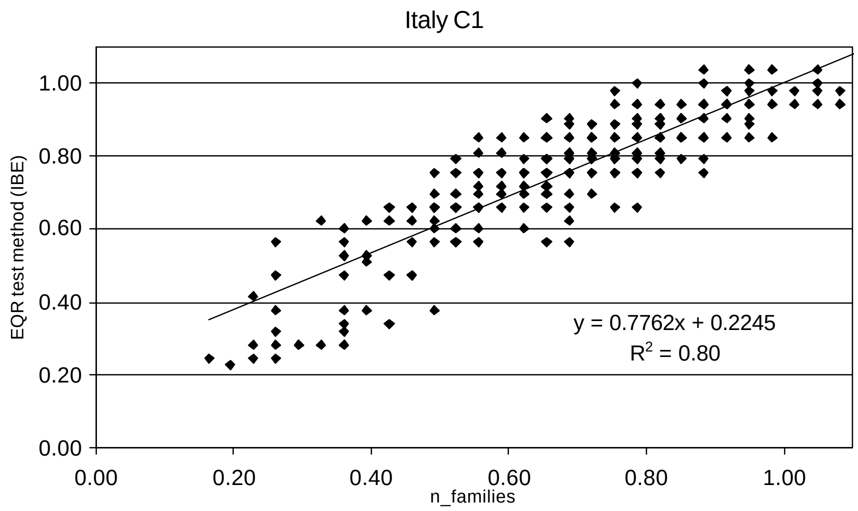


Figure 1G Number of families - R^2 = 0.80; p<0.001

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.

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.

Spaggiari, R. & S. Franceschini, 2000. Procedure di calcolo dello stato ecologico dei corsi d'acqua e di rappresentazione grafica delle informazioni. Biol. Amb. 14 (2): 1-6.

4.4.7 - The Netherlands C1

General features

Sites are located especially in pleistocene sand areas, up to 200m above sea level (Knoben & van den Berg, 2004).

Aim of collection, number of samples

For the central rivers intercalibration pilot an extended selection of RC-1 sites was selected from the national aquatic ecological database Limnodata Neerlandica. Data were provided by dr. Roel Knoben and Marcel van den Berg from RIZA, Royal Haskoning.

In each quality class 12-21 sites were selected, considering geographical coverage of the country (especially pleistocene sand areas, up to 200m above sea level) and the entire gradient of deterioration. Most recent samples were taken from sampling campaigns in spring. Samples were collected by 15 different regional waterboards, thus some deviation in methods may be expected to be present (Knoben & van den Berg, 2004).

Degradation factor

The Dutch metric is not directed to a specific pressure but indicate the distance from reference conditions by the loss of type specific species and negative dominating species. Main pressure is a combination of regulation, eutrophication and some saprobity. Many if not all sites will be designated as heavily modified because of discharge regulation. Real reference sites are considered not existing in the country (Knoben & van den Berg, 2004).

Sampling and sorting, criteria for abundance registration

Sampling is in most cases performed with standard macroinvertebrate handnet (w*h 30 * 20 cm; mesh size 0.5 mm) over 5 m meter (1.5 m^2) covering all microhabitats present. Live individuals are either picked from a white tray during 1-1.5 hour in the field or in the laboratory. All individuals are counted and recorded (including estimates in case of species in very high numbers s). Abundances are true numbers. Identification level is species level, with some exceptions for Oligochaeta (family) and Hydracarina.

Sites' classification

Taxonomic level of determination is in principle at species. Oligochaetes sometimes at family level only.

At this moment a new assessment system is available, unofficially known as the KRW-maatlatten (the Dutch metric). This system operates at species level. Final assessment result is discrete with steps of 0.1 ranging from 0 to 1, which is classified in 5 quality classes (Knoben & van den Berg, 2004).

Notes on classification

National: About 15% (12 on 79) of the samples are classified as 'high status' according to KRW method.

No Best Available Classification nor pressures based classification available.

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

Originally the normalization has been undertaken according to the maximum value of samples in high status class (no true reference sites). To keep comparability with the other datasets, ICMi and metrics was recalculated normalizing according to 75th percentile observed in the 'high status' samples. The minimum and maximum observed values for ICMi (in EQR) have been 0.17 and 0.15. Between ICMi and KRW, a regression coefficient of 0.18 was found (see Figure 1A).

The conversion of the class boundary values for the KRW method from the original boundaries to ICMi values is done according to Table 1. Original boundaries are provided by Knoben (pers. comm.)

In Table 2 is also reported the conversion of boundaries according to the original normalization, i.e. to the maximum value of the high status samples.

Table 1 KRW class boundaries conversion for C1 dataset. Normalization 75th percentile high status samples

	KRW score	KRW EQR	ICMi EQR
Limit high-good	1	1.000	1.034
Limit good-moderate	0.8	0.800	0.935
Limit moderate-poor	0.6	0.600	0.836
Limit poor-bad	0.4	0.400	0.737

ICM index = KRW EQR * 0.4944 + 0.5392

R^2=0.18; $p<0.001$

Table 2 KRW class boundaries conversion for C1 dataset. Normalization maximum high status samples

	KRW score	KRW EQR	ICMi EQR
Limit high-good	1	1.000	0.738
Limit good-moderate	0.8	0.800	0.674
Limit moderate-poor	0.6	0.600	0.611
Limit poor-bad	0.4	0.400	0.547

ICM index = KRW EQR * 0.318 + 0.4199

R^2=0.20; p<0.001

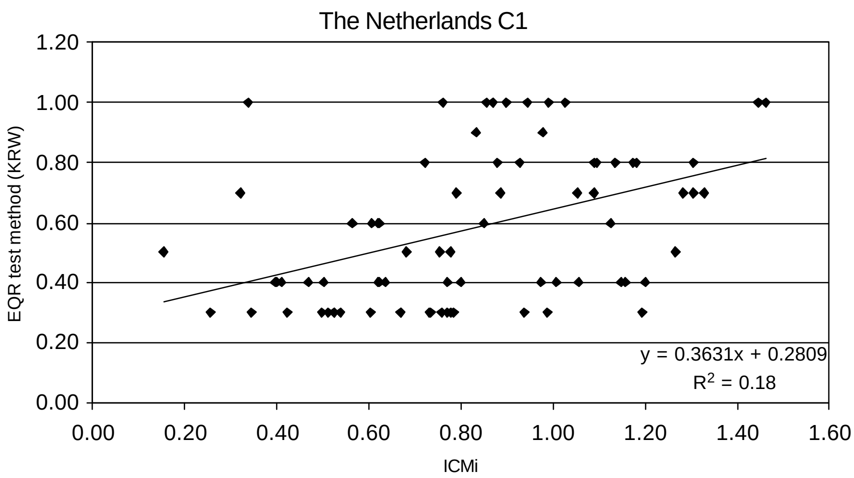


Figure 1 A ICMi - R^2 = 0.18; p<0.001

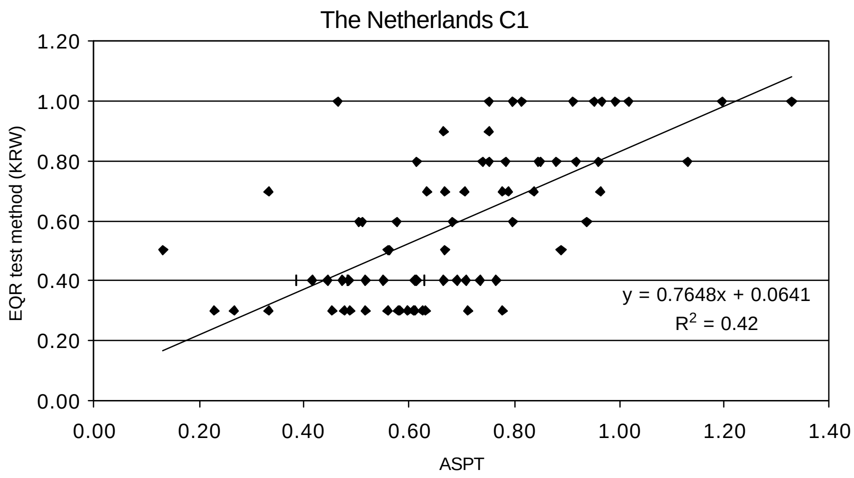


Figure 1B ASPT - $R^2 = 0.42$; $p<0.001$

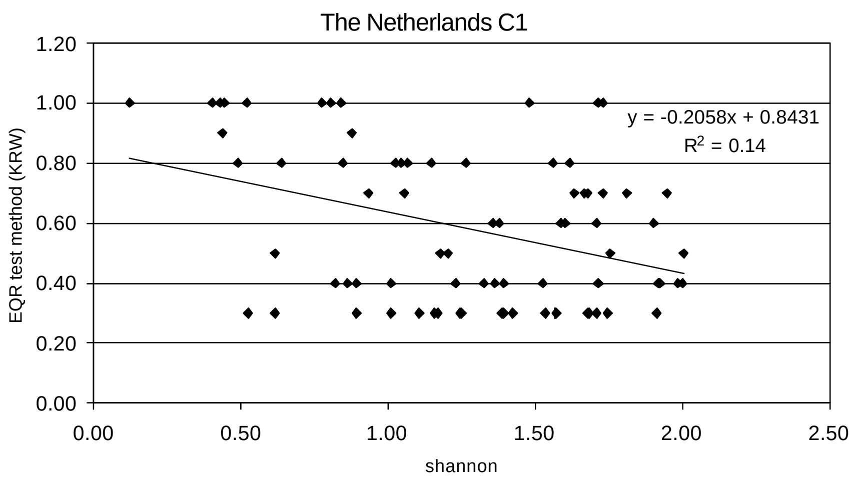


Figure 1C Shannon - $R^2 = 0.14$; $p<0.001$

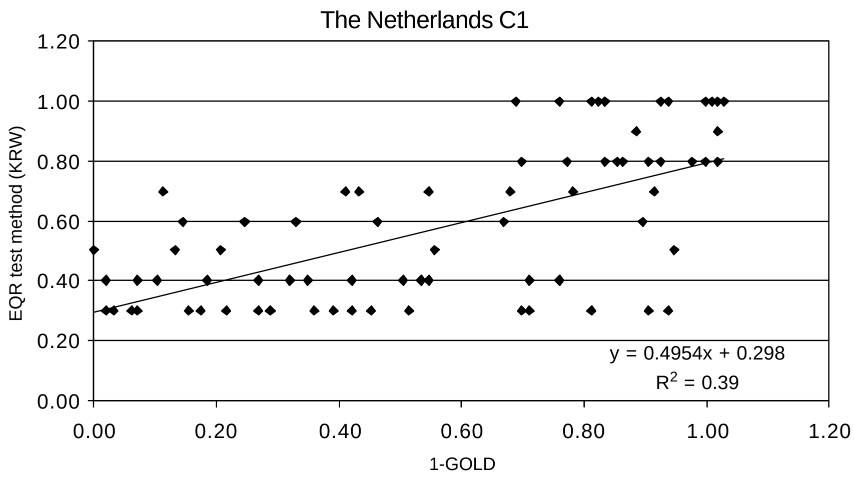


Figure 1D 1-GOLD - R^2 = 0.39; p<0.001

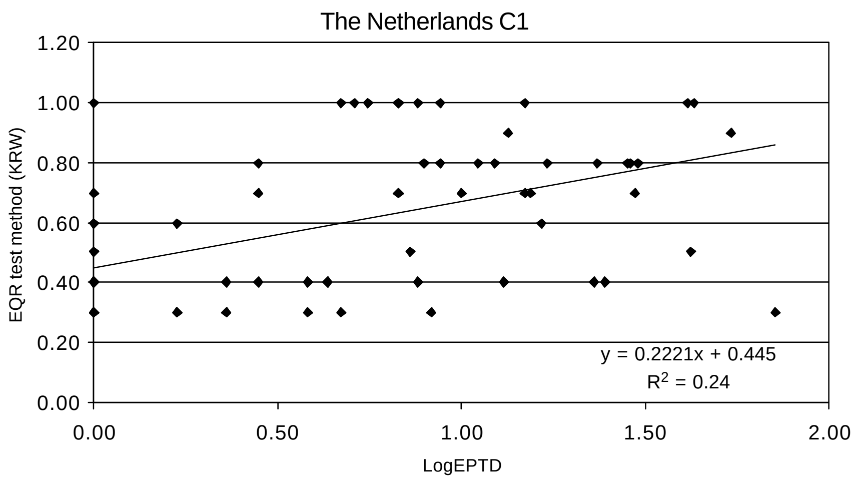


Figure 1E Log EPTD - R^2 = 0.24; p<0.001

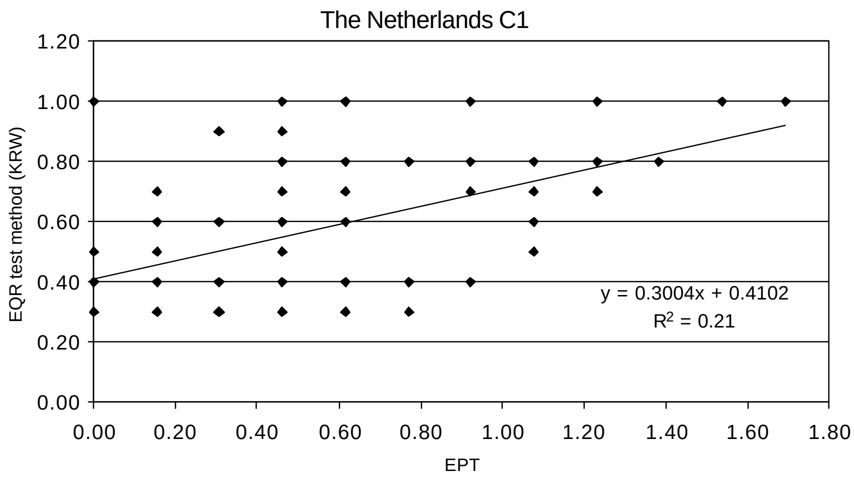


Figure 1F EPT - R^2 = 0.21; p<0.001

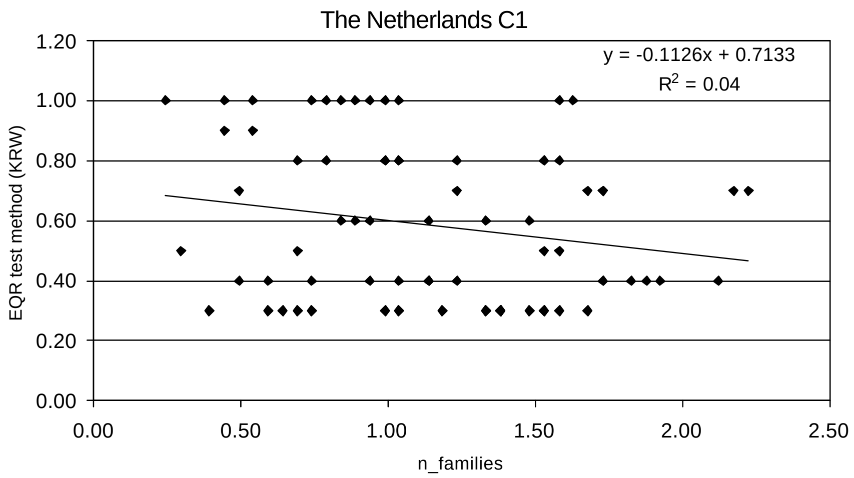


Figure 1G Number of families - R^2 = 0.04; p<0.001

General remarks, comments

From Knoben & van den Berg (2004).

The separate metrics from ICM often have a poor or no relation with the national measure and classification. The single ICMs that gives best correlation is the ASPT gave the best relation with national measure, but still there is considerable chance of false classification due to the large variation.

Most indices in the ICM seem not suitable for making a proper assessment of Dutch lowland rivers. Therefore we are doubtful about the chances of combining the indices in the ICM as bench market for the status of macroinvertebrates in the Netherlands.

Possible hypothesis to be considered when low correlations ICMi vs National method arise:

- are the samples collected with the same sampling method?
- are the same reference conditions present (e.g. data from the same area/ stream types)?
- does the test method properly describe the quality gradient?
- identification level: could some problems arise if the test method undertake identification (e.g.) to species level? Is it possible that only few macroinvertebrates' orders identified to species level are present?
- does the dataset represents the full quality gradient?

References related to the presented dataset

Knoben, R. & M. van den Berg, 2004. Report on Intercalibration pilot exercise Central GIG Rivers from The Netherlands. Pilot exercise report. 7 pp. November 2004.

4.4.8 - 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 was 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 organic pollution/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 dominant 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 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/log N
S = number of families
N = total abundance

Values are from 0 to an open end. Boundaries between classes are: High-Good 5.5; Good-Moderate 4; Moderate-Poor 2.5; Poor-Bad 1. Maximum observed value is 11.75, the minimum is 0.74.

Notes on classification

No Best Available Classification nor pressures based classification available.

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

For the calculation of the ICMi, metrics were normalized according to 75^{th} percentile of high status samples provided and classified using national method (see explanations in previous chapters). Final ICMi is re-normalized according to 75^{th} 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^{th} percentile.

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

The relationship of ICMi was reported both for BMWP-PL and Margalef (see Figures 1Aa and 1Ab). ICMi regression coefficient found is: ICMi vs BMWP-PL: 0.74; ICMi vs Margalef: 0.40.

For all the single ICMs the linear regression is reported only in relation to the BMWP-PL (Figures 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. Original boundaries are provided by Soszka (pers. comm.).

Table 1 Polish BMWP class boundaries conversion

	BMWP score	BMWP EQR	ICMi EQR
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

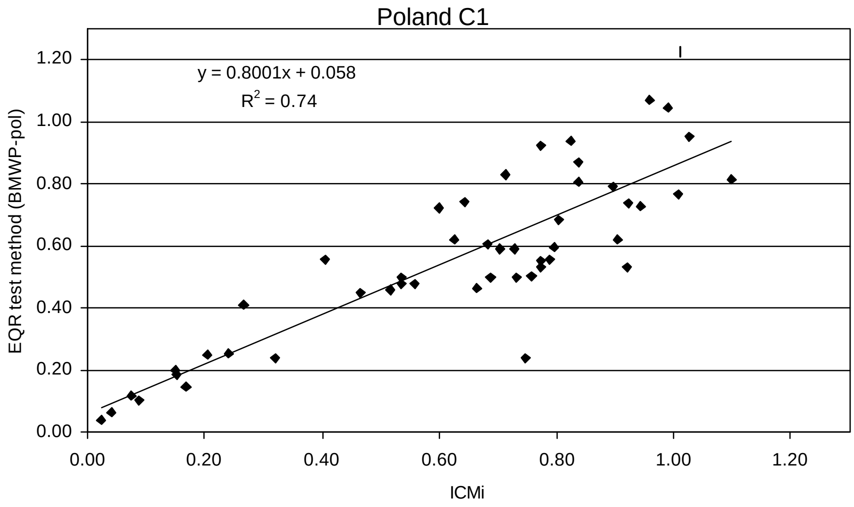


Figure 1Aa ICMi vs BMWP-PL - R^2 = 0.74; p<0.001

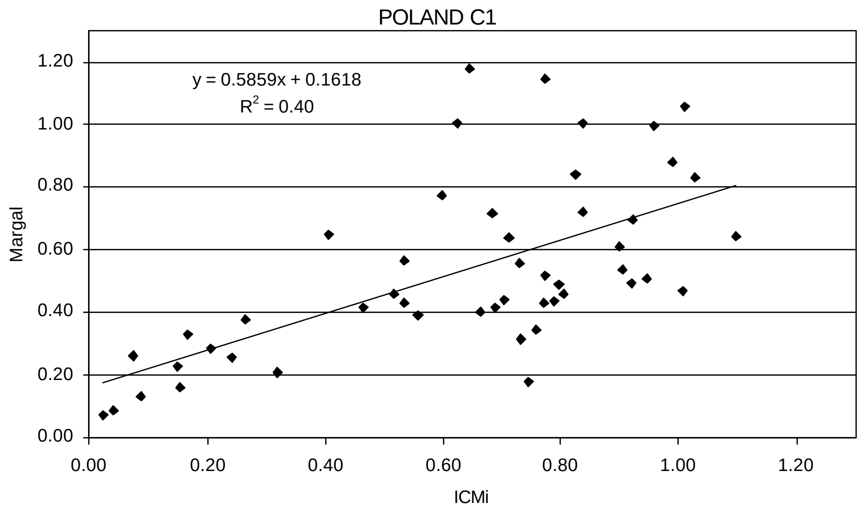


Figure 1Ab ICMi vs Margalef - R^2 = 0.40; p<0.001

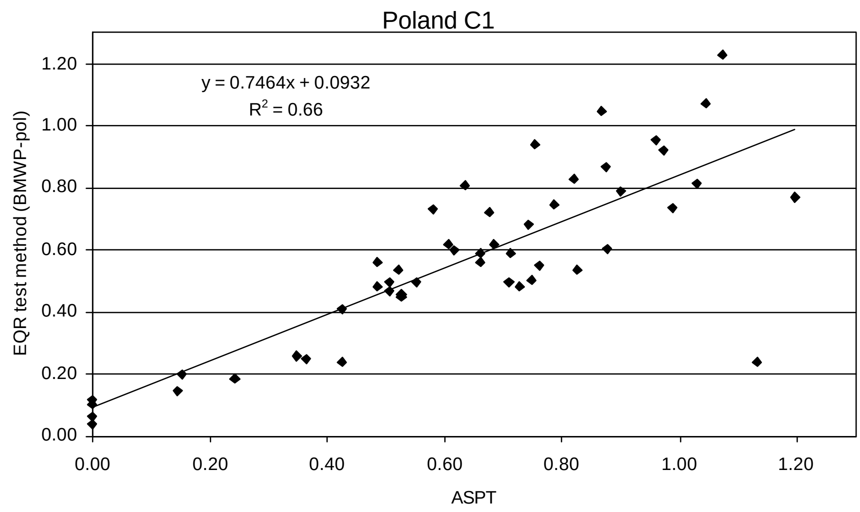


Figure 1B ASPT vs BMWP-PL - R^2 = 0.66; p<0.001

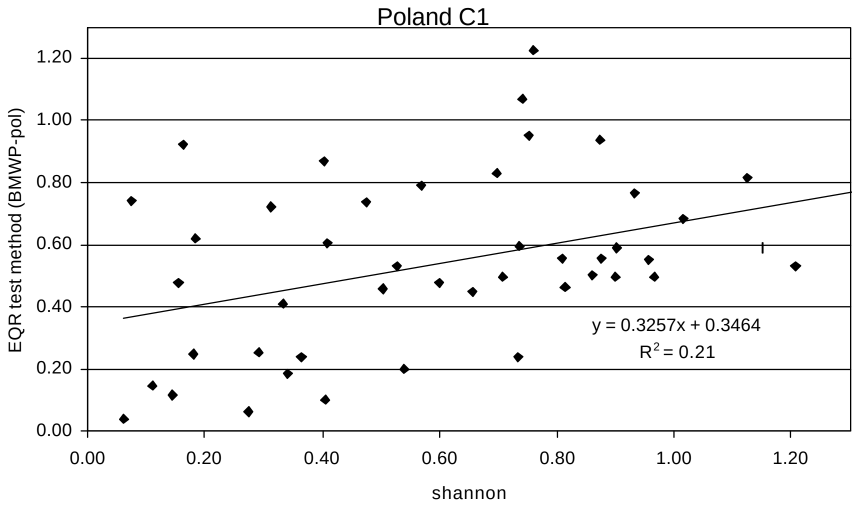


Figure 1C Shannon - R^2 = 0.21; p<0.001

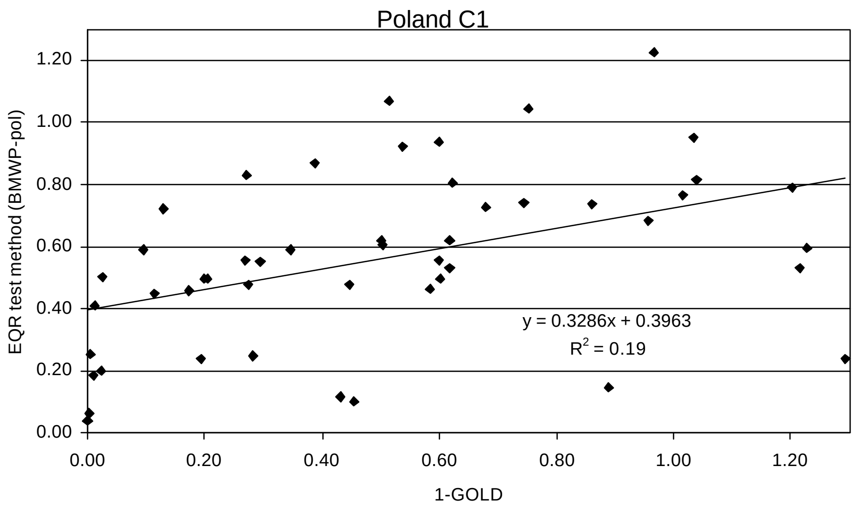


Figure 1D 1-GOLD - R^2 = 0.19; p<0.001

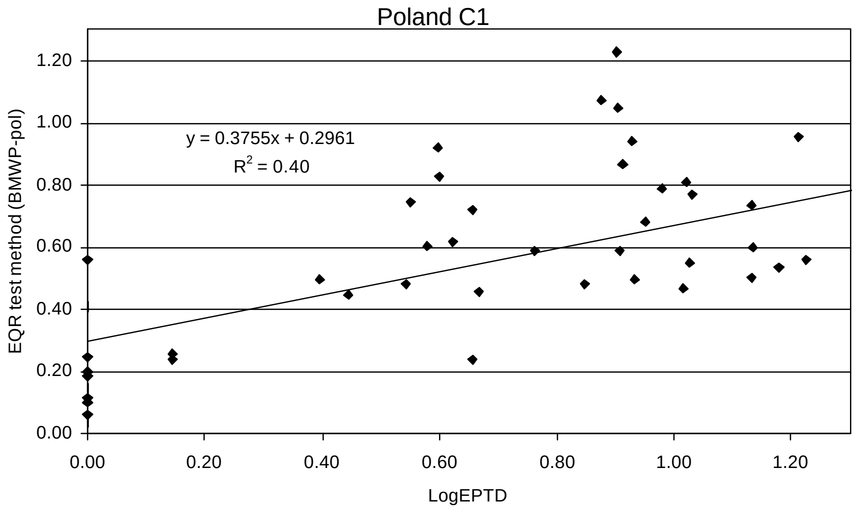


Figure 1E Log EPTD - $R^2 = 0.40$; $p<0.001$

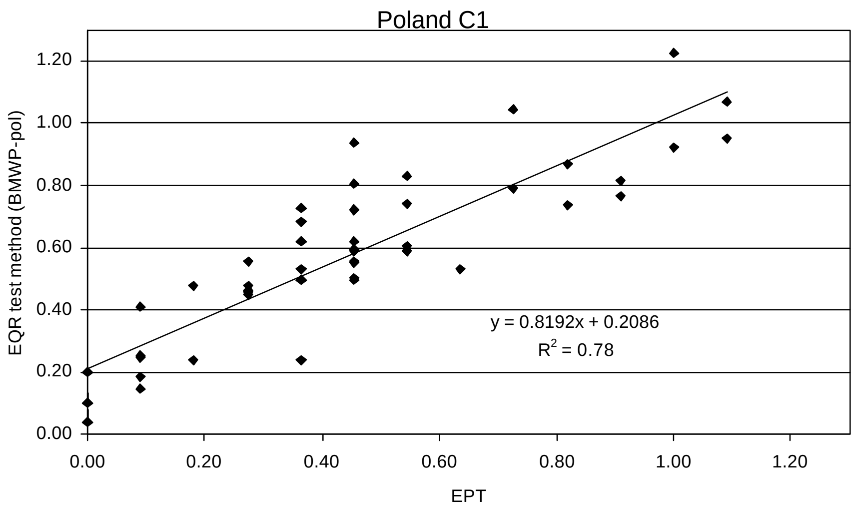


Figure 1F EPT - $R^2 = 0.78$; $p<0.001$

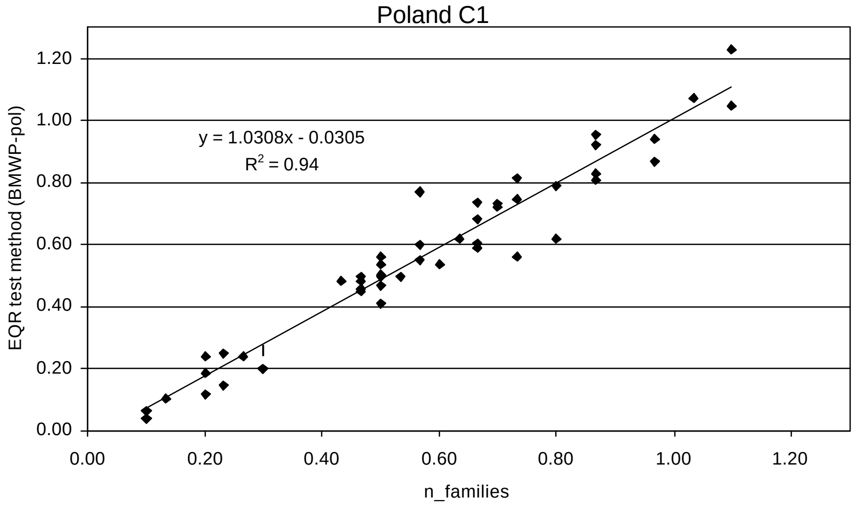


Figure 1G Number of families - R^2 = 0.94; p<0.001

Notes on dataset description

The content of the present description was verified by Dr. Hanna 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).

4.4.9 - UK C1

General features

This dataset comprises sites with small catchment area (less than 100km2), mean catchment altitude low (<200m), with medium or high alkalinity (0.4 - >2 meq/l). These sites have a predominantly siliceous geology. They correspond to sites on the UK's WFD River Type 1 (Rivers Task Team, 2004). They are predominantly from in the South West of England (particularly Cornwall), Kent/Sussex, around the New Forest in Hampshire, the Lake District, West Wales and Anglesey. The dataset includes very few sites from Scotland, where such sites they are found mainly in the Western Isles, Aberdeenshire, the coastal fringes of the far Northwest, the South West and parts of the West. Across UK, this type of stream covers approximately 11% of river length.

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 of data the values of EQI ASPT vary between 0.4 and 1.16; EQI N-Taxa ranges from 0.11 to 1.54. 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.

No Best Available Classification, nor pressures based classification were used for this dataset.

Comparison between the ICMi and National method, single ICM and National method

For the calculation of the ICMi, metrics was normalized according to 75th percentiles of High status samples (see explanations in previous chapters). Final ICMi is re-normalized according to 75th percentile. In the present set the values of the ICMi vary from 0.1 to 1.1.

Between ICMi and EQI ASPT a regression coefficient of 0.82 was found (see Figure 1A).

Results on linear regression between single ICM ASPT-EQI are shown in Figures 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 Figures 2 B-G.

The values of the metric ASPT contain a small mistake of calculation in the normalization. The influence of this is anyhow minimal on the goodness of the regression.

The conversion of the class boundary values for the EQI ASPT method and NFAM-EQI from the original boundaries to ICMi values is done according to Table 1 and Table 2. Original boundaries are reported in GQA, EA, 1997.

Table 1 EQI ASPT class boundaries conversion

	EQI ASPT score	EQI ASPT EQR	ICMi EQR
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 2 NFAM-EQI class boundaries conversion

	NFAM-EQI score	NFAM-EQI EQR	ICMi EQR
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			

