Development, Evaluation & Implementation of a Standardised Fish-based Assessment Method for the Ecological Status of European Rivers

A Contribution to the Water Framework Directive

WP 10

Comparison of the European Fish Index with the Standardised European Model, the Spatially Based Models (eco-regional and European), and Existing Methods (D 16 – 17)

FINAL REPORT

Paul Quataert, Jan Breine, Ilse Simoens

Institute for Forestry and Game Management

A project under the 5th Framework Programme Energy, Environment and Sustainable Management.

Key Action 1: Sustainable Management and Quality of Water

Contract n°: EVK1 -CT-2001-00094

Abstract

When developing a decision tool as a fish index it is crucial to document its performance in relation to its goals. An important requirement of the WFD is that the newly developed European Fish Index (EFI) can distinguish between a (nearly) pristine and disturbed status. Also the position of EFI with respect to existing national or regional fish indexes should be clear as some of them offer a long time series. Finally, comparison of the EFI with respect to other possible approaches gives insight in its relative merits and shortcomings.

To realize this evaluation, the central idea of this chapter is to think the fish index as a laboratory test to detect whether or not a site is disturbed. This analogy allows expressing the performance in terms of the detection capacity (sensitivity and specificity, impacted and non-impacted predictive value) and consistency. The cumulative distribution function of the new metric conditional on other indices turned out to be a powerful graphical tool to detect anomalies and to gain insight on a more profound level. These cumulative curves can be estimated easily from the data by the empirical distribution function.

Keywords

Cumulative distribution; Consistency; Absence of a Golden Standard; Pressure Status; Discriminating capacity; Sensitivity; Specificity; Impacted Predictive Value; Non-impacted Predictive Value; Ordinal Data; Fish Index

The tasks of WP 10

Originally the tasks were defined as follows:

(1) to compare the new FAME method with existing national indices,

(2) to document the strengths and weaknesses of the new method in a simple, understandable way, and,

(3) to suggest adaptations to specific regional or local conditions if necessary. Based on these results the scientific partners can select, in co-operation with Applied Partners, the most common method for each country.

The evolutions during the FAME-project made adaptations and extensions necessary:

(1) To cope with the different points of view, four different biotic indices were developed stemming from three different approaches. As a consequence, there was no single FAME method and it was necessary to broaden the task to position all the methods with respect to each other. This comparison offered a basis to make a definite choice.

(2) Also it became clear that the existing national indices cannot be considered as the golden standard. In fact no such absolute reference stick exists. Yet an important requirement of the Water Framework Directive (WFD) is that the fish index should be able to discriminate between undisturbed (reference) sites and impacted sites (EU Water Framework Directive, 2000). Hence the focus of the evaluation will be the assessment of the discriminating capacity of the index of the fish index with respect to this contrast.

Introduction

When developing a decision tool as the fish index (van Dijk et al., 1994, Strandberg, 1971) it is crucial to know the performance in relation to its goals. Transparency of quality can add much to an optimal use. No instrument can offer perfect information, but knowledge of the quality provides insight under which conditions and to which extent the instrument is reliable and guides on how and on when to apply the tool. Also the strengths and limitations, the advantages and disadvantages, the position and complementariness with respect to other instruments help to decide on whether to use it or to choose for an alternative.

A fish index (FI) reduces a complex reality into one single discrete measure (Karr 1981, Karr et al. 1986, Fausch et al. 1990, Lyons et al. 1996, Hughes & Oberdorff 1999, Hughes et al. 1998, Whittier & Hughes, 2001). It is not realistic to expect this number gives a complete picture of the ecological status and also stochastic variation is present (Fore et al. 1993). No absolute reference stick or golden standard exists to evaluate the ecological status of rivers. We only have plausible constructs on how to assess biotic integrity based on theoretical and empirical findings. Depending on the approach and the choices made the appraisals can be quite different. Yet, in spite of all these limitations, the fish index can be a useful tool and documentation of its quality can improve its use.

Within the FAME-project in total six indices evaluating the ecological status of the river were available. Table TTT1 gives an overview. These different approaches gave the opportunity to "cross-validate", i.e. positioning the outcomes of the different indices with respect to each other to get insight in the quality of the final choice of FAME, the European Fish Index.

Table TTT1. The different assessment methods of the ecological status

Method	Main characteristics	Origin / Author
IBI	Index of Biotic Integrity • Assesses the Ecological Status of a site based on biological characteristics (e.g. species composition) • According to the WFD the categories of the index range from 1 to 5 (high, good, moderate, poor and bad quality).	
FI	Fish Index. IBI for rivers based on the fish composition.	
EFI	European Fish Index • Differences in reference conditions are accounted by a regression model based on descriptors of a given site and its river group (≠ eco-region), predicting the pristine or reference distribution of the metrics • The sum (or average) of the p-values of the metrics with respect to their pristine distribution evaluate the biotic integrity on a continuous scale between 0 and 1. The EFI is a categorisation into the 5 (WFD) classes.	FAME 1: RCA = Reference Condition Approach (Pont, 2005)
SEM	Standardised European Model • The metrics are scaled by eco-region to make them comparable. A distinction is made between sites with a small or a large number species (in a pristine condition). • A discriminant model predicts the index value.	FAME 2: SEA = Standardised EU Approach (Böhmer, 2005)
SBM	Spatially Based Methods • Differences in expected reference conditions are taken into account by stratification (i.e. without modelling explicitly environmental variability). Hence, for each stratum there is a different model. • The selected metrics define a multivariate discriminant model determining (the probability of) the index value.	FAME 3: SBA = Spatially Based Approach
SBM-ER	Spatially Based Approach / Eco-regional • Within each eco-region a further stratification is made to arrive at more homogeneous strata. • Within each stratum a discriminant model predicts the fish index.	FAME 3a: SBA/E = eco-regional SBA (Schmutz, 2005)
SBM-EU	Spatially Based Approach / Eco-regional Clustered on a EU-level • To overcome the too small datasets of SBM-ER the strata are clustered based on the species composition. • Within each (clustered) stratum a discriminant model predicts the fish index.	FAME 3b: SBA/C = clustered SBA/E (Melcher, 2005)
ExM	Existing (local) methods (at a national or regional level) • Sometimes within a country a further stratification or modelling is used to standardise. • Different methods going from expert judgement to modelling are used to assess the class of biotic integrity.	FAME Ref 1: L = Local Methods Different Authors (see table TTT2)
PS	Pre-classification of the Pressure Status • Not a fish-based method! Based on an evaluation of the physical status (hydrology and morphology) and chemical characteristics (toxicity and nutrients). • The combination was categorised in 5 classes to comply with the WFD. The most important contrast is between 1-2 (high + good) and 3-5 (moderate + poor + bad)	FAME Ref 2: H = human impact Developed within FAME as reference for Model Building

The relation with the (pre-classification of the) Pressure Status (PS)

A first goal required by the Water Framework Directive (WFD) is that the EFI can distinguish between a (nearly) pristine and disturbed status (EU Water Framework Directive, 2000). Especially the power to discriminate between heavily impacted and reference sites should be large. To calibrate the models in function of this requirement, the FAME project classified the ecological status on physical (hydrology and morphology) and chemical (toxicity and nutrients) parameters. This classification independent of fish is named the pre-classification of the Pressure Status (PS) or before, the human impact (Goudie, 1993). It played a key role to develop the fish indices and is also used here to compare and evaluate the different methods with respect to each other.

The assessment of the pressure status was based on available data because lack of time and it was not easy to find a common denominator on a European level for the pressure or impact variables (Degerman, 2005). More fundamentally, the pre-classification lacks by definition the biological information the fish indices incorporate. In fact an index from the perspective of the fish community is developed to offer more than is possible by judgement of the environmental conditions. Because of this limited quality and the inherent shortcomings, the correspondence of the EFI with the PS certainly is not the only criterion. Yet it is an important orientation and at least the results of EFI and PS should be consistent with respect to each other.

The position with respect to the existing national or regional methods (ExM)

A second important point is the relation of the newly developed index with the existing national or regional methods (ExM). Possibly they are adapted better to the local situation and some of them offer a long time series. Knowing their relation with the new index is crucial to

benefit from the past and to be open the future. This opening for the future is important as the local methods are not golden standards even if they are developed in a context more close to the local situation. The risk of developing a method on a local scale is to loose the picture on the totality. Further, the rationale and quality of these methods varies very much, going from expert judgement to sophisticated statistical models. This hampers a comparison on a European level. For instance, an index value of 2 in a region as Flanders (Belgium) with many pressures probably has a very different meaning than in Lithuania. Table TTT2 gives an overview of the existing methods offered by the FAME-partners.

Table TTT2. The available existing national or local methods (ExM)

Country	N	Name	Method used to develop the model
Austria	483	Mulfa	Expert judgement
	166	National	Expert judgement
Belgium: Flanders	1043	IBI for upstream rivers & for Barbel and Bream zone	Index of Bio-integrity (Karr et al., 1986) adapted to the conditions of Trout and Grayling zone (upstream rivers) and of the Barbel and Bream zone in Flanders. (Breine et al. 2000).
Belgium: Wallonia	65	IBI for the Meuse & the wadable parts of a large river basin	Index of Bio-integrity (Karr et al., 1986) adapted to the conditions of the Meuse (France, Belgium and the Netherlands) + the wadable parts of a large river basin (Meuse and Scheldt) in Wallonia
France	1584		Reference Condition Approach comparable with EFI (Oberdorff & Hughes, 1992).
Lithuania	355		Index of Bio-integrity (Karr et al., 1986) adapted to the conditions of Lithuania.
Sweden	3544	FIX (Fish Index)	Index of biological status using fish Swedish Electrofishing RegiSter (SERS).
U.K.	203	Salmon Index	Based on the presence of Salmon
	203	Trout Index	Based on the presence of Trout

Correspondence with other approaches

In the end, to cope with different points of view, the FAME-project followed three different approaches resulting in four fish indices. This offered an excellent opportunity to test to which extent the different points of view lead to different appraisals of the sites. However, large differences are not very plausible. All methods are based on (about) the same dataset FIDES

(Beier, 2005) with pre-defined metrics and the models are calibrated with respect to the same variable (PS, the pre-classification of the pressure status). Also, apparently different statistical methods can have common mathematical roots relying on the same core properties of the data. Table TTT3 makes a conceptual comparison of the three FAME approaches on three dimensions. A generic formula representing the model building is:

$$FI = f(\text{metrics} \mid \text{environment}) \leftrightarrow PS$$

In words, the fish index is a (complex) function of the metrics taking into account the environmental variability calibrated with respect to the pre-classification.