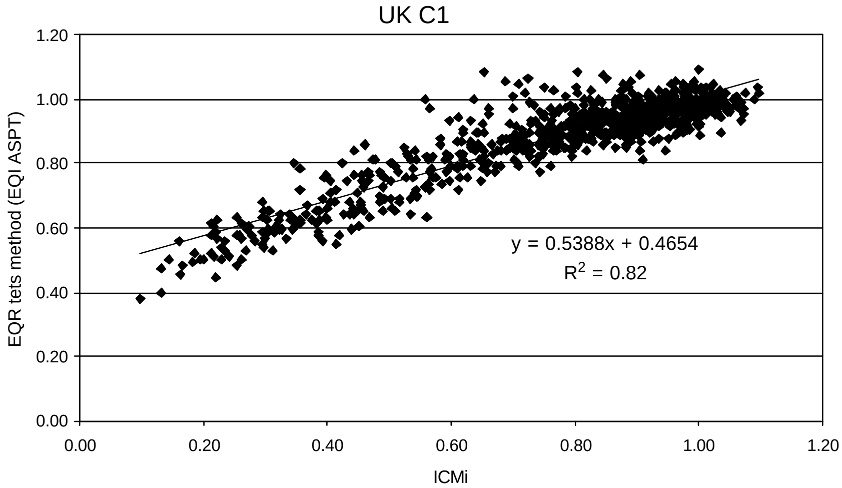


Figure 1A ICMi / EQI ASPT - R^2 = 0.82; p<0.001

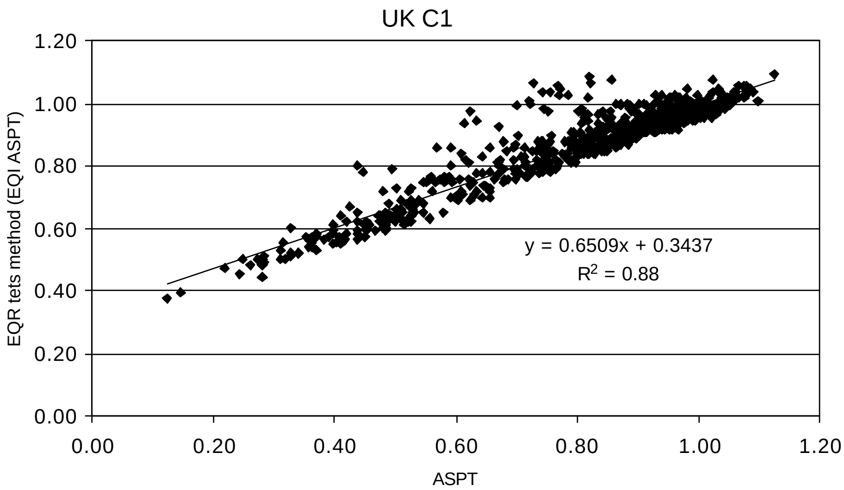


Figure 1B ASPT / EQI ASPT - R^2 = 0.88; p<0.001

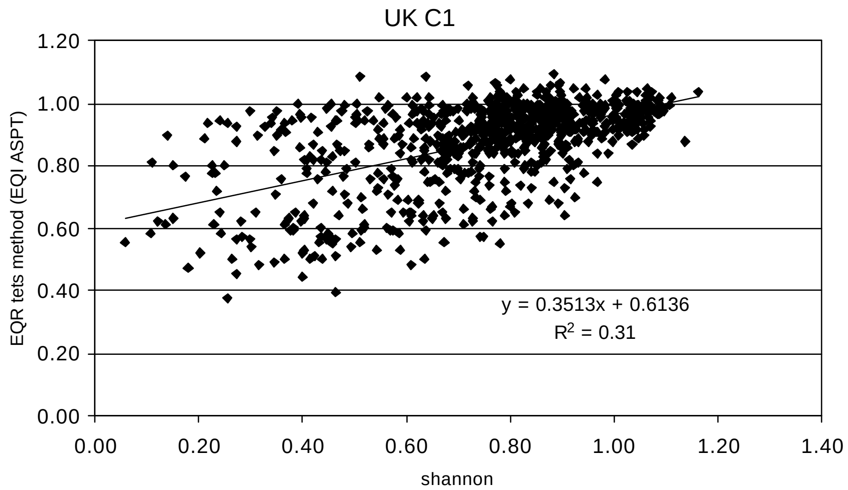


Figure 1C Shannon / EQI ASPT - R^2 = 0.31; p<0.648

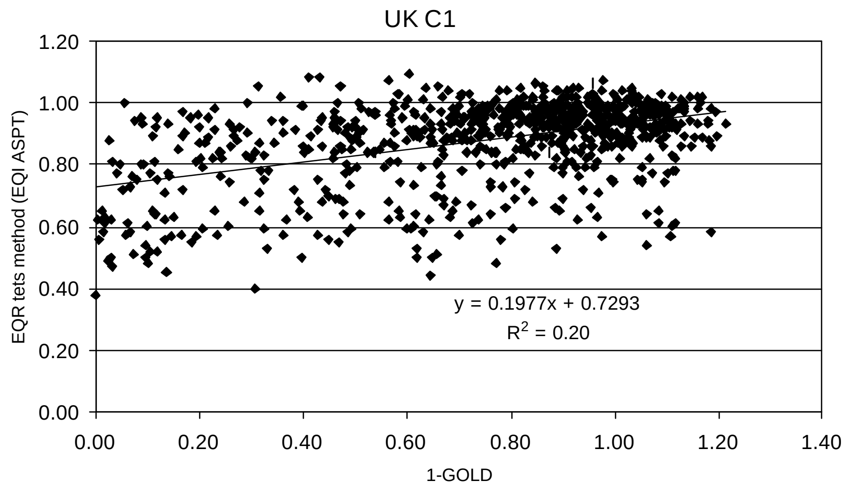


Figure 1D 1-GOLD/ EQI ASPT - R^2 = 0.20; p<0.001

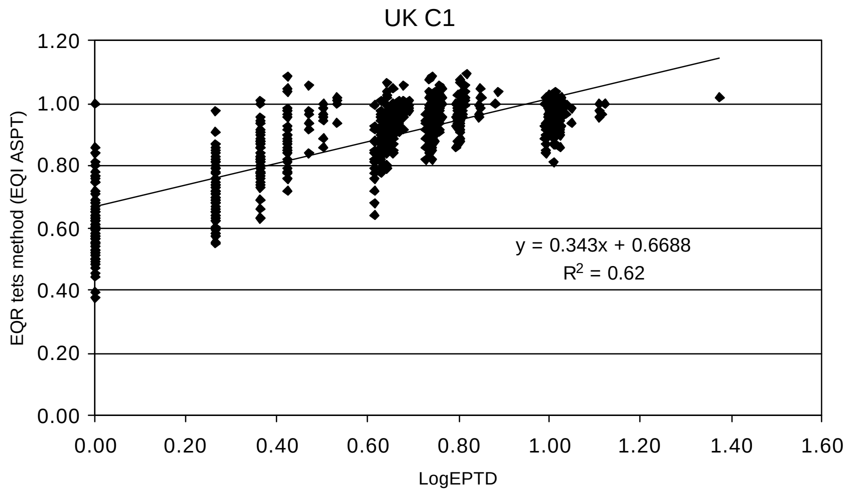


Figure 1E Log EPTD/ EQI ASPT - R^2 = 0.62; p<0.001

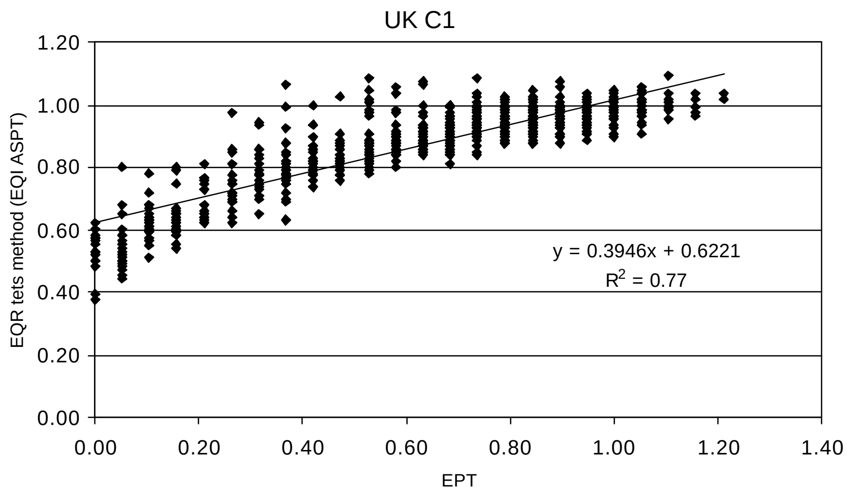


Figure 1F EPT / EQI ASPT - R^2 = 0.77; p<0.001

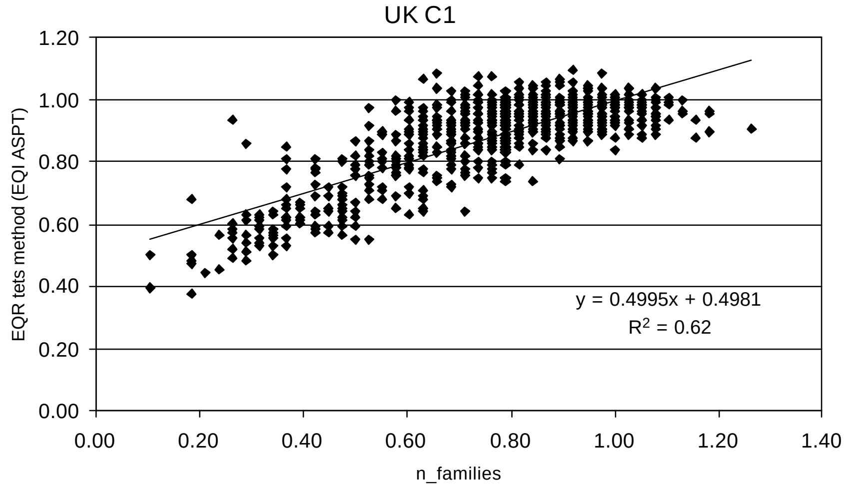


Figure 1G Number of families/ EQI ASPT - R^2 = 0.62; p<0.121

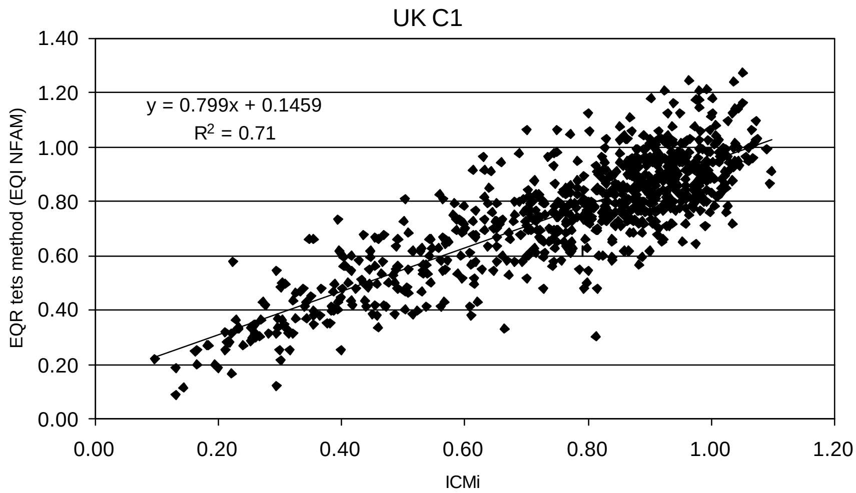


Figure 2A ICMi / NFAM-EQI - R^2 = 0.71; p<0.001

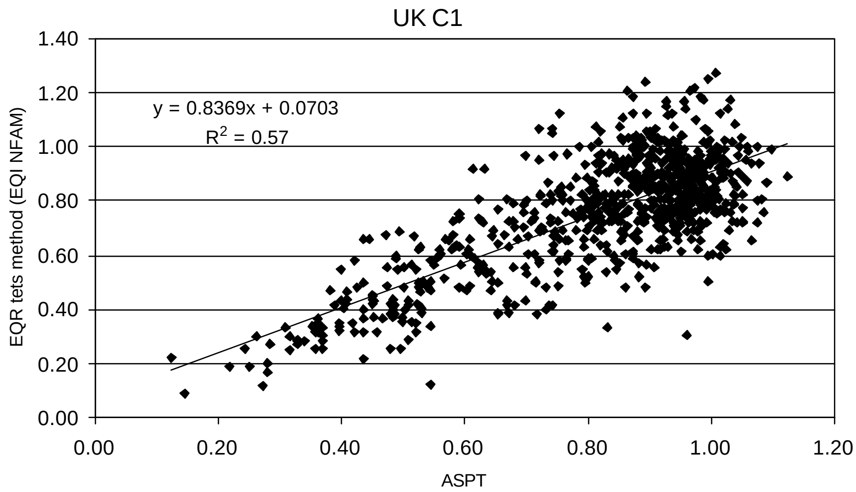


Figure 2B ASPT / NFAM-EQI - R^2 = 0.57; p<0.001

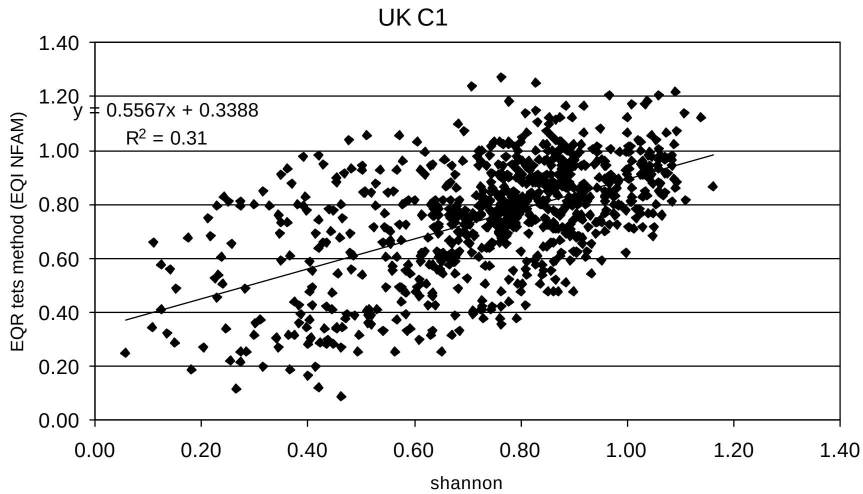


Figure 2C Shannon / NFAM-EQI - R^2 = 0.31; p<0.10

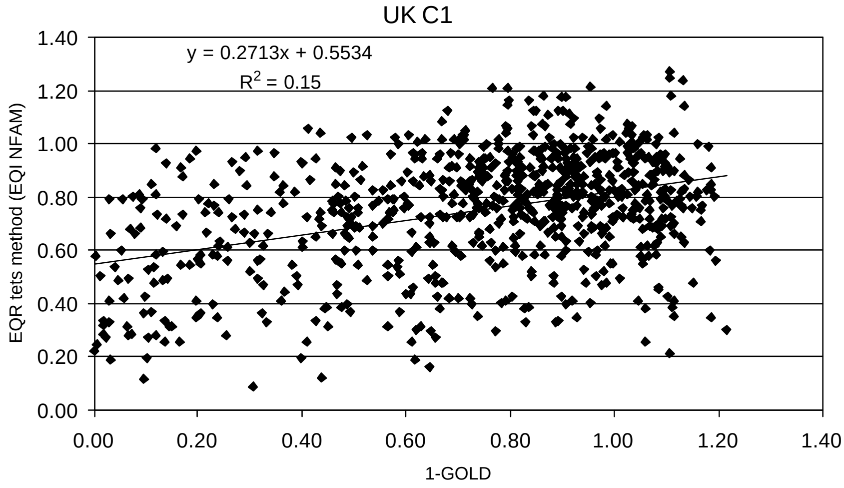


Figure 2D 1-GOLD/ NFAM-EQI - $R^2 = 0.15$; $p<0.001$

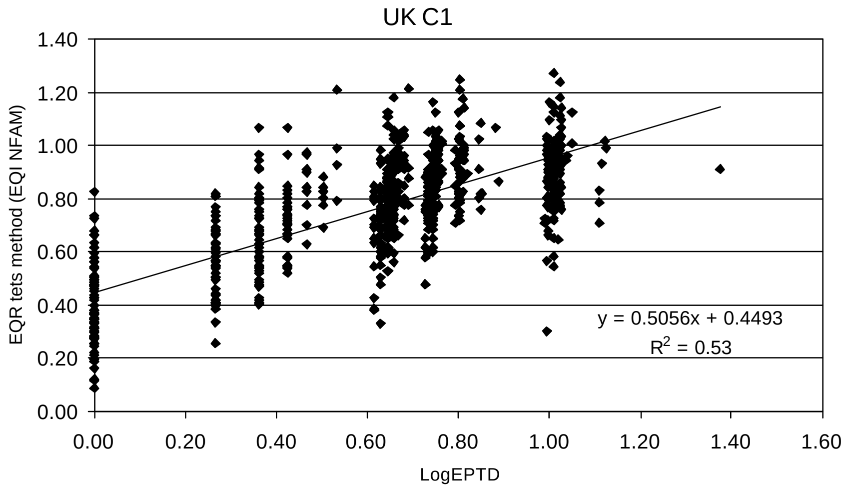


Figure 2E Log EPTD/ NFAM-EQI - $R^2 = 0.53$; $p<0.001$

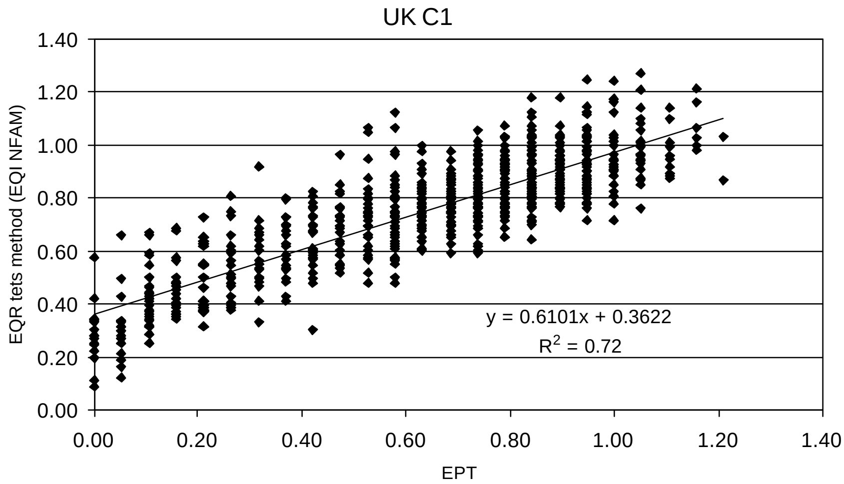


Figure 2F EPT / NFAM-EQI - R^2 = 0.72; p<0.001

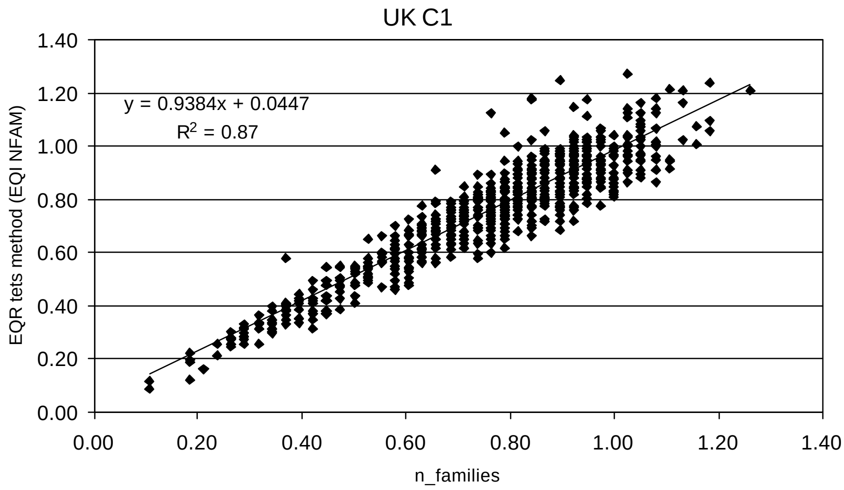


Figure 2G Number of families/ NFAM-EQI - R^2 = 0.87; p<0.001

N-families v EQI-NFAM shows the variation between site-specific predictions of reference condition provided by RIVPACS vs type-specific reference condition provided by 75%-ile. In this case, there is a small amount of variation caused by the type of calculation used. In fact, NFAM (y axis) is related only to the taxa included in BMWP-score, whereas N-families (x axis), calculated by AQEMrap software, includes all families. Small errors of this kind may be also present in other datasets.

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.

Rivers Task Team, 2004. Type Specific Reference Condition Descriptions for Rivers for Great Britain (v1. PR1 29.06.04) Report to UK Technical Advisory Group on the Water Framework Directive (TAG Work Programme 8a (02) Reference conditions for Rivers TAG2004 WP8a (02)).

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.

4.5 - IC type C2 (small lowland streams dominated by rocky substrates)

4.5.1 - 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, AFNOR, 1992). 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.

For each type, the classification is done following a WFD compliant procedure, according to the REFCOND Guidance recommendations, in relation to the type specific reference conditions. The Reference value (RV) for the normalization (EQR calculation) is the median of the IBGN values observed in reference sites. The H/G boundary is set at the 25th percentile of the values observed in reference sites. The G/M boundary is first calculated separately for the two metrics (H/G boundary minus 1 for the GFI, and 1/4 of the range below H/G boundary for the number of taxa), and the combination of the two metrics gives the IBGN G/M boundary.

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^{th} percentiles of the reference samples. According to this procedure, the boundary High/Good for IBGN in C1 type changed from 17 to 14 and boundary Good/Moderate changed from 13 to 12.

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

For the calculation of the ICMi, metrics were normalized according to the 75^{th} 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^{th} 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 Figures 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^{th} 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. Original boundaries are provided by Wasson (pers. comm.).

In Table 2 is also reported the conversion of boundaries according to the original normalization, i.e. to the median value of the reference samples.

Table 1 IBGN class boundaries conversion for C2 dataset. Normalization: 75th percentile High status samples

	IBGN score	IBGN EQR	ICMi EQR
Reference value	16	0.941	0.898
Limit high-good	14	0.824	0.785
Limit good-moderate	12	0.706	0.672
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 2 IBGN class boundaries conversion for C2 dataset. Normalization: median Reference samples

	IBGN score	IBGN EQR	ICMi EQR
Reference value	16	1.000	1.018
Limit high-good	14	0.867	0.855
Limit good-moderate	12	0.733	0.692
Limit moderate-poor	nd	nd	nd
Limit poor-bad	nd	nd	nd

ICM index = IBGN EQR * 1.2221 - 0.2045

R^2=0.80; p<0.001

Important notice

The reference values and class boundaries tested here are provisional, and may change due to ongoing work on reference sites selection and sampling.

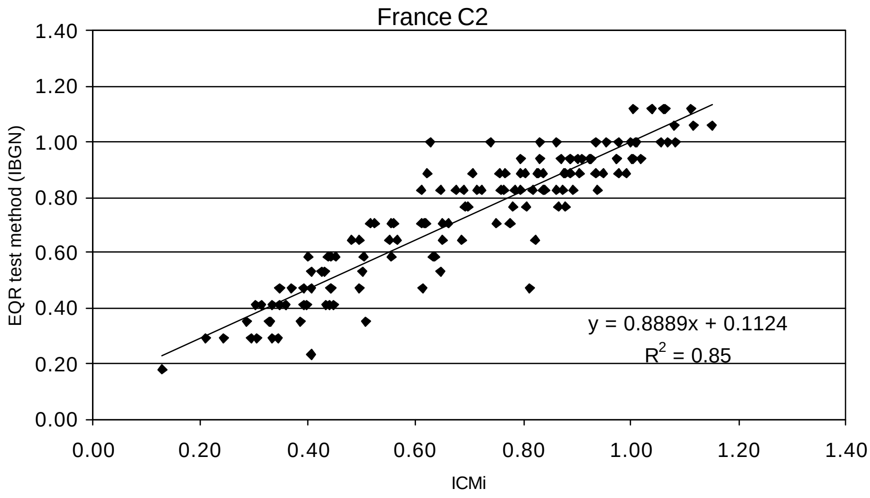


Figure 1A ICMi - R^2 = 0.85; p<0.001

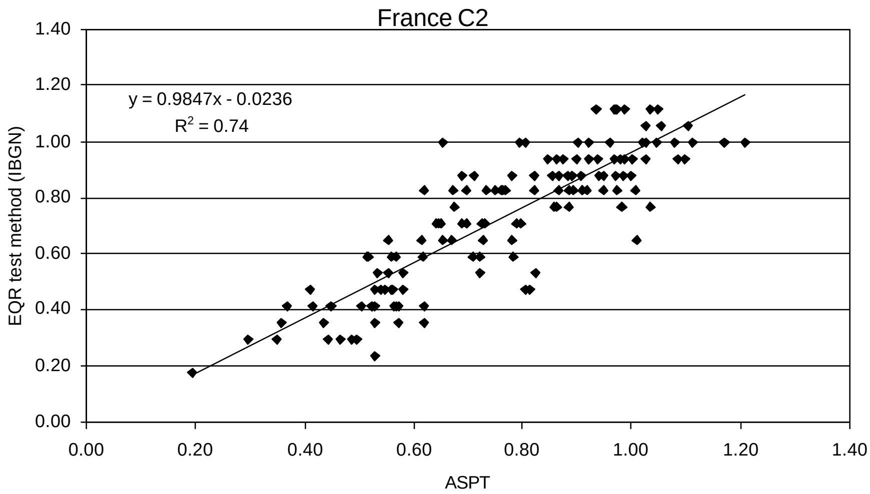


Figure 1B ASPT - R^2 = 0.74; p<0.001

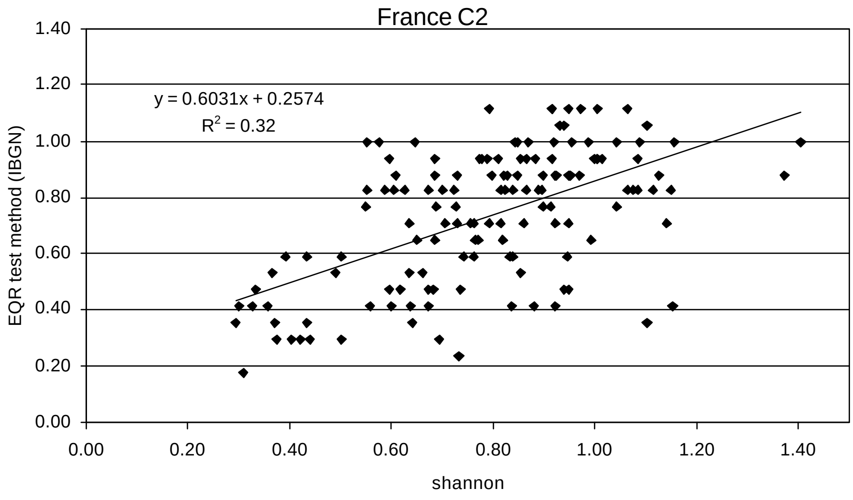


Figure 1C Shannon - R^2 = 0.32; p<0.001

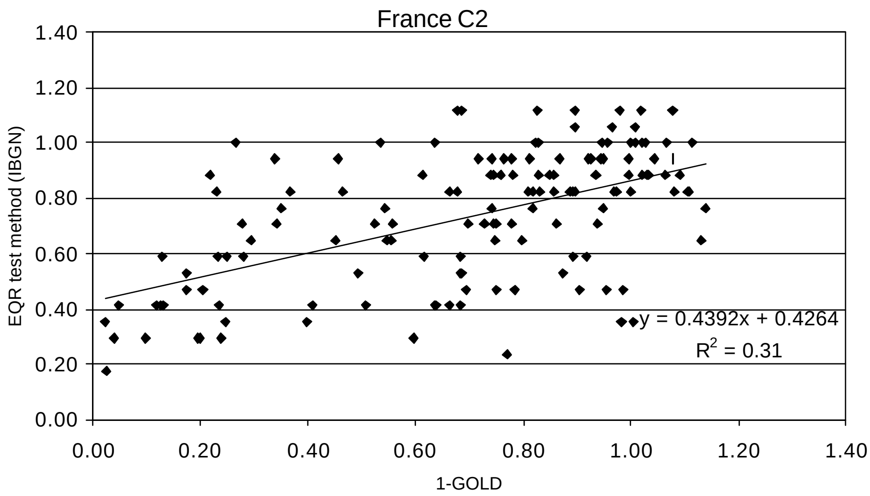


Figure 1D 1-GOLD - R^2 = 0.31; p<0.001

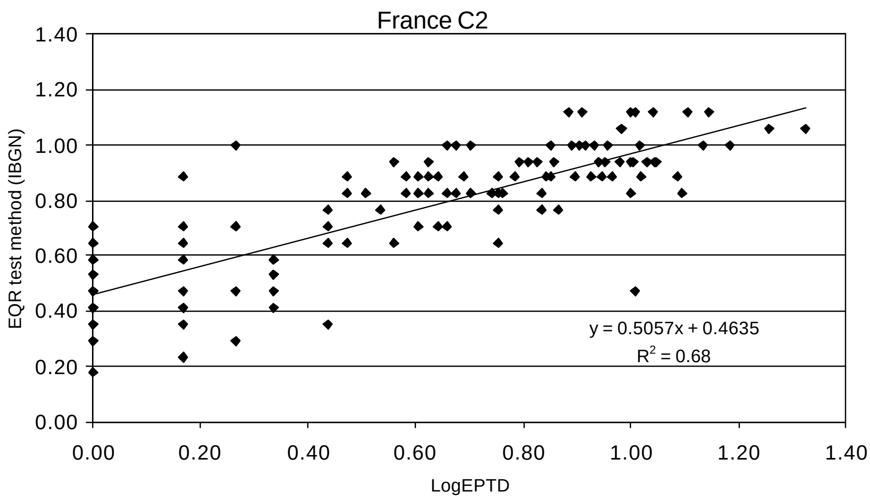


Figure 1E Log EPTD - $R^2 = 0.68$; p<0.001

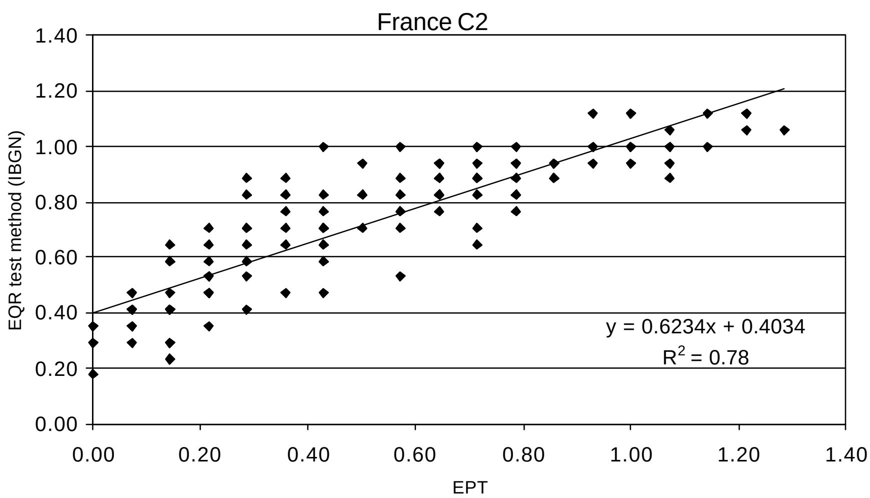


Figure 1F EPT - $R^2 = 0.78$; p<0.001

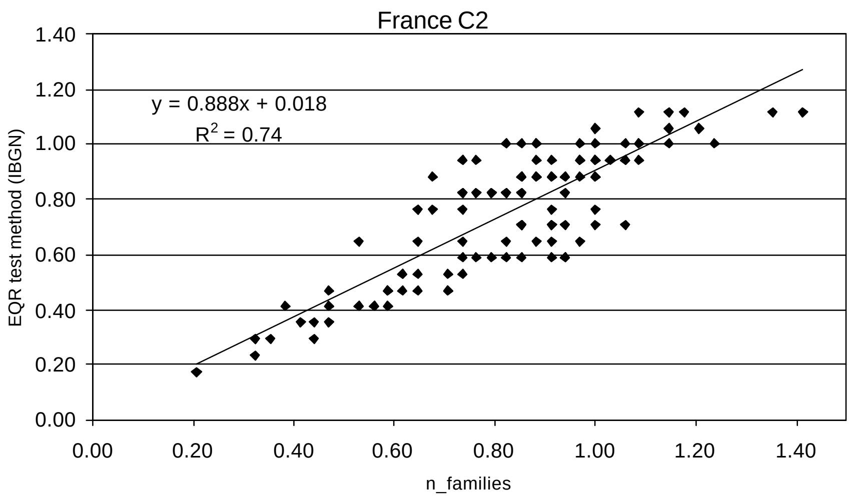


Figure 1G Number of families - R^2 = 0.74; p<0.001

General remarks

Remarks on the original calculation: normalization performed according to the median value of the 24 reference sites indicted by the data provider (Jean Gabriel Wasson). 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 75th 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

AFNOR (Association Française de Normalisation), 1992. Détermination de la qualité biologique des eaux courantes : Indice Biologique Global Normalisé. Norme NF T 90-350.

4.5.2 - 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. This version of the newly developed MMI is a prototype. Some changes may occur in the definitive version due to the ongoing work.

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 75th 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 75th percentile, are shown in Figures 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. Original boundaries are provided by Pardo (pers. comm.).

Table 1 Spanish MMI class boundaries conversion for C2 dataset

	MMI score	MMI EQR	ICMi EQR
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; $p<0.001$

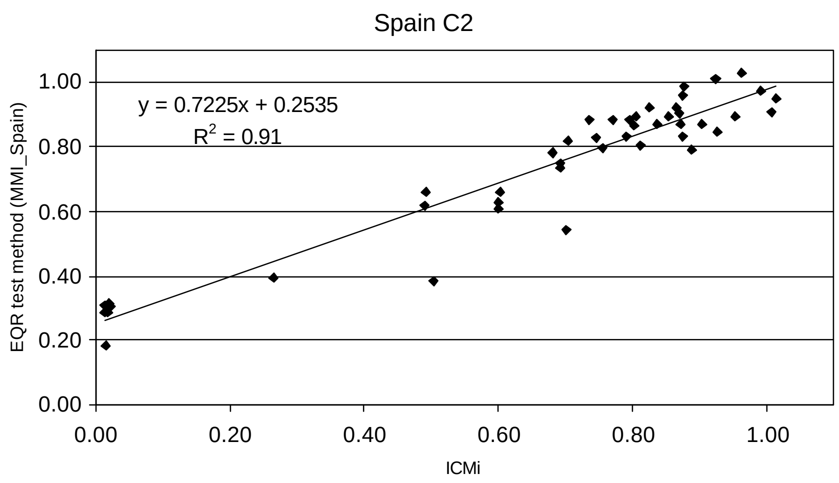


Figure 1A ICMi - R^2 = 0.91; p<0.001

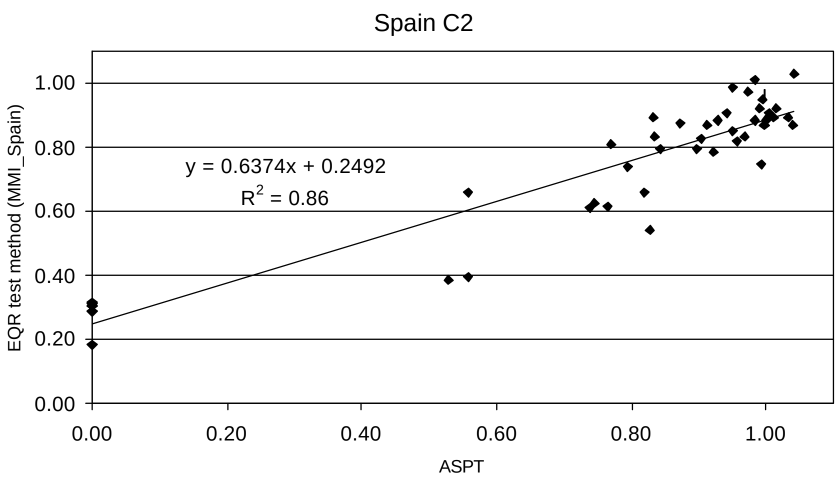


Figure 1B ASPT - R^2 = 0.86; p<0.001

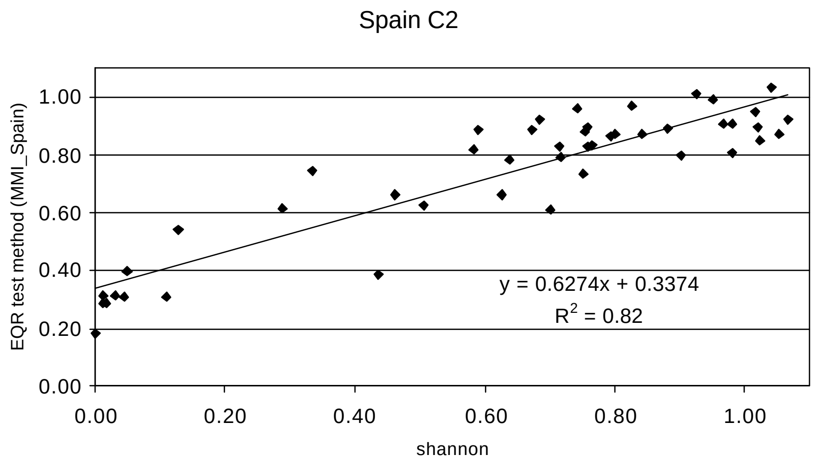


Figure 1C Shannon - R^2 = 0.82; p<0.001

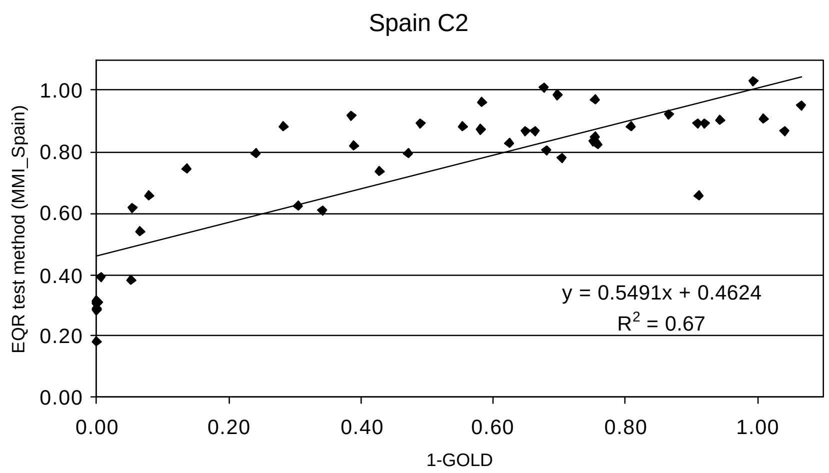


Figure 1D 1-GOLD - R^2 = 0.67; p<0.001

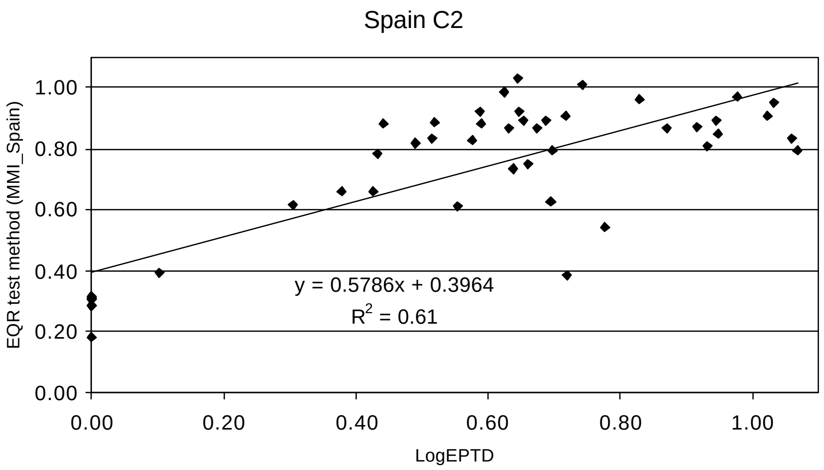


Figure 1E Log EPTD - R^2 = 0.61; p<0.001

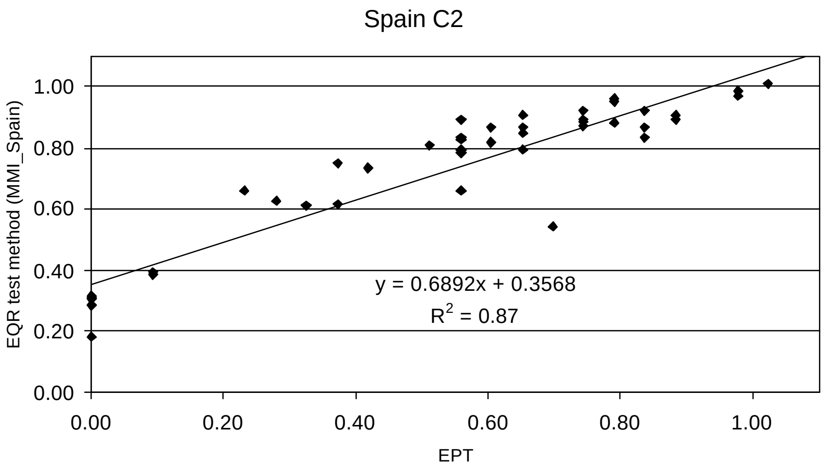


Figure 1F EPT - R^2 = 0.87; p<0.001

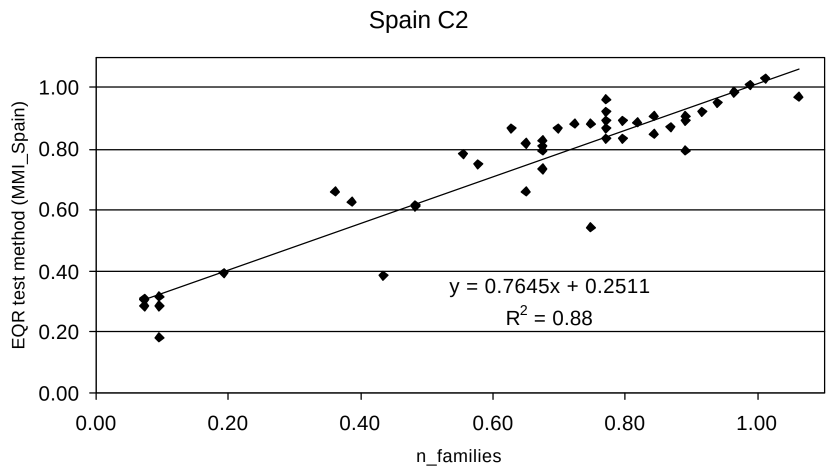


Figure 1G Number of families - R^2 = 0.88; p<0.001

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.

4.6 - IC type M1 (small mid-altitude streams highly seasonal regime)

4.6.1 - France M1

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 Normalisé, AFNOR, 1992). Sampling is carried out taking a number of 8 samples with a Surber sampler (base area 1/20 m^2). These samples are characterized by different fixed couple of substrate size 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 a 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.

For each type, the classification is done following a WFD compliant procedure, according to the REFCOND Guidance recommendations, in relation to the type specific reference conditions. The Reference value (RV) for the normalization (EQR calculation) is the median of the IBGN values observed in reference sites. The H/G boundary is set at the 25th percentile of the values observed in reference sites. The G/M boundary is first calculated separately for the two metrics (H/G boundary minus 1 for the GFI, and 1/4 of the range below H/G boundary for the number of taxa), and the combination of the two metrics gives the IBGN G/M boundary.

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 25th 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 75th 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 75th 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 Figures 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 75th 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.

In Table 2 is also reported the conversion of boundaries according to the original normalization, i.e. to the median value of the reference samples.

Table 1 IBGN class boundaries conversion for M1 dataset. Normalization: 75th percentile high status samples

	IBGN score	IBGN EQR	ICMi EQR
Reference value	17	0.986	0.918
Limit high-good	15	0.870	0.807
Limit good-moderate	13	0.754	0.695
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 2 IBGN class boundaries conversion for M1 dataset. Normalization: median reference samples

	IBGN score	IBGN EQR	ICMi EQR
Reference value	17	1.000	1.019
Limit high-good	15	0.882	0.894
Limit good-moderate	13	0.765	0.770
Limit moderate-poor	nd	nd	nd
Limit poor-bad	nd	nd	nd

ICM index = IBGN EQR * 1.56 - 0.0375
R^2=0.86; p<0.001

Important notice
The reference values and class boundaries tested here are provisional, and may change due to ongoing work on reference sites selection and sampling.

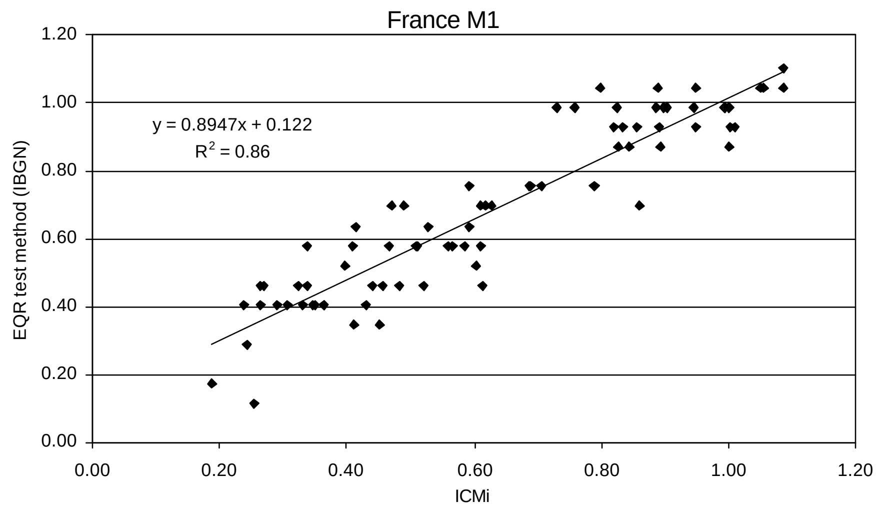


Figure 1A ICMi - R^2 = 0.86; p<0.001

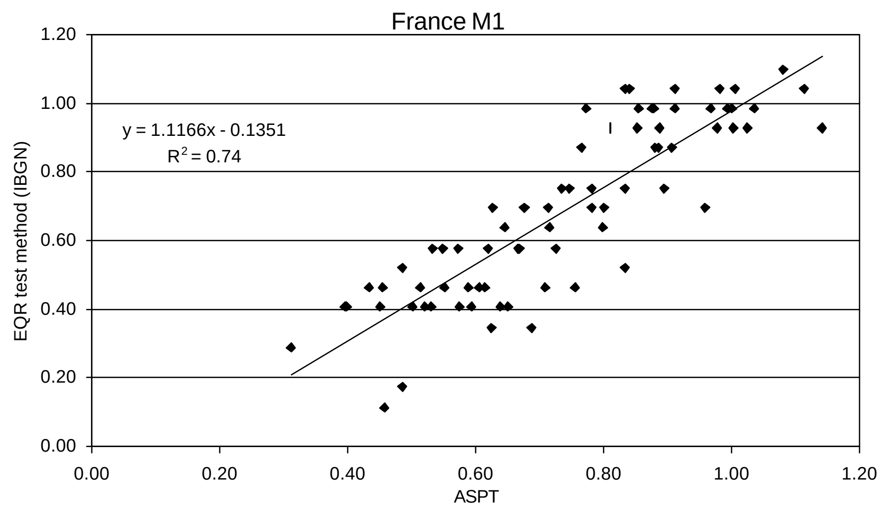


Figure 1B ASPT - R^2 = 0.74; p<0.001

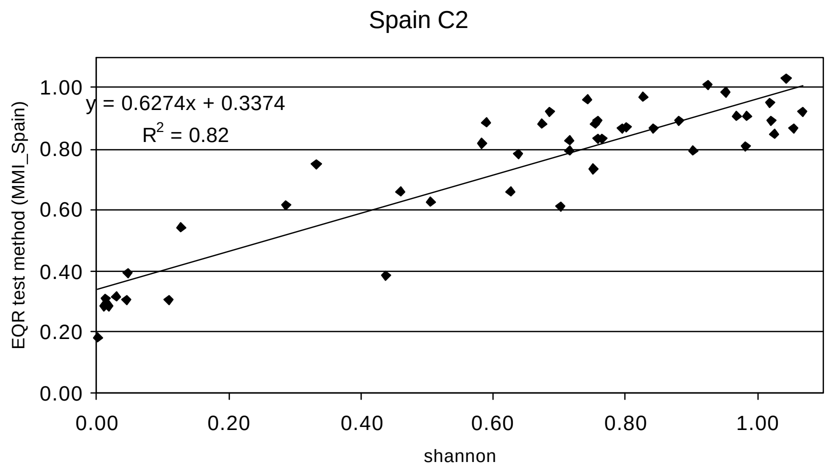


Figure 1C Shannon - $R^2 = 0.50$; p<0.001

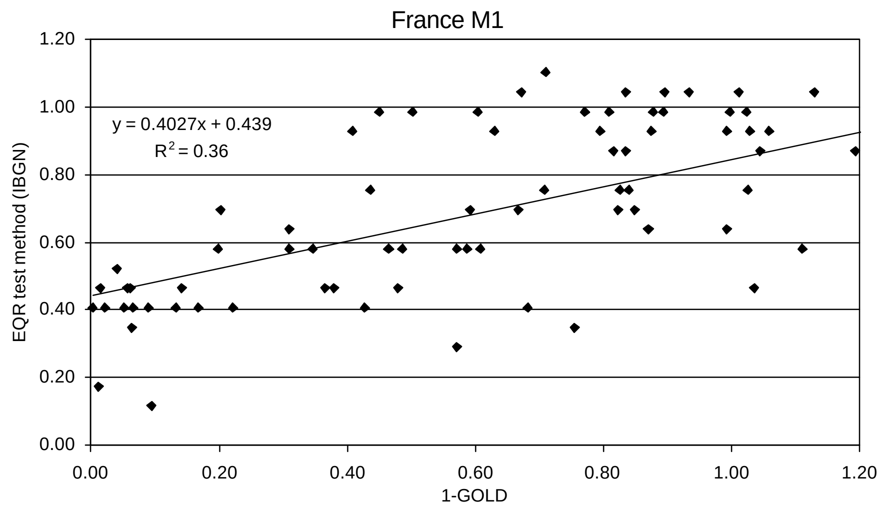


Figure 1D 1-GOLD - $R^2 = 0.36$; p<0.001

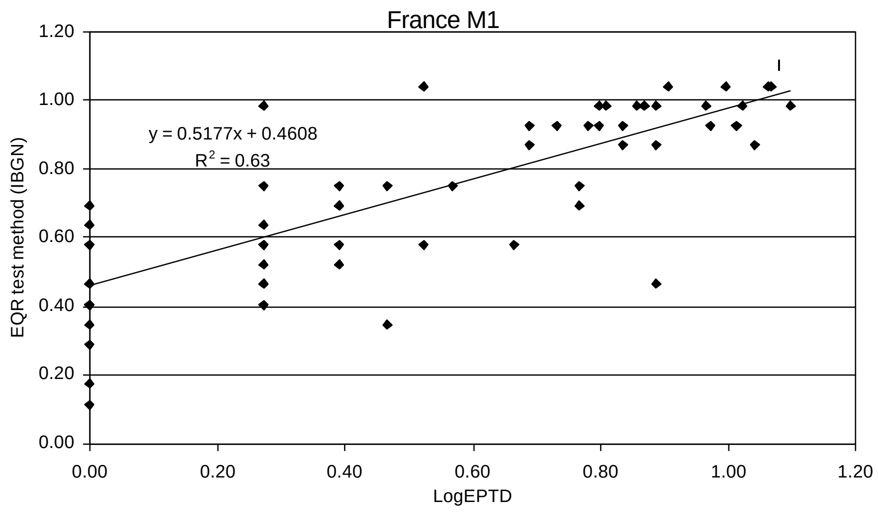


Figure 1E Log EPTD - R^2 = 0.63; p<0.001

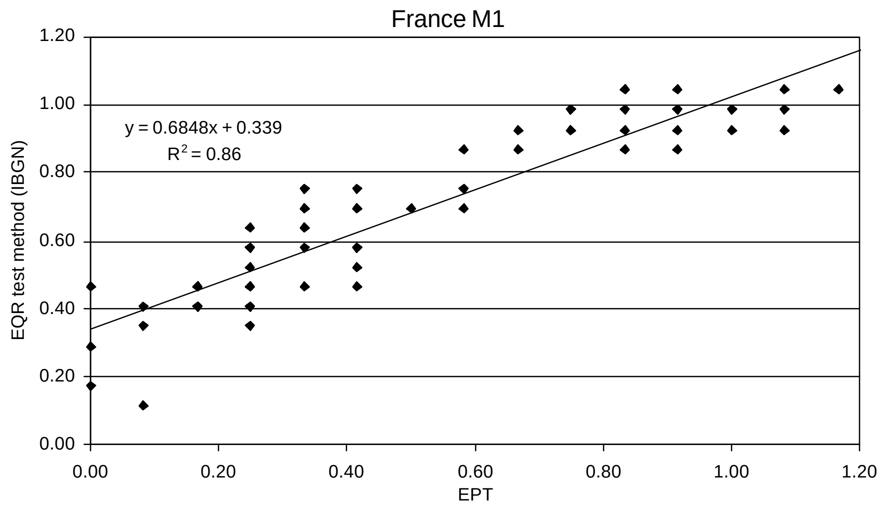


Figure 1F EPT - R^2 = 0.86; p<0.001

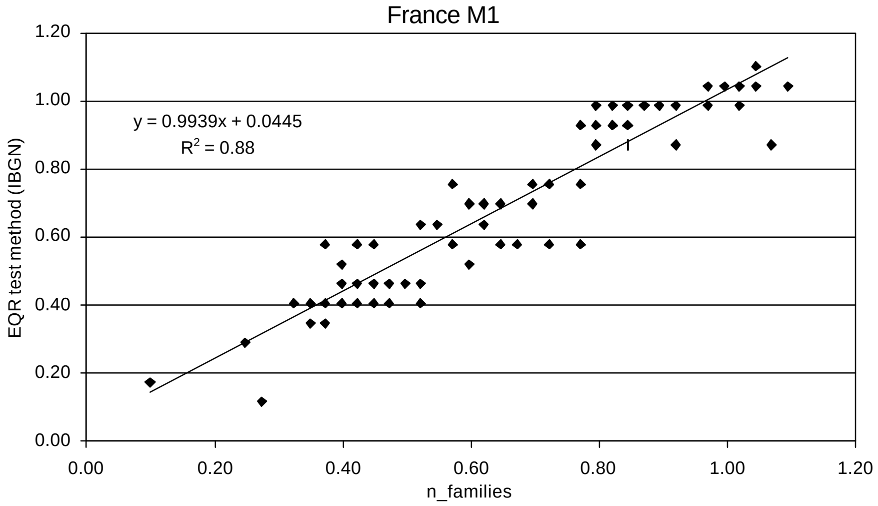


Figure 1G Number of families - R^2 = 0.88; p<0.001

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

AFNOR (Association Française de Normalisation), 1992. Détermination de la qualité biologique des eaux courantes : Indice Biologique Global Normalisé. Norme NF T 90-350.

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.

4.6.2 - Italy M1

General features

The sites are located in Southern Apennines (region Campania, see AQEM Consortium, 2002; Buffagni et al., 2004a; 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, 2004).

11 sites have been investigated 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, 2002), 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, 2004). 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 dataset 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^{th} 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. Original boundaries are provided according to APAT/IRSA-CNR, 2004 and Spaggiari & Franceschini, 2000 for the conversion in values.

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

Table 1 IBE class boundaries conversion

	IBE score	IBE EQR	ICMi EQR
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			

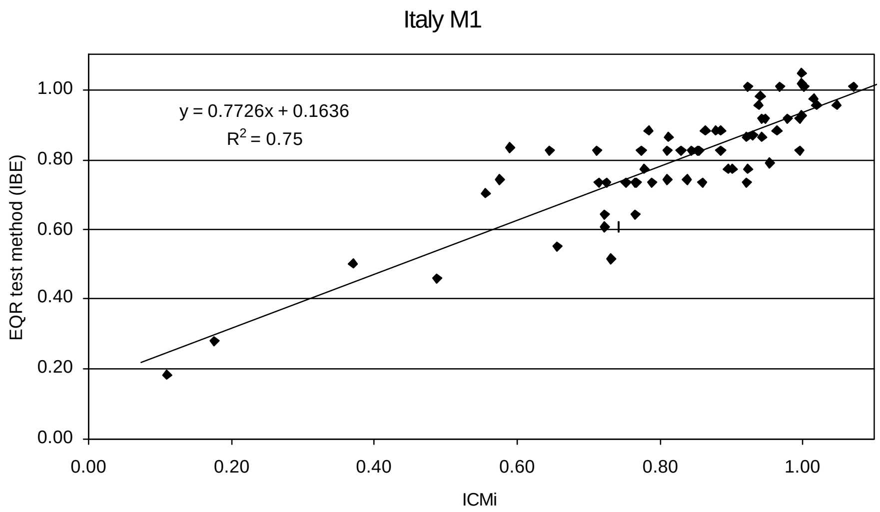


Figure 1A ICMi - R^2 = 0.75; p<0.001

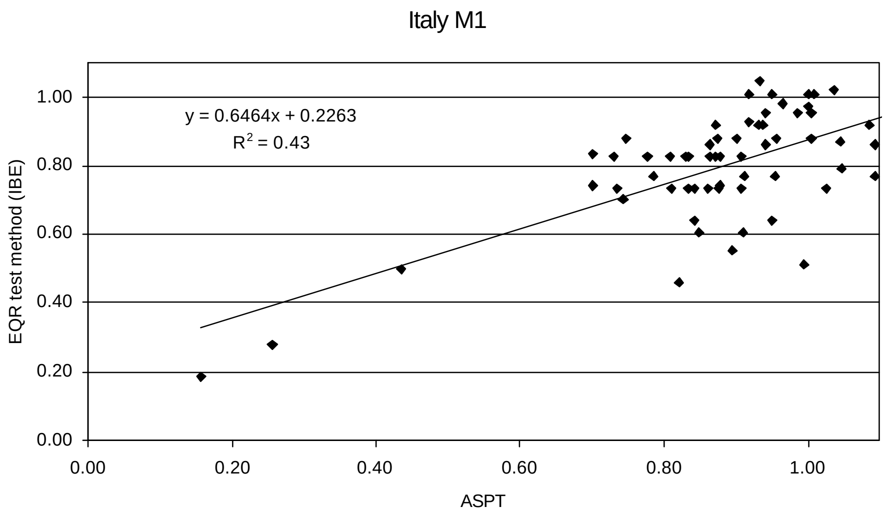


Figure 1B ASPT - R^2 = 0.43; p<0.001

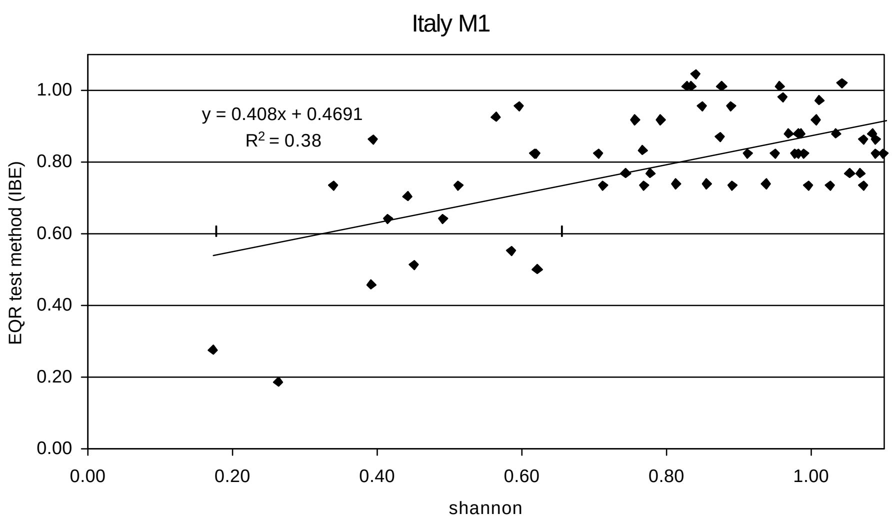


Figure 1C Shannon - R^2 = 0.38; p<0.001

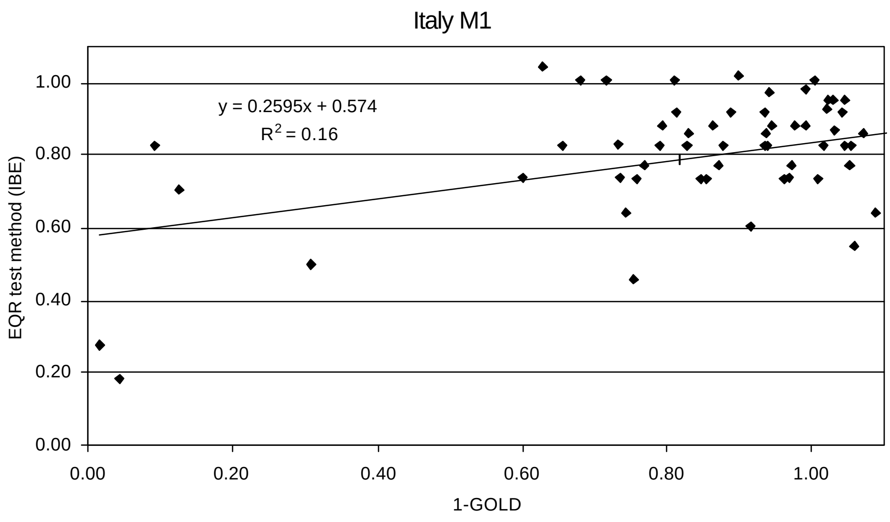


Figure 1D 1-GOLD - R^2 = 0.16; p=0.001

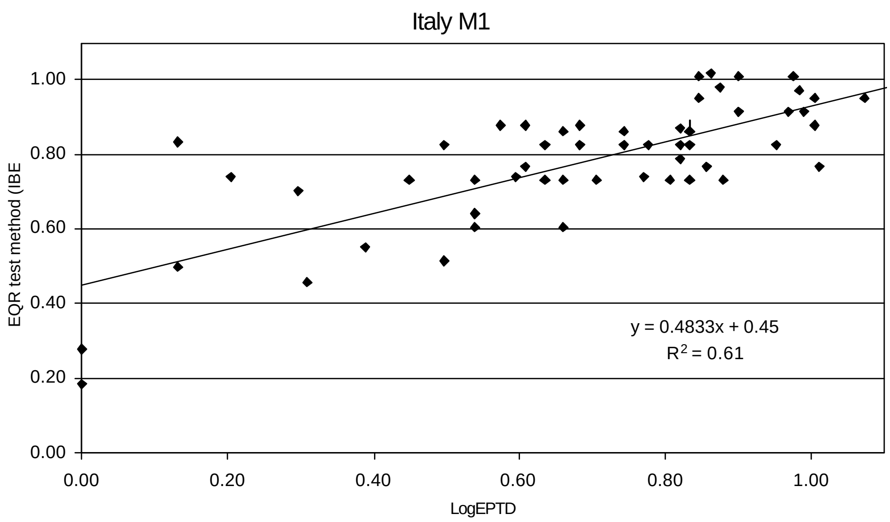


Figure 1E Log EPTD - R^2 = 0.61; p<0.001

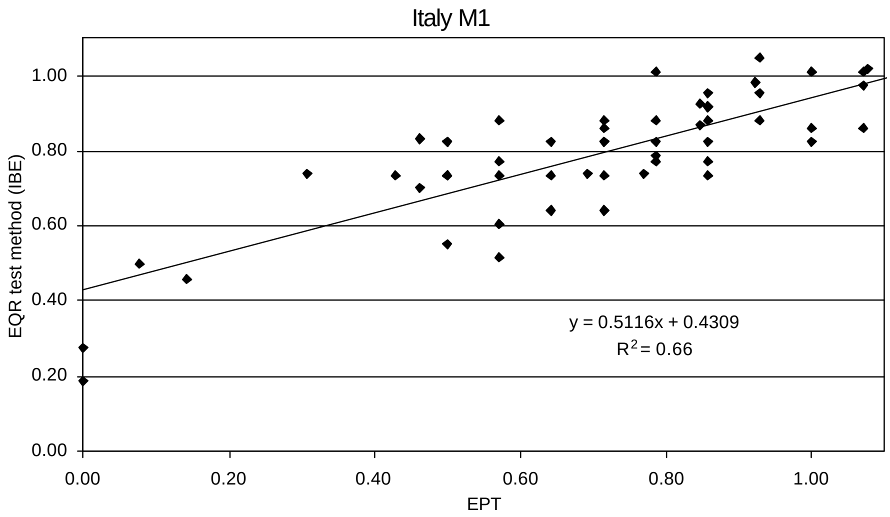


Figure 1F EPT - R^2 = 0.66; p<0.001

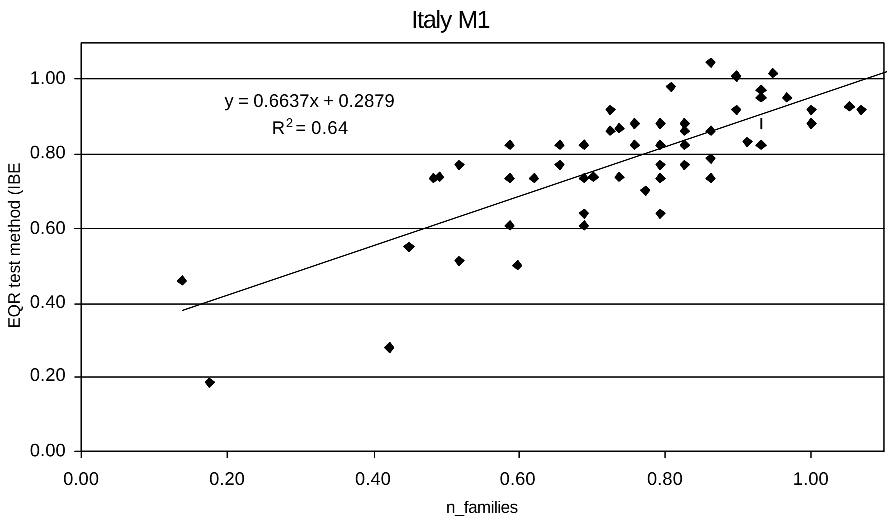


Figure 1G Number of families - R^2 = 0.64; p<0.001

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 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.

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, 2004a. 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.
Spaggiari, R. & S. Franceschini, 2000. Procedure di calcolo dello stato ecologico dei corsi d'acqua e di rappresentazione grafica delle informazioni. Biol. Amb. 14 (2): 1-6.

4.7 - IC type M5 (Temporary streams)

4.7.1 - Italy M5

General features

The sites are located in three areas of the region Sardinia (Buffagni et al., 2004b). 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.

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., 2004b). 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, 2002), 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, 2004). 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 database 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 Figures 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. Original boundaries are provided according to APAT/IRSA-CNR, 2004 and Spaggiari & Franceschini, 2000 for the conversion in values.

Table 1 IBE class boundaries conversion

	IBE score	IBE EQR	ICMi EQR
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; $p<0.001$

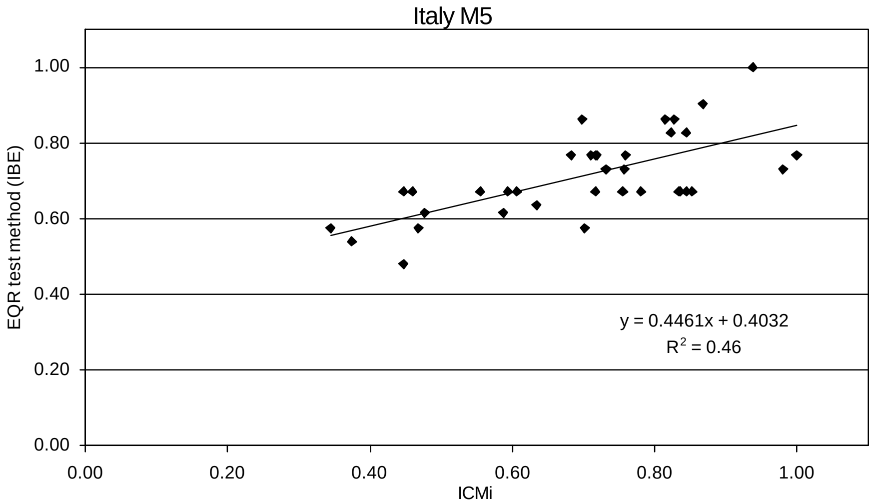


Figure 1A ICMi - R^2 = 0.46; $p<0.001$

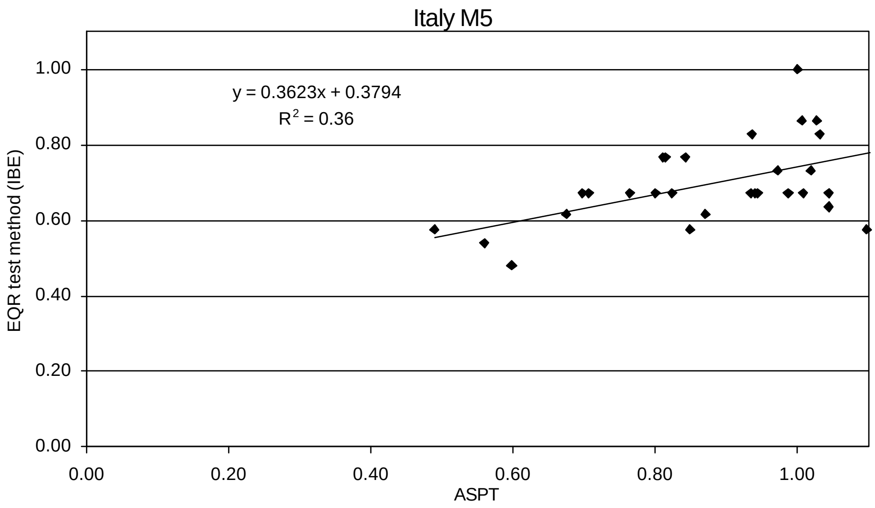


Figure 1B ASPT - R^2 = 0.36; p<0.001

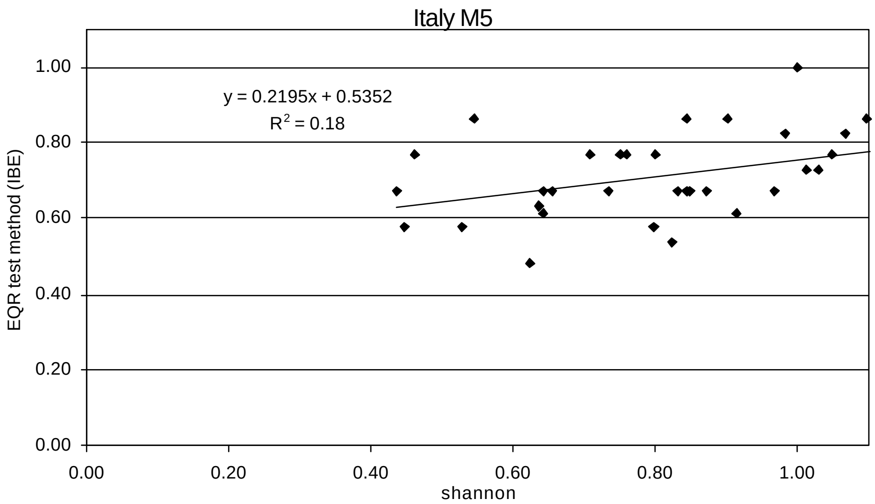


Figure 1C Shannon - R^2 = 0.18; p=0.009

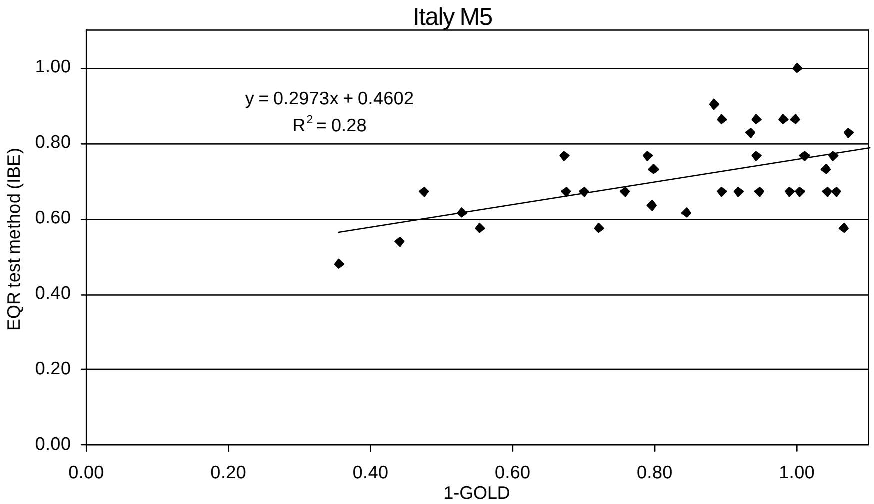


Figure 1D 1-GOLD - R^2 = 0.28; p<0.001

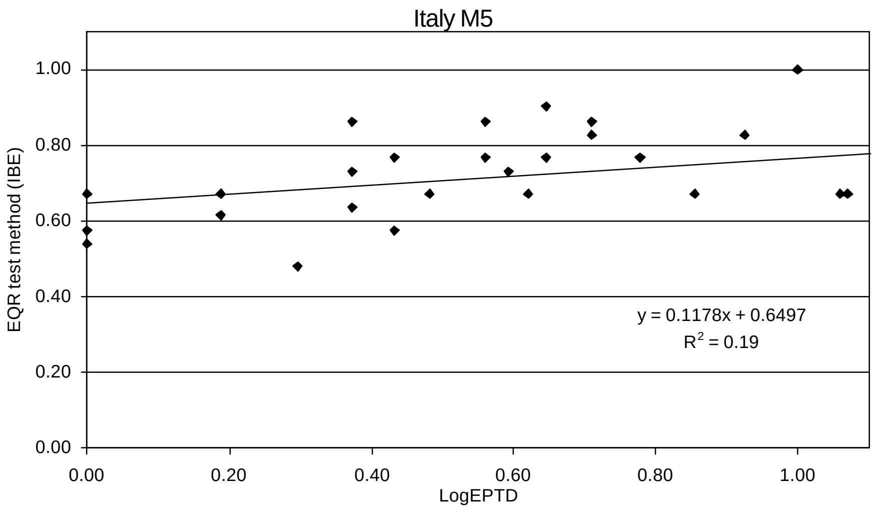


Figure 1E Log EPTD - R^2 = 0.19; p=0.008

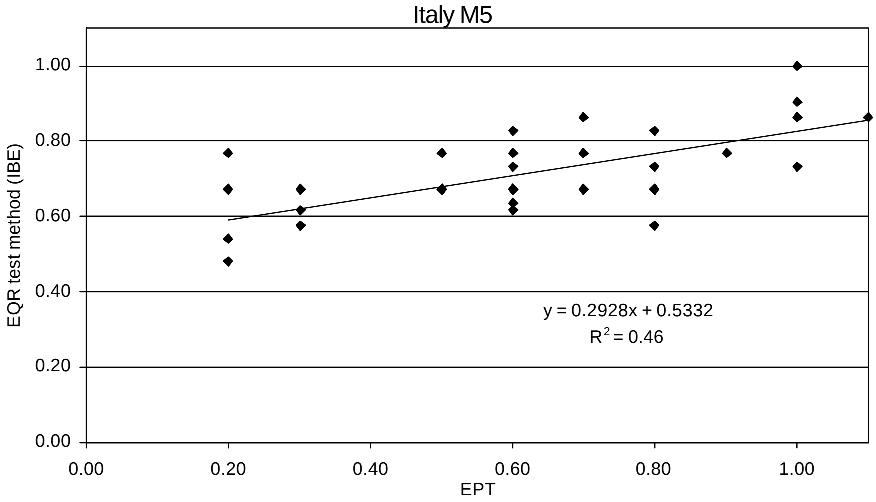


Figure 1F EPT - R^2 = 0.46; p<0.001

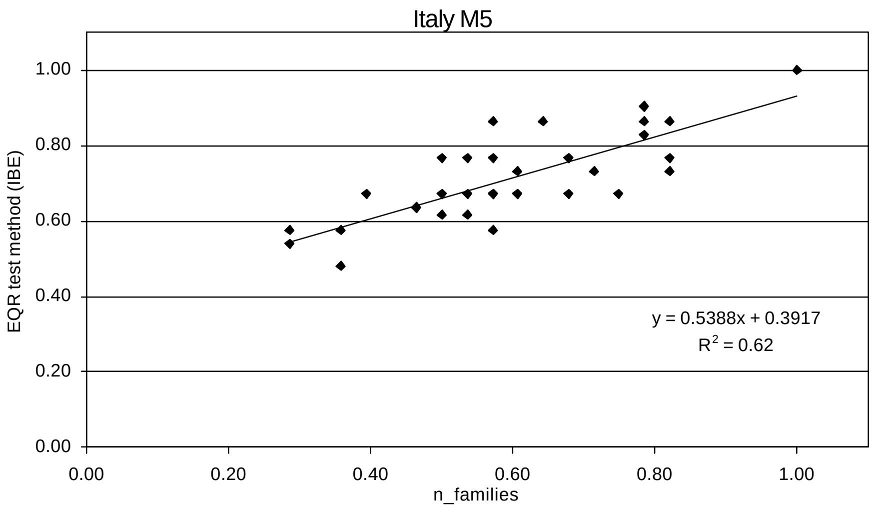


Figure 1G Number of families - R^2 = 0.62; p<0.001

General remarks

In about 20 years, the calculation of IBE index encountered several updates (Ghetti, 1986; 1995; 1997; APAT/IRSA-CNR, 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.

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

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.

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, 2004b. 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.

Spaggiari, R. & S. Franceschini, 2000. Procedure di calcolo dello stato ecologico dei corsi d'acqua e di rappresentazione grafica delle informazioni. Biol. Amb. 14 (2): 1-6.

4.8 - Summary tables for test datasets

In Table 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.

Table 4.1 Test datasets features

dataset name (country and stream type)	size class km^2	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)
B C1	10-100	<200	Flemish Environment Agency	G. Verhaegen - Flemish Environment Agency	Flemish Environment Agency	National monitoring program	70 ca.	3 years (2000-2)	208	10	nd
DK C1	15-100	<200	Regional Danish authorities (counties)	J. Skriver - NERI	NERI	National monitoring program	135 ca.	6 years, various seas.	346	72	17 high status sites according to expert judjement, inverts community and abiotic data
EST C1	<100-<1000	<200	Estonian Agricultural University	H. Timm - EAU	Estonian Agricultural University	National Estonian database	23	1 year, 1 seas.	23	9	nd
F C1	10-300	<100	Direction Regionale Environment	J-G. Wasson - Cemagref-Lyon	Database: Cemagref-Lyon	National monitoring network	20	10 years	139	59	24 reference sites according to pressures data
D C1	10-100	<200	Various regional authorities	S. Birk - UNI Essen	Umweltbund esamt and LAWA	Quality monitoring	38	Various years, 3 sesons	91	1	nd

continued

Table 4.1 (continued)

dataset name (country and stream type)	size class km^2	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)
I C1	10-100	<200	ARPA Lombardia	P. Genoni - ARPA Lombardia	ARPA Lombardia	Various ARPA surveys (monitoring, internal IC, method testing)	39	6 years, 4 seas.	361	84	nd
NL C1	10-100	<200	Various authorities	R. Knoben, M. van den Berg - RIZA	Database: Limnodata Neerlandica	Not specified	nd	nd	79	12	nd
PL C1	10-100	<200	Institute Environmet Protection, Warsaw	H. Soszka - Institute Environment Protection, Warsaw	Voivodships Inspectorates of Environmental Protection in Poland, Warsaw University	Various, pilot monitoring; scientific project	49	1 year, 1 seas.	49	11	nd
UK C1	small	lowl.	Environment Agency	J. Murray Bligh - Environment Agency	Environment Agency	Environment Protection	789	1 year, 2 seas. combined data	789	202	nd
total	**C1**						**>1160**		**2085**	**470**	

continued

Tab. 4.1 (continued)

dataset name (country and stream type)	main degradation factor	quality gradient	support data	classification method	min/max possible values	sampling method	sampling area m^2	calculation formulae
B C1	General	High-bad	Not available	Multimetric Index Flanders	0/1	3 minutes sampling from all available microhabitats.	Not specified	Multimetric Index (combination of 5 metrics)
DK C1	General	High-moderate (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.	1.25 ca.	Two entries table (2 metrics: indicator group and diversity group)
EST C1	General	High-poor	Hydro-chemical 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
F C1	General	High-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)
D C1	Morph.-general	High-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

continued

Tab. 4.1 (continued)

dataset name (country and stream type)	main degradation factor	quality gradient	support data	classification method	min/max possible values	sampling method	sampling area m^2	calculation formulae
I C1	General	High-bad	Main physical, chemicals, micro-biological variables (not available for all samples).	IBE (national method)	0/14	Riffle only, semiquant.	0.9 (estimated)	Two entries table (2 metrics: n° of taxa and Faunistic Group)
NL C1	General	High-bad	Not specified	KRW-maatlatten	0/1	All microhabitats present	1.5	Dutch metric
PL C1	Eutrophication	High-bad	Waterchemistry data.	BMWP-POL and Margalef div. Index	BMWP-POL:0/ open end Margalef: <1/open end	4 quantitative replicates from dominating substrates + 1 qualitative from all habitats.	1	Worst classification between BMWP-POL and Margalef div. index
UK C1	Organic	High-bad	Chemical monitoring data and pressures.	National GQA classification	0/>1	RIVPACS	Not specified	EQI ASPT (observed ASPT/RIVPACS predicted ASPT)

continued

Tab. 4.1 (continued)

dataset name (country and stream type)	size class km^2	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)
F C2	10-200	<150	Direction Regionale Environment	J. G. Wasson Cemagref-Lyon	Database: Cemagref-Lyon	National monitoring network	38	10 years	143	73	27 reference sites according to pressures data
E 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 seas.	46	7	7
total	**C2**						**63**		**189**	**80**	
F M1	10-100	200-800	Direction Regionale Environment	J. G. Wasson -Cemagref-Lyon	Database: Cemagref-Lyon	National monitoring network	32	6 years	77	28	nd
I 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 seas.	63	21	12
total	**M1**						**55**		**140**	**49**	
I M5	10-400	100-150	CNR-IRSA	A. Buffagni - CNR-IRSA	CNR-IRSA	National research samples	12	3 seas.	37	1	8
total	**M5**						**12**		**37**		

continued

Tab. 4.1 (continued)

dataset name (country and stream type)	main degradation factor	quality gradient	support data	classification method	min/max possible values	sampling method	sampling area m^2	calculation formulae
F C2	General	High-bad	National PC monitoring network.	IBGN	1/20, transformed in EQR	8 habitat samples characterized by substrate dimension and flow velocity.	0.4	Two entries table (2 metrics: n° of family and Faunistic Indicator Group).
E C2	Organic - nutrients	High-bad	Physico-chemistry, hydro-morphological, diatoms (not for 1 year data).	Multimetric index. Multiple regression analysis to select metric combination	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.
F M1	General	High-bad	National PC monitoring network.	IBGN	1/20, transformed in EQR	8 habitat samples characterized by substrate dimension and flow velocity.	0.4	Two entries table (2 metrics: n° of family and Faunistic Indicator Group).
I M1	General	High-bad	Main physical, chemicals, microbiological variables. Environmental indices: HMS, HQA, IFF.	IBE	0/14	Riffle only, semiquantitative.	0.9 (estimated)	Two entries table (2 metrics: n° of taxa and Faunistic Group).
I M5	General	High-bad	Main physical, chemicals, microbiological variables. Environmental indices: HMS, HQA, IFF.	IBE	1/14	Riffle only, semiquantitative.	0.9 (estimated)	Two entries table (2 metrics: n° of taxa and Faunistic Group).

4.9 - Summary of the biological assessment methods tested

Table 4.2 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. (next three pages).

Table 4.2 Considered assessment methods

method name (acronym)	method full name (English translation)	country	main degradation factor investigated	sampling method (in the present exercise)	sampling area (m^2)	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.
Sampl.: Swedish example, classif.: 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 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, Saprobic Index	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, but mainly organic pollution)	Not fixed n. of replicates, possibly collected along a representative transect in the *riffle* area.	Not fixed (estimated 0.5 to 1)	Semiquantitative sorting; abundance estimation in three class.
KRW	KRW (the Dutch metric)	The Netherlands	General degradation (stressor not specified)	Sampling of all microhabitats present, field or lab sorting for 1-1.5 hours	1.5	Real abundances
Polish method	BMWP-Polish & 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)	Sampling proportional 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 proportionally, both in *riffle* and *pool*.	Not fixed	Abundance classes, number of individuals.

continued

Table 4.2 (continued)

method name (acronym)	ID level	calculation formulae	min/max values	class boundaries	literature references
MIF	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
Sampl.: Swedish example, classif.: 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	AFNOR, 1992
DIN 38 410, Saprobic Index	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: total 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-CNR, 2004; Ghetti, 1997
KRW	Species/Family	Not specified, one metric	0/1	H-G: 1; G-M: 0.8; M-P: 0.6; P-B: 0.4	Knoben & van den Berg, 2004
Polish 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	Species/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

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, physico-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) Examples of potential benchmark datasets already existing (derived by E.U. co-funded projects): AQEM (invertebrates), FAME (fishes), STAR (invertebrates, diatoms, macrophytes, fishes, hydromorphology).

According to the general criteria described in the box above (see also paragraph 5.2), 10 datasets from 6 different European countries were included in the benchmark dataset used for the calculation in the present paper. Datasets are described below, divided in: AQEM Project datasets (5.6), STAR Project datasets (5.7) and extra AQEM/STAR datasets (5.8).

5.1 - The AQEM and STAR Projects datasets

At sites investigated for the Project EU-AQEM (Hering et al., 2003; 2004), samples were collected with the aim of developing and testing macroinvertebrate based assessment methods, which satisfy WFD requirements. Following the principles of the AQEM Project, the STAR Project (Furse, 2001) aims at developing 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 multi-habitat approach (derived from Barbour et al., 1999; see also Hering et al., 2004), with the collection of 20 sample units proportionally distributed among the micro-habitats present in the river. Additional data is available for all samples and sites regarding 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 catchments area, etc. (for a detailed description see AQEM Consortium, 2002). Additionally, extra environmental data is often available for single datasets. In particular, within the STAR project, data on hydromorphological features was collected by applying the River Habitat Survey protocol concurrently with the collection of biological data. Similar information on hydromorphology is obtained, in an even rougher and general way, from the 'site protocol'.

Each set of data collected for the AQEM and STAR Projects includes a set of reference sites, selected according to the demands of the WFD. Criteria for the selection of the reference sites are specified in Hering et al. (2003).

Three different steps of classification were provided during the AQEM Project for each of the study sites: firstly, a pre-classification was provided to give an overall idea of the degree of degradation of the site; after data collection, a post-classification was undertaken, usually based on multivariate techniques (or others statistical methods) run on benthic community and pressures data; lastly, the final-classification based on the definition of a multimetric index, which usually fits quite well with the post-classification of the site (see Hering et al., 2004; Nijboer et al., 2004).

5.2 - The Best Available Classification (BAC) concept

In the examples reported in this paper, the quality classification used for benchmark datasets is referred to as a Best Available Classification (BAC: see the box below).

The BAC corresponds, depending on the country, to the post or final classification used in the AQEM Project, sometimes related to which of the two is better represented by the quality gradient of the sites described by macroinvertebrates' community and related pressures. For the STAR datasets, the analyses are still in progress, thus the BAC is provided according to the analysis run until now. Details on the BAC and criteria to derive it, are reported analyses are still in progress, thus the BAC is provided according to the analysis run until now. Details on the BAC and criteria to derive it, are reported in the single sections describing the benchmark datasets. A summary is reported in Table 5.1. General criteria for the acceptance of a BAC are reported in the following paragraphs. A general definition of BAC classification is outlined in the box below.

A few more definitions

Best Available Classification (BAC)

The ecological 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 in action, 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 on the basis of the criteria outlined in the Guidance on the Inter-calibration 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 Inter-calibration 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. Part of the data produced by the two projects has been used for the comparison and harmonization exercises presented in this Paper.

National Standard Classification

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

Notes: (a) Each MS has 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.

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*., 2004a). 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*., 2004a 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, catchments characteristics, etc. (AQEM Consortium, 2002).

5.3 - Criteria to derive a BAC classification

- Evaluation of tolerance to pollution included.
- Richness/Diversity considered.
- Abundance considered.
- Type specific classification → the used classification system must be stream type adapted (i.e. type or site specific reference conditions).
- Pressure analysis is combined with biological information → abiotic or biotic classification only is not acceptable (e.g. multimetric systems alone: BAC; morphological or chemical classification alone is inadequate, or only pressures-based classification).
- Classification based on multivariate analysis is acceptable.
- 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 the scope of the present paper to argue on the procedure and protocols to be used to derive Best Available Classifications, these will be defined in the proper circumstances (e.g. European Commission, 2004).

BAC, Reference conditions and High Status sites

In the REFCOND Guidance, the concepts of Reference Conditions (and sites) often overlap with those of High status conditions (and sites). On this point, text was written to facilitate the clearness of the description of some of the major issues discussed. In the Guidance, it can be read: “RC equal high ecological status, i.e. no or only very minor evidence of disturbance for each of the general physico-chemical, hydromorphological and biological quality elements”.

For the European Inter-calibration process, the two concepts should be kept clearly distinct. In fact, the practical implications of basing any data handling on one or the other concept are central to the IC comparison of national results.

For High status sites/samples, in the present paper, we refer to sites/samples classified as belonging to the highest of five quality classes (i.e. highest quality) based on the national, standard method in use for aquatic invertebrates. More properly, it should be regarded as a ‘macroinvertebrates High Status’ (e.g. HS_m), not considering other BQEs or chemical information → it is then a narrower definition than the full High status concept, which covers a broader spectrum of information. For the IC process, it has been implicitly decided to proceed to an Inter-calibration of single BQEs one at a time, at least in the initial stage.

For Reference sites/samples, we contemplate those sites/samples that rise above screening criteria for the main acting pressures (meeting the requirements of the WFD) and have been at least partly validated by biological information (different from the standard invertebrate method). Any totally abiotic criteria to accept/refuse reference sites would comply with the WFD approach to water bodies classification.

For Best Available Classification (see box above), which necessarily includes assumptions on the reference conditions setting, we refer to a way of setting boundaries across quality classes and accepting/refusing sites as reference sites by using all the available information, as much as possible to reflect WFD requirements. The reference sites derived by a BAC should match up with the reference sites that will be finally accepted after fully WFD –compliant methods – for all BQEs, chemical compounds, etc. – are applied.