Table TTT3. Conceptual comparison of the approaches within FAME

Approach	EFI	SEM	SBM-ER/EU
Adjustments for the Environmental Variability	Environmental descriptors + correction term for the river group	Standardisation by Eco-region Stratification by Species Richness	Stratification by and within Eco-region Clustering of strata to increase precision
Use of the pre-classification to develop the model	Indirect calibration	Direct calibration	Direct calibration
Type of the classification model	Sum of p-values with respect to the predicted reference distribution	Discriminant model	Discriminant model

A first dimension is on how to adjust for the environmental variability of the metrics. As such a metric does not give direct information about the quality. One should take into account the context to interpret its indicator value. What is high or low depends on the environmental conditions. The two Spatially Based Methods (SBM) stratify by a fish typology, the Standardised European Model (SEM) scale the metrics by eco-region and the European Fish Index (EFI) use a regression model to predict the metrics in the reference situation as a function of environmental variables (hence RCA, the reference condition approach). The difference between the two SBM models is that SBM-EU clustered the fish typology on a European level, whereas SBM-ER worked within each eco-region. As a consequence some

strata were too small to get a reliable estimate of the parameters and similar types but in different eco-regions got a different model.

A second dimension is how the pre-classification is used by the different methods. SBM and SES optimized directly the model parameters to predict the pressure status. For EFI the optimisation is indirect. First the metrics are modelled in pristine conditions and accepted as candidates if the reference distribution had an appropriate shape. Only in the next step the remaining metrics were selected as a function of their discriminating capacity. The advantage of this two step approach is that metrics are selected first on their inherent quality in reference conditions which is important to arrive at a sound biological model. The disadvantage is that some good discriminators can be eliminated. However, as the PS is not the golden standard, a certain distance with respect to this criterion can be an advantage.

The third dimension is the classification model. For SBM and SEM it is a discriminant model predicting the class of the index directly. EFI on the other hand is a categorisation of a sum or average of ten metrics scores. These metrics scores are p-values giving the discrepancy of the observed metric values with respect to the reference distributions predicted by the regression models mentioned above. As p-values are confined to the interval 0-1, this will also be the case for the average. The thresholds for the subdivision in categories are at 0.67, 0.45, 0.28 and 0.19 so that between 1 and 0.67 EFI = 1, between 0.67 and 0.42 EFI = 2, and so on till a value of 5 for an average score between 0.19 and 0.

More specifically the metrics of EFI can be grouped two by two into 5 dimensions: trophic structure, reproduction guilds, type of habitat, migration and connectivity and disturbance in

general. This is stressed by the second part of the formula below (EFI_n = the continuous score behind EFI, EFI_{ci} = the components of the index, EFI_d = the dimensions of the index).

$$EFI_n = \frac{1}{10}\sum_{i=1}^{10} EFI_{ci} = \frac{1}{5}\sum_{d=1}^{5} \frac{EFI_{d1} + EFI_{d2}}{2} = \frac{1}{5}\sum_{d=1}^{5} EFI_d$$

The general analysis strategy

In total six indices were available. As should be clear from above, none can be considered as an absolute reference stick (Shoukri, 2004). Hence the solution was in a first step to compare all indices with respect to each other with special consideration of the relation to the PS and ExM. In a second step the focus of evaluation was EFI, the final choice of FAME.

Data & Methods

The selection of the data

As there is no golden standard, in a first step it was tried to position all indices with respect to each other and to motivate the choice for the European Fish Index. The second part of the analysis consisted in a more in depth evaluation of EFI, the final index of FAME. For each part a different part of FIDES was used.

Selection for the positioning of the indices

Only sites with all six indices available at the same time are included; in total 3946 fishing occasions. Although the size of this sample is quite large, the spread over Europe is bad; about 92 % stems from 3 countries: 36 % from Sweden, 36 % from France and 20 % from Flanders (Belgium). An alternative solution could be to compare the indices two by two and to take all data with the two indices available. However this hampers the comparison over all indices simultaneously as the geographical configuration changes for each analysis.

Selection for the evaluation of the EFI

With EFI as the focus, it was judged acceptable to change the selection from analysis to analysis to have more data and a better geographical representation. For instance, when comparing with the Pressure Status the size is 9876 with a better geographical spread: Sweden 24 %, UK 21 %, France 16 %, Belgium 13 %, the Netherlands 10 % and Austria 7 % (to give the most important ones). The size for the comparison with the existing methods is 5436 with a similar distribution.

Weak points

As already stated the geographical spread is not very good. Another problem is that for some sites there are many fishing occasions (up to 36) risking to give some situations too much weight, but restriction to one occasion per site would decrease the sample size too much (from 3946 to 1781). Another drawback is that the evaluation also includes calibration data. Again, exclusion of these data would reduce the size (especially because the calibration sets of the different methods were not equal) and further deteriorate the balance in the data. Finally one may not forget this is only an internal validation. New data are necessary to really validate the index.

The quality measures

The basic idea of this paper is to think of a fish index (FI) as a laboratory test designed to detect whether or not a site is disturbed. This analogy allows expressing the performance of a fish index in term of the detection capacity (sensitivity, specificity, impacted and non-impacted predictive value) and the consistency to discriminate between impacted and non-impacted sites.

Sensitivity and specificity

For a laboratory test two important quality measures are the sensitivity π and the specificity ρ (Motulsky, 1995). The sensitivity quantifies which percentage of impacted sites will be classified as impacted. Its complement, $1 - \pi$, is the percentage missed and classified as non-impacted. In statistical testing this is the type II error β and $\pi = 1 - \beta$ the power. Conversely, the specificity quantifies which percentage of the non-disturbed sites will be classified as non-impacted. The opposite is the percentage that will be classified wrongly as impacted. In statistical testing this is the type I error $\alpha = 1 - \rho$.

Very important to recognize is that sensitivity and specificity are negatively interrelated for a given configuration. If one increases the sensitivity, the specificity goes down. This is because increasing the sensitivity implies a faster reaction of the test also when the site is undisturbed. As a consequence the specificity will decrease. So it is not easy to optimize both of them.

Nevertheless, a test of high quality should have both a high sensitivity and specificity. As much as possible of the disturbed sites should be detected (π high) and at the same time the number of undisturbed sites classified as disturbed should be low (α low). Or π / α should be high. If of the sites classified as disturbed too many are not, the information of a test will not be very useful. To quantify this two alternative quality measure: the impacted and non-impacted predictive value.

The impacted and non-impacted predictive value

Two other quality measures of laboratory tests are the positive predictive value (PPV) and negative predictive value (NPV). Or applied in this context: the impacted predictive value (IPV) and non-impacted value (NPV). The impacted predictive value (IPV) is the percentage of sites classified as impacted by the test, which are effectively disturbed. The non-impacted predictive value (NPV) is the percentage of sites classified as non-impacted by the test, that are indeed not disturbed.

IPV and NPV evaluate a test from a different perspective than the sensitivity and specificity. For the former, the starting point is the outcome of the test and the question is how confident we may be on the message: e.g. if the test says the site is impacted, which percentage will be

impacted indeed (IPV). For the latter, the perspective is the type of the site: e.g. if a site is impacted, which percentage will be detected (sensitivity).

Table TTT4 illustrates the two perspectives. The rows indicate the situation of the sites: disturbed or reference; the columns the classification of the test: disturbed or not. If a site is disturbed and the test classifies it as impacted, then it is truly impacted (TI); if not, it is falsely non-impacted (FN). Similarly, for the reference sites, there are falsely impacted (FI) and truly non-impacted (TN) sites. Table TTT4 shows how to calculate the different quality measures from this information.

Table TTT4. Relation between sensitivity, specificity and the predictive values

	Test = Disturbed (D)	Test = Reference (R)	Quality from Site perspective
Site = Disturbed (D)	TI	FN	$\pi = \frac{\#\mathrm{TI}}{\#\mathrm{TI} + \#\mathrm{FN}}$
Site = Reference(R)	FI	TN	$\rho = \frac{\#\mathrm{TN}}{\#\mathrm{TN} + \#\mathrm{FI}}$
Quality from Test perspective	$\mathrm{PPV} = \frac{\#\mathrm{TI}}{\#\mathrm{TI} + \#\mathrm{FI}}$	$\mathrm{NPV} = \frac{\#\mathrm{TN}}{\#\mathrm{TN} + \#\mathrm{FN}}$	

The relation between sensitivity & specificity and the predictive values

Because all measures rely on the same table, their values are interrelated. A simple derivation shows that for the link of PPV and NPP with sensitivity and specificity, a third variable is involved, namely the prevalence PD, the proportion (or percentage) of disturbed sites in the

population studied (N = size of the population, #D or #R the number of disturbed or reference sites in the population).

$$P_D = \frac{\#D}{\#D + \#R} = \frac{\#D}{N}$$

Then # TI = π N P_D and # FI ~ (1-ρ) N (1-P_D) = α N (1-P_D). Imputation in the formula of table TTT4 gives:

$$IPV = \frac{\frac{\pi}{\alpha}}{\frac{\pi}{\alpha} + \frac{1-P_D}{P_D}}$$

For the non-impacted predictive value a similar derivation results in:

$$NPV = \frac{\frac{\rho}{\beta}}{\frac{\rho}{\beta} + \frac{P_D}{1-P_D}}$$

Both formulas show that the balance between sensitivity (1 – type II-error) and specificity (1 – type I-error) determine the predictive values and that it is important to have both sensitivity and specificity high.