5.4 - Basic characteristics of benchmark datasets

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

In general terms, the characteristics for each dataset are:

- taxalist to family level
- taxalist must include abundance for each taxon (at least estimated)
- preferably the sampling area should be known
- samples from reference sites must be present
- a wide quality gradient has to be embodied (i.e. possibly all 5 classes)
- criteria to classify reference 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 samples/sites/seasons are considered
- how wide is the quality gradient (e.g. from High to Moderate, from Good to Bad)
- river type
- ancillary data available (pressures, 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, semi-quantitative)
- 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 corresponds to the multimetric classification developed for AQEM Project. Thus, BAC corresponds to what in AQEM has been called final classification. Class boundaries were set using the 25^{th} % percentile of references and the 75^{th} % percentile of bad sites. This range was divided by three. Even if the BAC classification is in this case based on a multimetric system, it has to be stated here that this system was developed analysing pressures data in combination with biological ones. In fact, PCA and cluster-analysis were performed on abiotic parameters to confirm the significant gradients in the dataset and thus the multimetric classification.

Number of samples

Data were collected and provided by Dr. Otto Moog from BOKU, Wien. 12 sites have been investigated for two seasons. Total number of samples is 24, 7 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 ranging between 100 and 1000 km^2, 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 ranges from 'reference' 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, for both the investigated stream types, consists on the post-classification, based on combination of community structure and threshold of saprobic index. To define classes the upper class boundary of high class was supported by analyses of reference sites database collected originally for PERLA predictive system. Setting of remaining class boundaries was based on cluster analyses of taxa composition data combined with plotting metrics values in boxplots (clusters were categories). 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 ranging between 100 and 1000 km^2, siliceous geology and altitude between 200m and 500m.

Degradation factor

The main stressor observed is organic pollution. According to the Best Available Classification performed, the quality gradient quality classes from 'reference' 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. 11 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 ranging 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 'reference' 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 - Italian Benchmark datasets

Italy IBM101 (IO2)
Italy IBM102 (IO3)
Italy IBC101 (IO4)

Sites' classification: Best Available Classification

For all Italian benchmark datasets, the BAC corresponds 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. (2004a) 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, 2002), 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 investigated for three seasons: spring and autumn 2000 and winter 2001. Total number of samples within each stream type is 33. The classification of the samples may vary considering different seasons; on the total number of 99 samples 24 are classified as reference.

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., 2004a; 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 'reference' to 'bad' status.

Sampling and sorting notes

The samples consist of the 10 pool replicates, collected proportionally to microhabitat occurence, since the assessment system is developed on this area (Buffagni et al., 2004a). This 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 belong 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 'reference' to 'moderate' status.

Sampling and sorting notes

The samples refer to the 10 pool replicates, collected proportionally to microhabitat occurence. This 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 'reference' 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. Mike Furse from CEH. 13 sites have been investigated for two seasons (spring and autumn). Total number of samples is 70. 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 'reference' 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 Mike Furse from CEH. 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 100 and 1000 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 'reference' 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 for 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 CNR-IRSA. 11 sites have been investigated. For all sites, data from the summer sampling period is included. For few sites also winter and 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, collected proportionally. 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.

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. 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 chapter 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 chapter4)

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 chapter4)

General degradation is the main factor of alteration. The dataset covers all the range of ecological status, from ‘reference’ 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. The attribution to IC type according to EC, 2003 has to be considered tentative, since some AQEM and IC typology is not consistent.

Table 5.1 Selected benchmark datasets

dataset code	country IC type	AQEM/ STAR code	size class km2	altitude class m	data collector and owner	aim of the collection	# sites	# season	# samples	reference samples # (BAC)	sampling method
AB04	Austria, C3/4	A04	100-1000	200-800	BOKU-Wien O. Moog	EU AQEM Project	12	2	24	5	AQEM sampling protocol
CB01	Czech Republic, C3/4	C01	100-1000	200-500	Masaryk University K. Brabec	EU AQEM Project	12	2	24	2	AQEM sampling protocol
CB03	Czech Republic, C3	C03	100-1000	200-500	Masaryk University K. Brabec	EU AQEM Project	11	2	22	7	AQEM sampling protocol
IBC 101	Italy, C1	I04	10-100	<200	CNR-IRSA A. Buffagni	EU AQEM Project	11	3	33	9	AQEM sampling protocol
IBM 101	Italy, M1	I02	10-100	200-800	CNR-IRSA A. Buffagni	EU AQEM Project	11	3	33	8	AQEM sampling protocol
IBM 102	Italy, M2	I03	100-1000	200-800	CNR-IRSA A. Buffagni	EU AQEM Project	11	3	33	7	AQEM sampling protocol
IBM 103	Italy, M1	I06	10-100	200-800	CNR-IRSA A. Buffagni	STAR Project	12	1 (3 for 3 sites)	16	2	AQEM sampling protocol
UB01	UK, C1/2	U15	10-100	<200	CEH Dorset	EU STAR Project additional stream type	13	2	70	18	STAR sampling protocol and RIVPACS
UB02	UK, C4	U23	100-1000	<200	CEH Dorset	EU STAR Project core stream type	12	2	66	18	STAR sampling protocol and RIVPACS
FBM101	F, M1	-	10-100	200-800	Direction Regionale de l'Environment	national monitoring network, refcond setting	32	6 years, several seasons	77	17	IBGN
	Total						**137**		**388**	**93**	

continued

Table 5.1 (continued)

dataset code	main degradation factor	quality gradient	support/ pressures data	Best Available Classification criteria
AB04	Morphology	from reference to bad	main physic, chemicals, microbiological variables. AQEM site protocol	Multimetric classification. Range between 25th percentile of high status and 75th of bad divided by three. PCA and cluster-analysis performed on abiotic parameters to confirm the significant gradients in the dataset and thus the multimetric classification.
CB01	Organic pollution	from reference to poor	main physic, chemicals, microbiological variables. AQEM site protocol	Post-classification, community structure and thresholds for saprobic value
CB03	Organic pollution	from reference to poor	main physic, chemicals, microbiological variables. AQEM site protocol	Post-classification, community structure and thresholds for saprobic value
IBC101	General degradation	from reference to bad	main physic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between reference and good class along multivariate axis. Remaining classes equally spaced
IBM101	General degradation	from reference to bad	main physic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between reference and good class along multivariate axis. Remaining classes equally spaced
IBM102	Morphology	from reference to moderate	main physic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between reference and good class along multivariate axis. Remaining classes equally spaced
IBM103	General degradation	from reference to moderate	main physic, chemicals, microbiological variables. HMS, HQA, IFF	Ecological breakpoints between reference and good class along multivariate axis. Remaining classes equally spaced
UB01	Organic pollution	from reference to bad	main physic, chemicals, microbiological variables. STAR site protocol, HMS, HQA,	RIVPACS classification
UB02	Organic pollution	from reference to bad	main physic, chemicals, microbiological variables. STAR site protocol, HMS, HQA,	RIVPACS classification
FBM101	General degradation	from reference to bad	pressures from national monitoring network, detailed geological and landuse features	IBGN classification, with WFD compliant reference definition

6 - COMMON EUROPEAN METRICS: ICMS AND OTHERS

Three examples of possible sets of metrics for European Inter-calibration are presented in this chapter. The first set is represented by the ICMs (see Table 3.1), used in the Paper to illustrate the different Options for the European IC process. They were conceived and selected because they were quick and simple to use and apply. The 'rough' acknowledged identification level (Family) also deals with this aspect. 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 European stream types. They correspond, respectively, to the '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

The aim of the present paragraph is to illustrate the performance of the ICMs and ICMi in a range of European test datasets. In Table 6.1 the R^2 value between each national assessment method and the different metrics is reported. Additionally, some examples representing the variation of the ICM index with respect to the quality classes identified by the MS' method are reported for Estonia (R-C1), UK (R-C1) and Spain (R-C2); further examples (for Poland and Italy, R-C1) can be found in chapter 8.3.

In respect of the WFD, all methods tested consider tolerance and richness of the benthic community. To better fit the normative definitions, abundance of benthic taxa, even if only roughly estimated for some datasets, has been nevertheless considered in the present exercise. The methods that have defined type-specific, biological reference conditions are here (tentatively) considered as WFD-compliant.

In test datasets the fit of the ICMi is generally good with a mean value of 0.60. In most of the datasets the R^2 for the ICMi are very good (>0.70). The types for which the ICMi shows R^2 values lower than 0.35 (the Netherlands and Germany) need further investigation.

Table 6.1 R^2 between National assessment methods and ICMs values in test datasets *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
NLC1	0.42	0.14	0.21	0.04	0.39	0.24	0.18
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	0.65	0.27	0.57	0.52	0.28	0.39	0.60
Mean WFD compliant	0.67	0.28	0.56	0.52	0.30	0.57	0.62

One of the possible reasons for this poor relationship could be the absence of reference sites in the datasets and/or the short quality gradient investigated. As for the benchmark data (see chapter 6.2), the metrics 1-GOLD and Shannon generally show the worst correlation, while the best fits are observed for ASPT, EPT taxa and number of families.

The following box&whiskers figures (6.1-6.3) show as the ICM index adequately describes the quality gradients for the considered test datasets. In general, metrics exhibited a good ability to discriminate between High status and the other four status classes. One important thing to note is the separation between Good and Moderate status samples: the interquartile range is always well separated among the two classes.

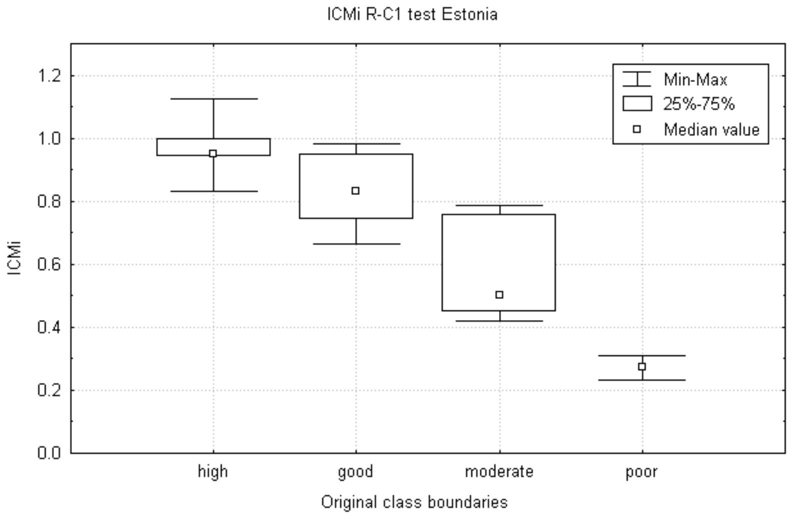


Figure 6.1 Box and whiskers representation for Estonia test dataset (R-C1 type). National quality classification vs ICMi

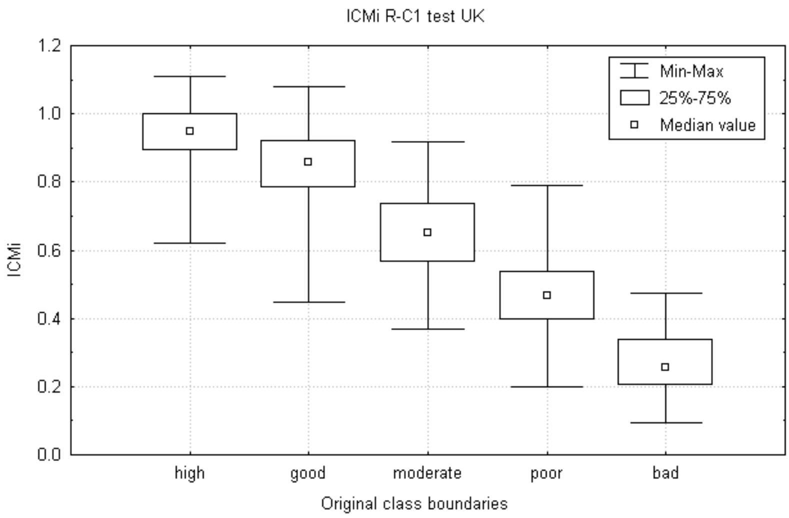


Figure 6.2 Box and whiskers representation for UK test dataset (R-C1 type). National quality classification vs ICMi

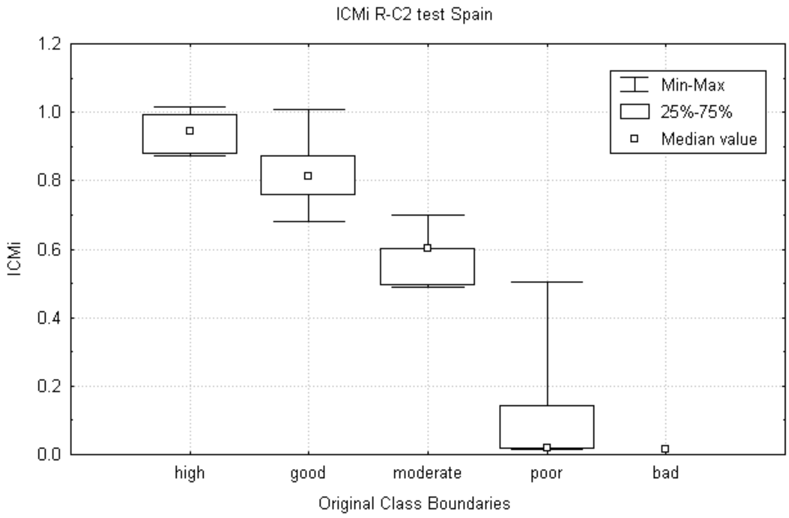


Figure 6.3 Box and whiskers representation for Spain test dataset (R-C2 type). National quality classification vs ICMi

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

As was done in the previous chapter, the performance of the selected metrics and ICMi is tested here with respect to the benchmark datasets. In the following table (see Table 6.2) 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 0.45 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 regression for the ICMi is good in all datasets (mean = 0.62), with the exception of the German datasets DB04 and DB01 which have a R^2 value respectively of 0.25 and 0.13.

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

	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
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
DB04	0.20	0.16	0.20	0.25	0.01	0.18	0.25
DB05	0.38	0.46	0.55	0.42	0.28	0.51	0.59
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
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

In this particular case the low regression is probably related to the short ecological gradient observed in these stream types for which the main anthropic stress is degradation in stream morphology. The analysis of the R^2 values has led to the inclusion of a selection of datasets in the benchmark that were used for comparison with test datasets (see chapter 8.3). Datasets showing a R^2 lower than 0.50 have been excluded from the benchmark. Despite its fair regression values, the dataset DB05 was excluded because no reference sites were present.

The relationship between ICMi and BAC is shown in box&whiskers representations. From such figures it is possible to consider 1) the variability of the ICMi values in each BAC class and 2) if the quality gradient expressed from the BAC is well represented by the ICMi.

Figure 6.4 considers the values of ICMi in all the benchmark datasets belonging to the IC types R-C (Different types are considered together: AB04, IBC101, CB01, CB03, UB01 and UB02).

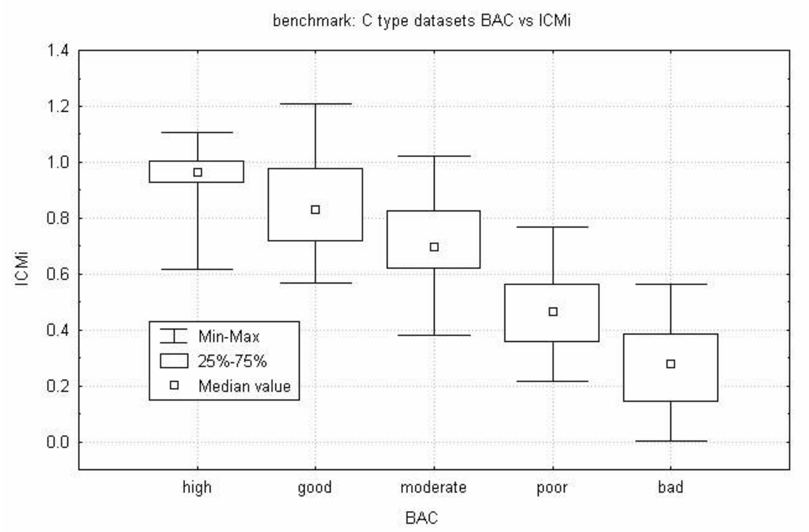


Figure 6.4 Box&whiskers representation for the benchmark datasets selected for R-C types. BAC vs ICMi

The trend of the median values shows a good fit with the quality classes expressed from the BAC. The classes Moderate, Poor and Bad result in being well separated. A slight overlap of the interquartiles can be observed for classes Good and Moderate and Good and High. Generally when different types and stressors are considered together, the variability of the biological and environmental systems is larger and slight overlaps are expected. It must be borne in mind however, that the aim for which the ICMi was developed was the straightforward comparison of different samples and datasets and not the mechanical classification of samples from spaced out areas. In this context, a slight overlap is fully acceptable.

In Figure 6.5, the same representation is provided for the benchmark datasets belonging to IC types R-M (IBM1, IBM202 and FBM101).

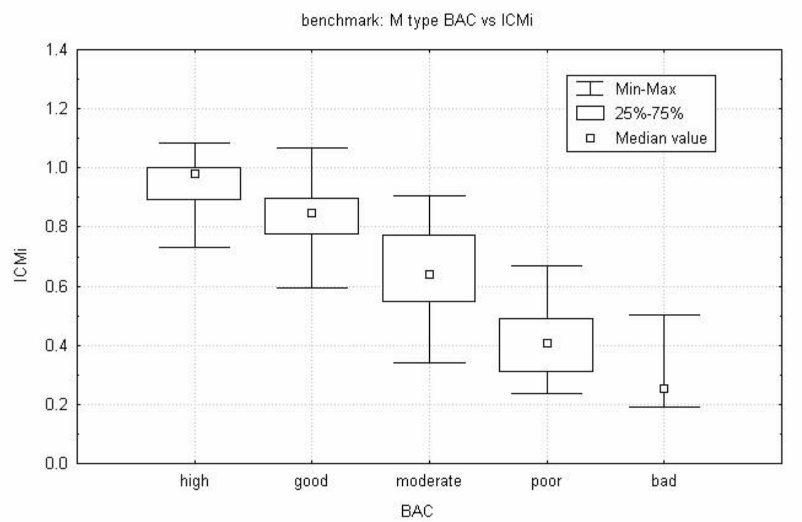


Figure 6.5 Box&whiskers representation for the benchmark datasets selected for R-M types. BAC vs ICMi

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 6.6 represents the combined results for all the selected benchmark datasets. 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 at any rate slight between quality classes, confirming the overall good performance of the ICM index even when considering together dissimilar river types and GIGs.

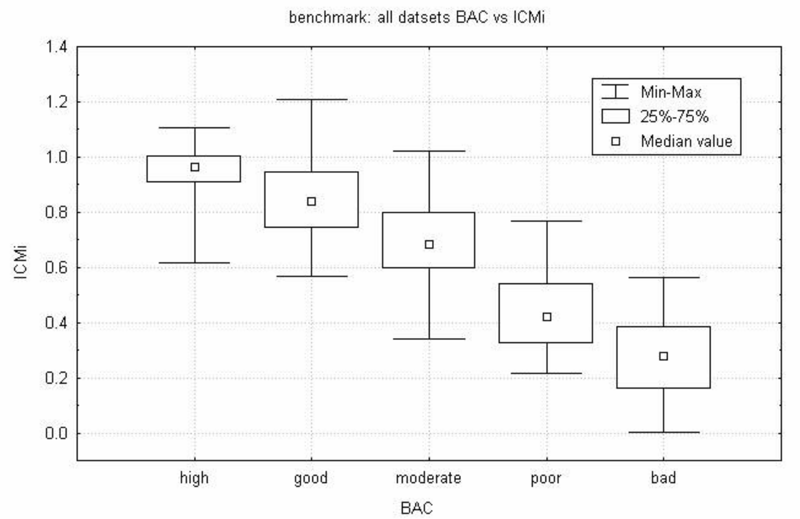


Figure 6.6 Box&whiskers representation for the benchmark datasets selected (all types included). BAC vs ICMi

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

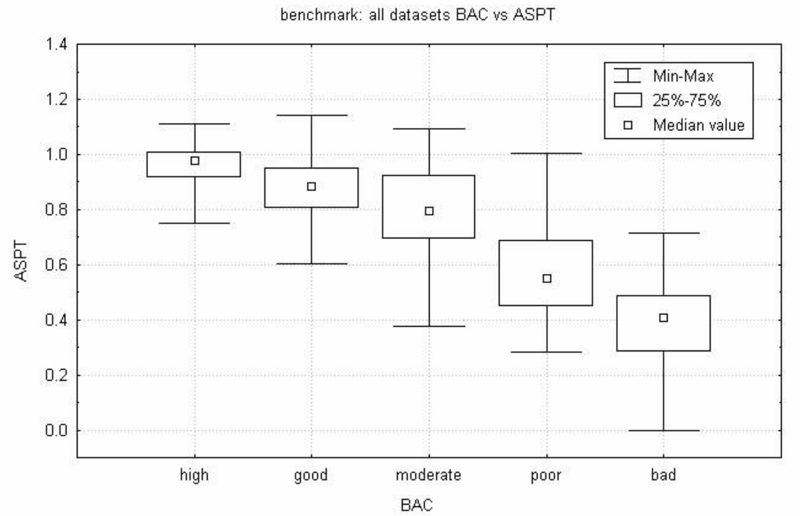


Figure 6.7 Box&whiskers representation for the benchmark datasets selected for all the types. BAC vs ASPT

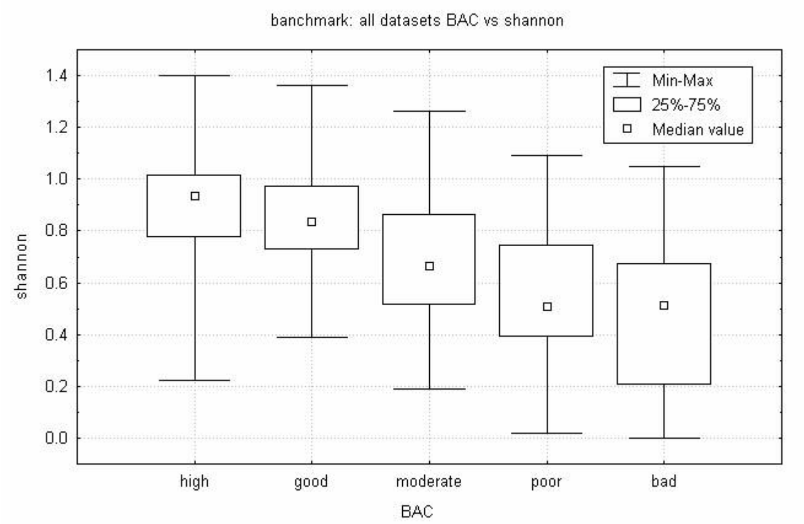


Figure 6.8 Box&whiskers representation for the benchmark datasets selected for all the types. BAC vs Shannon

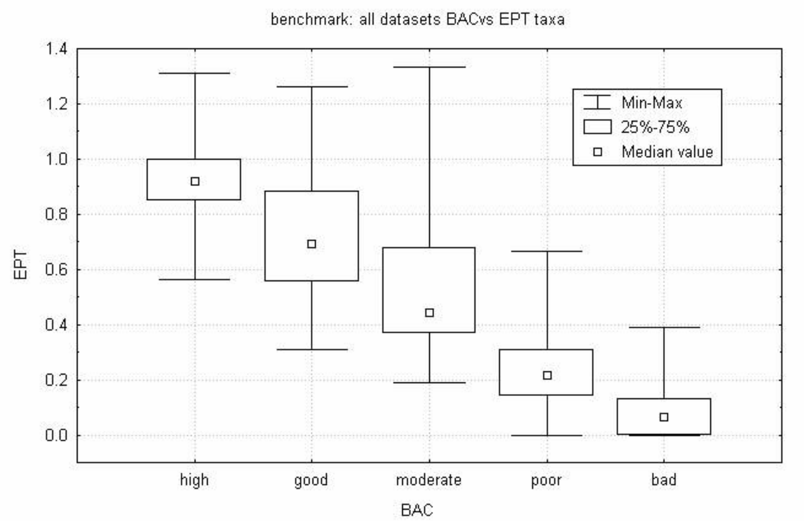


Figure 6.9 Box&whiskers representation for the benchmark datasets selected for all the types. BAC vs EPT

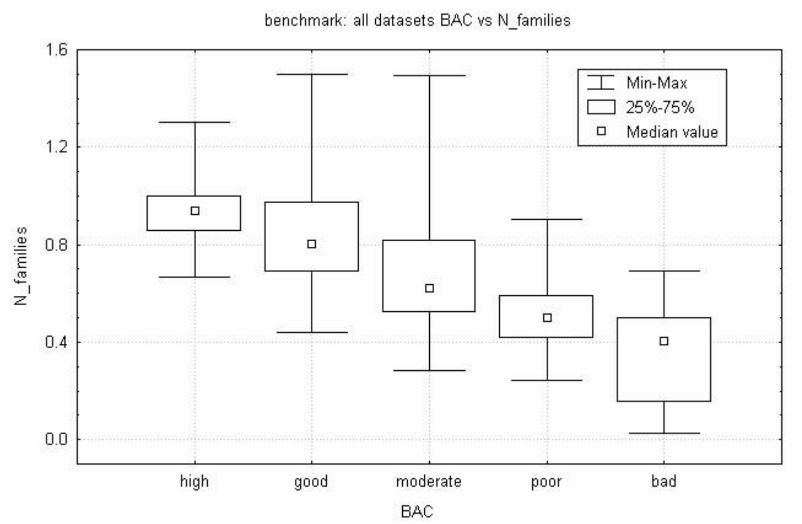


Figure 6.10 Box&whiskers representation for the benchmark datasets selected for all the types. BAC vs Number of families

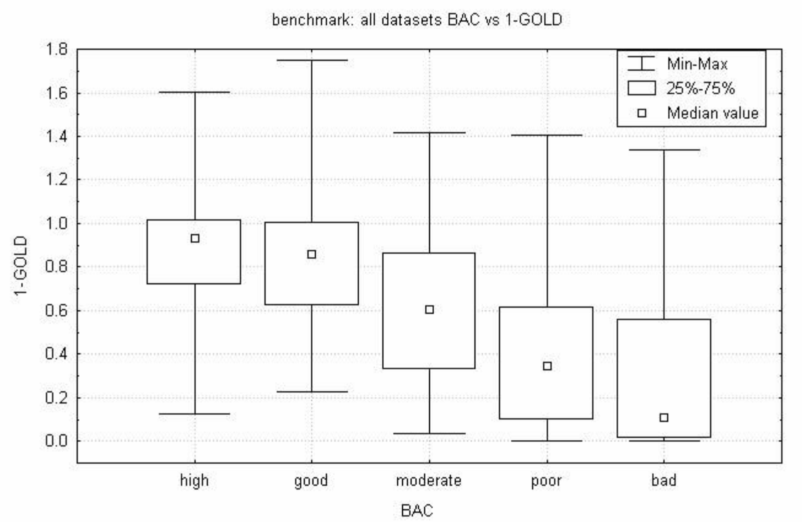


Figure 6.11 Box&whiskers representation for the benchmark datasets selected for all the types. BAC vs Log(1-GOLD)

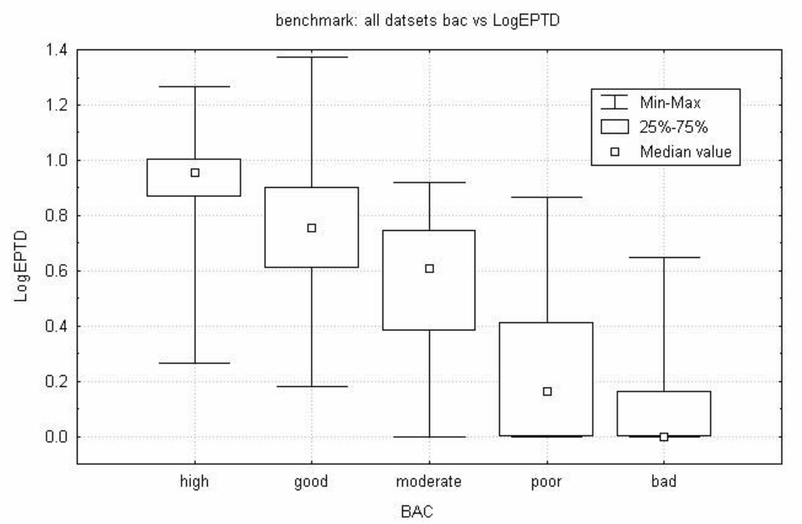


Figure 6.12 Box&whiskers representation for the benchmark datasets selected for all the types. BAC vs Log_EPTD

The trend of the median values shows a good response for all the metrics. For Number of families, EPT taxa, ASPT and LogEPTD the High and Good class are well separated, demonstrating the good performance of the single metrics. Also abundance based metrics perform well in the benchmark datasets. Some ICMs show a general overlap among classes, i.e. 1-GOLD and Shannon.

6.3 - Validation of ICMs and ICM index approach by pressures analysis

The potential of ICMs and ICM index in representing the ecological quality gradient as described by MSs' methods is evaluated in chapter 4. The aim of this paragraph is to present a concise illustration of the relationship between ICMs and pressures and to verify if ICMi provides a valid response in terms of general degradation.

For this purpose, data from two different stream types were considered: small mountain calcareous streams in southern Italy (R-M1 → based on STAR/AQEM benchmark dataset; see 5.6.3 and a part of test data see 4.6.2 sites from Campania) and small lowland, spring-fed streams in the Po valley, Italy (R-C1 → based on STAR/AQEM benchmark dataset; see 5.6.3). The pressures considered are: a) water pollution; b) general degradation of the site; c) morphological alteration; d) habitat degradation. They all refer to an on-site survey, i.e. no catchments scale data are being used for the examples. In addition, it must be noted how other pressures are acting on the investigated sites, such as e.g. hydrological disturbance at southern sites and pesticide contamination in the northern area. Nevertheless, we propose that the four aspects considered here can adequately account for a large portion of the stressors acting on the invertebrate community in those stream types. Water pollution (a) was evaluated through the quantification of the following parameters: O_2 (%) BOD5 (mg/l) *E. coli* (ufc/100ml) N-NO_3 (mg/l) N-NH_4 (mg/l) TP (µg/l). In particular, to represent water quality by means of a numerical indicator, according to the Italian legislation (D.L. 152/99) a score is given to each of the considered parameters, on the basis of its observed concentration (see also Buffagni et al., 2004a). The individual scores are then summed up into a single site index (here named D.L. 152).The general degradation of the site (b) was assessed through the application of an Italian index (Index of Fluvial Functioning, IFF: Siligardi et al., 2000). This index is useful in describing the overall quality of river sites (Balestrini et al., 2004), not assessing any specific aspects of river degradation. The degradation of stream morphology was also considered, by applying the River Habitat Survey method (RHS: Raven et al., 1998) and then calculating the Habitat Modification Score index (HMS)(c). Based on RHS as well, the Habitat Quality Assessment score (HQA) was also calculated, which embodies information on habitat diversification and quality (d). Each of the pressures was considered individually and in combination with the others (as an average).

In Table 6.3, a summary of the R^2 values for single ICMs and for the ICMi with respect to the considered pressures is reported for R-M1.

Table 6.3 R^2 coefficient for single ICMs and ICMi in relation to selected pressures for Italian R-M1 sites, in Southern Italy (n = 45)

	ASPT	Shannon	EPT	N_FAM	1-GOLD	LselEPTD	ICMi
chemical DL 152	0.39	0.37	0.48	0.32	0.36	0.35	0.47
general IFF	0.32	0.32	0.31	0.20	0.22	0.51	0.45
morphology HMS	0.05	0.08	0.04	0.02	0.01	0.18	0.09
Habitat HQA	0.39	0.46	0.37	0.29	0.26	0.54	0.54
Combined pressures	0.30	0.35	0.30	0.22	0.19	0.53	0.45
Combined pressures (excluding HMS)	0.42	0.45	0.43	0.33	0.30	0.63	0.60

The best performing metric is the Abundance of Selected taxa (LselEPTD), especially in respect of the HQA and IFF indices (R^2 of 0.54 and 0.51 respectively). EPT taxa are well correlated to chemical quality (R^2 =0.48), even with a correlation higher than ASPT (R^2 =0.39). The Shannon diversity index is well correlated with habitat quality (R^2 =0.48). Apart from the correlation values with HMS, all other R^2 are statistically significant ($p<0.05$). In general terms, the ICM index show higher R^2 values than single metrics in relation to both combined and single pressures. The relationship between ICMs and HMS is not apparent ($p>0.05$, for all the metrics, with the exclusion of Log sel EPTD) and for this reason the combined pressures were also considered excluding HMS. Possibly, in this stream type, more detailed methods based on invertebrates are needed to detect morphological degradation. In addition, the scarce or absent relationship with a specific pressure is not surprising, because the ICMi was developed for the detection of general degradation of a site.

In Table 6.4, the relationship between ICMs and pressures is reported for R-C1. On the whole, R^2 values are higher than 0.5 for all the metrics ($p<0.05$). On a single value for distinct ICMs, the best performing metric is 1-GOLD in relation to water chemical quality (R^2 = 0.79), together with the abundance of selected EPTD taxa in relation to combined pressures. The metric with the worst performance is Shannon diversity (with R^2 always lower than 0.4). The combination of the metrics into the ICMi shows a quite high R^2 value of 0.81 in relation to the combined pressures. The highest value for ICMi and single pressures is with respect to habitat quality and diversity (R^2 = 0.73), followed by chemical quality (R^2 = 0.67). In this river type, the relation of ICMs with morphological degradation is apparent as well. In fact, the alteration in stream morphology here, is often combined with e.g. the general degradation of the site, i.e. different pressures are strongly self-correlated.

Table 6.4 R^2 coefficient for single metrics and ICMi in relation to pressures for R-C1 sites in Northern Italy (n = 33)

	ASPT	Shannon	EPT	N_FAM	1-GOLD	LselEPTD	ICMi
chemical DL 152	0.67	0.38	0.54	0.64	0.79	0.47	0.67
general IFF	0.55	0.16	0.77	0.41	0.20	0.73	0.61
morphology HMS	0.45	0.26	0.55	0.49	0.43	0.52	0.55
Habitat HQA	0.67	0.23	0.80	0.50	0.50	0.77	0.73
Combined pressures	0.75	0.32	0.84	0.63	0.57	0.79	0.81

In conclusion, the proposed analysis of pressures confirms the overall good performance of the ICMs and ICMi approach in describing the environmental quality in the two examined stream types, which belong to different GIGs and show rather distant characteristics.

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

6.4.1 - 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 physico-chemical status and hydromorphological conditions. In particular for the Inter-calibration exercise, water bodies in large geographic areas need to be compared. Within many GIGs (Geographical Inter-calibration Groups), it was decided to compare the different assessment systems by means of 'Inter-calibration Common Metrics', which are suited to assess environmental degradation in a large variety of stream types, in accordance with the proposal arising from the STAR project (Buffagni & Erba, 2004). Though these metrics do not usually give 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 suitable 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 could in future be used:

- As 'Inter-calibration 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.

The aim of this chapter is to illustrate some analyses carried out on the STAR and AQEM data to evaluate the performance of different metrics in assessing different impact types over a wide range of types grouped into two main sub-groups. The better performing metrics could be those proposed as ICMs.

6.4.2 - Database and methods

Database

The analysis was restricted to the two largest Stream Type Groups defined for STAR, which represent many IC stream types and cover a wide geographical area: the "Central Lowland" and "Central Mountain" groups (see Table 6.5). 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 6.5 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	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 (see Table 6.10): i) physical-chemical measures (organic pollution/eutrophication), ii) hydromorphological parameters (hydromorphological/general degradation), and iii) land use parameters (organic pollution, general degradation). Table 6.6 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 6.6 Number of environmental variables used for the analysis of the main stressors. A complete list is given in Table 6.10

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 6.6.

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 6.6 and 6.10.

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.
3. 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.
4. 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 (Tables 6.11 and 6.12).

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. Tables 6.11 and 6.12 show the 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) (see Table 6.7). 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, apart from some datasets included in the benchmark dataset (see chapter 5) but not considered here.

In conclusion, the classification applied is mainly coherent to the "Best Available Classification" that is preliminarily used as a benchmark within the Inter-calibration exercise (see Chapter 5).

Table 6.7 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 (see Table 6.7). 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 (see Figure 6.15).

6.4.3 - Stream Type Group 1: "Central Lowland"

Environmental gradients (PCA)

Figure 6.13a 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.

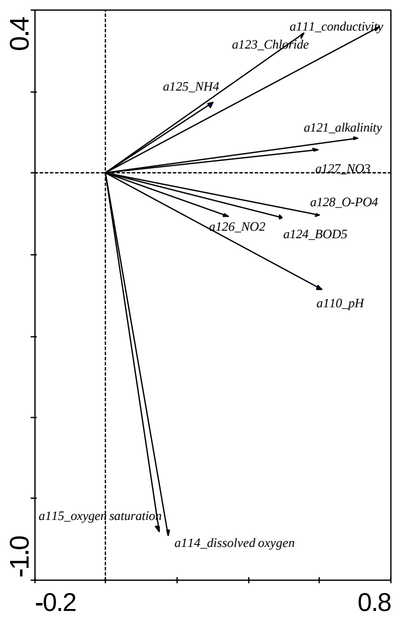


Figure 6.13a. PCA of eleven physical-chemical variables of Stream Type Group 1, axis 1 *vs*. 2. The PC1 gradient is represented by axis 1.

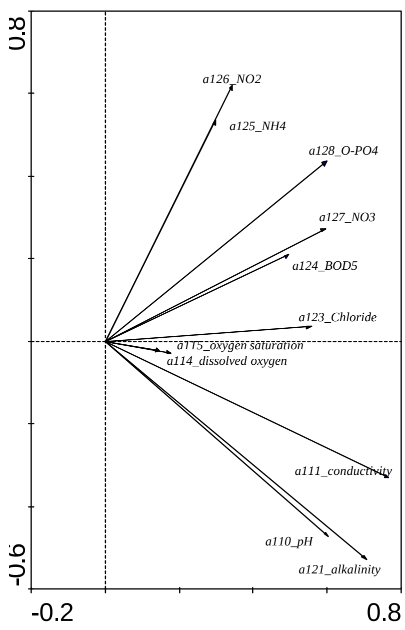


Figure 6.13b: PCA of eleven physical-chemical variables of Stream Type Group 1, axis 1 *vs*. 3. The PC1 gradient is represented by axis 1.

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 (see Table 6.8).

Table 6.8 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

	Axes				
	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

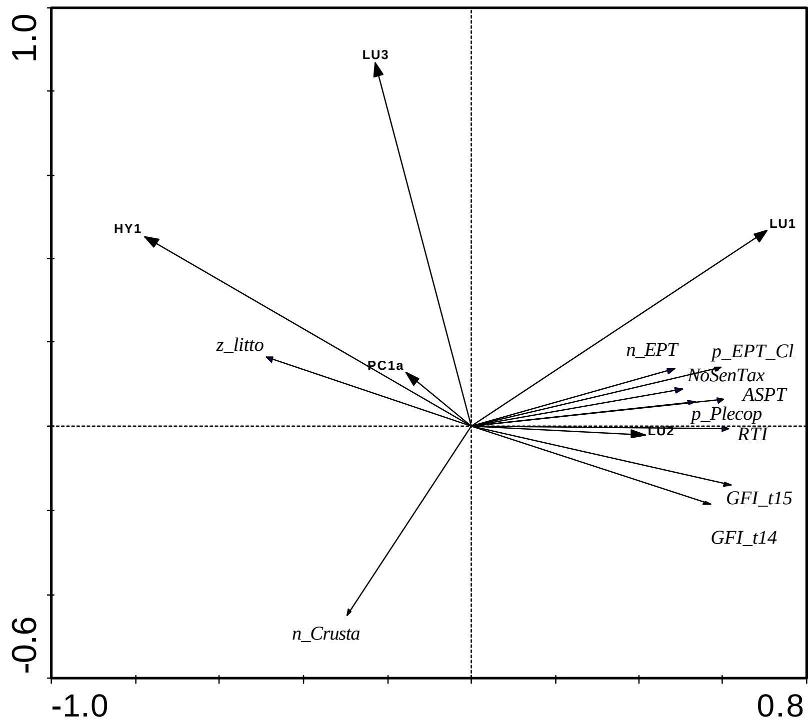


Figure 6.14 RDA biplot of 85 metrics, 5 gradients, and 387 samples of STG 1. For clarity, the 'species fit' was set to > 25 % in order to show the ten strongest metrics in the analysis. Metric codes: n_EPT = number of Ephemeroptera-Plecoptera-Trichoptera taxa; p_EPT_cl$^{(*)}$ = % EPT based on abundance 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 (see Figure 6.14). 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 Table 6.11.

6.4.4 - 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 (see Table 6.9). 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 6.9 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

	Axes				
	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 (see Figure 6.15). 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 (see Figure 6.15). 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.