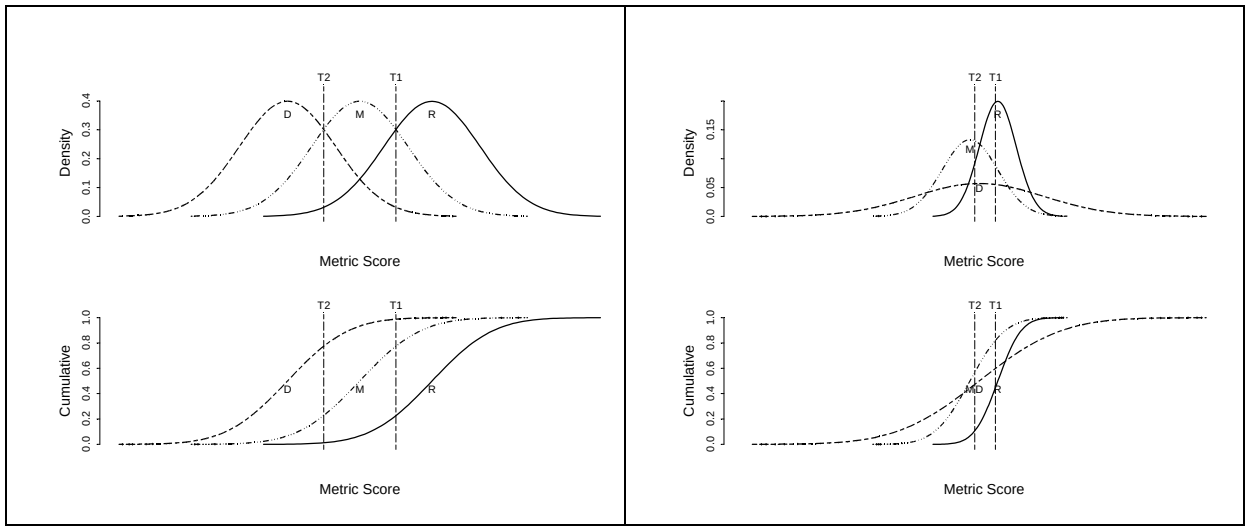
Consistency

EFI is a categorisation of a continuous variable which is the average (or the sum) of the p-values of the metrics with respect to their reference distributions predicted by regression models. As a consequence, the variable behind EFI is confined between 0 (totally degraded) and 1 (very high quality). The thresholds are: 0.67, 0.45, 0.28 and 0.19 so that between 1 and 0.67 EFI is 1, between 0.67 and 0.45 it is 2, and so on until 5 for a value between 0.19 and 0.

This is an interesting property as we can study the behaviour of the EFI in a continuous scale and see how the distribution varies as a function of other indices. Figure FFF2 shows the basic principle: the left part is an ideal situation; the panel on the right is a counter example. It

shows how the density and cumulative distribution of a continuous quality measure changes as the impact (as assessed by another index; e.g. a pre-classification based on abiotic variable) increases from reference R (no or small impact, pristine situation; categories 1-2), over moderate M (category 3) to disturbed D (high or very high impact; categories 4-5).

FFF2 Graphical evaluation of the consistency and discriminatory power of an index: the ideal situation (left) and a counter example (right)



In the ideal example (figure FFF2a) the distribution shifts from right to left as the impact increases. The ordering of the curves is consistent with the ranking by the other index. This property is necessary to have consistent indices. In the counter example (figure FFF2b) the ordering of the density curves is not uniform any more. Further the variance of the distributions differs, such that, although the curve for highly disturbed sites is located left from the curve of reference sites, still the probability of having a very high score is higher for disturbed sites than for undisturbed sited. This heteroskedasticity complicates the distinction between reference and disturbed sites very much.

Working with the cumulative distribution is most practical. In the ideal situation (left panel of FFF2) the curves do not cross and are ordered in a uniform way. In absence of consistency (the right panel) the curves cross. This cumulative representation is also easier to estimate it in practice (Chambers et al., 1983). Density curves need a lot of tuning to get the good shape, while a cumulative curve comes down to order the observations from small to large and assign to each of them the fraction of observations smaller or equal (the empirical distribution function).

A further advantage is that the cumulative representation allows assessing the discriminative capacity as it is the integral of the density function. Hence the cumulative graphs give the surface under the density function. The vertical lines in the figure represent two possible thresholds. The right one to distinguish between reference or not (i.e. the main contrast between 1-2 and 3-5); the left one to separate poor and bad status form the moderate to high status (the contrast between 1-3 and 4-5).

At the first contrast 1-2/3-5 (left panel of figure FFF2) about 20 % of the reference sites are below the threshold and will be misclassified as disturbed. Or the specificity is 80 %. On the other hand, nearly all (strongly) disturbed sites have a score below the threshold. The sensitivity is close to 100 %. Also, for the moderately affected sites the sensitivity remains high: about 80 %. For the second threshold nearly no reference sites are misclassified (specificity close to 100 %) and only 25 % of the moderately impacted ones (specificity about 75 %). Still the sensitivity is 80% for D.

As already stated for the counter-example (right panel of figure FFF2) the ranking of the curves is not as expected and the standard deviation of the curves is not equal. As a

consequence, the cumulative distributions cross. It is very hard if not impossible to get a good threshold. With T1 the specificity is low, and the sensitivity for moderately impacted sites is higher than for severely affected sites. T2 has a better specificity at the expense of sensitivity. This technique will be applied to compare the EFI with the other indices.

The statistical properties of an index

According to the WFD each index here consists of five classes range from very high quality (class 1) to very low quality (class 5): 1 (high) > 2 (good) > 3 (moderate) > 4 (poor) > 5 (bad). Important to recognize is that an index is an ordinal variable which statistical properties lie somewhere between a continuous and a class variable (Sokal 1995; Agresti, 2002; McCullagh & Nelder, 1983). In contrast to a "simple" categorical variable the ordering of the classes has a meaning: 1 > 2 > 3 > 4 > 5. However, the numerical scores cannot be used for arithmetic calculations (and hence, the mean and standard error can in principle not be calculated). The fundamental reason is that the distance between the scores is not defined; e.g. the difference between high and good status is not necessarily the same as between good and moderate and indeed it will be demonstrated that a better numerical scale is $1 \approx 2 > 3 > 4 \approx 5$ as the differentiation between 1 & 2 and 4 & 5 is not very consistent.

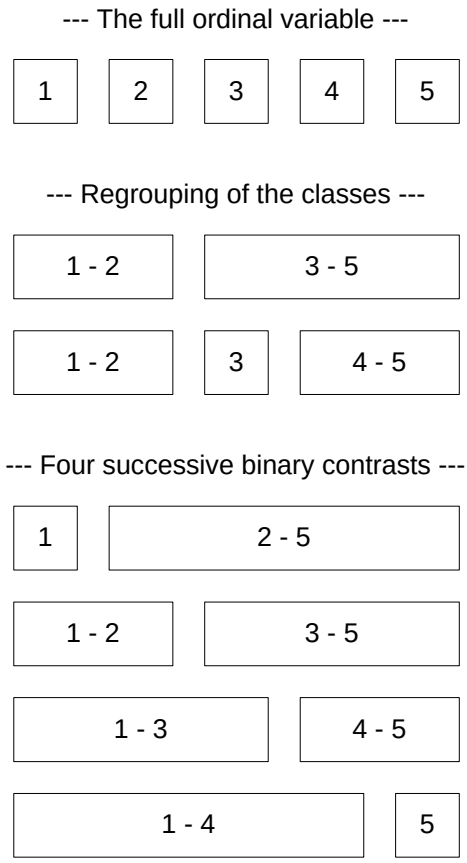
A way to exploit the ordinal character of the index is to group neighbouring classes in different ways. In this way the analysis is simplified without loosing information. Figure FFF1 shows the two possible approaches.

The first is inspired by the WFD which requires that a good distinction is made between (close to) reference and disturbed sites. This comes down to discriminate between classes 1-2 and 3-5 which is a binary variable easy to analyse. In the next step one can study the position

of moderately impacted sites by splitting 3-5 into 3 and 4-5, again a binary variable. Finally the composition of 1-2 and 4-5 is studied. This successive analysis at three levels of detail (cx; with x = 2, 3, 5 indicating the number of classes) simplifies (van Belle, 2002) the analysis and still (nearly) all information is used.

An alternative approach is to create four (one less than the levels of the ordinal variable) binary variables by moving the threshold. Each new variable makes the contrast between undisturbed and disturbed at a different point. This change of threshold will be used in this paper many times to control if the threshold changes the conclusions or not.

FFF1 Meaningful contrasts of the (fish) index



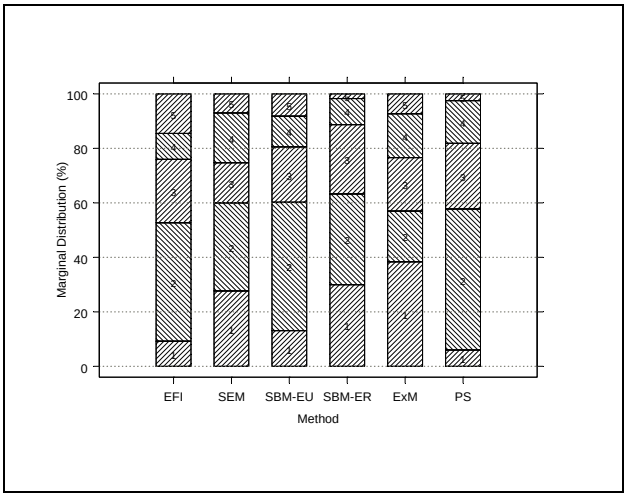
Results

A. POSITIONING of the INDICES WITH RESPECT TO EACH OTHER

The marginal distributions (figure FFF3)

At the main contrast 1-2/3-5 the marginal distribution is about equal for all indices. About 60 % of the sites is classified as reference (i.e. 1-2). Also at 1-3/4-5 the classification is about equal (80 % in 1-2), with the exception of the SBM-ER (90 %). However, the subdivision of 1-2 as well as class 4-5 is not consistent. The percentage in class 1 ranges from about 10 % (EFI) till 45 % (ExM). Similarly in class 5, the percentage ranges from 2 % (PS) to more than 10 % (EFI). In general PS and SBM-ER give the mildest quotation, while EFI is most severe.

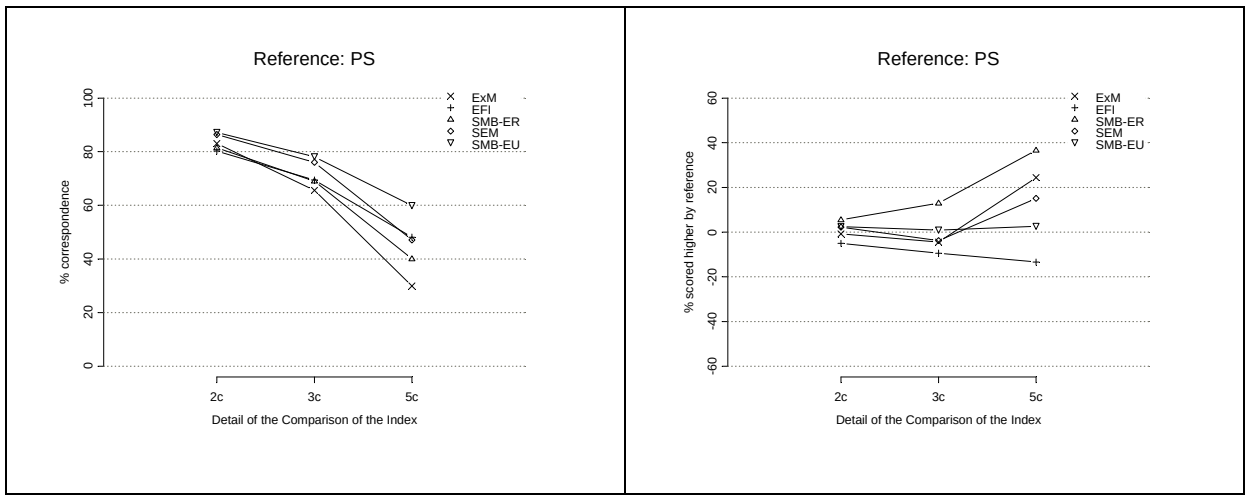
FFF3 Comparison of the marginal distributions



Correspondence as a function of the level of detail (figure FFF4)

Figure FFF4 shows as a function of the level of detail (2c = 1-2/3-5; 3c = 1-2/3/4-5; 5c = full index) the correspondence between PS and the fish indices in terms of matching and symmetry. Matching is defined as the percentage of cases with an equal classification. The symmetry is judged by the difference of the percentage of cases where the reference index is higher and where it is lower. A positive value indicates that on average the reference index is more severe, a negative value that it is less severe.

FFF4 Correspondence of the fish indices with respect to PS as a function of detail



As expected, the correspondence decreases with the level of detail. For the main contrast (2c) the matching ranges from 80% to 90% and the asymmetry remains smaller than 10 %. For three classes the match decreases to 60 – 80 % and for the full index to 35 – 60 %. The asymmetry becomes only pronounced for the full index.

SMB-EU has the highest match at each level of detail. At the main contrast (c2) it is as high as 90 % and it remains 80 % at c3, to drop to 60 % at full detail (c5). Also the asymmetry

remains low (a difference less than 5 %). Specifically, these qualities are superior to SBM-ER. Its match is about 10 % to 20 % lower. This proves that the strategy of clustering the strata at European level has been successful. The behaviour of SEM is similar to SBM-EU and is only worse at full detail. The matching of EFI is somewhat less. As indicated by the negative asymmetry, the reason is EFI scores systematically higher than PS or is more severe.

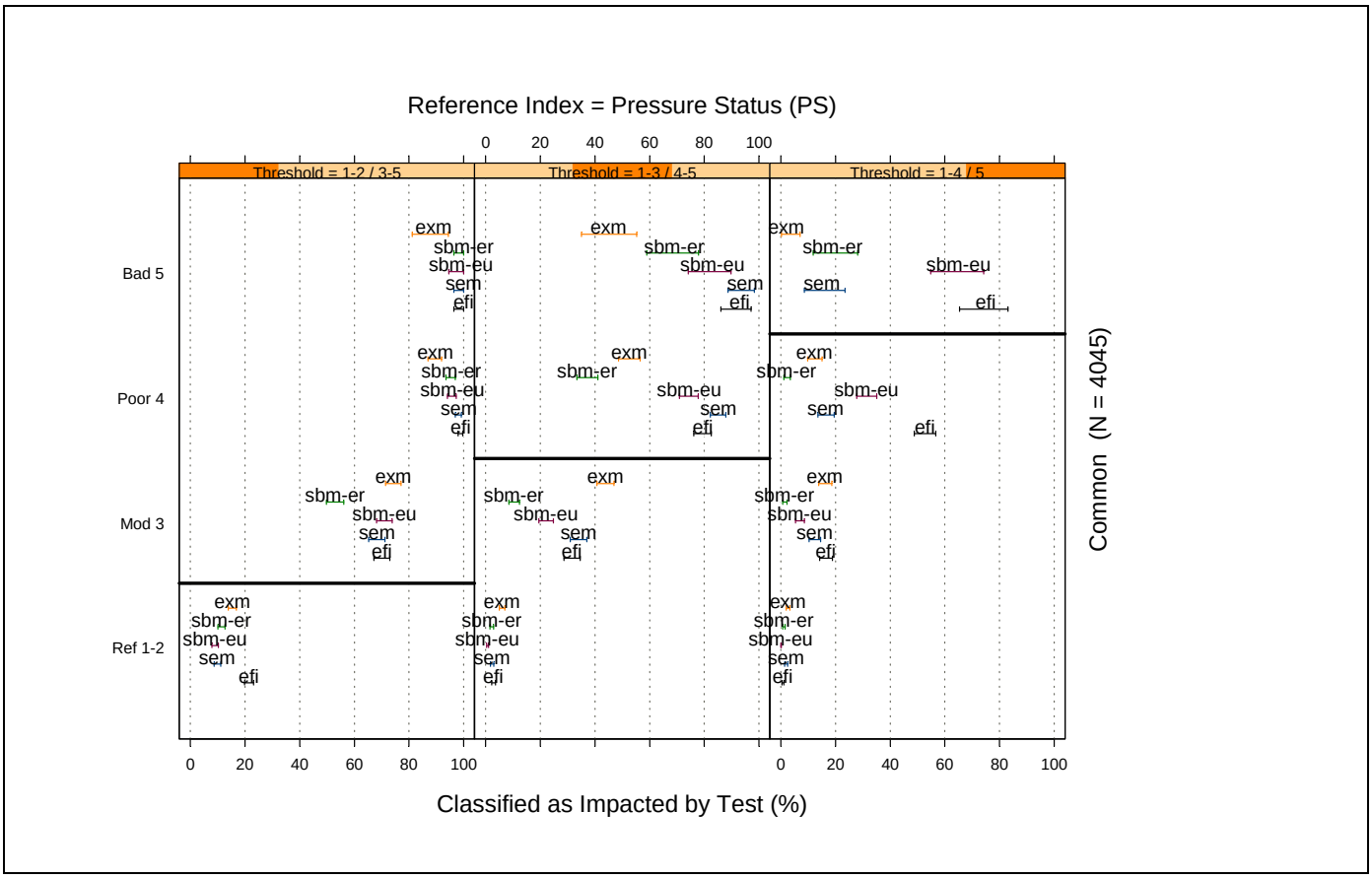
The correspondence ExM with PS is quite good. It is only at full detail both the matching and the symmetry go down very much. This is an important result. The existing methods did not optimize with respect to the Pressure Status, but were derived independently from it.

Sensitivity and Specificity (figure FFF5)

The panels of figure FFF5 show the detection rate, i.e. the percentage of sites classified as impacted by the different fish indices as a function of the impaction threshold (the three panels) and the pre-classification by PS (the vertical axis). As expected the general pattern is that (a) in each panel the detection rate increases as the level of PS increases from reference to bad (i.e. going from bottom to top) and that (b) increasing the threshold (i.e. going from left to right), the detection rate decreases.

The interpretation of the detection rate depends on the position in the graphs. Below the horizontal line it is an estimate of 1-specificity (preferably a low value) and above it is sensitivity (preferably a high value). For instance, the bottom left part of the figure is the percentage of reference sites misclassified as impacted by the other indices. Hence, it is an estimate of 1-specificity. Just above, it is shown which percentage of the sites with moderate PS is correctly classified as impacted, hence it is the sensitivity.

FFF5 Sensitivity and Specificity of the fish indices with respect to PS



At the first threshold 1-2/3-5 (panel 1), there is a monotone increase of the detection rate for increasing PS. For PS = 1-2 the detection rate ranges form 10% to 20% (hence a specificity of 80% - 90%), for PS = 3 it ranges from 50% to 75% and for PS = 4 & 5 it is close to 100%. There two specific situations to mention. In the reference situation (PS = 1-2) EFI has a higher detection rate (20 %) than the others (10% - 15%). It is the most severe index. For moderate PS (3), SBM-ER has a clearly smaller value than the others.