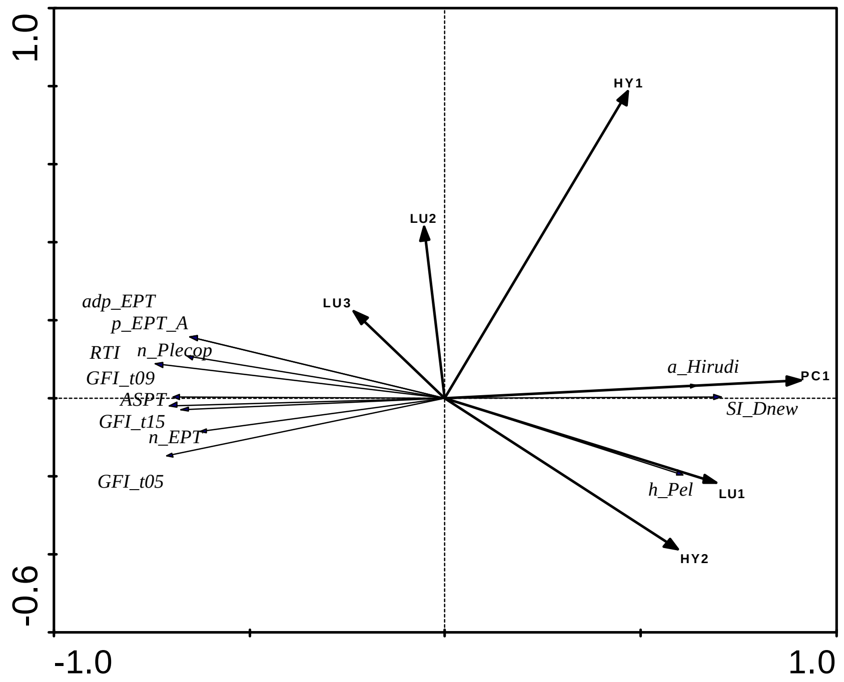


Figure 6.15 RDA biplot of 102 metrics, 6 gradients, and 295 samples of STG 2. For clarity, the 'species fit' was set to > 40 % in order to show the twelve strongest metrics in the analysis. Metric codes: adp_EPT = number of Ephemeroptera-Plecoptera-Trichoptera taxa, adjusted; p_EPT_A(*) = % EPT Austrian 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 Table 6.12.

6.4.5 - Conclusion

The data evaluation has confirmed 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 apparent that specific individual metrics can react differently to different stressors. While many metrics based on species level strongly correlate to environmental gradients, some family-based metrics have a comparatively good performance, too. Because for the Inter-calibration exercise ICMs are likely to be selected, these results attest to a sound scientific basis for their potential use. In addition, as is evidently demonstrated elsewhere in the present Paper, it must be borne in mind that the combination of single metrics into an ICM index clearly increases the performance in describing the quality gradients. Since the AQEM/STAR dataset is the first pan-European benthic invertebrate data set, some ICMs for the Central and Baltic GIG and, if feasible, also for the Nordic GIG, might be selected from Tables 6.10 and 6.11. These should mainly be restricted to family-based metrics, to also allow for data comparison from those countries, which do not have datasets on species level. Overall, the ICMi selection might address the inclusion of metrics reacting on different stressors. As a general conclusion the analyses carried out on the STAR and AQEM data related to the two groups 'Central lowland' and 'Central Mountain' confirm the good performance of the ICMs selected for the pilot Inter-calibration exercise. ASPT shows good relationships both with hydromorphological quality ($r = -0.58$) and organic pollution ($r = -0.73$); its performance is comparable to that of the metrics for which species identification is needed. 1-GOLD resulted as one of the best performing metrics for the detection of organic pollution, while the abundance of selected EPTD taxa is one of the best metrics for the detection of hydromorphological degradation.

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 nearby 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 but owning the expertise of species level identification.

Table 6.10 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. Next 3 pages

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	+	+

continued

Table 6.10 (continued)

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 [% length]	HY	+	+
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	+	+

continued

Table 6.10 (continued)

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	+	+

Table 6.11 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

continued

Tab. 6.11(continued)

Order	Metric short	Metric name	Metric with HY1		Metric with pre-/post-classification		Metric with HY1: stream type-specific		
			r	p	r	p	r mean	r max	r min
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 preferences (Moog, 1995)	0,47	< 0,001	-0,61	< 0,001	0,47	0,58	0,29
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

continued

Tab. 6.11 (continued)

			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
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
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

Table 6.12 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

continued

Tab. 6.12 (continued)

			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
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
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

continued

Tab. 6.12 (continued)

			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
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
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, whose assessment systems are different in terms of taxonomic resolution and general approach, the Inter-calibration using 'Inter-calibration Common Metrics' (ICM) (Buffagni & Erba, 2004) is a suitable procedure. However, it might result in being difficult to explain to water managers and the general public over the short period of the WFD IC process. Thus, we outline an alternative, which is based on a simple comparison of assessment results from national assessment systems, without using the ICM-tool. The 'direct comparison approach' can be used within trans-boundary river catchments and could also serve as an alternative or validation of the pan-European ICM approach.

In this section, the 'direct comparison approach' is exemplified on the basis of four case studies comprising assessment methods of a total of nine countries, using benthic invertebrates and macrophytes and covering the common Inter-calibration 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 1 and System 2) are calculated with a number of samples. The results are compared by a regression which leads to a "conversion formula" from System 1 into System 2 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 (see Table 1 and 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 European Commission, 2003c.

2. ***Calculation of index values of all methods included in the comparison for each sample in the dataset.***

   Benthic Invertebrates: For intercalibration stream type R-C3 six and for intercalibration 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 intercalibration stream types R-C3 and R-C4 three assessment indices are calculated (see 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 in this study the high|good and good|moderate boundary values are expressed as EQR values, following the WFD requirements. This also allows for integration of assessment methods directly 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. Next 2 pages

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	Swedish Environmental Protection Agency 2000
				DSFI (SE) – Danish Stream Fauna Index applied in Sweden	

continued

Table 1 (continued)

		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

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 = - 0.34 (Slovak SI and British ASPT) to r = 0.86 (Austrian SI and German SI). 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 (see 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 (see Table 3). To compare the quality classes the boundary values of all indices are converted into values of the ASPT-scale (see Figure 1).

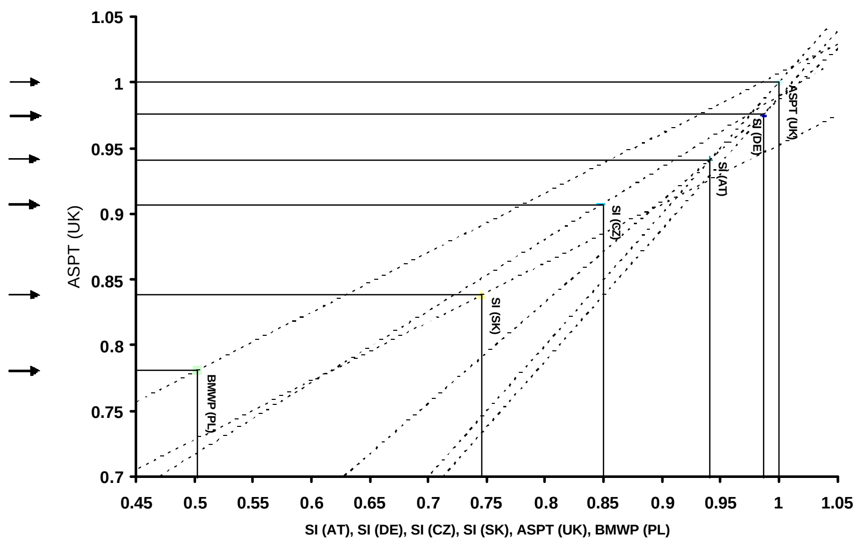
For the high|good status boundary the largest deviation amounts to >0.2 ASPT-EQR units between ASPT (UK) and BMWP (PL). The smallest difference is between ASPT (UK) and SI (DE) (0.025 ASPT-EQR units).

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 Values of high|good and good|moderate boundary values derived by regression analysis

R-C3			SI (AT)	SI (DE)	SI (CZ)	SI (SK)	ASPT (UK)	BMWP (PL)
benchmark for comparison	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**

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



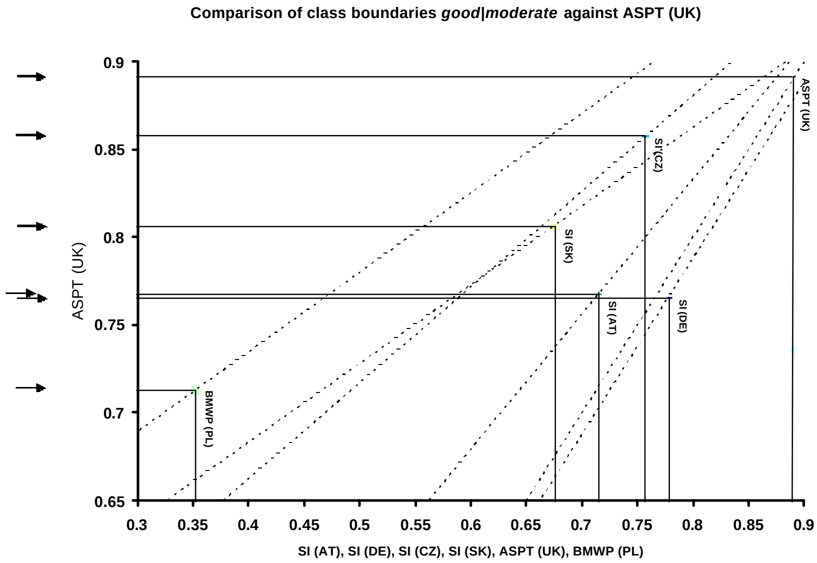


Figure 1 Class boundary comparisons through conversion of national boundaries into ASPT-EQR units using regression lines. Diagrams depict dotted regression lines between normalised values of various national indices (abscissa) and ASPT (ordinate), regression formulae are specified in the annex. Solid lines represent the position of class boundaries on national (abscissa) and benchmark scale (ordinate). The latter are marked by small arrows

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

Correlation and regression - Spearman correlation coefficients range from -0.75 (German SI and DSFI) to 0.79 (ASPT 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) (see 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
95^{th} 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) (see Table 5 and 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 Values of high|good and good|moderate boundary values derived by regression

R-C4			**SI (DE)**	**ASPT (UK)**	**ASPT (SE)**	**DSFI (DK)**	**DSFI (SE)**
benchmark for comparison	**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**

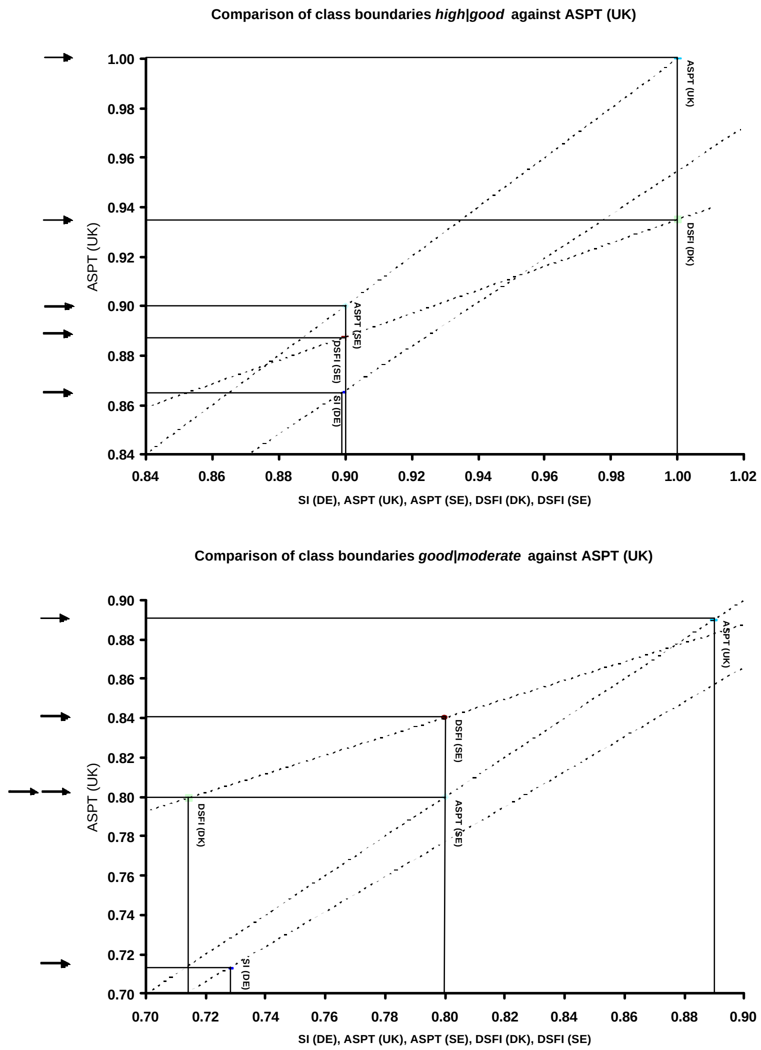


Figure 2 Class boundary comparisons through conversion of national boundaries into ASPT-EQR units using regression lines. Diagrams depict dotted regression lines between normalised values of various national indices (abscissa) and ASPT (ordinate), regression formulae are specified in the annex. Solid lines represent the position of class boundaries on national (abscissa) and benchmark scale (ordinate). The latter are marked by small arrows

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 0.78 (German RI and British MTR) and 0.93 (British MTR and French IBMR). 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 (see 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 (see Table 7 and 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 Values of high|good and good|moderate boundary values derived by regression analysis

R-C3			**MTR (UK)**	**IBMR (FR)**	**RI-Moose (DE)**
benchmark for comparison	**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**

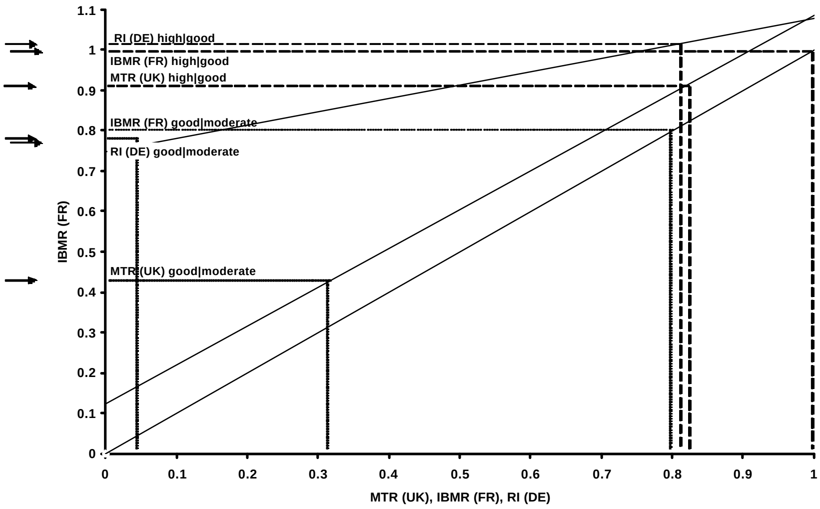


Figure 3 Class boundary comparisons through conversion of national boundaries into IBMR-EQR units using regression lines. The diagram depicts regression lines between normalised values of various national indices (abscissa) and IBMR (ordinate), regression formulae are specified in the annex. Dashed lines represent the position of class boundaries on national (abscissa) and benchmark scale (ordinate). The latter are marked by small arrows

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 (see 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 (see Table 9 and 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 Values of high|good and good|moderate boundary values derived by regression analysis

R-C4			MTR (UK)	IBMR (FR)	RI (DE)
benchmark for comparison	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**

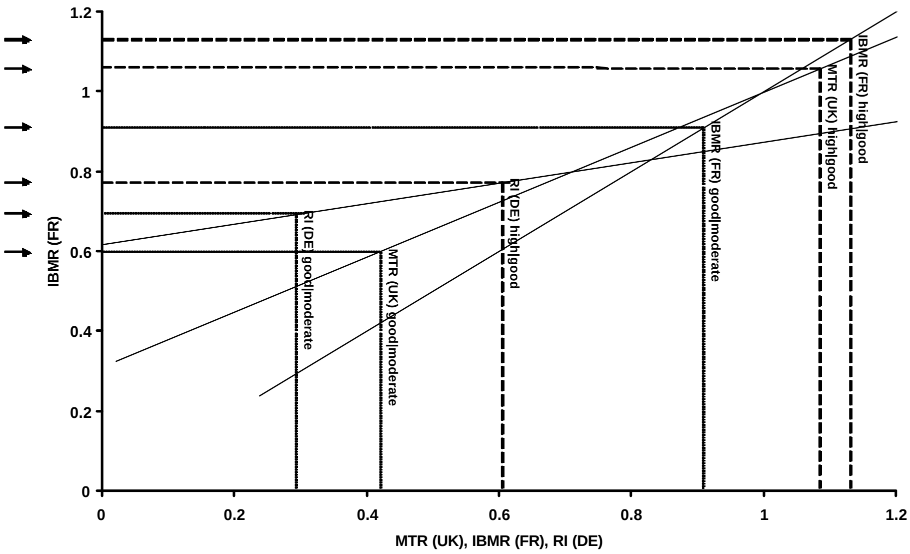


Figure 4 Class boundary comparisons through conversion of national boundaries into IBMR-EQR units using regression lines. The diagram depicts regression lines between normalised values of various national indices (abscissa) and IBMR (ordinate), regression formulae are specified in the annex. Dashed lines represent the position of class boundaries on national (abscissa) and benchmark scale (ordinate). The latter are marked by small arrows

Discussion

Reference values - In the Inter-calibration exercise class boundaries expressed as EQR values are compared. The pre-requisite for Inter-calibration 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 from 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 7.4). Austria defines the median of the Saprobic Index of all available reference sites of a certain type as the saprobic basic condition (i.e. reference value). In Germany saprobic reference values have been derived by taking the 9oth percentile of all available sites (minus double * 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 facilitates the comparison of EQR class boundary values, even if no nationally defined method references are available. Furthermore, the comparison is based on homogeneously derived reference values. Nevertheless, the calculated boundaries have to be considered tentative, as they are in most cases calculated based on the pre-classification of sites and they will still need 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.

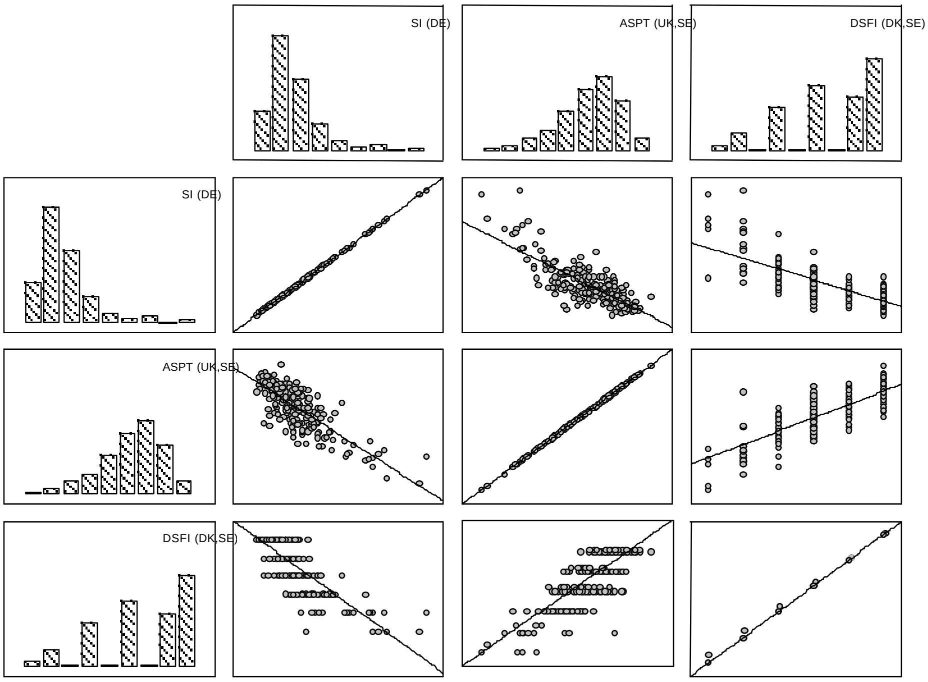
The example of the 'direct comparison approach' reveals major differences between class boundary settings of the methods included. Nevertheless, the significance of discrepancies between the individual methods needs to be specified. Additional analyses have to consider e.g. the level of confidence resulting 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 for more than 20 years. The applicability of WFD for Inter-calibration purposes is shown in this study. Particularly, if EQR values based on reference conditions are used, the National Methods can easily be compared with 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 harmonized 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.

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)	0.00	1.00	0.26	0.88	1.04	0.49	0.78	0.61	3.07	-0.21	2.26	0.00
SI (DE)	0.30	0.79	0.00	1.00	1.01	0.47	0.94	0.47	3.24	-0.24	2.27	0.00
SI (CZ)	-0.79	1.26	-0.84	1.34	0.00	1.00	0.08	0.85	3.36	-0.30	2.17	-0.01
SI (SK)	-0.53	1.20	-0.17	1.03	0.64	0.65	0.00	1.00	2.91	-0.21	2.08	0.00
ASPT (UK)	10.29	-2.17	11.28	-2.85	8.28	-1.24	8.31	-1.14	0.00	1.00	4.14	0.02
BMWP (PL)	304.94	-94.48	342.73	-121.07	217.94	-54.59	219.11	-49.86	-121.80	40.24	0.00	1.00

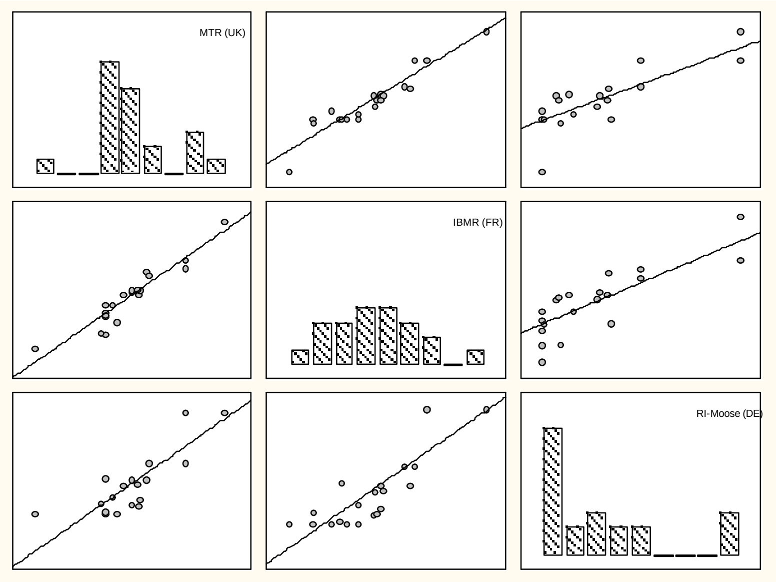
R-C3: Regression formulae (EQR values; based on 95th percentile of high status sites)

	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)	0.00	1.00	0.08	0.88	0.37	0.61	0.29	0.68	0.35	0.58	0.66	0.28
SI (DE)	0.21	0.80	0.00	1.00	0.43	0.58	0.45	0.53	0.29	0.69	0.65	0.33
SI (CZ)	-0.08	1.02	-0.16	1.07	0.00	1.00	0.16	0.76	0.19	0.69	0.55	0.34
SI (SK)	-0.09	1.07	0.02	0.92	0.26	0.72	0.00	1.00	0.37	0.54	0.65	0.26
ASPT (UK)	0.22	0.77	-0.01	1.00	0.44	0.55	0.50	0.45	0.00	1.00	0.55	0.45
BMWP (PL)	-0.37	1.26	-0.71	1.61	0.00	0.91	0.10	0.74	-0.61	1.51	0.00	1.00

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 MTR (UK), IBMR (FR), RI (DE)

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)	0.00	1.00	3.35	-0.24	3.35	-0.24	2.70	-0.13	2.70	-0.13
ASPT (UK)	11.06	-2.65	0.00	1.00	0.00	1.00	3.21	0.47	3.21	0.47
ASPT (SE)	11.06	-2.65	0.00	1.00	0.00	1.00	3.21	0.47	3.21	0.47
DSFI (DK)	14.24	-4.41	-2.68	1.41	-2.68	1.41	0.00	1.00	0.00	1.00
DSFI (SE)	14.24	-4.41	-2.68	1.41	-2.68	1.41	0.00	1.00	0.00	1.00

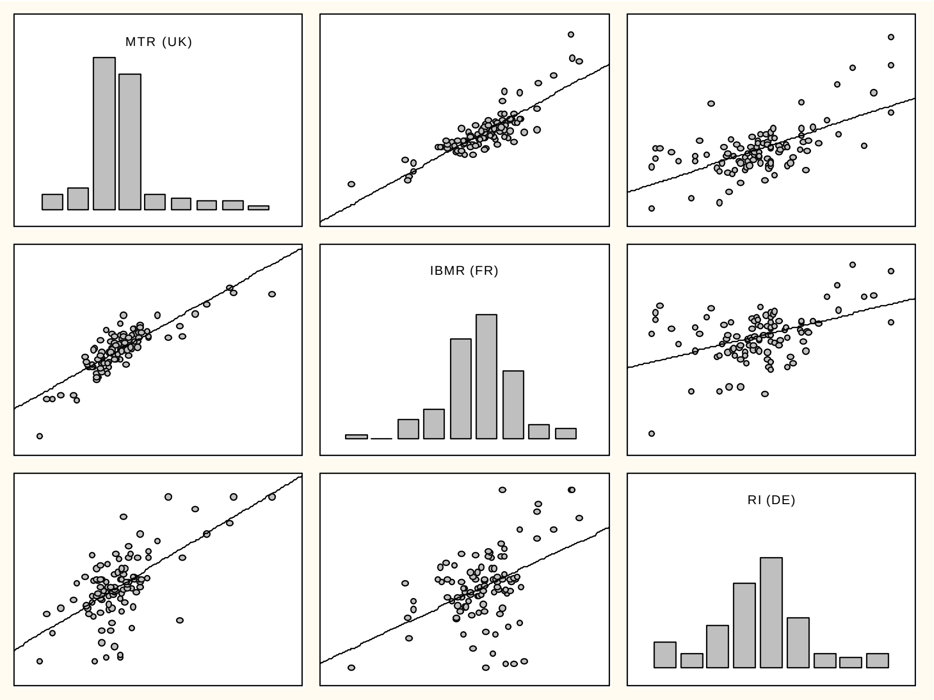
R-C4: Regression formulae (EQR values; based on 95th percentile of high status sites)

	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)	0.00	1.00	0.28	0.70	0.28	0.70	0.55	0.39	0.55	0.39
ASPT (UK)	0.07	0.89	0.00	1.00	0.00	1.00	0.46	0.47	0.46	0.47
ASPT (SE)	0.07	0.89	0.00	1.00	0.00	1.00	0.46	0.47	0.46	0.47
DSFI (DK)	-0.49	1.47	-0.38	1.40	-0.38	1.40	0.00	1.00	0.00	1.00
DSFI (SE)	-0.49	1.47	-0.38	1.40	-0.38	1.40	0.00	1.00	0.00	1.00

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

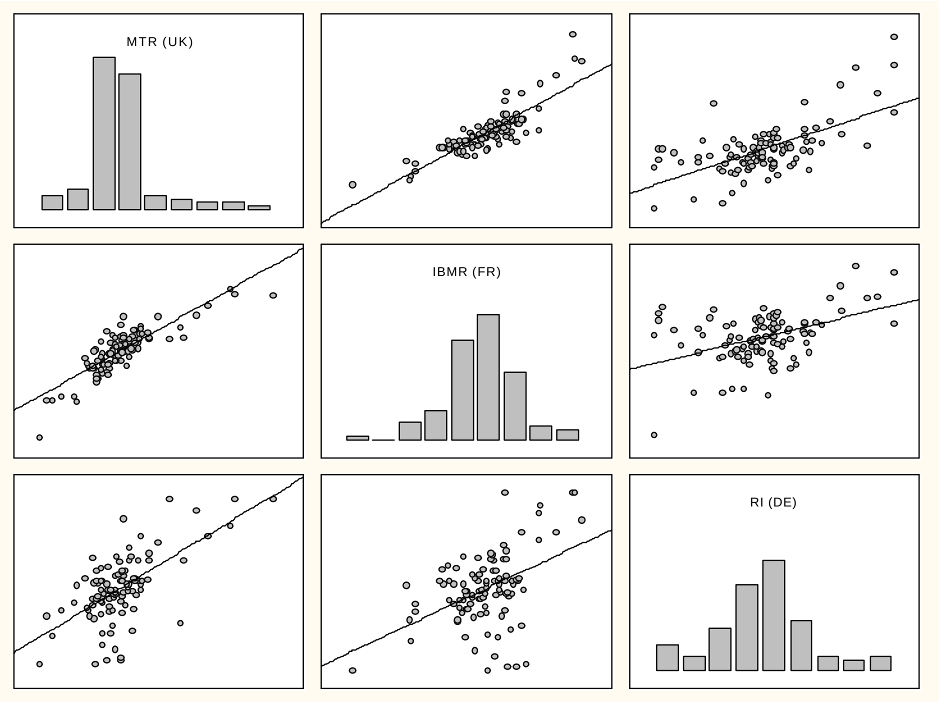
R-C3: Regression formulae (EQR values; based on 95th percentile of all AQEM/STAR samples)

	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 of National Methods' class boundaries through conversion in ICMi value: intra- & inter-GIG

The results of the conversion of the class boundaries of the national assessment methods in the ICMi are discussed here (see also the Test datasets description, Chapter 4). The first part of the paragraph presents the results for all the specified datasets listed below (see Table 7.10) in the comparison phase.

Table 7.10 Characteristics of the tested methods and their compliance with the WFD requirements

MS	IC type	Classification method	Type specific adaptations	tolerance	abundance	richness	available criteria for reference condition
Belgium	C1	MIF	Y	Y	Y	Y	nk
Denmark	C1	DSFI	Y	Y	Y	Y	N
Estonia	C1	ASPT	N	Y	Y/N	Y	N**
France	C1, C2, M1	IBGN	Y	Y	Y/N	Y	Y
Germany	C1	SI(DE) & GD(DE)	Y	Y	Y	Y	Y
Italy	C1, M1, M5	IBE	N	Y	Y/N	Y	N
The Netherlands	C1	KRW	nk	nk	Y	nk	N
Poland	C1	BMWP & Margalef div. ind.	N	Y	Y/N	Y	nk
Spain	C2	MMI-Spain	Y	Y	Y	Y	Y
Uk	C1	EQI-ASPT & EQI-NFAM	Y*	Y	Y/N	Y	Y

* = site specific prediction; ** = under development

7.2.1 - Intra-GIG comparison within single stream types

The compared Test datasets belong here to IC stream type R-C1 (9 countries, see Table above). 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 the 75th percentile. In the same way, values of the National Methods 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 the following graphs the different types and countries are represented by a letter; in Table 7.11 the correspondence between letters and countries is reported. A summary of the analyses done for each dataset is also reported, in particular it is indicated if the type was simply compared to the other datasets or if it was also harmonized (see chapter 8).

For the comparison of the different types it is important to check if the ICMi shows a good correlation with the National Method for all the countries under consideration. Summary results of the R^2 values for the regressions used, considering ICMi as dependent variable (y axis) and National Method as independent variable (x axis), are shown in Figure 7.5; the results are ordered considering decreasing R^2 values. For countries considering two indices for the classification, the shown R^2 value is the mean between the R^2 of the two systems.

Table 7.11 Correspondence between letter and countries for the interpretation of the graphs. *indicates WFD compliant methods

	IC River type	Country	Compared	Harmonized (via median)
A*	R-C2	France	Y	Y
B*	R-C2	Spain	Y	Y
C	R-M1	Italy	Y	Y
D	R-C1	Estonia	Y	Y
E*	R-C1	France	Y	Y
F	R-C1	Italy	Y	Y
G	R-C1	Poland	Y	Y
H*	R-C1	UK	Y	Y
I*	R-C1	Germany	Y	Y
L	R-C1	Denmark	Y	Y
M*	R-M1	France	Y	Y
N	R-M5	Italy	Y	No (National method has to be refined)
O*	R-C1	Belgium	Y	not yet
P	R-C1	the Netherlands	Y	No (further investigation needed)

For most of the methods, as was seen in the separate consideration of each type in chapter 4, 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 The Netherlands (P) with a R^2 value of 0.18 (see chapter 4.3.7 for possible explanations).

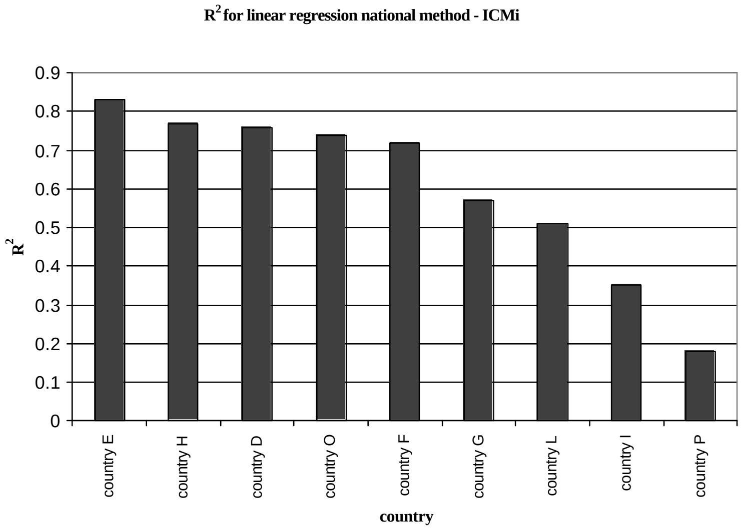


Figure 7.5 R^2 for linear regression, C1 test datasets. p<0.001 for all data

The mediocre result for Denmark (L) (R^2 = 0.52) can be explained with the low values that some ICMs (i.e. Shannon index and 1-GOLD) have, at some presumably slightly impacted sites. Also, some taxa selected for the metric Log_EPTD may occur rarely in the considered streams. Additionally, in this country it has been difficult to find reference data for C1 streams. Furthermore, a short quality gradient is observed, since only about 4% of sites fall into Poor and Bad classes. Also, it must be noted that the DSFI index is not continuous and this can affect the regression, implying poor R^2 results. A detailed description and comments on regressions can be found in the paragraph 4.3.2.

A list of possible hypotheses for consideration when low correlations of ICMi *vs* National method values arise, are presented below.

As a first point, the structure of the data should be checked, in order to avoid the following conditions:

- Differences in the sampling method within the same dataset may occur. In this case, the subsets of data should be normalized separately.
- 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 causes for poor correlation:

- For particular stream types or for not yet validated methods, attention should be paid to the capacity of the method in describing in a correct 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 macroinvertebrate taxa 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 for consideration of low correlations between ICMi and National Method in countries with stressor specific assessment modules (e.g. Germany):

- 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 invertebrate community (e.g. organic pollution). This would lead to overall higher values for most biological metrics corresponding to low values for the National assessment system.
- When invertebrates are identified to species level and data refer to large geographic areas, it can be that natural variability among communities is high (at least comparatively higher than that observed for data at the family level).
- If none, only one or a few sites belong to High/Good status classes and a few are classified as Bad status as well, 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 (for a single BQE), 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 fairly lower) 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 equally broad.

In Table 7.12 the results of the conversion of the boundaries of the National Method into ICMi values obtained by means of linear regression for the R-C1 types are reported (see also each Test dataset description).

Table 7.12 Results of the conversion of National Method boundaries into ICMi

	R-C1											
	BE	DK	EE	FR	DE		IT	PL		NL	UK	
					GD	SI		BMWP	Marg		ASPT	#Fam
Tot # sites	208	347	23	132	91	91	361	49	49	79	789	789
# High	10	29	9	64	1	2	84	11	19	12	317	328
ICMi H/G boundary	0.84	0.96	0.89	0.82	0.88	0.85	0.84	0.83	0.65	1.03	0.86	0.83
ICMi median H/G boundary							**0.84**					
# Good	26	178	6	29	15	77	176	16	15	12	269	259
ICMi G/M boundary	0.62	0.76	0.68	0.71	0.77	0.58	0.63	0.61	0.54	0.94	0.69	0.67
ICMi median G/M boundary							**0.67**					

For the High/Good boundary, it is possible to identify a group of countries with a boundary value around 0.9 (EE and DE), another with values around 0.85 (B, FR, IT PL–BMWP, DE-SI and UK) and a third one with values close to 1 (NL and DK). The high values observed in the Netherlands and in Denmark can be explained by the absence of reference sites for this stream type. In general terms, the actual ICMi values observed at High status sites are comparatively low (i.e. no true reference sites available) with respect to other countries and this leads to higher EQR values after normalization. All the Inter-calibration procedures tested in the present Paper need data from reference sites (observed or reconstructed by models) and the results from such circumstances must then be used as a guide for further studies and comparison. The Margalef index in Poland showed the lowest observed value.

The same results reported in Table 7.12 are summarized in a Box&Whiskers diagram (see Figure 7.6). The values of the boundaries expressed as ICMi after the conversion from the National boundary value are shown for R-C1 type. In the same Figure, the resulting boundaries for Moderate/Poor and Poor/Bad are also indicated. In Table 7.13, the statistical descriptors for the same ICMi values and boundaries are reported.

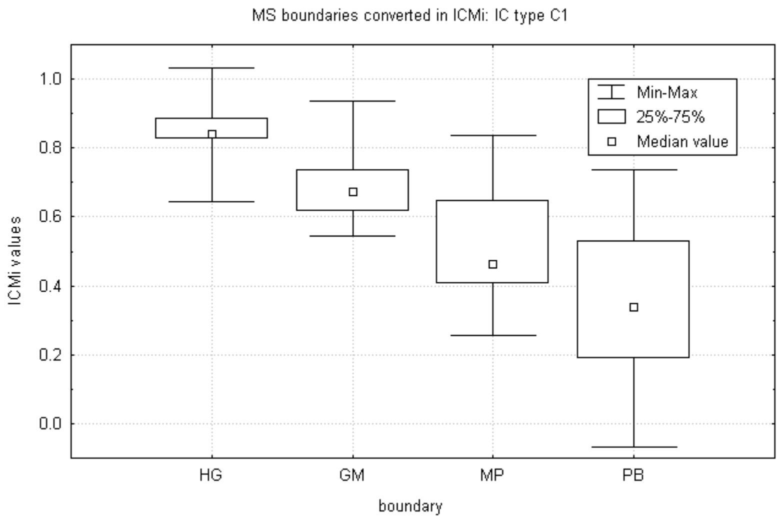


Figure 7.6 Box and whiskers for class boundaries converted in ICMi in R-C1 type (9 countries)

Table 7.13 Values of minimum, 25th percentile, median, 75th percentile, maximum, mean and standard deviation of class boundaries in ICMi for the R-C1 type (9 countries)

	min	max	25th %ile	75th %ile	median
HG	0.645	1.034	0.826	0.886	0.841
GM	0.543	0.935	0.617	0.738	0.671
MP	0.255	0.836	0.405	0.648	0.464
PB	-0.067	0.737	0.189	0.531	0.340

Figure 7.6 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 shows no overlap among classes, while maximum values tend to show a relatively larger variability for lower class boundaries. Not surprisingly, Mann Whitney-U test (Tab. 7.14) shows significant differences among all the four boundaries (seven countries). Nevertheless, this shows how the conversion to ICMi, with related normalization, maintains the boundary values for the quality classes clearly separated.

Table 7.14 p-level for Mann-Whitney U test: R-C1 datasets, MSs' existing boundaries (9 countries)

	HG	GM	MP	PB
HG		0.002	<0.001	<0.001
GM	0.002		0.008	<0.001
MP	<0.001	0.008		<0.001
PB	<0.001	<0.001	<0.001	

IMPORTANT WARNING

Country L (Denmark) has only few reference sites for the R-C1 type. The normalization of ICMi values and metrics might thus have lead to an inconsistent range of values (i.e. not the full quality gradient covered by the samples available).The same situation can be found in other countries (e.g. The Netherlands) and stream types. This might partly explain the high positive differences between the observed values and the median values (for WFD-compliant methods). In other words, the actual boundaries set in Denmark might not be as over-protective as they appear.
Also, the Danish system contains only a few categorical index classes (i.e. it is very unlike a continuous system). The calculation option here used to derive the boundary values is based on the minimum obtainable value, not on the average between values from two adjacent classes. Thus, most of the difference could be due to an artefact, because the Danish system has fewer categorical values than the other methods considered.

7.2.2 - Inter-GIG comparison: concurrent comparison of all stream types and methods

Class boundary values expressed as ICMi derived from National boundaries can also be compared considering different types and GIGs. Thus, in this paragraph the discussion of the results of the conversion of the class boundaries of the National assessment methods to the ICMi refers to different GIGs and types. The considered datasets include types M1, M5, C2, for a total of 5 datasets/countries. For type C1, three datasets were selected randomly and a mean value was considered. In Chapter 4, the description of such datasets and the regression between ICMi and National Methods is reported. Similarly to the previous paragraph, the R^2 values for the linear regression between National Methods and ICMi were analyzed (Fig. 7.7).