The next panel for the threshold (1-3/4-5) has a more diffuse character. Still there is a clear monotone increase of the detection rate, but there are considerable differences between the fish indices. While for reference sites (PS = 1-2) the detection rate is uniformly low (resulting

in a high specificity), for the sites with moderate PS the misclassification ranges from about 10 % (SBM-ER) to 45 % (ExM). For poor and bad sites according to PS, the sensitivity of both SBM-ER and ExM is rather low with respect to the other three indices.

For the third contrast 1-4/5 (panel 3) the detection rate is small even for sites classified as bad by PS. EFI attains a sensitivity of 75 % for PS = 5. The price (supposing PS is the standard) is that 55 % for poor PS are classified as bad. SBM-EU does nearly as good for PS = 5 (70 %), but keeps the misclassification lower for poor sites (35 %). Important to notice is that the sensitivity of SEM is poor although its performance was similar to SBM-EU at the other thresholds.

In conclusion

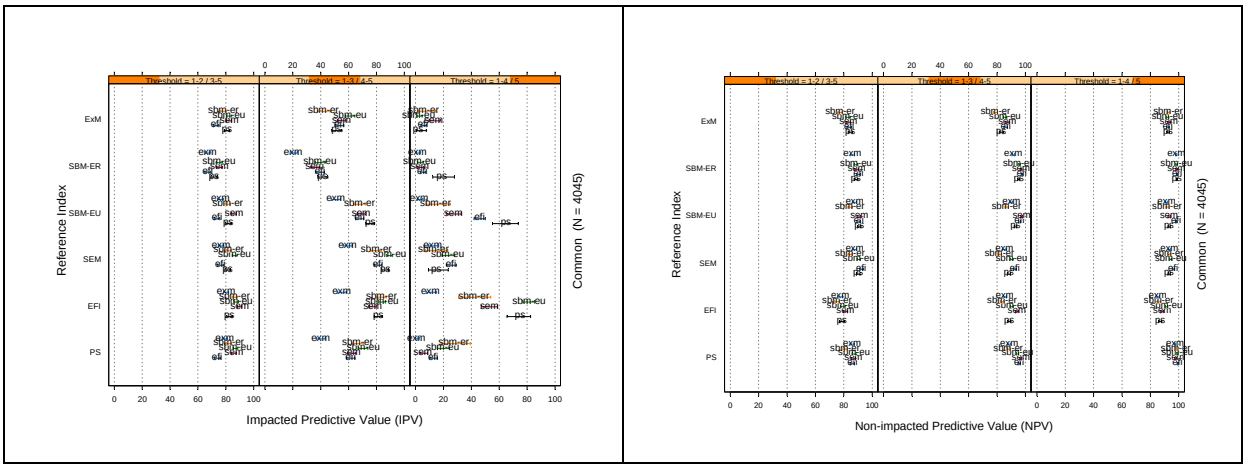
Extreme differences are seldom observed. EFI, SBM-EU and SEM show a very similar and consistent behaviour. In most cases EFI is the more severe method, while SEM has a small sensitivity at 1-4/5. With respect to the main contrast (1-2/3-5), the behaviour of ExM is very close to these FAME methods. At higher contrast this similarity disappears. Yet this result is interesting as for ExM no optimisation is done with respect to PS. Finally, quite consistently SMB-EU performs better than SMB-ER, from whichever perspective the assessment is made.

The Impacted (IPV) and Non-impacted (NPV) Predictive Value (figure FFF6)

In figure FFF6 shows the impacted and non-impacted predictive values of all indices with respect to each other. The label in the vertical axis specifies the reference index and the labels in the graphs the test index. The lower part (PS) is a summary of FFF5 making a balance between sensitivity and specificity.

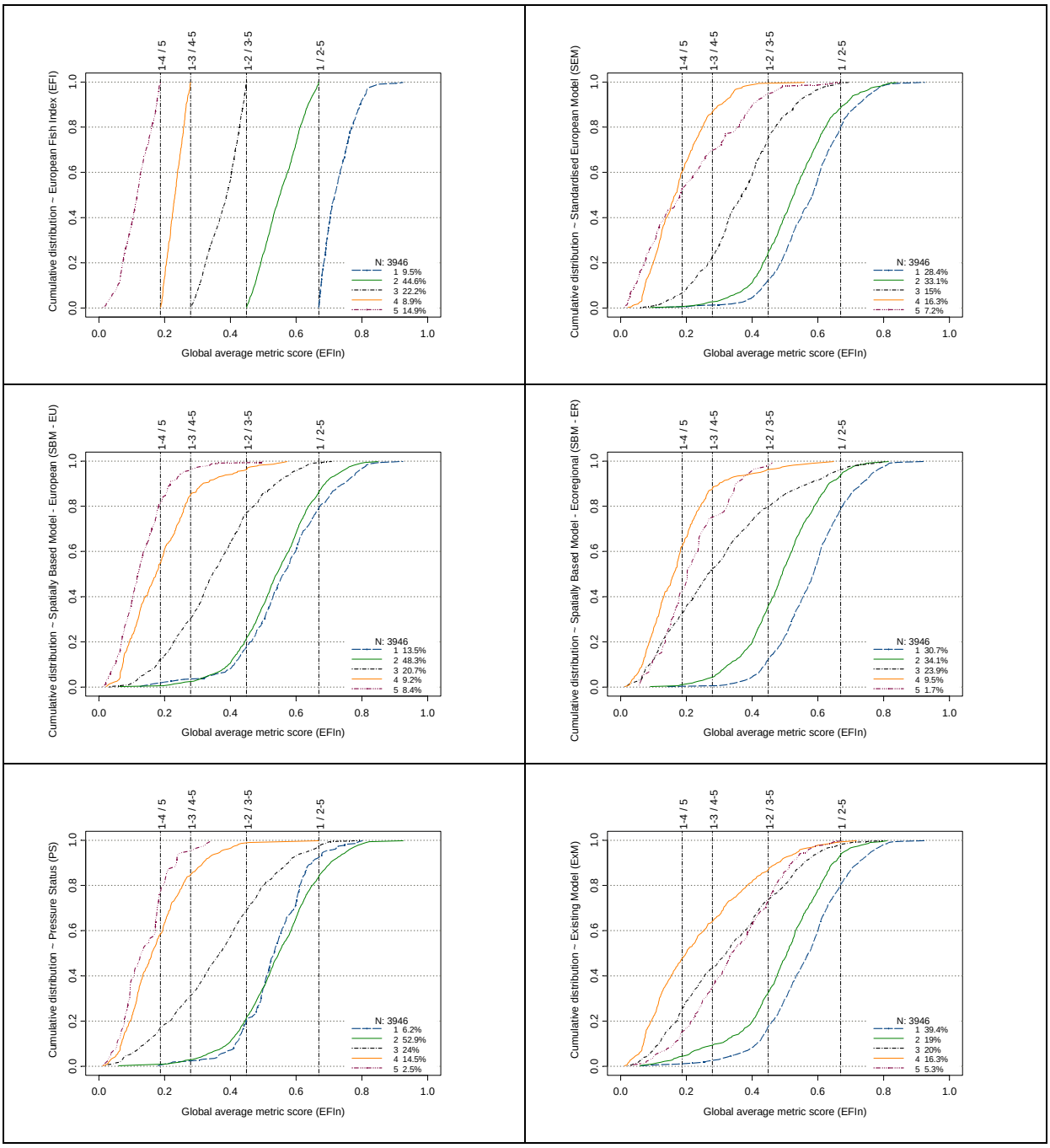
NPV does not change very much. For any combination the NPV is about 80 % (i.e. if a site is declared non-impacted, in 20 % of the cases it is impacted). This uniform picture is not true for the IPV. At the higher thresholds in general it becomes rather small indicating a large percentage of the impacted cases are misclassified.

FFF6 Impacted (IPV) and Non-impacted Predictive Value (NPV) of all indices with respect to each other.



For PS, at 1-2/3-5, IPV is about 80 %. EFI has a lower value because its specificity is lower: a relatively high number of reference sites is classified as disturbed. At the next threshold 1-3/4-5, the behaviour of ExM is deviant. At this level it is a poor predictor of the PS. Also for the other indices, IPV decreases (to 60 %). In the next step IPV becomes too small to be of any practical value.

FFF7 Consistency of the indices with respect to average metric score behind EFI



B. EVALUATION OF THE EUROPEAN FISH INDEX

Because behind the European Fish Index there is a continuous metric score it was possible to explore its behaviour based on the cumulative distribution. Also, as EFI has become the final proposal of FAME, it is necessary to make a more in depth evaluation.

Consistency of the EFI with the other indices (figure FFF7)

The first panel compares EFI with itself. It shows how the average of the metric scores of the EFI (the normalized EFI, EFI_n), is subdivided into five categories with thresholds at 0.67, 0.45, 0.28 and 0.19 so that between 1 and 0.67 EFI = 1, between 0.67 and 0.42 EFI = 2, and so on till a value of 5 for an average score between 0.19 and 0. So by definition there is no overlap or a perfect separation. Still, it is interesting to see that the distribution of EFI_n is uniform for the classes 2 to 4: the cumulative distributions are nearly straight lines.

The next five panels show cumulative distributions shifting to the left as the value of the index increases from class 2 to 4. However, for class 5 (bad status) the behaviour is deviant for SEM, SBM-ER and ExM. For PS and SBM-EU the ranking is good, but the distance is small between class 4 and 5. Also the distinction between class 1 and 2 is rather small. For PS the lines are even intertwined. This is not a surprise as for EFI no distinction was made between class 1 and 2 to build the model.