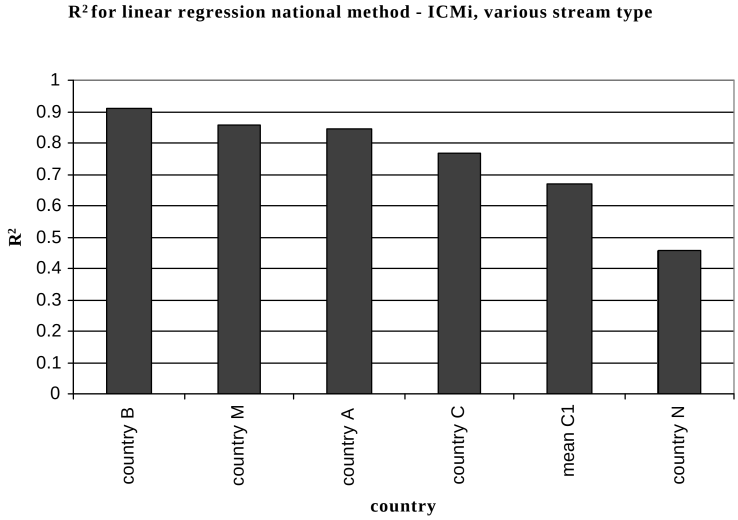


Figure 7.7 R^2 for linear regressions, test datasets from various IC types ($p<0.001$ for all datasets)

All the methods, except one, show a R^2 higher than 0.60, in four cases higher than 0.70. Thus, the ICMi fits well with National Methods in very different stream types. The lowest correlation is observed for type R-M5 in Italy (country N) (R^2= 0.46). In this instance, due to the peculiar 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 types (Buffagni et al., 2004b). Besides, specific approaches, methods and metrics are likely to be needed for special categories of stream types, such as temporary rivers. The applicability itself of National Methods should be carefully checked for their use in such stream types. In this context, a discussion on the more appropriate ICMs to be used in type R-M5 is at the moment in progress among Mediterranean GIG partners.

In Table 7.15 the values of class boundaries expressed as ICMi derived from linear regression with national standard methods are given. The boundary High/Good is similar for Spain (R-C2) France R-M1 and Italy (R-M1 and R-M5). Comparatively lower values are instead found for France R-C2. The boundary

Good/Moderate seems similar in France and Italy (for R-M1 and R-M5) while it is a little bit lower in R-C2 type.

Table 7.15 Results of the conversion of National Method boundaries into ICMi

	R-C2		**R-M1**		**R-M5**
	ES	FR	FR	IT	IT
Tot # sites	46	143	77	63	37
# High	7	73	28	21	1
ICMi H/G boundary	0.92	0.79	0.87	0.90	0.91
# Good	25	17	4	38	16
ICMi G/M boundary	0.62	0.67	0.75	0.72	0.72

The distribution of the values of the class boundaries converted in ICMi is shown in a Box&Whiskers representation (see Figure 7.8), with the related results, referring to the main statistical descriptors, in Table 7.16. Italy R-M5 data were excluded from the calculation, due to the particular feature of the stream type (triangles in Fig. 7.8).

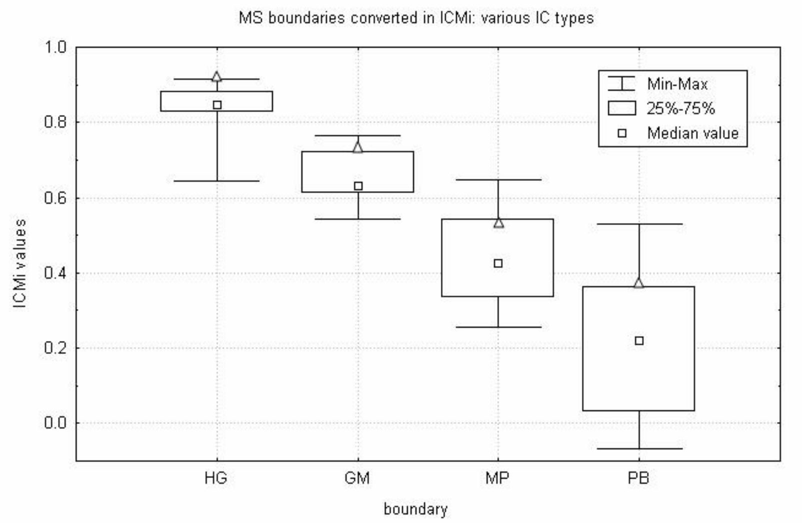


Figure 7.8 Box and whiskers for class boundaries converted in ICMi, various stream types (7 countries). In triangles: Italy R-M5 data

Table 7.16 Values of minimum, 25th percentile, median, 75th percentile, maximum, mean and standard deviation of class boundaries for ICMi, various stream types (7 datasets)

	min	max	25th %ile	75th %ile	median
HG	0.645	0.915	0.827	0.884	0.846
GM	0.543	0.766	0.612	0.722	0.631
MP	0.255	0.648	0.334	0.543	0.426
PB	-0.067	0.531	0.032	0.364	0.220

Even considering data from different stream types and GIGs, the interquartile ranges show nearly no overlapping between boundaries. Thus, quite obviously Mann Whitney U test shows significant differences among the four boundaries.

Table 7.17 p-level for Mann-Whitney U test: various IC type (7 datasets)

	HG	GM	MP	PB
HG		<0.001	<0.001	<0.001
GM	<0.001		0.002	<0.001
MP	<0.001	0.002		0.03
PB	<0.001	<0.001	0.03	

7.2.3 - Inter-GIG comparison for WFD-compliant methods

The compliance to WFD for the methods under consideration will not 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, required in any case for the aim of the present exercise and for fully WFD-compliant methods, has not been stringently considered because none of the methods take it carefully into account. An indication of WFD-compliance can be found in Table 7.10. For technical reasons, Belgium's data was not included even if WFD-compliant.

As in the above paragraphs, in Table 7.18 the values of class boundaries expressed as ICMi derived from linear regression with National Standard Methods, are given for methods considered WFD-compliant. The boundary High/Good is close to the median value for all the countries, with the exception of Spain for which the highest boundary value (0.92) was found. Comparatively lower values are found instead for France (R-C2)(see discussion on normalization options for single French datasets). A higher distance from the median value is observed for the boundary Good/Moderate. The values closest to the median out of all methods are for the UK, FR (R-C2) and Spain. A difference of around 0.10 is instead found for the other boundaries.

Table 7.18 Results of the conversion of National Method boundaries into ICMi for WFD compliant methods

	R-C1					**R-C2**		**R-M1**
	FR	DE		UK		ES	FR	FR
		GD	SI	ASPT	#Fam			
Tot # sites	132	91	91	789	789	46	143	77
# High	64	1	2	317	328	7	73	28
ICMi H/G boundary	0.82	0.88	0.85	0.86	0.83	0.92	0.79	0.87
ICMi median H/G boundary				**0.85**				
# Good	29	15	77	269	259	25	17	4
ICMi G/M boundary	0.71	0.77	0.58	0.69	0.67	0.62	0.67	0.75
ICMi median G/M boundary				**0.68**				

The distribution of the values of the class boundaries converted in ICMi for WFD-compliant methods is also showed in the Box&Whiskers representation of Figure 7.9, with the main statistical descriptors in Table 7.19.

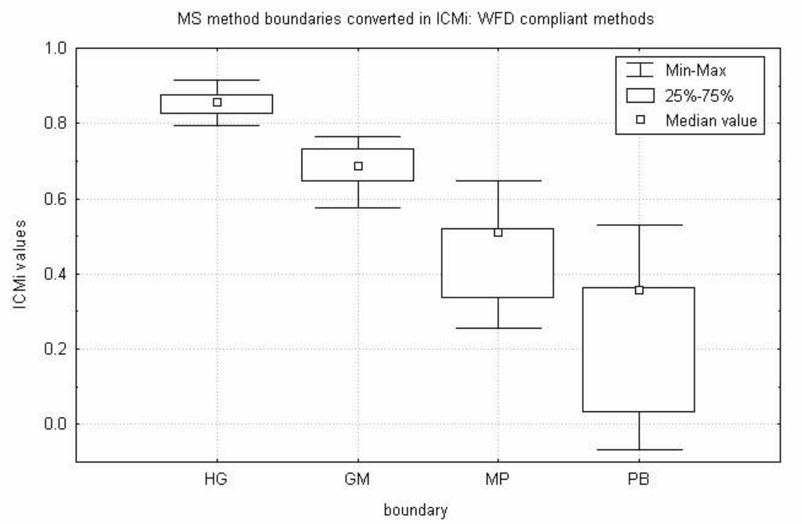


Figure 7.9 Box&whiskers for class boundaries converted in ICMi, WFD-compliant methods

Table 7.19 Values of minimum, 25th percentile, median, 75th percentile, maximum, mean and standard deviation of class boundaries in ICMi, various stream types

	min	max	25th %ile	75th %ile	median
HG	0.794	0.915	0.824	0.875	0.855
GM	0.577	0.766	0.644	0.732	0.687
MP	0.255	0.648	0.334	0.521	0.511
PB	-0.067	0.531	0.032	0.363	0.357

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.

Table 7.20 p-level for Mann-Whitney U test: WFD compliant countries

	HG	GM	MP	PB
HG		0.001	0.003	0.003
GM	0.001		0.008	0.003
MP	0.003	0.008		0.347
PB	0.003	0.003	0.347	

IMPORTANT WARNING FOR FRENCH DATASETS

All the reference values and class boundaries tested for the French datasets in all types are provisional. Changes may occur due to ongoing work on reference sites selection and sampling. Furthermore, in the examples reported here the normalization was carried out considering the 75th percentile of High status samples in order to have a full comparability with countries that do not yet have defined reference conditions.

7.3 - Overall comparison of National data to an International, benchmark dataset

7.3.1 - Comparison of R-C1 data

The ICM index was calculated for the benchmark dataset (see Chapters 5 and 6) and for test datasets (see Chapters 4 and 6). The procedure of indirect comparison *via* benchmark and ICMi requires the contrast of a test dataset against a benchmark one, through the values of the ICM index. The purpose of this section is to illustrate the results of such comparison for some selected river types.

The values of the ICM index were compared among Good status classes of benchmark dataset and Good status classes of test datasets. In Figure 7.10, the variation of the ICMi in Good status classes for some tested datasets (R-C1) in comparison with benchmark values (left part of the graph) is presented. As it has been done for Good status, the comparison of ICMi values was also carried out for High status (see Figure 7.11). Lower median values are observed for Italy and Poland. The value reported for Germany refers to the only one High status sample, for that stream type.

In the Figures, the different countries are indicated by a letter. Table 8.21 indicates the correspondence of letters with the name of the country. Furthermore, in the tables the analyses done for each river type are also indicated, specifying if they were simply compared to the benchmark or if they were also harmonized on the basis of the results of such comparison (see Chapter 8).

Table 7.21 Correspondence between letter and countries for the interpretation of the Box & whiskers graphs

	IC River type	Country	Compared	Harmonized (via benchmark & ICMi)
A	R-C2	France	Y	Y
B	R-C2	Spain	Y	Y
C	R-M1	Italy	Y	Y
D	R-C1	Estonia	Y	Y
E	R-C1	France	Y	Y
F	R-C1	Italy	Y	Y
G	R-C1	Poland	Y	Y
H	R-C1	UK	Y	Y
I	R-C1	Germany	Y	Y
L	R-C1	Denmark	**not yet**	
M	R-M1	France	**Not possible (only 4 good samples)**	
N	R-M5	Italy	**No (National method has to be refined)**	
O	R-C1	Belgium	**not yet**	
P	R-C1	the Netherlands	**No (further investigation needed)**	

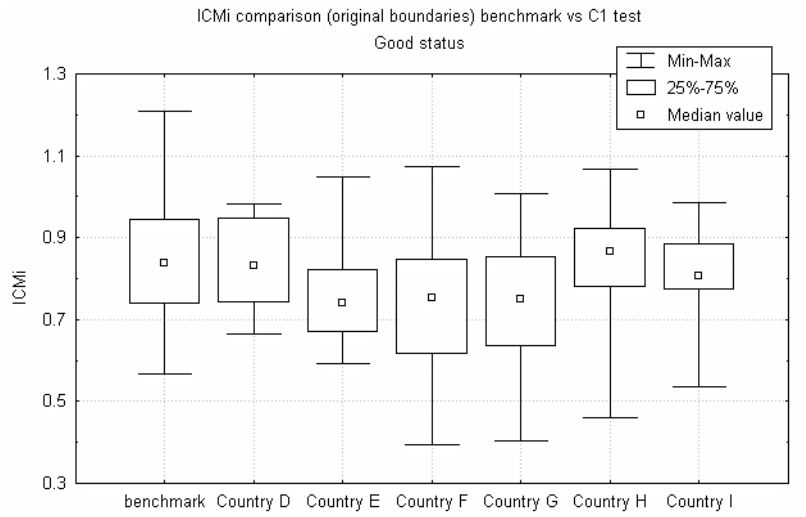


Figure 7.10 Variation of the ICM index for Good status class (R-C1) in the different datasets (benchmark *vs* National Methods/datasets). For test data, National standard boundaries were considered

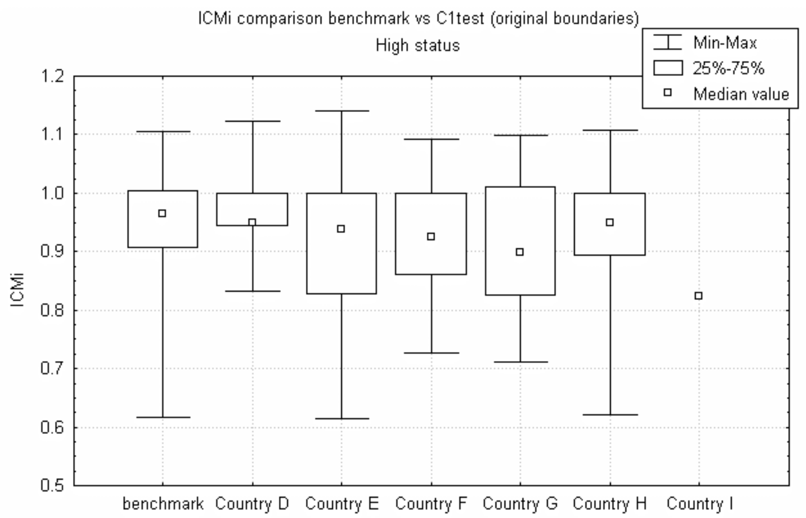


Figure 7.11 Variation of the ICM index for High status class (R-C1) in the different datasets (benchmark *vs* National Methods/datasets). For test data, National standard boundaries were considered

The further step is the statistical comparison among each test dataset and the benchmark one, in order to see if any significant differences are present. Firstly, the Good status samples were tested. The application of the Mann Whitney test reveals differences for Good status between benchmark and test data for France ($p=0.0008$), Italy ($p<0.00005$) and Poland ($p=0.035$). The statistical comparison of high status shows significant differences for France and Italy (p of 0.036 and 0.04 respectively).

In Table 7.22 the summary results of the statistical comparison are reported.

Table 7.22 Summary of the results of the statistical comparison of the R-C1 datasets with benchmark data

				R-C1					
			benchmark	EE (D)	FR (E)	IT (F)	PL (G)	UK (H)	DE (I)
Original MSs Boundaries	HIGH Status	# TOT	398	23	132	361	49	789	91
		# high	105	9	64	84	11	202	1
		ICMi median value	0.96	0.95	0.94	0.92	0.90	0.95	0.82
		p-level (MS vs bench)		0.87	0.036	0.04	0.22	0.165	--
	GOOD Status	# good	103	6	29	176	15	345	17
		ICMi median value	0.84	0.83	0.74	0.75	0.75	0.86	0.80
		p-level (MS vs bench)		0.89	0.0008	$< 5*10^{-5}$	0.035	0.621	0.31

7.3.2 - Comparison of R-C2 and R-M1 data

In this section, the comparison and statistical test for some examples derived from R-C2 and R-M1 test datasets - in the same way as previously done for R-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 (France). Figures 7.12 and 7.13 show respectively the variation of ICMi for samples in High and Good status before harmonization, accordingly to the original National boundaries. The lowest median values are observed for France (R-C2), for High and Good status classes with a significant difference with the benchmark ($p=0.0008$ for High status and $p= 0.00002$ for Good status). It has to be noted that the normalization rule used in this test (75th percentile of high class) might give lower EQR values for the H/G and G/M boundaries than the WFD compliant National classification based on reference sites. For these examples the normalization according to 75th percentile of High status samples was maintained in order to guarantee a full comparability with other countries.

In Table 7.23 the summary results of the statistical comparison are reported

Table 7.23 Summary of the results of the statistical comparison of the R-C2 and R-M1 datasets with benchmark data

				R-C2		R-M1
			Bench.	FR (A)	ES (B)	IT (C)
Original MSs Boundaries	HIGH Status	#TOT	398	143	46	63
		# high	105	73	6	21
		ICMi median value	0.96	0.89	0.94	0.97
		p-level (MS vs bench)		0.0008	0.72	0.81
	GOOD Status	# good	103	17	25	38
		ICMi median value	0.84	0.66	0.81	0.82
		p-level (MS vs bench)		$2*10^{-5}$	0.36	0.28

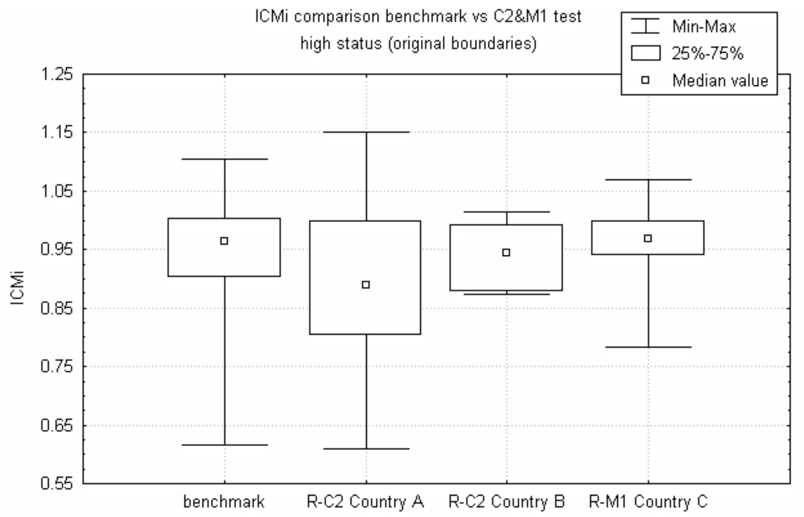


Figure 7.12 Variation of the ICM index for High status class (R-C2 and R-M1) in the different datasets (benchmark *vs* National Methods/datasets). For test data, National standard boundaries were considered

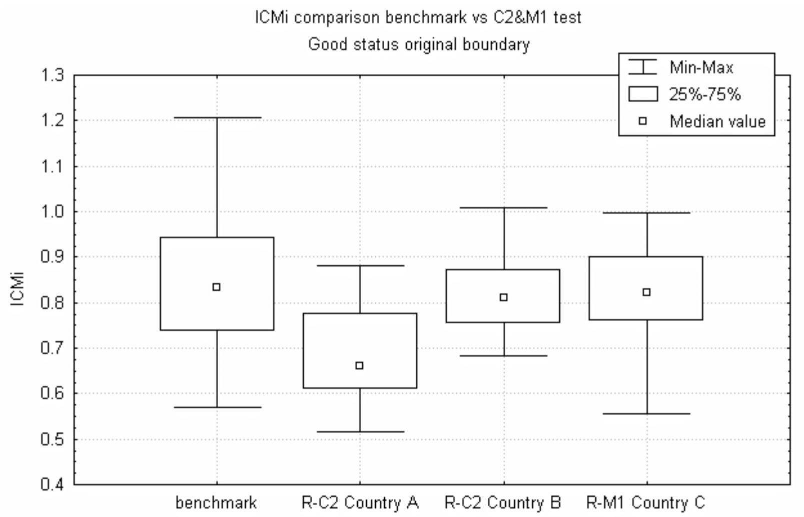


Figure 7.13 Variation of the ICM index for Good status class (R-C2 and R-M1) in the different datasets (benchmark *vs* National Methods/datasets). For test data, National standard boundaries were considered

7.4 - General consideration

In the present Chapter, three comparison Options were presented.

The first Option deals with the calculation of different National classification *formulae* on the samples belonging to a single dataset. This means that the final step of classification according to the different assessment methods i.e. the calculation of the index value and the attribution to a quality class, was carried out. National formulae only are applied, the sampling protocol being related to one field method only. On the basis of such approach, boundaries were compared. The method seems particularly useful when the comparison is done for systems that are based on the same concepts, as when two different versions of the same method are used in different countries e.g. Saprobic system in Germany and Austria, BMWP-based systems in UK and Spain. When basic differences exist for the compared methods the validity of the results is questionable, because part of the differences may be due to the dissimilar approach at the basis of the assessment systems.

Another Option considers the comparison of class boundaries of National Methods through their direct conversion into ICMi values. This was carried out as follows: 1) on the invertebrate sample correctly collected with each National sampling protocol the ICMi is calculated and a regression formula is derived for each dataset and method 2) the actual National boundary is translated into an ICMi value. Only slight differences among National boundary values - especially for the High/Good boundary - were generally found. For the Netherlands and Denmark the highest boundary values in terms of ICMi were found. This is probably in relation to the presence of sites that, even if classified as High status, are characterized by low values of ICMi (probably because of the absence of true reference sites), which determine higher values after normalization. The clear non-continuity of some indices (e.g. DSFI) can further disturb the comparison (with all the Options used). The Good/Moderate boundary showed larger differences among countries, with discrepancies up to about 10% found in many cases. In general, when differences are found in terms of ICMi boundary values, a way to assess their significance has to be established. Another important detail to note is that such methods of comparison and the consequent results are strongly dependent on the number and kind of datasets included.

Both the two former options are based on the concept of averaging class boundaries. Consequently, an important requisite to be included for methods and data when deriving reference value is their acknowledged WFD-compliance, a factor which is a long way from being reached by most European assessment systems.

The last Option, is the comparison of the ICMi values obtained for test data belonging to the different quality classes to ICMi values obtained for an external (benchmark) dataset. In this last case it was possible to run statistical tests in order to check if any differences observed among the two sets of data were statistically significant. For most of the tested datasets, methods and countries, no significant differences were found.

8 - HARMONIZATION

In this Chapter, examples of possible approaches to harmonization of class boundaries are provided, based on AQEM/STAR data only (8.1) and on both MSs' and AQEM/STAR datasets (8.2 and 8.3).

8.1 - Bilateral harmonization

8.1.1 - Averaging class boundaries of national methods - Same sample, different formulae (no ICMi)

Introduction -Based on the results of the 'direct comparison approach' this section presents a simple procedure to obtain harmonized 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 includes the results of the comparison analysis of 247 AQEM/STAR samples from Inter-calibration stream type R-C4. Table 8.1 lists converted boundary values for the Good status of the indices SI (DE), ASPT (UK), ASPT (SE), DSFI (DK) and DSFI (SE). Harmonization is done by averaging all EQR boundary values per national index.

Table 8.1 Class boundaries and their average values per national method derived by regression analysis

R-C4			**SI (DE)**	**ASPT (UK)**	**ASPT (SE)**	**DSFI (DK)**	**DSFI (SE)**	**average**
benchmark for comparison	**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 8.1 displays the average EQR boundary values per assessment index. Figure 8.1 shows the effect of harmonization on the distribution of quality classes in the test dataset. Before harmonization of both reference conditions (95th 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 harmonization 44 percent of samples are of equal quality status (High, Good, or Moderate and Worse).

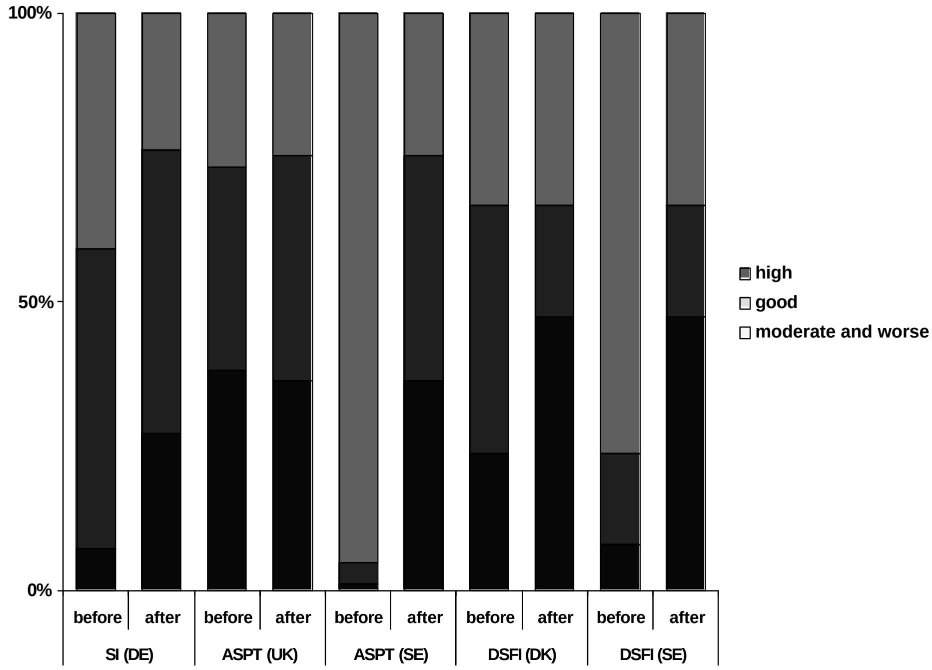


Figure 8.1 Distribution of quality classes before and after harmonisation via averaging of class boundary values (AQEM/STAR benthic invertebrate dataset; n=247)

8.2 - Harmonization of national boundaries through conversion to ICMi values

In section 7.2 the results of the comparison phase are presented. For the comparison, the boundaries of the different National assessment methods were converted by linear regression to an ICMi value. The aim of this section is to illustrate a possible Option of harmonization, which is an additional step after the results of that comparison.

Within each IC type, a very simple harmonization Option might 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 the original boundary of their methods if current values are below the obtained median. If the boundary is higher than the median value, no change in boundary value should be expected, because this would mean that the MS has already adopted more strict criteria e.g. in order to adapt to the WFD requirements. If the boundary is equal to the median a change should not be expected.

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 a reference conditions definition.

IMPORTANT WARNING

The option of averaging class boundary values of assessment methods (or to use a median value) is only applicable when all the considered biological methods are demonstrated as fully WFD-compliant. Other options to select a reference value against which harmonize boundaries may be evaluated (e.g. the highest or lowest value occurring within the GIG)

In addition, the use of this option is acceptable if all MSs contribute to 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).

8.2.1 - Inter-GIG harmonization: WFD-compliant methods

In this paragraph examples of harmonization for WFD-compliant methods are given. Figures 8.2 and 8.3 show the deviation from the median of the boundaries respectively for High/Good and Good/Moderate for countries presenting a WFD-compliant method. For countries that have two indices for the classification (e.g. Germany (I) and UK(H)) two different histograms are reported. The histograms are ordered according to increasing values. The 0 represents the median value of all the boundaries.

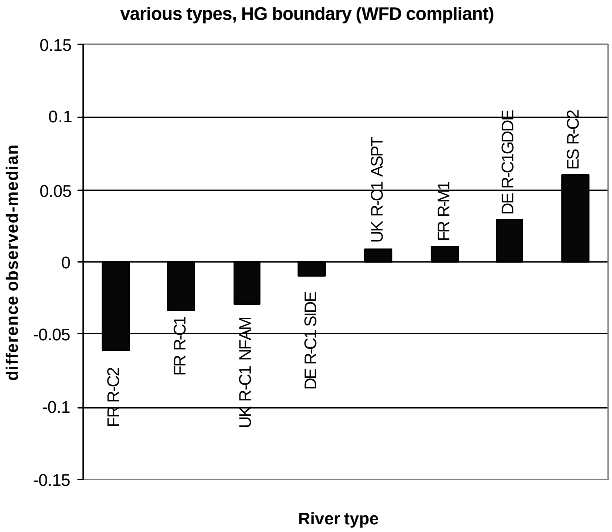


Figure 8.2 Difference between the National and the median value of WFD-compliant methods (EQRs for the ICMi) for the High/Good boundary

In four cases the National boundary is higher than the median value, with the maximum difference for Spain. Because of the positive divergence, a repositioning of the boundary is not expected for these four countries. The remaining countries show values lower than the median, with a difference of approximately 7% for Germany and the saprobic system. In these cases, the H/G boundary might be thought to be moved to an ICMi value of 0.85, which represents the median value of the High/Good boundary for WFD-compliant methods.

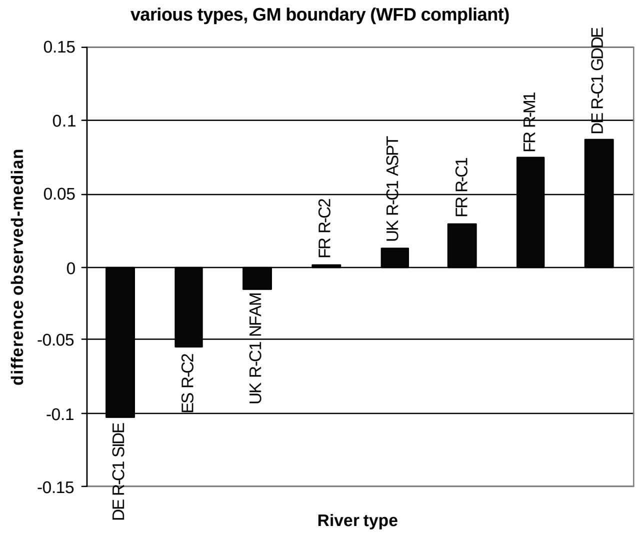


Figure 8.3 Difference between the National and the median value of WFD-compliant methods (EQRs for the ICMi) for the Good/Moderate boundary

For the Good/Moderate boundary, in five cases National EQR boundaries are higher than the median value and for this reason no adjustments are expected. The remaining countries show values lower than the median, with a difference up to around 10% for the saprobic system in Germany and to around 5% for Spain. These countries might consider moving their boundary – according to this approach - to a higher value of 0.68 units of the ICMi EQR, which represents the median of the Good/Moderate boundaries of WFD-compliant methods. The UK boundary, even if lower than the median, is very close and might not need any modification.

In general, where two indices concur to the definition of the quality class (e.g. Germany and UK), they usually show a different response:

- for Germany (I), boundaries may have to be refined for method saprobic system. For High status the difference is very small, while it is more evident for Good status samples. For method GD, boundaries are higher than the median. The method GD is a newly developed multimetric index for the detection of general degradation. In this case a boundary change is not expected. If the two methods are always applied concurrently and the one-out all-out principle is in use, the GD module will nevertheless guarantee an adequate site classification, in this stream type. In this case, the saprobic module could stay as it is and only be used to supply specific information on the organic pollution status of the site (i.e. being used as an advisory tool and basically not for classification).

- something similar happens for the UK, where the boundary is raised for method 'b' only (Number of families).
- this latter situation was also observed in the case of Poland (R-C1, see the next paragraph), where a Polish version of BMWP and Margalef's index was used, with a one-out all out principle. The second index never determined the Worst quality class, being thus basically used - with its current boundaries – as a source of supplementary information.

In Table 8.2 the summary results of the harmonization are reported. Countries that have higher value than the median were not harmonized. In the table, the harmonized boundaries expressed in terms of EQR for the National Method were calculated by linear regression from the median ICMi value of all the WFD-compliant boundaries.

Table 8.2 Harmonization of class boundaries for WFD-compliant methods

R-C type	country	boundary	ICMi (Original boundary)	ICMi median	MS_EQR (Original boundary)	MS_EQR harmon.
C1	DE, GD	High/Good	0.88	0.855	0.80	--
	DE, SI		0.85		0.85	0.85
	FR		0.82		0.82	0.86
	UK ASPT		0.86		0.94	--
	UK #FAM		0.83		0.83	0.96
C2	FR		0.79		0.82	0.90
	ES		0.91		0.93	--
M1	FR		0.87		0.87	--
C1	DE, GD	Good/Mod	0.77	0.687	0.60	--
	DE, SI		0.58		0.66	0.73
	FR		0.71		0.71	--
	UK ASPT		0.69		0.83	--
	UK #FAM		0.66		0.64	0.83
C2	FR		0.67		0.71	0.71
	ES		0.62		0.70	0.75
M1	FR		0.75		0.75	--

8.2.2 - Inter-GIG harmonization: non WFD-compliant methods

For the harmonization of non WFD-compliant methods, the median values of the High/Good and Good/Moderate boundaries, obtained for WFD-compliant methods were used. The resulting harmonization could consist of making all the boundaries equal to these median values. In figures 8.4 and 8.5 the comparison of boundaries for a non WFD-compliant method is shown, with respect to the difference from the median value of WFD-compliant methods.

Only countries Poland (R-C1) and Italy (R-C1) show an ICMi value below the WFD-compliant methods median both for High/Good boundary and for Good/Moderate one. Thus, only those two Countries – according to this harmonization approach – should adjust their HG and GM, boundaries. All the other countries show values higher than the median of WFD compliant methods and thus they do not need to reposition the boundary because it is apparently more restrictive.

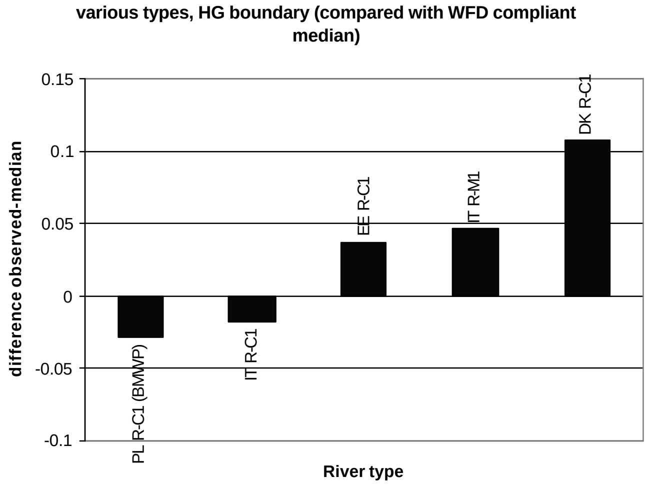


Figure 8.4 Difference between the value of National, non WFD-compliant methods and the median value of WFD-compliant methods (EQRs for the ICMi) for the High/Good boundary

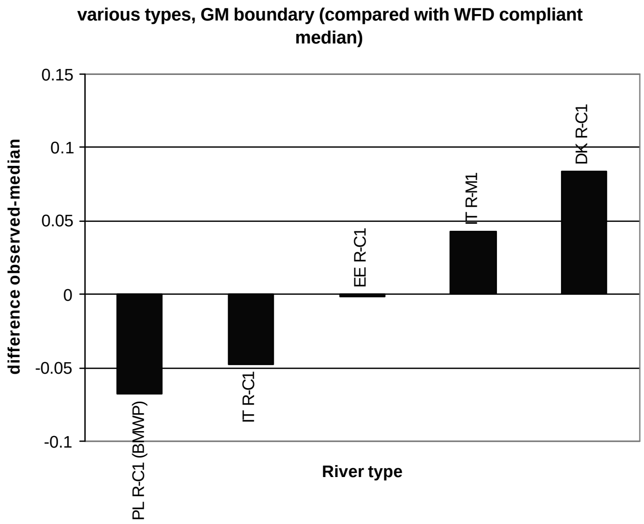


Figure 8.5 Difference between the value of National, non WFD-compliant methods and the median value of WFD-compliant methods (EQRs for the ICMi) for the Good/Moderate boundary.

The summary of the results of the process of harmonization with the indication of boundary values before and after harmonization is reported in Table 8.3 both in terms of ICMi EQR and National Method EQR. In the case of Poland, because Margalef's index never determined the final classification of the sites, it was not considered in the harmonization.

Table 8.3 Harmonization of class boundaries for non WFD-compliant methods

	country	boundary	ICMi (Original boundary)	ICMi median	MS_EQR (Original boundary)	MS_EQR harmon.
C1	DK	High/Good	0.96	0.855	1.00	--
	EE		0.89		0.93	--
	IT		0.84		0.91	0.92
	PL BMWP		0.83		0.78	0.81
	PL Margalef		0.65		0.55	
M1	IT		0.90		0.88	--
C1	DK	Good/Mod	0.76	0.687	0.71	--
	EE		0.68		0.78	0.78
	IT		0.63		0.72	0.76
	PL BMWP		0.61		0.54	0.62
	PL Margalef		0.54		0.40	
M1	IT		0.72		0.70	--

8.3 - Harmonization of class boundaries via ICMi with the use of a benchmark dataset

In this chapter some examples of harmonization of class boundaries 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 R-C1 (6 examples), R-C2 (2 examples) and R-M1 (1 example) are described. To summarize the whole procedure, a detailed example is given hereafter for Italy and Poland (R-C1).

8.3.1 - Specific examples from Italy and Poland (R-C1)

In this section, two specific examples of harmonization of class boundaries based on the comparison of National datasets (test datasets: R-C1 Italy and Poland) against a trans-National classification - based on standard samples from many European areas and in respect of the WFD requirements (benchmark datatset: STAR/AQEM data) - are reported.

Italian IBE harmonization (R-C1)

In Figure 8.6 the variation of the ICM index in the five classes of the Italian IBE index for the sites belonging to the Inter-calibration type R-C1 is shown (n=361).

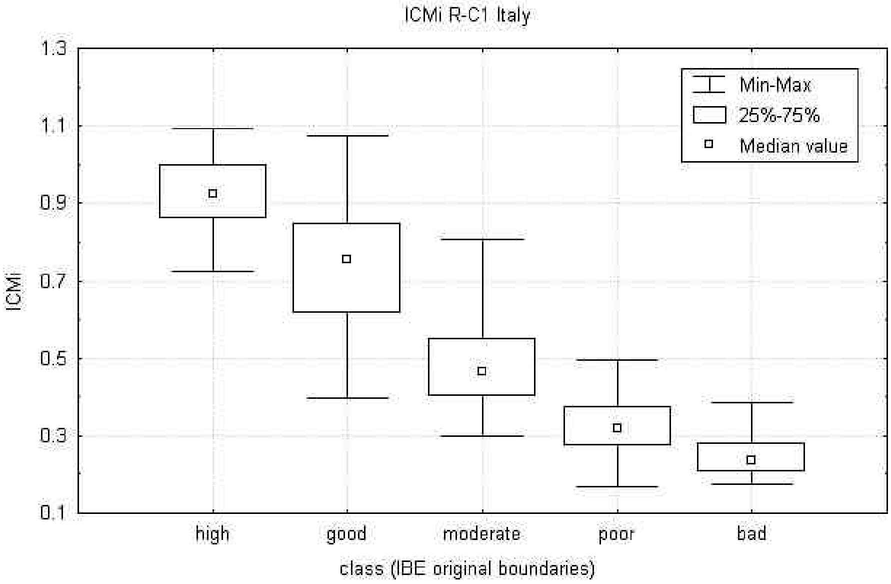


Figure 8.6 Variation of the ICM index within the IBE index quality classes for R-C1 (Italy)

The ICM index reflects well the quality classes derived from the IBE method, even if some overlay between classes Good and Moderate is apparent. Nevertheless, no overlap between the interquartile range of High, Good, Moderate and Poor classes exists. The results of the Tukey test confirms that all classes are statistically different ($p<0.00005$).

For the harmonization of class boundaries, the values of the ICMi obtained for the test dataset have to be compared with those obtained for the benchmark samples. The values of the ICMi for the Good (first) and High (next) classes are therefore statistically compared to the equivalent calculated from the benchmark dataset, in order to see if any differences exist. So, the Good status samples are tested first. If differences exist, and the test data shows lower values than the benchmark one, the Good/Moderate boundary has to be moved up. In the example, the comparison of the ICMi values of test and benchmark datasets revealed statistically significant differences for the Good status class ($p<0.00005$). The next step of the harmonization process now involves the repositioning of the boundary – and consequent sample exclusion from the dataset - until no more differences are found for ICMi values by statistically comparing the two datasets. The IBE Good/Moderate status boundary was moved up, because the median value was lower in the test dataset than in the benchmark dataset. The threshold value was repositioned step by step (i.e. from 7.6 to 8, from 8 to 8.4, etc.), until no more differences between the values of the ICMi according to the benchmark (STAR/AQEM) and IBE classifications were found. Statistical differences were found until the boundary was moved to 8.6, after which they were no longer significant ($p= 0.053$). The new Good/Moderate boundary was thus fixed at 8.6. After having compared and tested ICMi values for the Good status class, the High status class was compared and tested. The result of the Mann-Whitney test shows a significant difference ($p=0.040$) for High status samples between the benchmark and test datasets. Because of this difference, the boundary High/Good was moved up step by step as was done for the Good/Moderate boundary. To remove differences, it was enough to move the High/Good boundary from an IBE value of 9.6 to 10, when a non-significant p value of 0.09 was found.