Both elements suggest that a more reliable classification is to pool the extreme classes: 1-2, 3, 4-5. For this configuration class 3 has an intermediate position. For instance, focusing on the relation with PS (panel 5 of figure FFF7), shows that at the threshold 1-2/3-5 the sensitivity to detect a poor or bad status is close to 100 %, for the moderate status it is still 70 % and this for

a specificity of 80 % (20 % of the good sites are classified as disturbed). At the threshold 1-3/4-5 the specificity is about 100 % for class 1 and 2 (or nearly no undisturbed sites are classified as poor or bad) and 70 % for class 3 (30 % of these moderately impacted sites are given a higher, less good score). Still the sensitivity for class 4 and 5 is more than 80%.

Consistency between the Existing Methods and EFI (figure FFF8)

The previous section considered the existing national or regional methods (ExM) as a whole. Here the consistency is checked for each local method separately.

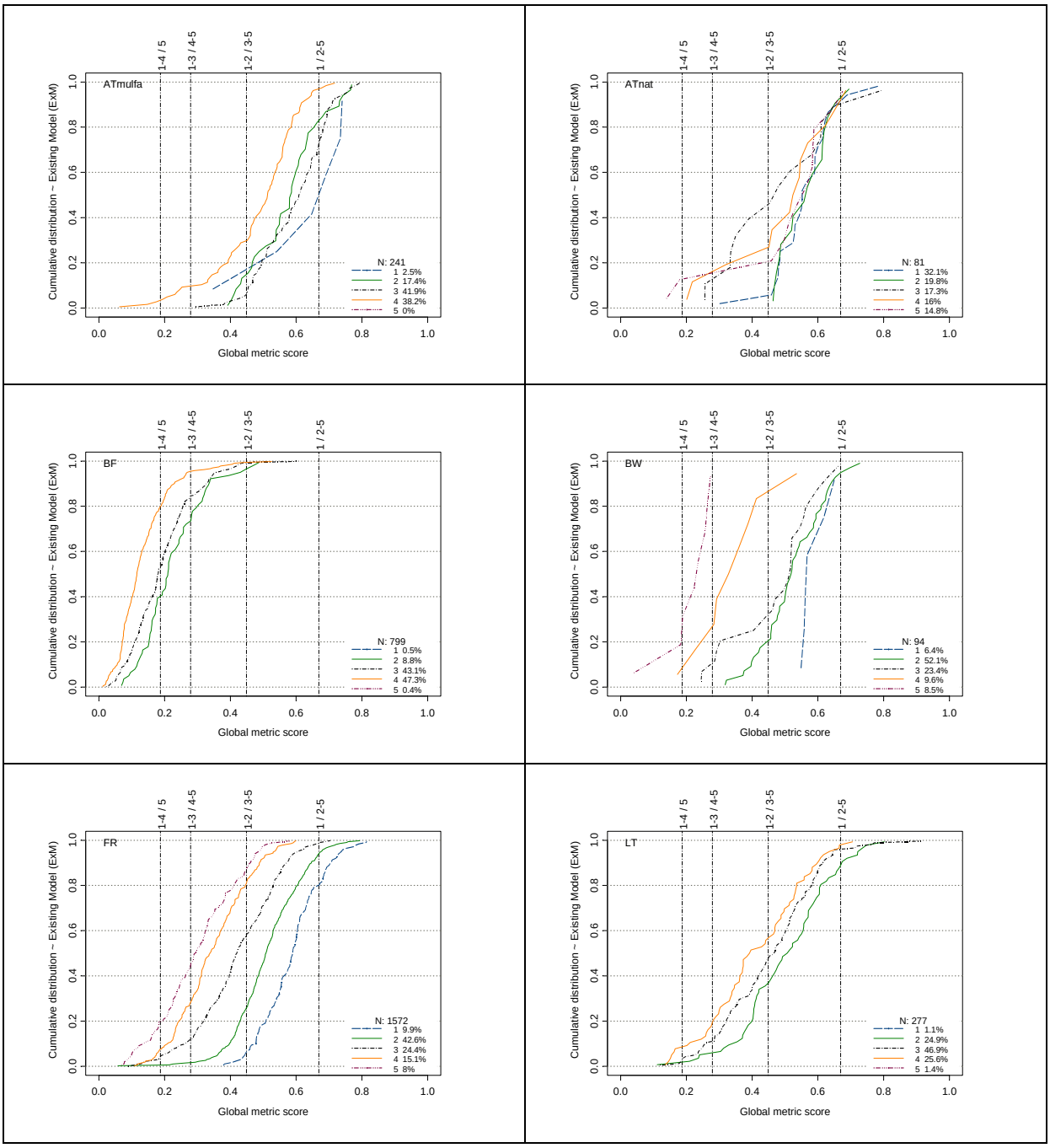
Overview

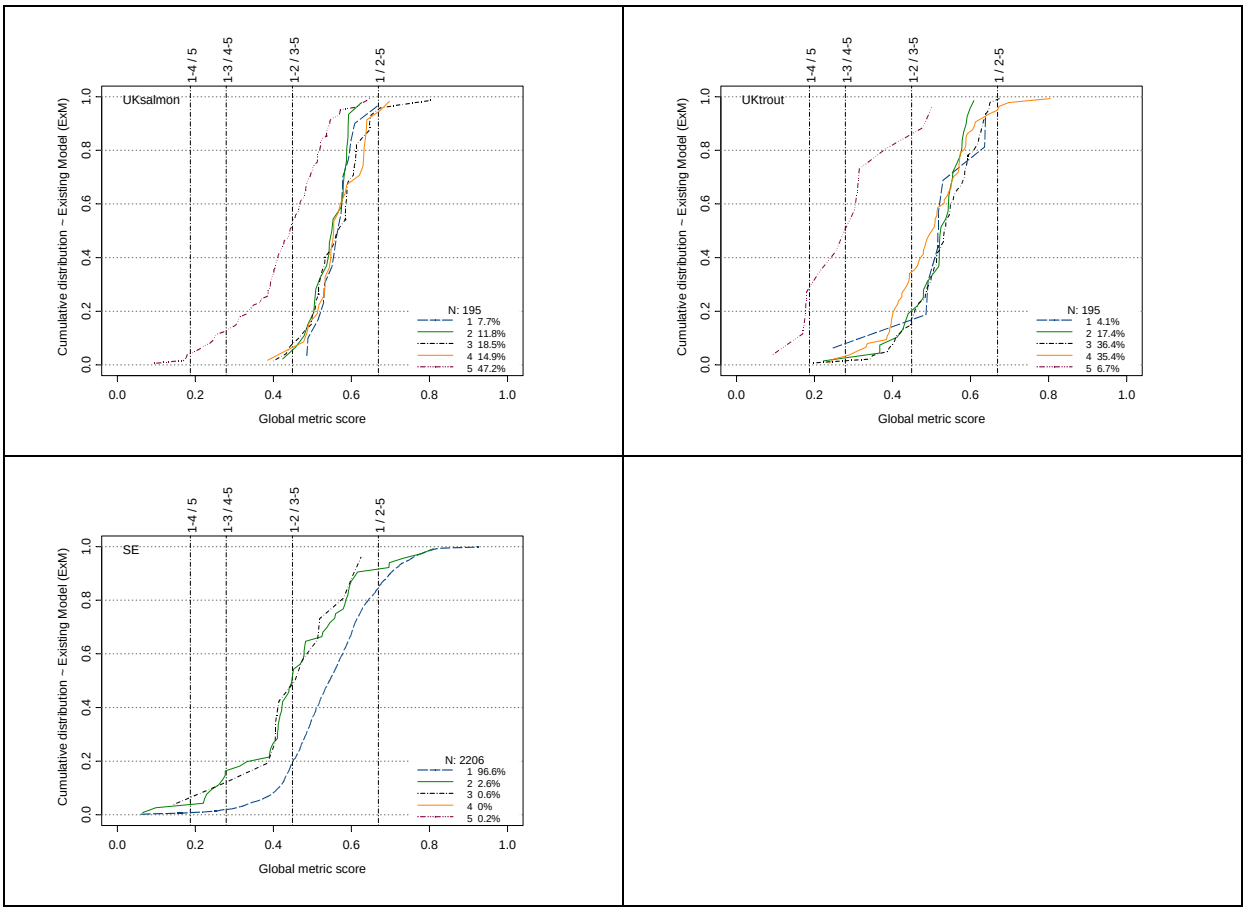
In Austria, there are two methods. For the national method (results not shown) the number of available data is too small (N = 81). For the Mulfa method (N = 241) the curves are ordered in the appropriate way. However the cumulative curves cross at the main threshold 1-2/3-5 and anyway the sensitivity is low.

For Flanders in Belgium (N = 799) nearly all sites are of class 3 and 4 and only a few of class 2. There are nearly no sites in the sample of class 5 because the fishing occasions with no fish are excluded. EFI makes no real distinction between class 2 and 3 and nearly all of them are classified as 4 or 5. The differentiation made at the local level disappears from a European perspective and also the judgement is more severe. What seems at a local level good or an improvement, is not at a European scale.

In the Walloon part of Belgium (results not shown) the number of sites is too small (N = 94) for an analysis but the graphs show a good ranking similar to France.

FFF8 Consistency of the existing methods (ExM) with respect to the score behind EFI





For France (N = 1572) the cumulative distributions of EFI are totally consistent with the local method. This was to be expected, as the local method for France is based on the same principle as the EFI. The discriminatory capacity at the contrast 1-2/3-5 is concordant with the scoring of the local method. However, at the next threshold 1-3/4-5 a lot the discriminating capacity is lost.

In Lithuania (N = 277) the ordering of the curves is good, however the discriminatory power is small. There will be many false classifications. In fact EFI does not make the difference between class 2 to 4 seen from the local level.