Only after the statistical testing, the samples excluded from the High status following the new boundary setting were included in the Good status class (and box of Fig 8.7) and the samples left out from the Good class were moved to the Moderate status class. In Figure 8.7, the variation of ICMi values within the IBE classes after harmonization is shown. The interquartile ranges of Good and Moderate classes are well separated. After having harmonized the Good/Moderate boundary, the Moderate status samples give the impression of being closer to the Good status ones, because some samples were moved from the Good to the Moderate class, thus pushing up the highest values for the Moderate class. The same is true in relation to the moving of samples from the High to the Good class. In general, after harmonization the values of ICMi in the High, Good and Moderate classes are then higher with respect to the original classification. For the Good status the 25^{th} percentile is shifted up from approximately 0.6 to 0.7 (see also Table 8.7).

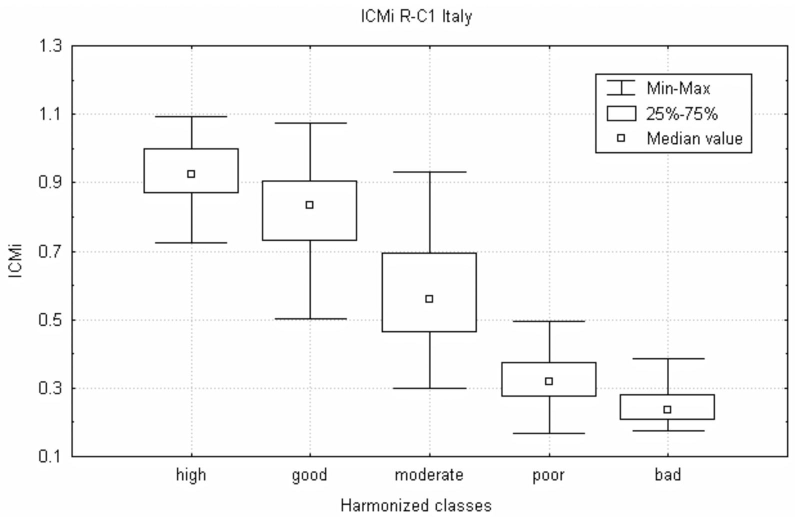


Figure 8.7 Variation of the ICM index within the IBE index classes after harmonization for R-C1 (Italy)

Polish BMWP harmonization (R-C1)

A second example for which it is necessary to shift boundaries because of the difference between test and benchmark data is R-C1 Poland. In general, Figure 8.8 confirms the results of Chapter 4.3 indicating that the ICMi follows the ecological gradient quite well, even if the Tukey test does not indicate a significant difference between the Good and Moderate classes ($p=0.46$). Significant differences were also not found between High and Good status classes ($p=0.15$). The use of a larger dataset might presumably improve the effectiveness of the statistical testing for this stream type in Poland.

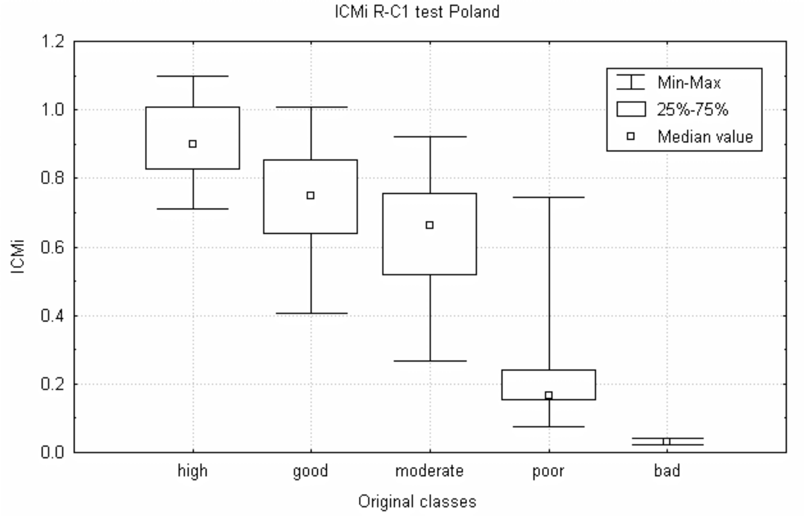


Figure 8.8 Variation of the ICM index within the Polish BMWP score classes before harmonization for R-C1 (Poland)

For R-C1 in Poland, significant differences were found for Good status samples between test and benchmark values (p=0.035). In Poland, site classification is derived from the combination of BMWP and Margalef's index classification, on the basis of the 'one-out all-out' principle. For the samples included in the test dataset, BMWP always determined the final class. The re-positioning of the boundaries was thus done by moving up the threshold Good/Moderate for the BMWP score. As a rule, the boundary was moved up by 5 scores (from 70 to 75). The p level for the comparison between the test data with the newly defined BMWP Good class and benchmark data was 0.12, indicating that differences no longer existed. With regard to the comparison of High status samples, the test did not reveal any differences (p=0.22). After the harmonization of boundaries (see Figure 8.9), the variability of ICMi within the Good status class is reduced, due to the moving of some samples to the Moderate class. Good and Moderate classes then overlap to less extent.

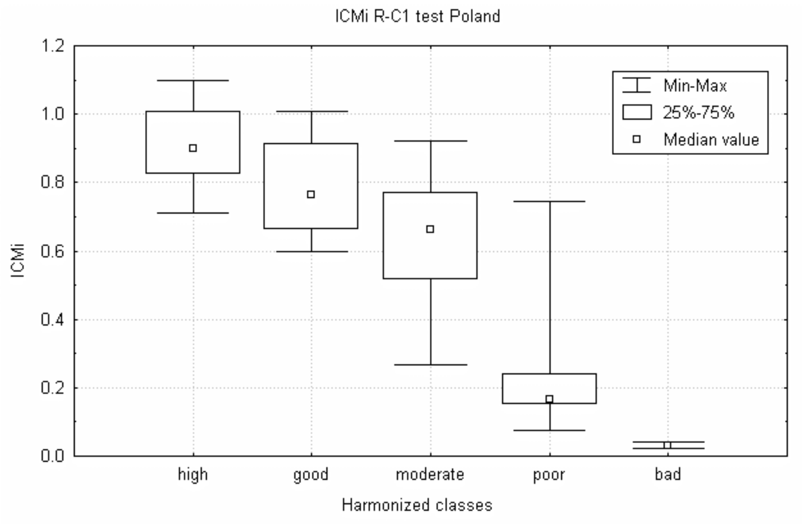


Figure 8.9 Variation of the ICM index within the Polish BMWP score classes after harmonization for R-C1 (Poland)

Overall considerations for Italy and Poland (R-C1)

In Table 8.4, the boundaries before and after harmonization for the Italian IBE and the Polish BMWP for R-C1 type are reported.

Table 8.4 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 8.10 and 8.11 show the percentage of sites belonging to the different quality classes before and after harmonization in Italy and Poland, for R-C1. For Italy the number of samples moving from Good to Moderate status is 82, corresponding to 22%. According to the current IBE boundaries, 72% of the samples would not require a restoration action (including High and Good samples), while after harmonization the samples that do not require rehabilitation – according to the macroinvertebrate, BQE - represent 50% of the total. This re-setting of the H/G boundary offers a more realistic picture, because the sites of the R-C1 river type under examination are located in a highly urbanized, agricultural area (the area neighbouring the town of Milan).

However, some further considerations are necessary. The Italian legislation (D.L.vo 152/99) requires, for the derivation of a final site classification, a comparison of the biological and chemical classification, by finally classifying a site according to the one-out all-out principle. The final classification will then be determined by the worst of the two. Thus, it may happen that some of the sites classified in Good status for biology (i.e. macroinvertebrates), move to Moderate status because of chemistry and *vice versa*. Further analysis should be addressed to the comparison of chemical and biological data.

Different samples from the same site often get a different quality class, depending on the year/season of investigation. For High status, approximately 25% of the sites get a stable classification while for Good status the percentage is around 34% and, for Moderate status, it is around 38%. To summarize, when more samples are collected at a site all year round (e.g. seasonal collection), a certain degree of variability is observed, especially due to variations of water quality in different periods of the year, which follow e.g. agricultural practices, discharge variations or temperature-related processes. Depending on the distribution of classification index values within each of the quality classes (i.e. closer to the upper or lower class boundary), the newly set boundaries might impact more or less than expected by simply looking at single seasonal samples on the final site classification. Hence, a more in-depth investigation of data is required to properly assess the consequences of the potential refinement of the Good/Moderate boundary.

In general terms, the percentage of samples to be moved from the Good to the Moderate status is smaller (6%) for Poland than observed for Italy.

After the harmonization, the overall percentage of samples belonging to High or Good status, which would not require any restoration according to the macroinvertebrate information, is nevertheless quite similar for the two countries: 50% for Italy and 47% for Poland. Before harmonization, the two countries showed, respectively, 72% and 54%.

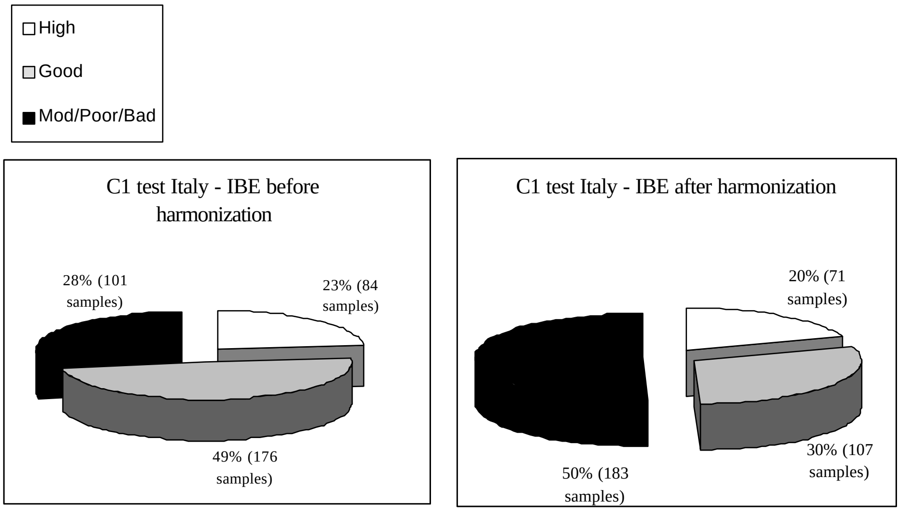


Figure 8.10 Sample distribution according to the national classification before and after harmonization (R-C1, Italy – IBE method)

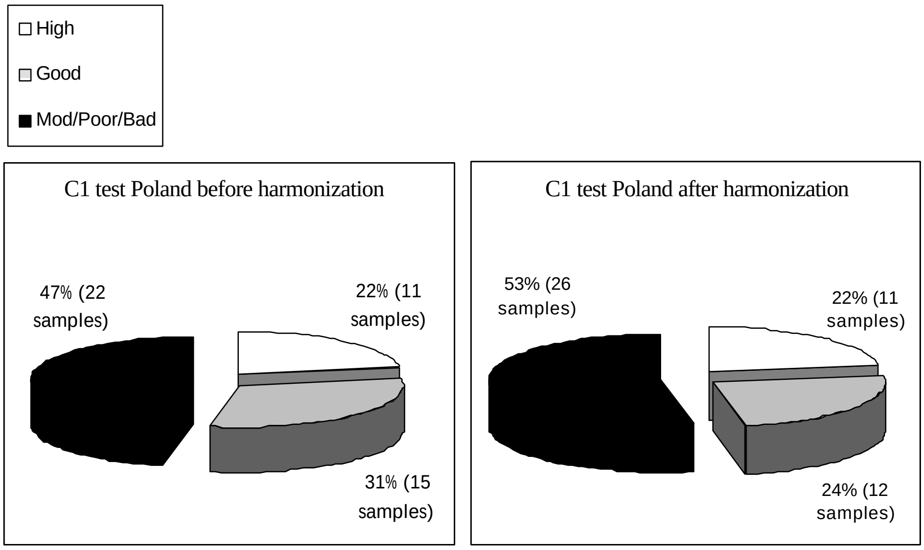


Figure 8.11 Sample distribution according to the national classification before and after harmonization (R-C1, Poland – BMWP method)

Direct comparison between the Italian and Polish datasets and boundaries

As a further test, a direct comparison (without comparing to a benchmark dataset) can be carried out between these two datasets. The Mann-Whitney U test found no differences, either for High status samples (p=0.52), or Good status samples (p=0.84) between the two datasets.

By indirectly comparing *via* the benchmark dataset (see the two paragraphs above), boundaries resulted in being adjusted for both countries – even though in different ways - in order to eliminate the differences between benchmark and test datasets classifications. It must be noted that during the indirect comparison *via* the benchmark dataset, the High status boundary resulted in being modified for Italy, but not for Poland. Indirectly, this tells us that the two boundaries are different. By direct comparison, no difference resulted. Thus, the comparison with an external, invariable dataset is recommended, especially 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, solves the problem of not having methods and assessment systems that fulfil WFD requirements.

8.3.2 - General outcome of the harmonization of boundaries for R-C1 type

The process of harmonization *via* ICMi with a benchmark dataset requires a statistical comparison between test datasets and benchmark datasets, in order to see if there are any significant differences (see 7.3.1). Firstly, as seen in the previous section, Good status samples are tested. If differences are found, the process requires the repositioning of the appropriate boundary for the assessment method under test in order to eliminate these differences. The process is than repeated for the High status boundary.

In the following Figures, the different countries are indicated by a letter and Table 8.5 indicates the correspondence between letters and country. An indication of the type of analyses done i.e. comparison and/or harmonization, is also reported.

Table 8.5 Correspondence between letters and countries for the interpretation of the Box&Whiskers graphs

	IC River type	Country	Compared	Harmonized (via benchmark & ICMi)
A	R-C2	France	Y	Y
B	R-C2	Spain	Y	Y
C	R-M1	Italy	Y	Y
D	R-C1	Estonia	Y	Y
E	R-C1	France	Y	Y
F	R-C1	Italy	Y	Y
G	R-C1	Poland	Y	Y
H	R-C1	UK	Y	Y
I	R-C1	Germany	Y	Y
L	R-C1	Denmark	**not yet**	
M	R-M1	France	**Not possible (only 4 good samples)**	
N	R-M5	Italy	**No (National method has to be refined)**	
O	R-C1	Belgium	**not yet**	
P	R-C1	the Netherlands	**No (further investigation needed)**	

According to the results of the comparison for Good status samples from R-C1 type, statistical differences were found for France, Italy and Poland. For these countries the boundary Good/Moderate was thus moved upwards in order to eliminate these differences (see 8.3.1 for detailed examples). The statistical comparison of High status shows significant differences for France and Italy requiring that the High/Good boundary to be moved upwards. In general, the repositioning of the boundaries does not imply large adjustments. The detailed results of the harmonization for Poland and Italy are shown in the previous section. For France, the harmonization resulted in a move of the Good/Moderate boundary value from IBGN 12 to IBGN 13 and of the High/Good boundary from 14 to 15. In Figure 8.12, the variation of the ICMi after harmonization for Good status class is presented, including the examples for which no differences were found from benchmark,. In Figure 8.13, the range of ICMi values observed in the

High status class after harmonization is shown. Table 8.6 represents the results of harmonization reporting the number of samples for test datasets before and after the harmonization, including p-levels. The number of samples for the benchmark dataset are also indicated, as well as the median values of ICMi in benchmark and test datasets.

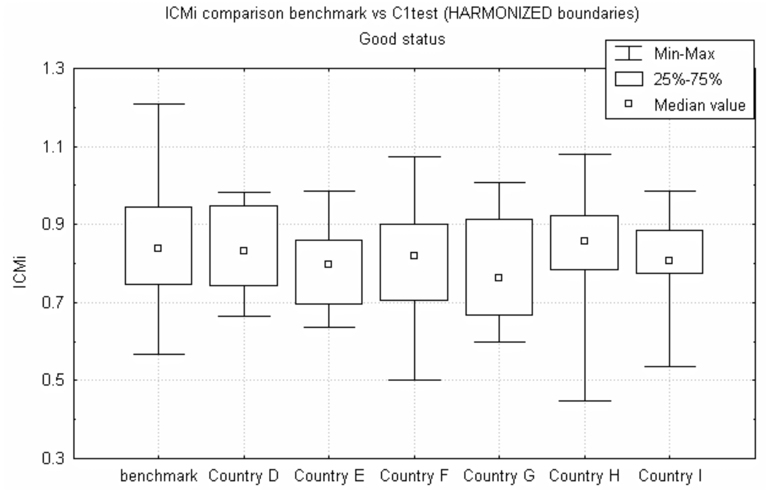


Figure 8.12 Observed range of ICMi values in the Good status class after harmonization for the R-C1 type

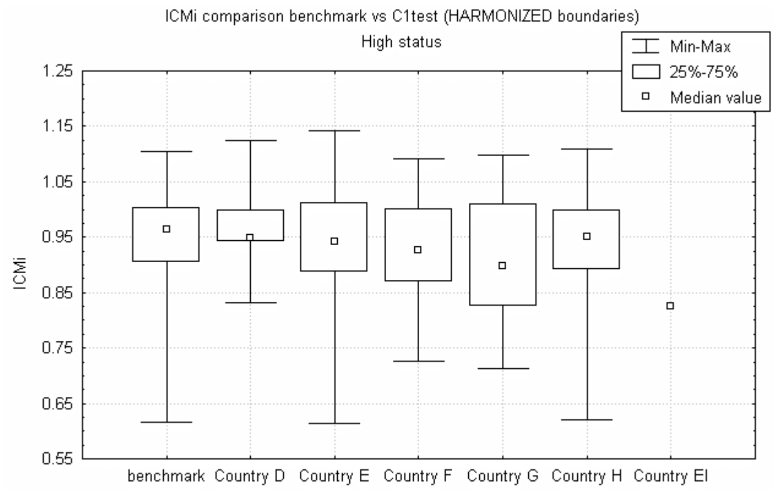


Figure 8.13 Observed range of ICMi values in the High status class after harmonization for the R-C1 type

Table 8.6 Summary results of the Harmonization process for R-C1 types

				R-C1					
			bench.	EE (D)	FR (E)	IT (F)	PL (G)	UK (H)	DE (I)
Original MSs Boundaries		# TOT	398	23	132	361	49	789	91
	HIGH Status	# high	105	9	64	84	11	202	1
		ICMi median value	0.96	0.95	0.94	0.92	0.90	0.95	0.82
		p-level (MS vs bench)		0.87	0.04	0.04	0.22	0.17	--
	GOOD Status	# good	103	6	29	176	15	345	17
		ICMi median value	0.84	0.83	0.74	0.75	0.75	0.86	0.8
		p-level (MS vs bench)		0.89	0.0008	$< 5*10^{-5}$	0.035	0.621	0.31
After Harmonization	HIGH Status	# high		9	50	71	11	202	1
		ICMi median value	0.96	0.95	0.94	0.93	0.90	0.95	0.82
		p-level (MS vs bench)		0.87	0.27	0.09	0.22	0.17	0
	GOOD Status	# good		6	25	107	12	345	17
		ICMi median value	0.84	0.83	0.80	0.82	0.76	0.86	0.80
		p-level (MS vs bench)		0.89	0.07	0.16	0.12	0.62	0.31

8.3.3 - General outcome of the harmonization of boundaries for R-C2 and R-M1 types

In these examples, differences were found for both High and Good status classes, for one of the two countries belonging to R-C2 (France). The process of harmonization then involved the move upwards of the Good/Moderate (IBGN from 12 to 14) and High/Good boundaries (IBGN from 14 to 15) for France in order to eliminate these differences. The results of the process of harmonization are presented in Figures 8.14 and 8.15, in which it can be seen, as for France, that in Good status samples, variability diminished after harmonization. The shifting of the boundary involved a 12% of sites moving from Good status to Worst condition. For France, in the proposed examples (R-C1 and R-C2), a normalization according to the 75th percentile of high status samples was used, in order to guarantee comparability with other datasets and with countries whose assessment systems are not based on well defined reference conditions. The results of the present analyses have to be confirmed by calculating normalized ICMi values according to the e.g. the median value of reference sites (defined on WFD-compliant protocols). This is generally true for all the systems and results compared here, which should be considered as a methodological guide to the technical phase of the Inter-calibration process. Nevertheless, for selected cases and circumstances (e.g. the Italian boundaries refinement), they might be profitably utilized in a more optimum fashion.

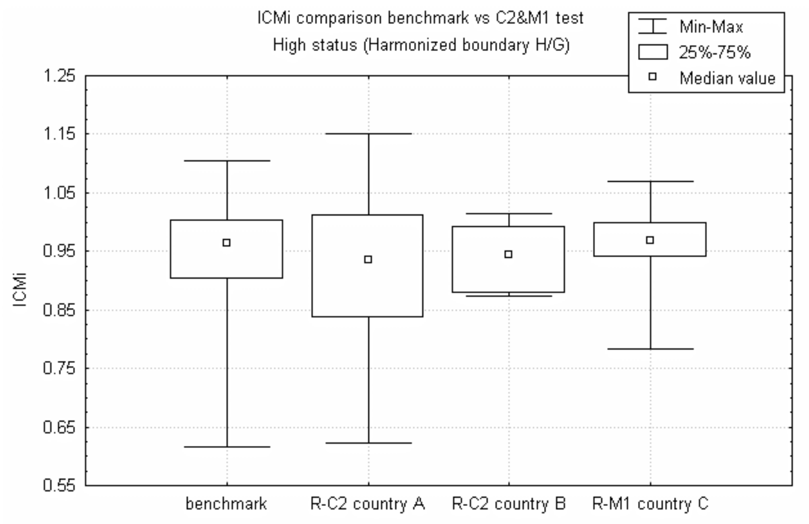


Figure 8.14 Observed range of ICMi values in the High status class after harmonization for the R-C2 and R-M1 types

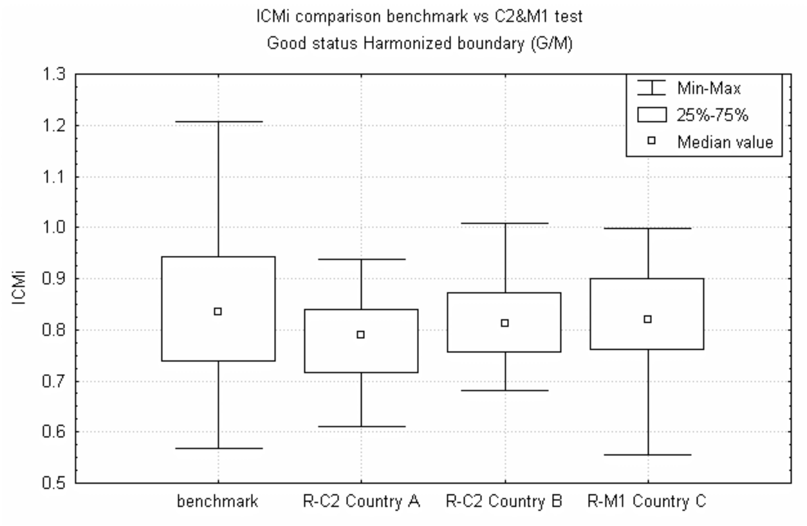


Figure 8.15 Observed range of ICMi values in the Good status class after harmonization for the R-C2 and R-M1 types

8.4 - Summary of harmonization results via ICMi

In this section the implications of the different Options of harmonization considered above are evaluated in terms of number of samples involved in the movement between classes.

8.4.1 - Harmonization of National Methods' class boundaries through conversion in ICMi value

Firstly, a summary of the results for the process of harmonization on the basis of the Option that implies the calculation of ICMi from National boundaries, by means of linear regression (Figures 8.16 and 8.17) is reported. In the Figures, results for R-C1 type are given separately from those for R-C2 and R-M1. The classification before harmonization is undertaken according to the original National boundaries. For the harmonization, the MS' boundaries have been recalculated from the ICMi boundary value (median value of WFD-compliant methods) according to the linear regression formulae. Boundaries higher than the median value of WFD-compliant methods were not harmonized. In the Figures, when higher boundaries were observed they are also noted. An indication of systems assumed to be WFD-compliant is also given.

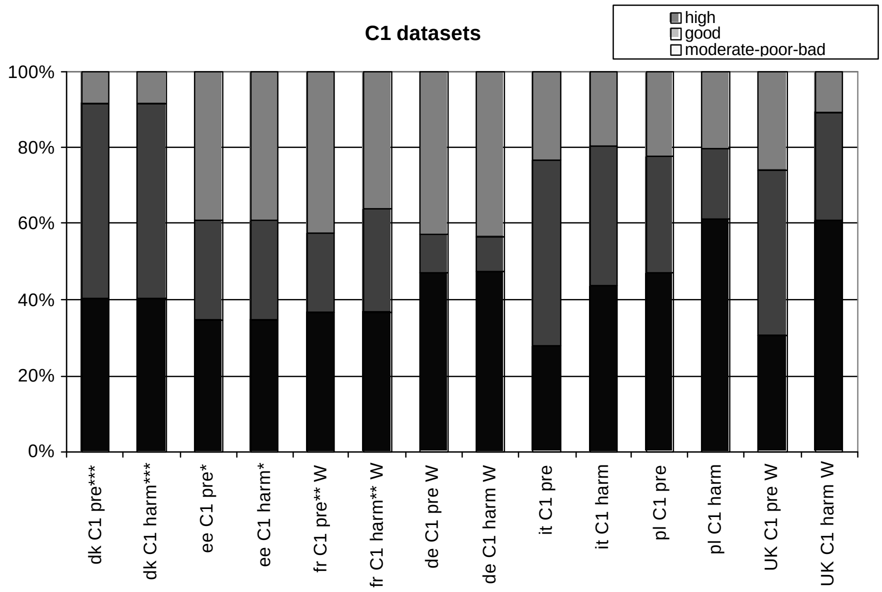


Figure 8.16 Percentage of R-C1 samples in quality classes according to original MS' classification and after harmonization of boundaries (*via* ICMi, without benchmark). (*: boundary High/Good higher than median; **: boundary Good/Moderate higher than median; ***: boundaries High/Good and Good/Moderate higher than median; W: method considered to be WFD-compliant)

Results show that the amount of samples being reallocated to a Worst class is in most cases lower than 10%, with the exception of Italy and UK, which reach a respective percentage of 20% and 30% of samples involved in a movement between classes implying restorative actions.

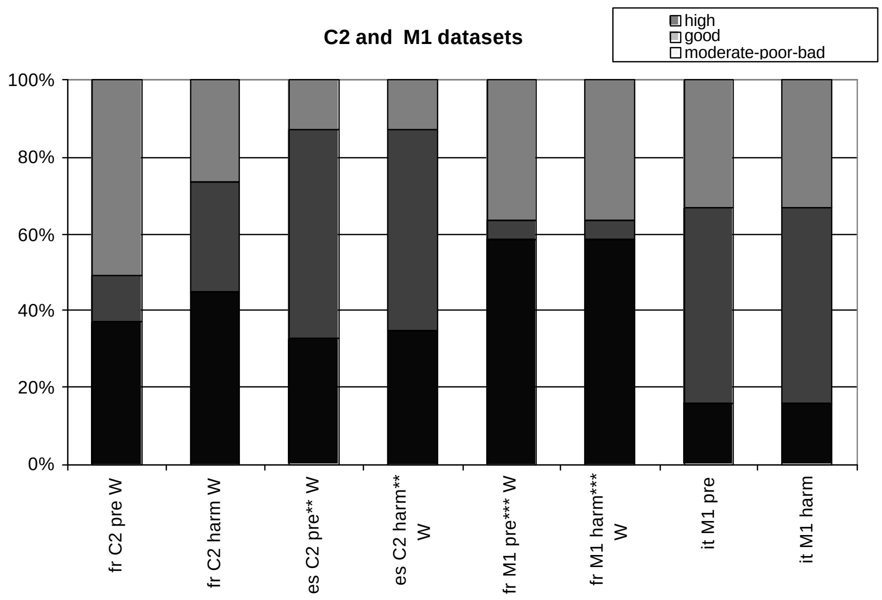


Figure 8.17 Percentage of R-C2 and R-M1 samples in quality classes according to original MS classification and after harmonization of boundaries (via ICMi, without benchmark). (*: boundary High/Good higher than median; **: boundary Good/Moderate higher than median; ***: boundaries High/Good and Good/Moderate higher than median; W: method considered to be WFD-compliant)

8.4.2 - Harmonization via ICMi of National data against an International, benchmark dataset

The effects of the harmonization *via* benchmark were evaluated both in terms of countries that show differences from benchmark and in terms of percentage of samples involved in a movement between classes. In Figure 8.18, the median value for samples belonging to the High status class before and after harmonization is presented for all types under consideration. The black line represents the median value for the benchmark High status class. As a general result, it can be said that out of all the examples considered only 37% of countries seem to need to adjust the High/Good boundary (3 countries out of 8, Fig. 8.18). 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 the benchmark dataset but significant differences are found only for France (R-C2 and R-C1) and Italy (R-C1). The repositioning of the H/G boundary involves very minor changes for R-C1 types (E and F in Figure 8.18). R-M1 type (Italy, letter C) is the only country to have a slightly higher median value than that of benchmark. Poland (G) presents quite a low median value but no significant differences were found. This is probably due to the variability of the ICMi in the dataset, with a high maximum value (Tab. 8.7), but also to the number of tested samples, which is quite low (i.e. 11 High status samples). In general terms, it has to be stated that probably, when more samples are included, the results of the statistical test might change, determining the need for a slight harmonization. In general, the percentage of samples moving from Good to Moderate status is around 10% for most of the examples considered (with the exception of C1 Italy).

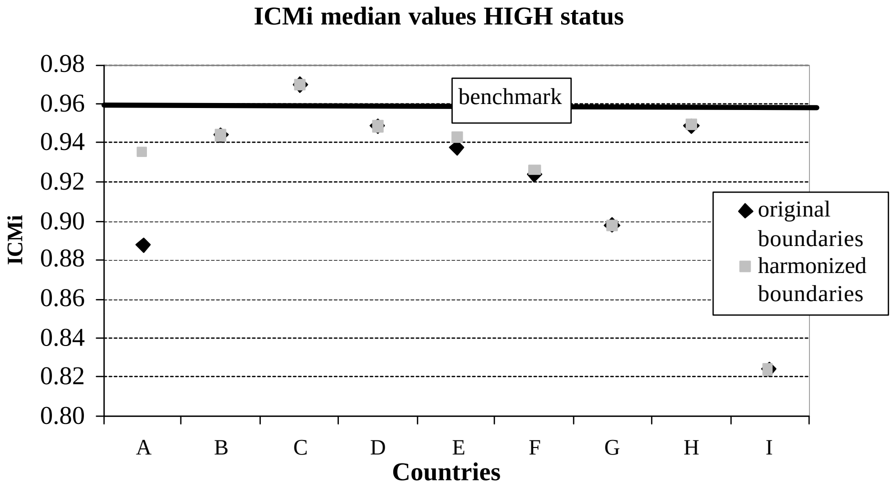


Figure 8.18 Median value (High status) of ICMi for the benchmark dataset (black line) compared to the values obtained for test methods and datasets for the same quality class (before and after harmonization)

In Fig. 8.19 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 status class within the benchmark dataset. It can be seen that all the samples have lower median values than the benchmark data with the exception of the UK (letter H). For the Good/Moderate boundary, 4 out of 9 countries might have to move their boundaries, because in these cases significant differences were found between test and benchmark data. The median values of almost all countries come quite close to the median value of benchmark data after harmonization, with the exception of Poland (letter G), whose median remains the lowest and corresponds to 0.76.

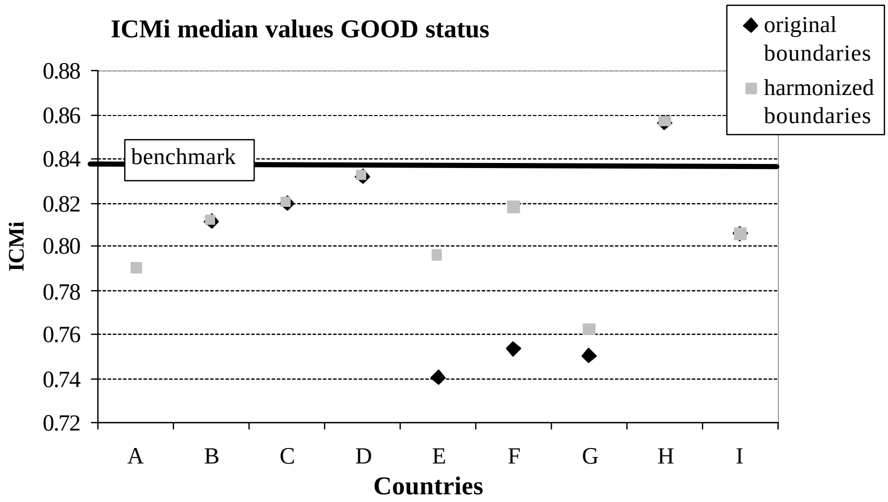


Figure 8.19 Median value (Good status) of ICMi for the benchmark dataset (black line) compared with the values obtained for test methods and datasets in the same quality class (before and after harmonization)

In Table 8.7 (above), a summary of the basic statistics for the examples tested and harmonized and for the benchmark dataset is reported.

Table 8.7 Summary of the basic statistics for the test – before and after harmonization - and benchmark datasets.

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		
25%ile	0.905	0.803	0.877	0.940	0.942	0.827	0.859	0.825	0.896		
75%ile	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		
25%ile	0.905	0.897	0.877	0.940	0.942	0.887	0.868	0.825	0.892		
75%ile	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	
25%ile	0.743	0.611	0.755	0.760	0.741	0.669	0.616	0.630	0.782	0.771	
75%ile	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	0.985	
25%ile	0.736	0.713	0.755	0.760	0.741	0.859	0.702	0.663	0.782	0.767	
75%ile	0.944	0.839	0.872	0.900	0.948	0.859	0.900	0.913	0.922	0.873	
median	0.839	0.790	0.811	0.820	0.832	0.796	0.818	0.763	0.856	0.795	

For all the countries whose Good/Moderate boundary showed differences from the benchmark value, the implications in terms of samples actually moved to the Moderate status class look minor (see Figure 8.20).

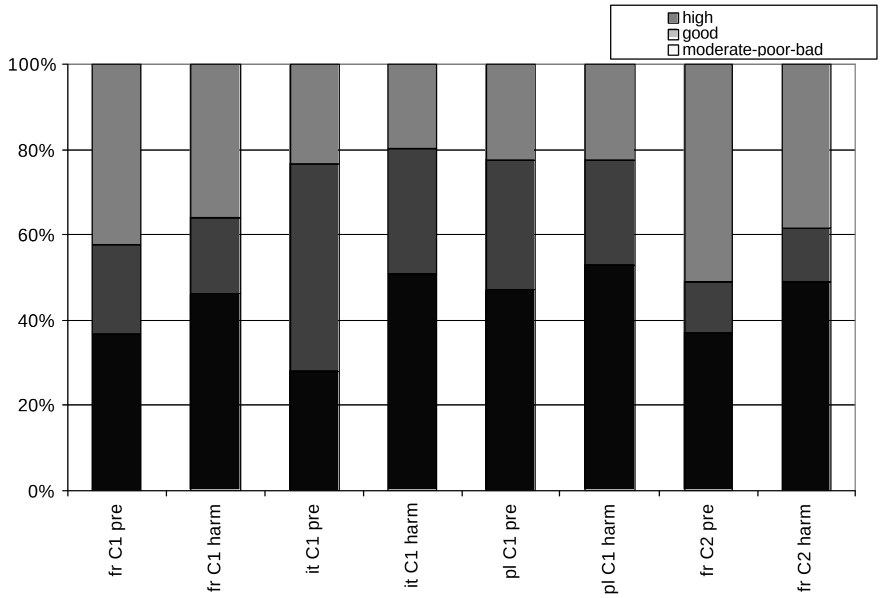


Figure 8.20 Percentage of samples (all river types) in quality classes according to original MS classification and after harmonization of boundaries (via ICMi, with benchmark)

8.5 - Discussion & general considerations

Some general considerations are reported here in relation to the three options of harmonization tested, regardless of whether or not they are based on the calculation of an ICMi.

8.5.1 - Averaging class boundaries of national methods - no ICMi

The pre-requisite for the harmonization of class boundary values via averaging is the inclusion of all assessment methods used to evaluate the quality of the respective common Inter-calibration stream type in the GIG. The methods must also be demonstrated to be WFD-compliant. This chapter presented only an example of the Option. The performance of harmonization is demonstrated by the increase of equally classified samples from 18 to 44 percent (see 8.1.1). For two reasons, high percentages of conformity between individual classifications cannot be expected:

(1) The more assessment methods that are involved in the process of harmonization 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.

In general, harmonization 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 the class boundary values of all assessment methods applied to a common stream type can represent an alternative to the harmonization approach using benchmark datasets (cfr. 8.3 and the following paragraphs). Both direct and indirect comparison approaches may serve as basis for this harmonization 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.

8.5.2 - Harmonizing via ICMi using the median value of WFD-compliant methods

For the Option of harmonizing via ICMi median value of WFD-compliant methods (Chapter 8.2), the results are strongly dependent on the included datasets. The inclusion or exclusion of one or more datasets has an obvious influence on the median value and thus on the final boundary re-definition. Thus, such exercises should include sets of data from all participating countries, whose systems should also be demonstrated to be WFD-compliant. Following the overall indications of the Inter-calibration guidance (EC, 2004), this should be done for every stream type within a GIG. If suitable, the exercise should be extended to a broader scale, e.g. an entire GIG, or even across GIGs. This Paper, which aims to illustrate possible procedures of European harmonization, has shown that in most cases the differences in boundaries expressed as ICMi value are very small.

One problem that has emerged during the analysis related to this Option was the different classification derived at by converting the National boundaries into ICMi values by linear regression. The classification according to the MS' original boundaries was compared to the classification derived from National boundaries converted in ICMi values by regression. In the ideal situation, i.e. with an R^2 value equal to 1, the two classifications coincide. On the contrary, depending on the characteristics of the dataset and on the distribution of the samples around the regression line, even when having high R^2 values (>0.80) some samples classified in a specific class by the National method can be differently classified when considering the boundary expressed in ICMi value. Three circumstances are possible:

1. after the conversion of the boundaries to ICMi EQR the samples are classified in the same way as considering the boundaries expressed in terms of EQR of the National method;

2. the results obtained by using ICMi EQR give a more optimistic view of the classification (i.e. higher quality) compared to the National classification;

3. the results obtained by using ICMi EQR give a more pessimistic view of the classification (i.e. lower quality) compared to the National classification.

These differences in classification were evaluated both considering the original boundaries of the National assessment systems and the boundaries after the harmonization carried out considering the median value of WFD-compliant methods. The following Figures (8.21 to 8.27) show the different percentages of classification for High and Good status samples for selected datasets before and after harmonization in the three cases explained above. The examples were selected in order to represent countries and types for which different R^2 values were observed and to present a different number of samples. The white histograms show the classification according to the original boundaries provided by MS while the grey ones represent the corresponding classification after harmonization.

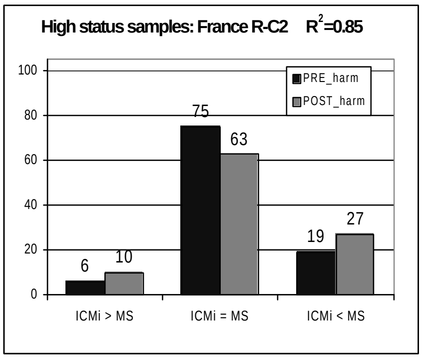


Figure 8.20. Percentage of samples classification for High status according to ICMi and MS method, before and after harmonization for France R-C2. (original N=73). Use of ICMi by linear regression with the National method (no benchmarking).

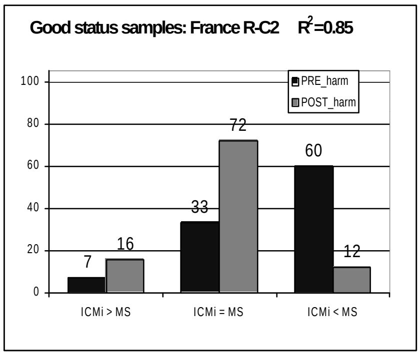


Figure 8.21. Percentages of samples' classification for Good status according to ICMi and MS method, before and after harmonization for France R-C2. (original N=17). Use of ICMi by linear regression with the National method (no benchmarking).

For the dataset 'France R-C2' (Fig. 8.20- 8.21; R^2 0.85), it is possible to observe that before harmonization the percentage of agreement between the two classification options is 75% for High status. The 6% of samples are classified in High status according to ICMi, while IBGN gives a Worst classification. 19% of samples are classified as High status by the MS' method while the ICMi boundary indicates a Good status class for the same samples. For classification in Good status (Fig. 8.21), the percentage of agreement is quite low before harmonization (33%), with 60% of samples classified in a status lower than Good according to ICMi. After the harmonization, differences are less important with a classification agreement equal to the 72%.