In Sweden (N = 2236) most of the observations are in class 1 (96.6 %) and 2 (2.6 %). This is the opposite situation of in Flanders. About 20 % of the sites in class 1 in Sweden are 2 for EFI and about 50 % of class 2, receive a higher score. So the EFI is more severe.

For UK (N = 195) there are two methods used on the same fishing occasions. For both we see the same pattern (only one is shown). There is no separation for 1-4 and a very important discrimination with respect to class 5. Although there is a differentiation in quality from 1-4 by the local method, EFI classifies all of them as 2. Of class 5 the appreciation by EFI varies from 2 till 4.

Summary

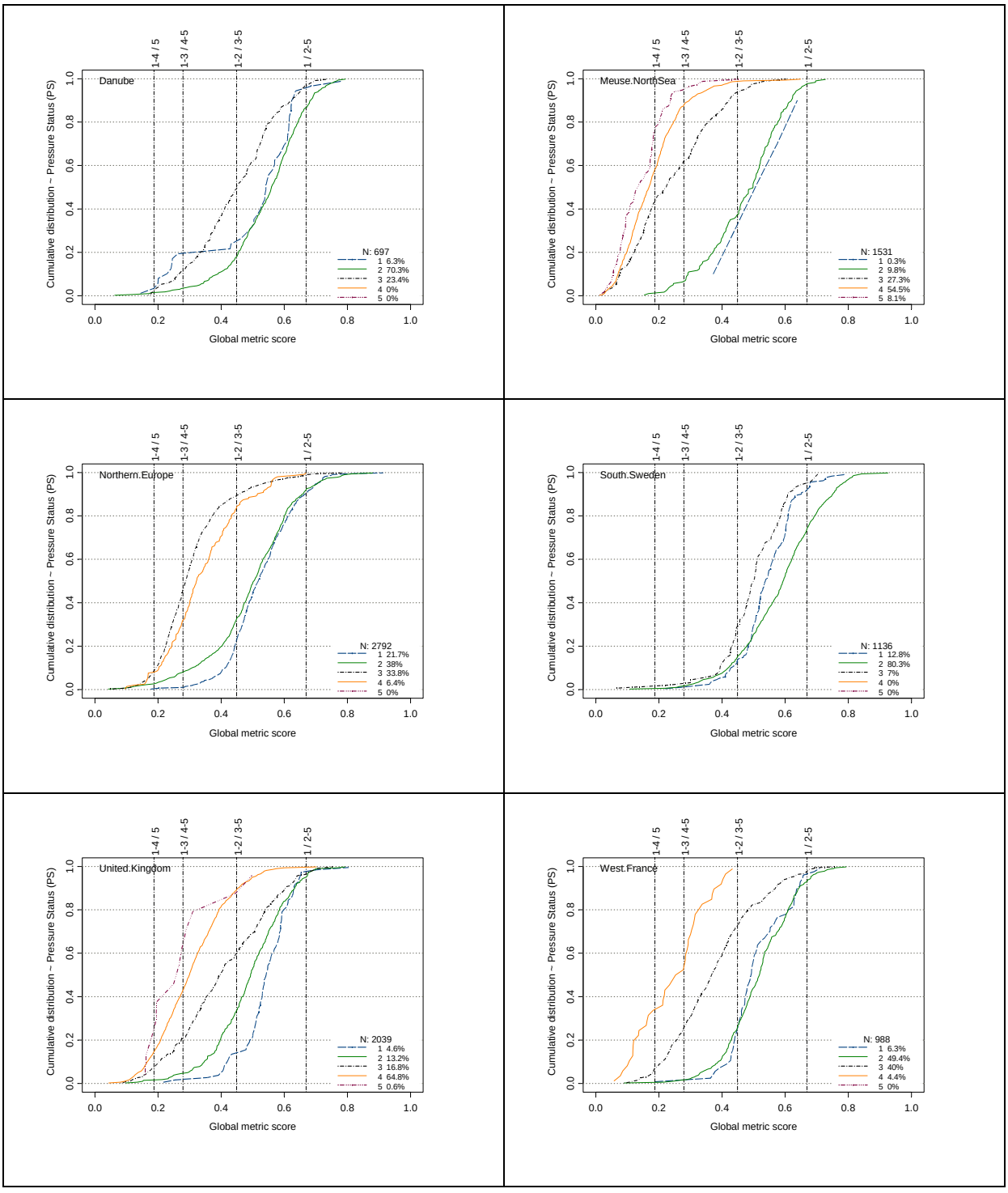
In most cases the number of fishing occasions is too small to arrive at definitive conclusions. Further, the distribution over the different impact classes is sometimes very unbalanced. For instance, for Flanders (Belgium), more than 90 % of the sites have an impact of 3 to 5; while in Sweden, 95 % of the fishing occasions are present in class 1. In fact, only for France there are a sufficient number of sites, quite well distributed over the different categories of impact. So the results should be interpreted very carefully.

The general impression is that the ordering of the curves is not in contradiction with the EFI; but the separation between the curves and hence the discriminatory power is low. In fact only for France we see excellent properties. This is as expected as the existing method for France stems from a similar approach.

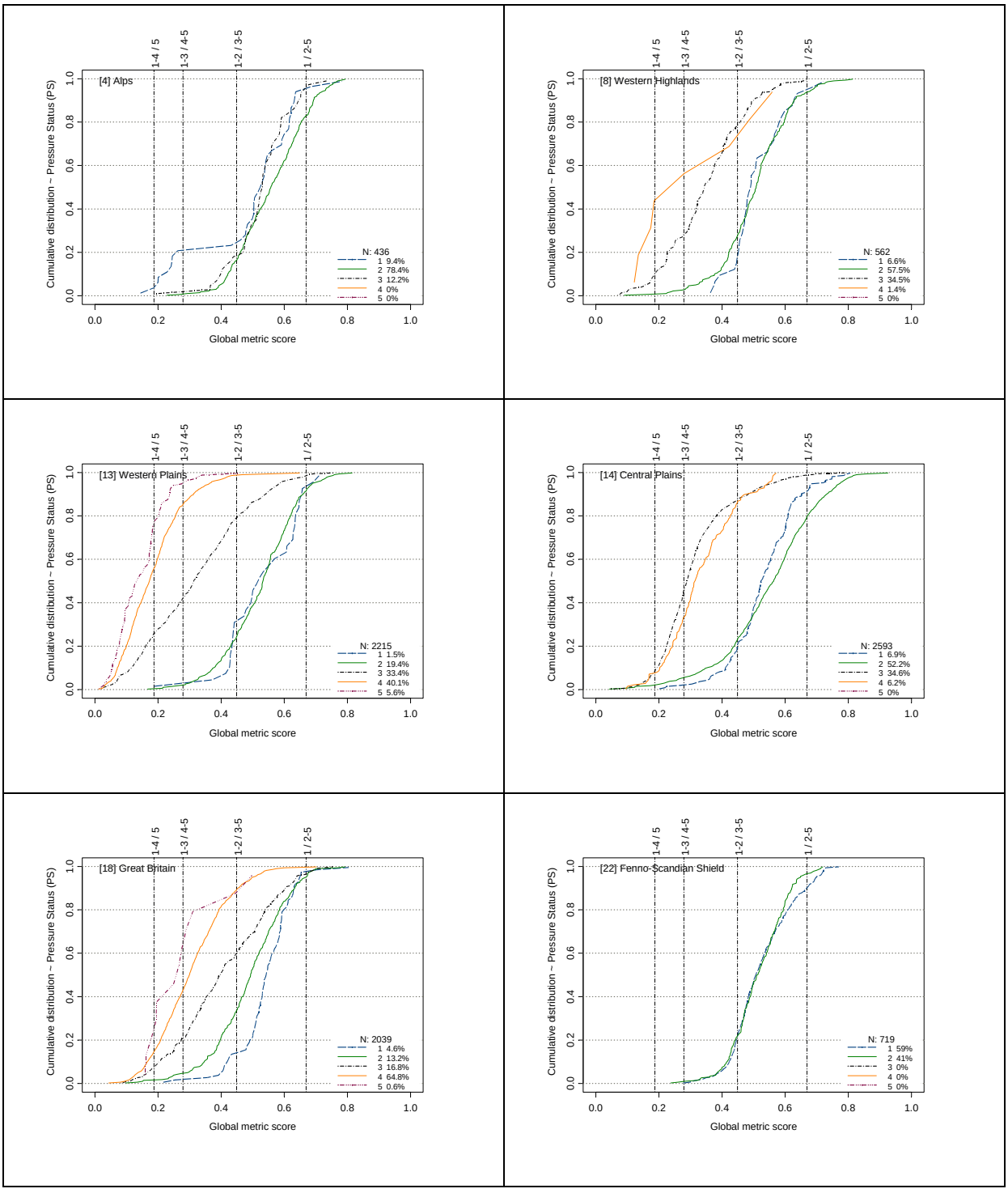
Consistency between EFI and PS stratified by Region

As seen before, the overall relation between EFI and PS is good. The question is if this still holds after splitting the dataset by region. This will be tested for two variables. First the exercise is made for River Group, a variable present in the prediction model of EFI. Then the same is done for Eco-region which is an external variable.

FFF9 Consistency of the pressure status (PS) with the score behind EFI by River Group



FFF10 Consistency of the pressure status (PS) with the score behind EFI by Eco-region



Consistency between EFI and PS evaluated by River Group (figure FFF9)

For Danube (N = 697) none of the sites are pre-classified (PS) higher than 3. The ordering of the EFI-curves is consistent with the ranking of PS. The distinction between PS 1 and 2 is poor, and their distinction from class 3 is moderate.