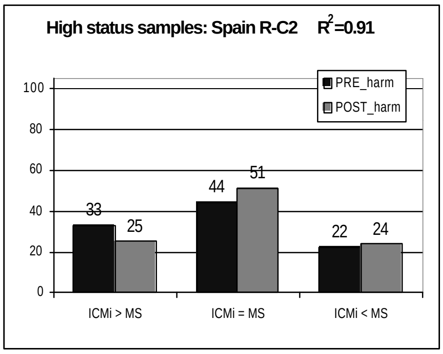


Figure 8.22. Percentages of samples' classification for High status according to ICMi and MS method, before and after harmonization for Spain R-C2. (original N=7). Use of ICMi by linear regression with the National method (no benchmarking).

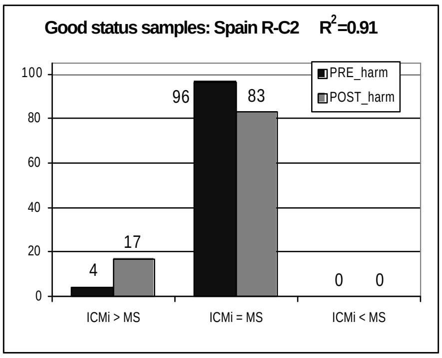


Figure 8.23. Percentages of samples' classification for Good status according to ICMi and MS method, before and after harmonization for Spain R-C2. (original N=25). Use of ICMi by linear regression with the National method (no benchmarking).

For the dataset 'Spain R-C2' (Fig. 8.22 and 8.23; R^2 0.91), the percentages of agreement among the two classifications for the High status samples is relatively low before (44%) and after harmonization (51%). In contrast, for Good status the agreement is almost absolute both before and after harmonization. The higher percentage of agreement with respect to France is presumably related to the higher R^2 value (i.e. best fit of National and ICMi methods) but also to the lower number of samples in the dataset

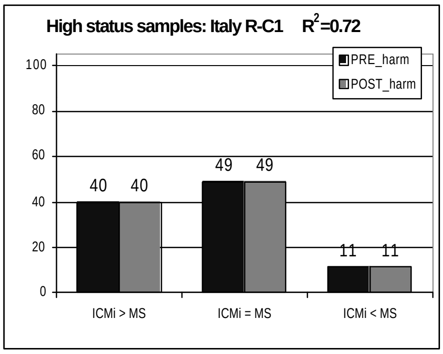


Figure 8.24. Percentages of samples' classification for High status according to ICMi and MS method, before and after harmonization for Italy R-C1. (original N=84). Use of ICMi by linear regression with the National method (no benchmarking).

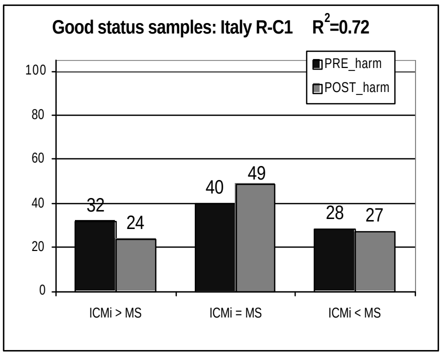


Figure 8.25. Percentages of samples classification for Good status according to ICMi and MS method, before and after harmonization for Italy R-C1. (original N=176). Use of ICMi by linear regression with the National method (no benchmarking).

For the dataset 'Italy R-C1' (Fig. 8.24 and 8.25; R^2 0.72), the percentage of agreement is around 50% both for High and Good samples before and after harmonization. For this dataset the percentage of samples classified in a higher status class by ICMi is higher than the opposite case (especially for High status samples).

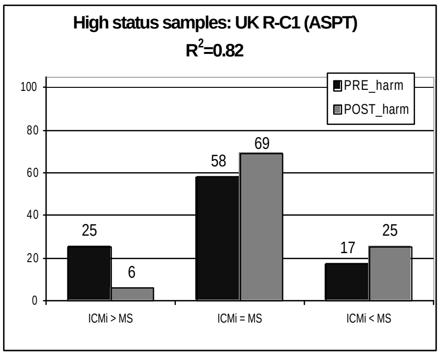


Figure 8.26. Percentages of samples classification for High status according to ICMi and MS method, before and after harmonization for UK R-C1. (original N=317). Use of ICMi by linear regression with the National method (no benchmarking).

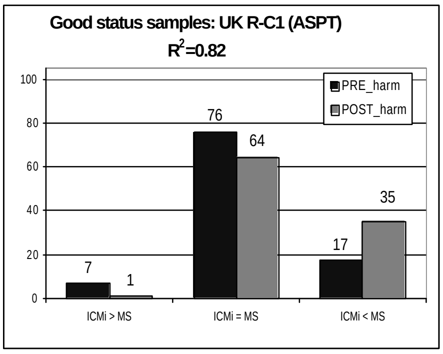


Figure 8.27. Percentages of samples classification for Good status according to ICMi and MS method, before and after harmonization for UK R-C1. (original N=269). Use of ICMi by linear regression with the National method (no benchmarking).

For the dataset 'UK R-C1' (Fig. 8.26 and 8.27; R^2 0.82), the percentage of agreement is around 60% both for High and Good samples before such as after harmonization with a value of 76% of agreement before harmonization for Good status samples. Here the ASPT classification is reported only as an indicative example because in-fact the classification used in the UK is the result of the one out all out among ASPT and Number of families.

In these examples, a good number of samples are classified in the same status by ICMi and MS' method, but the percentages of disagreement are noteworthy in all cases. The amount of disagreement is dependant on the distribution of samples and it may be argued that, even when R^2 is high, the disagreement may be high too. At least partly for the High status class, this can be due to a relative lack of samples of very good quality (i.e. high index values in the High status class), that would clearly increase the percentage of concordance between the two methods for this class.

These considerations show the actual possibility of making mistakes in the attribution of quality classes after an exercise of harmonization is carried out according to this procedure. The re-calculation of the class boundaries from ICMi to MS method, and viceversa, by means of linear regression, leads itself to a misclassification compared to the initial National classification. This might be due to the possible non-linearity of the relationship between ICMi and National methods and to the information they provide, which might be different i.e. determining a relatively large variability around the regression line (on this point, see also 8.5.1).

8.5.3 - Harmonizing via ICMi by using an external, WFD-compliant benchmarking system

The option of comparing the classification results derived by a MS' standard method and those 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 the three stream types from the Central and Mediterranean GIGs. This comparison was based on the values obtained for each sample by calculating the ICM index. The general results for the studied stream types highlighted how, when needed, only a small refinement was usually sufficient to set new boundaries in the quality classes of the National method, to fit the ICMi values found in the benchmarking dataset (WFD-compliant, BAC-based clustering of samples).

As expected, the situation found across Europe is not fully homogeneous, with some countries and boundaries not being statistically different from the benchmarking dataset for any classes, some with discrepancies for one of them and others with differences in both the relevant boundaries (H/G and G/M). 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 status) is too generous for the studied stream type.

A slightly larger percentage (0-12%; in one circumstance 22%) of the samples initially classified into quality class II (Good status) by the national classification schemes had to be moved to class III (Moderate status).

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 presumably lead to restoration measures in a sub-set of the sites only. The procedures applied here and the illustrated results can support the involved MSs in the revision of class boundaries to bring their methods in line with WFD requirements.

On a larger scale, as 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, in a previous example of the application of benchmark, data were derived from a single eco-region (Italy: Buffagni & Erba, 2004, see Annex III), here we used a trans-National, inter-GIG database, which is expected to include rivers quite dissimilar in general character. In IC applications the boundaries for a stream type are expected to be re-set no lower than at the GIG scale. Based on the examples provided here, we presume that a trans-GIG Inter-calibration might be tentatively adopted.

Even if exclusively comparing metrics based on a high taxonomic resolution (i.e. to the Family level), the variation between 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: R-C1) has a counterpart in the benchmark dataset, which nevertheless contains more data from 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 test datasets and stream types considered in the present Paper, only one of the types resulted in being hardly comparable with others. Not surprisingly, it was R-M5, which corresponds to south European temporary rivers. Apart from this single 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

9.1 - Overall considerations

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 Inter-calibration of class boundaries of biological assessment methods have been outlined and preliminarily tested. Each can be potentially applied depending on the type and amount of data available, the proximity of methods to be harmonized and the availability of reference sites etc.

The approaches to Inter-calibration proposed here mainly refer to Option 2 and hybrids of the IC Guidance. These Options are related to the use of Common Metrics specifically developed to run the IC exercise.

The main features and use of common metrics for the IC process (ICMs) have been described and applied to a number of questions and situations in Europe, covering a wide geographical range.

The comparison of a relatively high number of European MS' datasets has been performed comparing National methods directly or by using a simple ICM index (ICMi) to make them fully comparable. The general outcome indicates that the ICMi approach is suitable for comparing rivers and invertebrate communities in a wide range of circumstances.

To test the applicability of the ICMi procedure, 14 datasets from 10 countries were collated and analyzed. In a relatively short time, it was possible to apply the whole procedure to all dataset concurrently. Data provided by MSs were used - rough as well as re-elaborated - at a 'central' level. A great contribution to this phase was provided by the different MSs in actively giving comments on the results produced for their countries. Particular situations and specific topics were highlighted within the different analyses (e.g. the problem of setting reference conditions and the performance of metrics among different stream types).

Data was also collected and collated from EU co-funded projects (AQEM and STAR) in order to: a) establish a benchmark dataset and b) to test further comparison and harmonization procedures. In the future, the idea is to have an external dataset against which to compare the results of the ICMi calculation from test datasets (National data)(a) and to have samples collected with the same field approach to increase comparability (b).

The main features of such a benchmarking dataset deal with its full WFD-compliance and with the eminent comparability of data collected. Related to the use of benchmark data is the concept of Best Available Classification (BAC), which represents the ecological classification obtained by applying a WFD-compliant procedure and all the available, relevant information on a site.

9.2 - Metric-related aspects

The response of many biological metrics, including ICMs, was analyzed for groups of test stream types along observed pressure gradients.

- Metrics were selected in order to fulfil WFD requirements and be applicable over a wide geographical range. While a few metrics performed sometimes slightly better (i.e. species-level metrics), ICMs and ICMi demonstrated a very good general attitude, especially considering that they are based on a Family level identification.

- ICMs and ICMi were stressed against a high number of European biological assessment methods and very high correlations resulted. (i.e. they are able to describe the quality gradient actually detected by the methods presently in use). In general, the metrics selected were proven to adequately reflect the gradient as expressed by National assessment methods, being in most cases able to discriminate well among the different quality classes (expecially Good and Moderate).

- ICM metrics and index were also directly tested, for two example stream types, in relation to pressures. Results show a very good performance, especially when combining different pressures.

The ICMi approach supports the use of existing datasets directly collected by MSs, which can guarantee the 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.

9.3 - Harmonization and comparison options

The comparison exercise between European class boundaries and assessment systems led to different results for different stream types and Options used, but showed how systems and boundaries can be actually compared in the short term.

Examples of harmonization have been presented following three different approaches, which can be used individually or combined in different GIGs and European areas. All the results have to be considered preliminary, especially in respect of the fact that the procedure to set reference conditions is still under development in many countries. Moreover, many countries are currently working on their assessment system(s) in order to make them compliant with the WFD requirements.

9.3.1 - Harmonization via direct comparison of National methods

- The direct comparison approach (i.e. not using ICMs) has been used to demonstrate apparent discrepancies between MSs' assessment systems boundaries.
- These differences can be due to the fact that the existing methods have different sampling strategies and laboratory procedures, and are also based on different concepts. Jointly with the fact that fauna can differ for bio-geographical reasons even in similar physical contexts (and methods were locally developed to take these differences into account), this sometimes makes the option of applying different methods to the same samples not truly applicable.
- Yet, this direct 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. When testing different methods, the percentage of sites that might require a shift from the High and Good status to the Moderate status is in most cases around 30%.
- The option of averaging the values of class boundaries of MSs' assessment methods is only applicable when all the biological methods considered are demonstrated as fully WFD-compliant.

9.3.2 - Harmonization via ICM index – No external benchmarking

- The ICMi approach, by using reference conditions and data normalization set within each of the datasets under comparison, allows a large variety of Intra- and Inter-GIG comparisons (i.e. it supports a large pan-European comparability).
- The differences among boundaries observed with comparison *via* ICMi were lower than those obtained by direct comparison – even comparing datasets from different GIGs - ranging from 0 to around 10% (usually lower than 5%). The implications in terms of river restoration are quite low. In fact, in most cases less than 15% of samples had to be moved from High-Good status to a Worst condition. Only in one case (UK) 30% of samples may have to be moved from the High-Good status to a Worst condition.
- The option of averaging (or using the median) of the ICMi values for the class boundaries of MSs' assessment methods is only applicable when all the biological methods considered are demonstrated as fully WFD-compliant.
- The use of the regression formula to convert national boundaries into ICMi values, necessary for the Option of averaging the values of class boundaries of MSs' assessment methods *via* ICMi, leads to percentages of non-matching classification ranging from 4% up to 48%. High percentages of different classifications are found even when the R^2 value for the regression model is high (e.g. >0.8). This means that, even when starting from the original classification provided by MSs, some sites/samples move from their original classes after the simple conversion into ICMi.

9.3.3 - Harmonization via ICM index – Use of external benchmarking

- By using an entirely external benchmarking system (in the present Paper, the AQEM/STAR WFD-compliant dataset), the ICMi can be used to harmonize class boundaries within and between GIGs, achieving a full comparability and lack of ambiguity in results.
- The starting point of comparison and harmonization *via* ICMi using a benchmark dataset is the original classification provided by MSs. The ICMi values for all samples belonging to a quality class defined by the MS' method are compared within values obtained for the same class in

the benchmark dataset. This option (a full Option 2) overcomes the problems related to the use of regression *formulae*.

- The use of a 'hybrid' Option 2 – e.g. by using ICMi approach and not an external, supra-National benchmarking system - seem fully suitable for comparison purposes, while for the harmonization phase (i.e. the potential re-setting of class boundaries) a full Option 2 appears to be a more appropriate solution.
- In fact, the procedure of e.g. averaging class boundaries (directly or by using an ICMi) gives MSs greater opportunity to look for a rather pure 'political' agreement on boundaries. On the contrary, the comparison to an external and agreed benchmarking dataset supports a more objective (i.e. scientifically sound) definition of classes, being less dependant on a single MS' view of boundary setting. In any case, when boundaries of different assessment systems are alike, the problem of averaging is overcome. The true problem with any averaging approach is the definition of the acceptable deviation: how big does the difference have to be for it not to be considered acceptable?
- The harmonization approach via ICMi and external benchmarking does not require fixed class boundaries to be defined. Moreover, it allows a step by step adaptation of National method boundaries until no more differences are observed between National samples and benchmark samples, included in the Good and High quality classes. A substantial advantage of this approach is the possibility of performing a statistical comparison against a common and invariable dataset.
- The complete harmonization exercise provided here *via* ICM approach (in the full IC Option 2 application) led 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 a benchmark dataset handles the problem of not having fully WFD-compliant systems presently available for all MSs. If the comparison of the tested datasets with benchmark datasets does not show significant differences, it means that the tested method can be considered to provisionally fulfil WFD requirements on how class boundaries are set.

9.4 - General warnings and problems

A few important warnings in general, including problems that can be highlighted.

- A general limitation to the validity of the results for all the methods tested is the number of samples available. For the comparison and harmonization *via* averaging or by median value of WFD-compliant methods, the method is dependant on the number of datasets included. The inclusion of data from all MSs involved highly favours the success of the IC process. Also, for all methods, when there are very few samples are available, it is possible that even where differences exist they will not be revealed by statistical comparison.
- An important requirement for a successful application of most of the described procedures is the availability of datasets covering the whole degradation gradient. Besides, non-continuity in the sample range can cause problems in the interpretation of the regression model and in the aptitude of testing the outputs from different classification Options.
- For some stream types, low R^2 values between ICMi and National method were found. This can be related to the fact that the two methods provide information of a different kind. Actually, this mainly happens when the ICMi is compared to highly focused systems (e.g. saprobic system), because ICMi was developed to detect the general degradation of river sites.
- A situation of poor regression with ICMi was also found for some German datasets that could be potentially included in the benchmark dataset. Here, specific impact types were acting on the sites (e.g. hydromorphological alteration). In this case, the ICMi does not seem appropriate for representing in detail the ecological quality gradient among sites, better described by e.g. species level multimetric systems. Nonetheless, two examples for which ICMi works well even in relation to a specific impact type, are provided for Italian and Austrian rivers.
- In a few datsets, a high dispersion of data was observed in the regression models. In some cases, this can be related to the presence in the same dataset of samples/sites probably not strictly belonging to the same stream type. On this point, some partners have highlighted the need for a better definition of river types. It should be stated that the purpose of a broad definition of types within the IC process is to allow countries to find suitable sites and data for the IC exercise, e.g. according to the National geographic and monitoring situation. If different 'valid' sub-

types can be found within a single IC type or when the sampling technique used is different, data have to be normalized according to these sub-types and sampling methods.

- In general terms, a major advantage of the ICMi approach is its suitability for making different assessment systems comparable and not in its greater power to discriminate among different quality classes (→ comparability as a starting point). It is therefore suggested that an attempt be made to retain the same ICMi for different types and GIGs in order to guarantee a full pan-European comparability, even if for specific datasets and types it does not perform as well as for others. Indeed, even if the ICMi approach has proven to be satisfactory in representing the ecological gradient, further refinements may be needed for specific stream types.
- If calculated on the basis of MS' biological protocols only, a 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 does not guarantee conformity to WFD.
- When a MS' classification system is based on two different metrics (e.g. UK and Poland), used according to the 'one-out all-out' principle, additional difficulties appear. In such situations, the method for defining a single, final boundary should to be clarified, in order to suitably compare the MSs' system to that of others using a simpler system (or that use an averaging option between different metrics). This aspect is especially important when directly comparing methods without a benchmark.

- For some stream types/countries reference sites are not available (e.g. in the Netherlands). This is a problem that has to be addressed. A possible solution could be the use of reference values from neighbouring countries.
- A common restriction on all the possible procedures in the IC process is the scarce availability of data from reference sites.
- Apart from the availability of reference sites, a further question to be faced is the fact that not all countries have well established reference condition criteria. It is recommended that for the formal Inter-calibration exercise reference conditions will be defined. For this preliminary phase of the process, a surrogate for reference condition values could be the use of the 75^{th} percentile of High status samples/sites.

- When comparing different methods *via* ICMi without using a benchmark, if a MS has characterized the High status class by including samples of somewhat low quality (i.e. close to the lower boundary of the class) – e.g. because those samples are the best available ones - the normalization can lead to inconsistent anchor values for the High status class compared to other countries. This fact, which can have important implications on further steps of the Inter-calibration process, may not be as apparent as assumed and may strongly limit a correct comparability to other datasets/methods.
- With regards to the problem of the definition of reference sites/samples, it is our recommendation that reference conditions be defined on the combined analysis of pressures and biological data. This means that on the simple basis of a pressures estimation, a site should not be accepted as a reference site. A validation through analysis of biota is fundamental to ascribe the site to a natural or nearly natural condition (i.e. reference site/sample).

10 - NOTES FOR WATER MANAGERS

Based on the experience performed during the STAR Project, GIGs activity, ECOSTAT discussion, etc., partly summarized in the present report, it is possible to outline a few points that might be especially useful to water managers.

Some of the issues are reported below.

1. Is it really possible to make the results of the biological classification systems comparable across Europe?

Based on the pilot exercises performed within a number of Geographic Inter-calibration Groups and the STAR Project, it seems realistic to compare the data and classification results of different countries and stream types, provided that there is agreement on how to derive reference conditions. In detail, the approaches tested in the present paper demonstrated that selected biological metrics – named Inter-calibration Common Metrics (ICMs) – can be advantageously used for the purpose. They were consistently applied to a wide range of European stream types and datasets. ICMs, once combined in a simple multimetric index (ICMi), are able to cover most European stream types. Notably, they can be locally adapted/modified to particular conditions e.g. across GIGs, for temporary rivers. Other approaches can also be used to fit local situations better.

2. Can we be reasonably sure that the method(s) used to Inter-calibrate will adequately illustrate the environmental gradients described by existing National methods?

Details are provided for a relatively high number of countries and biological methods including the Central-Baltic and Mediterranean areas of Europe. Although the performance of the metrics and methods tested can vary slightly in different contexts, the overall performance of some approaches (e.g. the ICMi approach) is good. In most cases, the ICMi is able to satisfactorily describe the environmental gradient fit of the National methods (R^2 usually comprised between 0.5 to 0.7).

3. Are there any constraints related to specific impact types in the application of the method?

It has been demonstrated that ICMs and ICMi reflect well the ecological gradients described by MSs assessment methods and some examples of good performance were provided, even in relation to specific impact types (e.g hydromorphological degradation). Important requirements in correctly adopting the procedure are the agreement on reference conditions and the availability of data from the whole quality gradient. In addition, the idea that we are now proceeding with the harmonization of class boundaries related to the overall ecological quality (i.e. general degradation) of a site must be borne in mind.

4. Can we use the monitoring data we have been collecting for years for the purpose of the European Inter-calibration process?

Yes, the tested Options and approaches are based on the assumption that the IC process will be mainly performed on standard monitoring data and methods. Nevertheless, this does not mean that MS can not Inter-calibrate newly proposed methods or those under development, set to fit WFD requirements. One MS can simultaneously compare more than one method. In addition, even if a very small number of sites have to be statutorily included in the formal IC register, to give scientific foundation to the IC process, much more information and sites should be considered. This will offer a good opportunity to make use of and disseminate the existing datasets.

5. Is the collection of information gained by applying detailed, National protocols (e.g. species level identification, quantitative sampling, well developed predictive systems) going to be superfluous after the European Inter-calibration is concluded?

No, because IC is only one of the key phases of WFD implementation. The amount of information required for the IC process, while being exhaustive in some respects, is undoubtedly smaller than the knowledge needed to properly monitor, manage and restore European water bodies. Even for single monitoring activities, which are the first step to satisfactorily put the WFD into practice, information as detailed as possible is essential e.g. to detect the causes of degradation or to have large-scale changes of aquatic ecosystems under surveillance.

6. Some MSs and the European Commission itself funded a lot of research activities to support the WFD implementation or to directly develop fully WFD-compliant methods. Can the information obtained be used for the IC process?

Yes, this information is essential for the IC process. Some innovative issues put forward by the WFD, such as the type-specific approach, the need for setting measurable reference conditions, the use of supporting hydro-morphological elements, the search for a common understanding of ecological concepts, a combined pressures analysis, etc., do need scientific support to be effectively encompassed in the IC process. For instance, the so-called 'Boundary setting protocol', which will be the central focus of the whole IC process and WFD implementation, will have to take care of research projects results. Ideally, all available high-quality data – such as the STAR/AQEM datasets and other

European or MS' sources – should be used to set a common benchmarking system for testing National methods and classification.

7. Are all differences currently existing between classification systems going to disappear after the IC process?

An imperative point of the IC process is that any final decisions will be discussed and agreed upon as part of the socio-political background. Furthermore, WFD demands for major differences among classification systems to be harmonized and will consider allowing minor dissimilarities, which are inherently linked to methods in use, tradition or geographic distinctiveness.

8. Are there any serious risks that the actual class boundaries of Member States will result in being very different from one country to another?

The results presented in this Paper, which we are confident are representative of the larger European scenario, encourage an optimistic view of the current situation and of the changes that might be required to make all MSs systems equivalent. In addition, where relevant differences were observed, it often appeared that the comparison was influenced by a lack of knowledge (e.g. difficulty in setting calculation options and little data from reference sites), the upgrading of which would probably lead to a reduction in such differences.

9. What impact would a possible European Inter-calibration have on the actual distribution of river sites among quality classes?

Out of all the comparisons made, only in a few situations the Good/Moderate class boundary appear statistically different between MSs and external benchmarking systems. Even in those cases, harmonization of classification scales is not likely to have an unreasonalble impact on national assessment systems. According to the approaches tested here for invertebrates, the percentage of sites moving from Good to Moderate status, varies between 0% (ca half of the MSs systems) and 30% (maximum observed with ICMi approach, one occasion), for a range of 11 MSs datasets tested.

10. Will the IC process constrain MS to abandon their assessment systems?

No. The WFD demands the development/adoption by MSs of suitable methods to assess ecological quality. If a MS already has a WFD-compliant method in use, there is no reason why it should be changed for the IC process. Nevertheless, most European countries do not have a full set of WFD-compliant methods and it is likely that MSs – as a minimum measure – will modify and adapt their methods to fit the WFD better and scrupulously run the IC exercise.

11. Is the Inter-calibration process going to press MSs to adopt any common biological assessment systems?

No. The methods being used as common systems for comparison (e.g. ICMi) in the IC process are not designed to substitute existing classification methods. Nonetheless, if they can be demonstrated to perform better and likewise for existing methods where no standard methods are in use, they can be easily utilized to classify river sites accordingly to pan-European rules.

Finally, what the IC process is aiming for, is to make class boundaries comparable across Europe, especially for Good/Moderate and High/Good limits. It is not a direct comparison of methods, but samples and boundaries are matched to make the results of the classification analogous across Europe. In any case, the comparison phase can be usefully kept distinct from the harmonization phase.

12. Is the IC process accessible for newly associated countries or MS with a short tradition in biomonitoring (e.g. with small datasets available or superficial investigation only)?

Yes. The approaches tested and Options proposed relate to the simplest possible solutions for running the IC process. For instance, the family level identification for invertebrates was adopted to make most data from around Europe operative. When only small datatsets are available, a few additional collections can support a quick integration into the IC process.

13. How much time will be necessary to perform the comparison between the full gamut of river types present in Europe?

It strongly depends on data availability. If enough data is available, in terms of BQE as well as pressure data, the main constraints are linked to the discussion and agreement phases of common protocols and procedures. The 'technical' IC process *per se* might not be as long as formerly believed. For BQEs, for which assessment methods are still under development (e.g. fishes, macrophytes), the IC exercise will be likely start at a further phase of the process.

14. Are MSs expected to inter-calibrate class boundaries for all Biological Quality Elements in all stream types and for all the kinds of pressure?

No. BQEs should be considered for Inter-calibration only when its employment in terms of bio-indication is contemplated because of its usefulness to detect impact. Thus, only selected combinations of BQEs, stressor types and water body types have to be regarded. For rivers, the macroinvertebrate component can be constructively Inter-calibrated for a wide range of pressures and will provide important elements to guide the IC process during its first phase.

15. To aid the IC process, what kind of new data would be especially useful?

Usually, data to characterize reference conditions is scarce. Thus new collections might encompass field activities at potential reference sites, for different river types. Additionally, data on the full range of degradation of river sites in a given area would greatly increase the likelihood of properly describing the environmental gradients and setting reliable class boundaries. It is important to state that in the collection of new data for any aspect under study, a well defined, standard procedure across the whole MS' region is of invaluable benefit.

16. Can we make use of the proposed approach(es) with other BQEs or water body types?

The examples provided mainly deal with the invertebrate community (i.e. macrophytes in one case only). However, the approach of selecting, developing and using common Inter-calibration metrics (ICMs) to describe in a shared way the ecological gradient and status can be quite easily extended to all the BQEs for which data are available. For water bodies other than rivers, the applicability of the whole procedure or that of single steps, strongly depends on the datasets currently available. Nonetheless, this same constraints act in more general terms on the whole WFD implementation process.

11 - SHORT GLOSSARY

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.

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.

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.

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).

BMWP. Biotic index: Biological Monitoring Working Party score (Armitage et al. 1983). In Spain, a modified version (BMWP') is used for river classification.

BQE. Biological Quality Element (Water Framework Directive).

BBI. Belgian Biotic Index (De Pauw & Van Hooren, 1983). Biotic index in current usage in Belgium to classify rivers based on macroinvertebrates.

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.

CIS. European Common Implementation Strategy for the Water Framework Directive.

Class boundary. The EQR value representing the threshold between two quality classes.

DL 152/99. Decreto Legislativo 152/99. Decree of the Italian legislation that rules on actions for water protection. It contains all the normative indications for the monitoring of water bodies.

DSFI. Danish Stream Fauna Index (Skriver et al., 2000). It is used as a standard basis in Denmark to classify rivers based on aquatic invertebrates.

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).

ECOSTAT. CIS Working Group 2 A dedicated to the Ecological Status of surface water bodies within the implementation of the Water Framework Directive.

EPT. Total number of taxa belonging to the Insect Orders of Ephemeroptera, Plecoptera and Trichoptera

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.

GIGs. Geographical Inter-calibration 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.

GOLD. Total number of taxa belonging to the Orders of Gasteropoda, Oligocaeta and Diptera.

HMS. Habitat Modification Score. Environmental index derived from data collected with RHS protocol (Raven *et al*., 1997). It allows an evaluation of the morphological degradation of river channel due to human activities (e.g., bank reinforcement, channel re-sectioning, culverting, number of weirs, etc.).

HQA. Habitat Quality Assessment. Environmental index derived from data collected with RHS protocol (Raven et al., 1997). It assesses the habitat diversity and quality at a site through an appraisal of different features such as number of flow types, number of substrates, naturalness of land use, etc.

IBE. Biotic index: Indice Biotico Esteso (Ghetti, 1997; APAT/IRSA-CNR, 2004). Used as standard in Italy to classify rivers based on aquatic invertebrates according to national legislation (DL 152/99).

IBGN. Biotic index. Indice Biologique Global Normalisé (AFNOR, 1992). It is used as a standard basis in France to classify rivers based on aquatic invertebrates.

IC. European Inter-calibration Process for the WFD.

IFF. Indice di Funzionalità Fluviale (Siligardi et al., 2000) Environmental index widely applied in Italy. It was developed originally to asses the efficiency of a river in cycling organic matter. It allows the assessment of the overall ecological quality of a river site and the potential functioning of the river ecosystem.

Inter-calibration (Harmonization). The process by which the class boundaries of MS National methods should be accommodated to correspond to a common understanding of ecological status trans-National benchmarking. It must be preformed for High/Good and Good/Moderate status borders.

Inter-calibration Common Metric (ICM). A biological metric widely applicable within a GIG, which can be used to derive comparable information among different countries/stream types.

Inter-calibration 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 Inter-calibration exercise in order to ensure that the class boundaries that are established are consistent with the normative definitions in Section 1.2 and are comparable between Member States (WFD 1.4.1 (iv)).

Inter-calibration 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 eco-region in the Community; these sites will form an Inter-calibration network. The network shall consist of sites selected from a range of surface water body types present within each eco-region. 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 Inter-calibration process.

MMI. Multimetric Index. The multimetric approach uses a number of single (biological) metrics to assess environmental degradation (Karr et al., 1986). Different metrics, or metric groups, are assumed to provide distinct information on the different aspects of river structure and function.

Multimetric assessment systems have been applied in several circumstances, firstly in the USA (e.g. Karr et al., op. cit.; Barbour *at al.*, 1996) and recently in Europe (e.g. AQEM consortium, 2002).

MSs States members of the European Union.

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

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.).

REBECCA. 'Relationships between ecological and chemical status of surface waters' EU co-funded project within 6th Framework Programme Specific Targeted Research or Innovation Project – EU contract number: SSP1-CT-2003-502158.

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'.

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 that would be found in the absence of anthropogenic disturbance (from REFCOND guidance 14/06/2002).

RHS. River Habitat Survey (Raven *at al.*, 1997). Protocol of hydromorphological river assessment designed to characterize and assess, in broad terms, the physical structure of freshwater streams and rivers. The field survey has been designed, tested and improved as a result of extensive use on rivers in the UK since 1994. Recently, an adapted version for South European streams has been developed (Buffagni & Kemp, 2002)

Saprobic system. One of the first biological assessment systems developed in Europe. It is focused on species presence in relation to organic pollution (Kolkwitz & Marsson, 1902, 1908/9; Liebmann, 1962). It was extended and reviewed by several authors (see Knoben et al., 1995). A Saprobic Index is used in Germany (Friedrich & Herbst, 2004) and Austria (BMLF, 1999) to classify rivers based on aquatic invertebrates.

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.

Test data. Data derived by standard monitoring according to MS legislation and tradition. They refer to a stream type.

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.

12 - REFERENCES

AFNOR (Association Française de Normalisation), 1992. Détermination de la qualité biologique des eaux courantes: Indice Biologique Global Normalisé. Norme NF T 90-350.

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.

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. 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. 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.

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.

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, 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.

Belfiore, C., 1995. Description of *Electrogena calabra* n. sp., a new species from southern Italy (Ephemeroptera: Heptageniidae). Annls Limnol. 31: 29-34.

Belfiore, C & A. Buffagni, 1994. Revision of the Italian species of the *Ecdyonurus helveticus-group*: taxonomy of the nymphs (Ephemeroptera, Hepatgeniidae). Mitt. Schweitz. Ent. Ges. 67: 143-149.

Belfiore, C. & F. Desio, 1995. Taxonomy and distribution of *Electrogena ujhelyii* (Sowa, 1981) (Insecta: Ephemeroptera: Heptageniidae). Ann. Naturhist. Mus. Wien 97: 151-154.

Belfiore, C., G. Scillitani, O. Picariello & A Cataudo, 1997. Morphological and Electrophoretic Evidence for a New Species of *Electrogena* from Central Italy: Description of *E. lunaris* sp. n. (Ephemeroptera: Heptageniidae). Aqatic Insects 19: 129-140.

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

BMLF, 1999. BUNDESMINISTERIUM FÜR LAND UND FORSTWIRTSCHAFT (ed.) - Richtlinie zur Bestimmung der saprobiologischen Gewässergüte von Fließgewässern. Wien (BMLF).

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.

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., 1997. Mayfly community composition and the biological quality of streams. In: Landolt, P. & M. Sartori (eds), Ephemeroptera & Plecoptera: biology-Ecology-Systematics, MTL, Fribourg: 235-246.

Buffagni, A., 1998. *Heptagenia longicauda*, nuova per l'Italia, nel fiume Po (Ephemeroptera: Heptageniidae). Boll. Soc. ent. ital., Genova 130: 13-16.

Buffagni, A. & C. Belfiore, 1994. Recenti sviluppi delle ricerche tassonomiche e faunistiche sugli efemerotteri italiani (Ephemeroptera). Atti XVII Congr. naz. ital. Entomol., Udine 1994: 175-178.

Buffagni, A. & F. Desio, 1998. *Ephemera zettana* Kimmins, 1937 (Insecta: Ephemeroptera: Ephemeridae) in Italia Nord-Orientale. Gortania 20: 167-171.

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. & 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, 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, 2004a. 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, 2004b. Carattere lentico-lotico dei fiumi mediterranei e classificazione biologica di qualità. Quad. Ist. Ric. Acque 122: 165-189.

Buffagni, A., S. Erba & M.T. Furse, (2005). A procedure to harmonize class boundaries of assessment systems at the pan-European scale. (Submitted to Water Research).

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

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.

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

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

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

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

Environment Canada, 1993. Guidelines For Monitoring Benthos In Freshwater Environments.

Erba, S., A. Buffagni & J. G. Wasson, 2004. STAR ICM index step-by-step procedure for the Inter-calibration pilot exercise. Technical notes on the STAR ICMi procedure for the European Inter-calibration. STAR Project, CNR-IRSA, Brugherio (MI), Italy. September,17th 2004. 5pp.

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, 2003a. 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, 2003b. 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, 2003c. 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. 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.

Feld, C. K. & D. Hering, submitted. 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.

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

Friedrich, 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).

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.

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.

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.

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

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., 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.

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.

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.

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.

Knoben, R. A. E., C. Roos & M. C. M. van Oirschot, 1995. Biological assessment methods for watercourses. UN/ECE task force on monitoring & assessment, Lelystad, Volume 3: 1–86.

Knoben, 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.

Kolkwitz, R. & M. Marsson, 1902. Grundsätze für die biologiche Beurteilung des Wassers nach seiner Flora und Fauna. Mitt. Ausd. Kgl. Prüfungsanstalt für Wasserversorgung u. Abwässerbeseitigung 1: 33–72.

Kolkwitz, R. & M. Marsson, 1908. Ökologie der planzlichen Saprobien. Ber. dtschen. bot. Ges. 26: 505–519.

Kolkwitz, R. & M. Marsson, 1909. Ökologie der tierischen Saprobien. Int. Rev. Hydrobiol. 2: 126–519. London, 244 pp.

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

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.

Lenat, D. R. & V. H. Resh, 2001. Taxonomy and stream ecology– The benefits of genus and species-level identification. J. N. Am. Benthol. Soc. 20, 287-298.

Liebmann, H., 1962. Handbuch der Frischwasser und Abwasserbiologie. Band I, R. Oldenburg, Munich, 588 pp.

Lorenz, 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.

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

Meier, C., D. Hering, R. Biss, J. Böhmer, C. Rawer-Jost, A. Zenker, P. Haase & F. Schöll, 2004. Weiterentwicklung und Anpassung des nationalen Bewertungssystems für Makrozoobenthos an neue internationale Vorgaben. Berlin, Umweltbundesamt.

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.

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

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

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

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.

Nijboer, R. C., R. K. Johnson, P. F. M. Verdonschot, M. Sommerhäuser & A. Buffagni, 2004. Establishing reference conditions of European streams. 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: 91–105.

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.

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.

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.

Podraza, P., H. Schuhmacher & M. Sommerhäuser, 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.

Raven, P. J., P. J. A. Fox, M. Everard, N. T. H. Holmes & F. D. Dawson, 1997. River Habitat Survey: a new system for classifying rivers according to their habitat quality. In Boon, P. J. & D. L. Howell (eds), Freshwater Quality: Defining the Indefinable? The Stationery Office, Edinburgh: 215–234.

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.

Rekolainen, S., 2003. “Relationships between ecological and chemical status of surface waters.” REBECCA Project – EU contract number: SSP1-CT-2003-502158.

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.

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.

Rolauffs, P., D. Hering, M. Sommerhäuser, S. Rödiger & S. Jähnig, 2003. Entwicklung eines leitbildorientierten Saprobienindexes für die biologische Fließgewässerbewertung. Berlin, Umweltbundesamt.

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

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.

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

Schmutz, S., 2001. "Development, Evaluation and Implementation of a Standardised Fishbased Assessment Method for the Ecological Status of European Rivers. A Contribution to the Water Framework Directive." FAME Project - EU contract number: EVK1 -CT-2001-00094.

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.

Shannon, C. E. & W. Weaver, 1949. The Mathematical Theory of Communication. The University of Illinois Press, Urbana, IL.

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.

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.

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.

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

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.

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.

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

Ter Braak, C. J. F. & P. Smilauer, 2002. *CANOCO reference manual and CanoDraw for Windows user's guide version 4.5.* Biometris – Plant Research International, Wageningen and Èeské Bud jovice.

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

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

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.

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

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.

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.

Wasson, J. G., A. Chandesris, H. Pella, & L. Blanc, 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.

Woodiwiss, F. S., 1964. The biological system of stream classification used by the Trent River Board. Chemistry and Industry 14: 443-447.

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

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).

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

13 - ACKNOWLEDGMENTS

We are especially grateful to Jean-Gabriel Wasson (CEMAGREF, Lyon, France), who participated in all the major discussions and meetings that led to the preparation of the present Paper and who provided a considerable amount of invertebrate data, even though he was not part of the STAR Consortium.

We are also particularly grateful to Pietro Genoni (ARPA Lombardia, Italy), who provided the test data systematically used in the preparatory phases (i.e. for the ECOSTAT Discussion paper) and in the overall analysis itself, which consequently resulted in the description of the ICMi harmonization procedure.

We would also like to thank all Central and Mediterranean GIG delegates who provided the data presented in the Paper, as well as the STAR partners who contributed to the discussion on Inter-calibration.

We would also like to specifically thank the following contributors whose data and valuable comments have been included in the present report: Karel Brabec (Masaryk University - Brno, CZ), Maria Helena Alves (INAG, PT), João Manuel Bernardo (Uni Evora, PT), Marcel van den Berg (RIZA, Institute for Inland Water Management and Waste Water Treatment, NL), Wim Gabriels and Gaby Verhaegen (Flemish Environment Agency, B), Pietro Genoni (ARPA Lombardia, I), Malgorzata Golub (Institute of Environmental Protection, Warsaw, PL), Roel Knoben (Royal Haskoning, NL), Nicolas Mengin (CEMAGREF, Lyon, France), Isabel Pardo (UNI Vigo, Spain), Kristjan Piirimae (Tallinn Technical University, EE), Sandra Poikane (JRC, I), Jens Skriver (NERI, DK), Henn Timm (Estonian Agricultural University, EE), Manuel Toro Velasco (CEDEX, E) and all our other European colleagues who have participated in numerous discussions over the last two years on the IC process and its possible Options.

We are indebted to Jan Turvey, who promptly and efficiently corrected the main texts of the Paper and finally, we would like to acknowledge the enormous effort contributed by CNR-IRSA colleagues and students in the production of this report.