For the rivers of Meuse / North Sea (N = 1531) the distribution over the different pressure classes (PS) is good, except that nearly no sites belong to class 1 (0.3 %). The ordering of the cumulative curves is consistent with the ranking and also the separation is good between 1-2 and 3-5. However at the main threshold (1-2/3-5) 40 % of the mildly affected sites (PS = 2) will be classified as disturbed.

In Northern Europe (N = 2792) only class 5 of PS is not present. The discriminatory power at the main threshold 1-2/3-5 is good. The ordering of the curves is not consistent with PS for class 3 and 4 and no distinction is possible between 1 and 2.

South Sweden (N = 1136) contains no sites with PS larger than 3. The ordering of the curves is not consistent with the ranking. PS = 1 lies between class 1 and 3. The discrimination is poor at the main contrast 1-2/3-5.

For United Kingdom (N = 2039) the PS classes are well represented except for class 5 (0.6%). The ordering of the cumulative EFI-curves is as expected and the discriminatory power is good at the two main thresholds (1-2/3-5 and 1-3/4-5).

For West France (N = 988) except for class 5 all other classes of PS are sufficiently large. The ordering of the cumulative EFI-curves is consistent with the a priori ranking based on PS. Class 1 and 2 are not distinguished well. At the main contrast 1-2/3-5 the discrimination is good.

Consistency between EFI and PS evaluated by Eco-region (figure FFF10)

In the Alps (N = 436), all observations are classified by PS in classes 1-3 and there is no distinction possible between these classes. In the Pyrenees (results not shown) the results are similar.

For the Western Highlands (N = 562), the information about classes 4 and 5 is poor. The ordering of the curves is consistent with the ranking of PS, but there is no discrimination between 1 and 2. At the threshold 1-2/3-5 the discrimination capacity is good. A similar pattern is observed in the Central Highlands (results not shown).

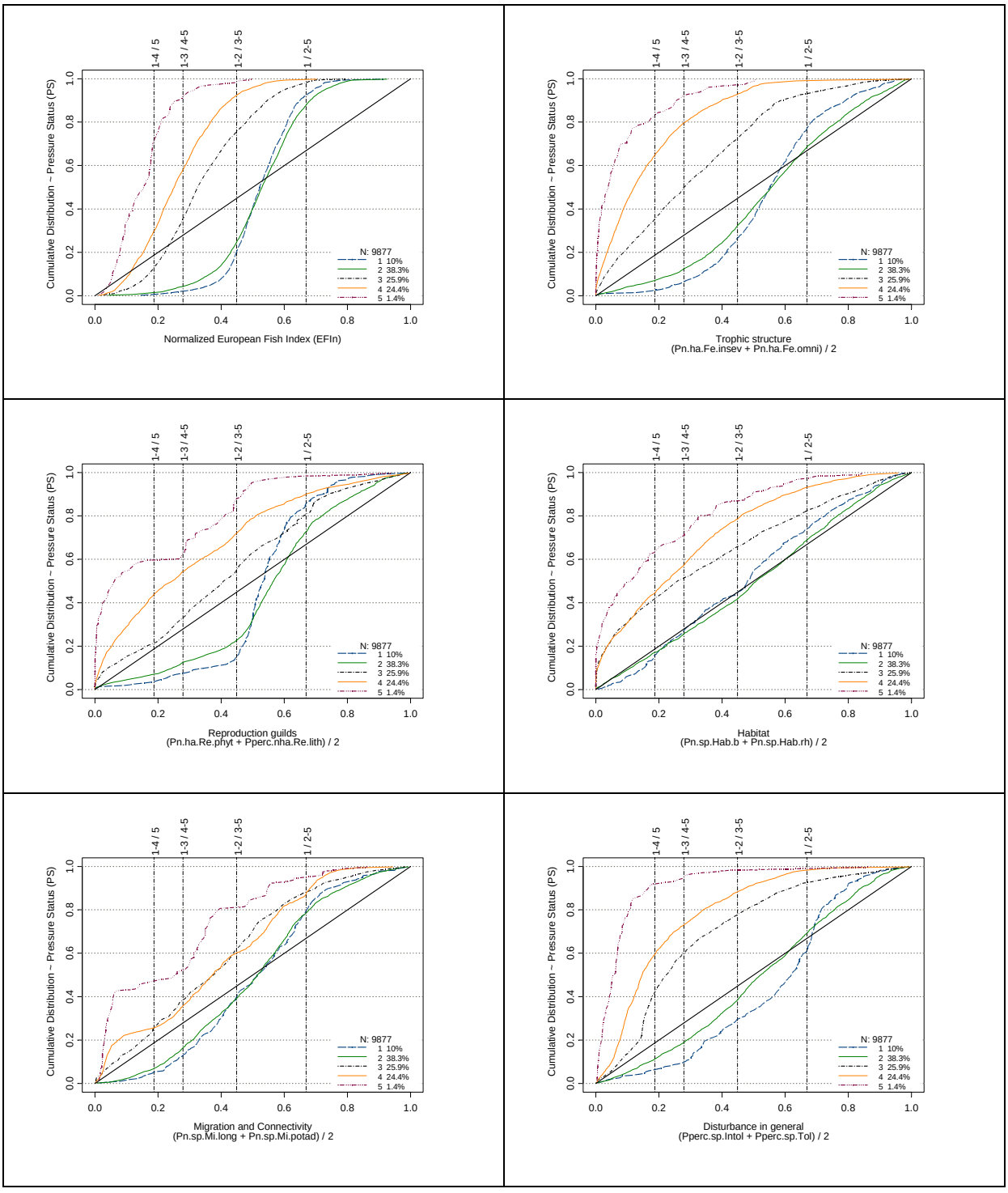
The distribution over PS in the Western Plains (N = 2215) is good. Only class 1 is somewhat underrepresented (1.5 %). The ordering of the EFI-curves is consistent with PS and the separation is high for the groups 1-2, 3 and 4-5. In the Central Plains (N = 2593) the pattern is similar. A very good contrast is present at the main threshold 1-2/3-5.

For Great Britain (N = 2039) for the distribution over PS only class 5 is underrepresented. The ordering of the curves is consistent with PS and the separation between the curves is high for the groups 1-2, 3 and 4-5 (as for the Western Plains).

Summary

With respect to the main contrast 1-2/3-5 in most cases the results are good. However the distinction in classes 1-2 and 4-5 is poor and not always 3 can be distinguished for 4-5.

FFF11 Consistency of the pressure status (PS) with the basic components of the EFI



The Consistency of the Components of the European Fish Index (figure FFF11)

The same method as in the previous sections we can apply to the individual components of the EFI. In fact the (normalized) continuous EFI_n is the average of 10 different metric scores which can be grouped two by two into 5 dimensions: trophic structure, reproduction guilds, type of habitat, migration and connectivity and disturbance in general.

Trophic structure (Pn.ha.Fe.insev & Pn.ha.Fe.omni)

For this combination the curves are ordered very nicely. Only the distinction between PS = 1 and 2 is somewhat blurred. Also the separation is excellent. If we compare with the full index we see that the sensitivity to detect a disturbed site is even larger than the full index. However this is at the expense of the specificity. More undisturbed sites are classified as disturbed than by the full index. Yet it is intriguing that with only two metrics such a clear cut result is obtained. This asks for further investigation.

Reproduction guilds (Pn.ha.Re.phyt & Pperc.nha.Re.lith)

From the perspective of reproduction guilds the curve for PS = 1 has a different behaviour from expected, especially in the higher regions. With respect to the main threshold 1-2 / 3-5 this is not a problem, but the determination of class 1 is totally unreliable: it will be a mixture of many different classes. All other curves are in the good order. However their distance is not very large, leading to a relative small sensitivity. Yet the specificity is good. Only about 20 % of the reference sites will be classified as disturbed.

Physical habitat (Pn.sp.Hab.b & Pn.sp.Hab.rh)

For the combination of the physical habitat metrics the specificity is low. A lot of undisturbed sites will be classified as disturbed as the curves for PS = 1 & 2 resemble closely a uniform

distribution. On the other hand the sensitivity is high: more than 80 % of the sites with PS = 4 or 5 will be classified as disturbed on the threshold 1-2 / 3-5 and this remains 60 % at the threshold 1-3 / 4-5.

Migration and connectivity (Pn.sp.Mi.long & Pn.sp.Mi.potad)

The metrics indicating connectivity problems have a low specificity. At the main threshold 1-2 / 3-5 40 % of the undisturbed sites will be classified as disturbed. At the same time the sensitivity is small. Only 60 % of the sites with PS = 3 or 4 will be detected. Only for PS = 5 it is 80 %.

General disturbance (Pperc.sp.Intol & Pperc.sp.Tol)

Finally for the disturbance in general the ordering of the cumulative curves is consistent with the pre-classification. Only at the end there is some change of the order of PS = 1 and 2 without any consequences as this is past the thresholds. For all situations the specificity is high. The problem is the specificity: nearly 40 % of the sites with PS = 1 will be classified as disturbed at the main threshold 1-2 / 3-5.

Summary

From this exercise we can derive how each metric contributes to the final index. For none of them a really deviant behaviour was observed. However, sometimes the discriminatory power can be low and the question is a little bit if all of them are really necessary. In fact the two metrics related with trophic structure seem to contribute most to the discrimination capacity: at the same time the specificity and sensitivity is high. Also the metrics related to general disturbance seem to offer a high sensitivity but the specificity is low. On the other hand the

metrics related to reproduction seem to offer a good specificity without scarifying too much of the specificity.

So in conclusion, we do not argue that two metrics are sufficient. Yet the suggestion is that probably less than 10 metrics suffice if we look from the perspective of sensitivity and specificity. Also this approach allows for an alternative selection procedure of the metrics, adding successively that metric that increases the sensitivity and specificity and hence the discrimination capacity most.

Discussion and Conclusions

A. EVALUATION of the INDICES WITH RESPECT TO EACH OTHER

As there is no golden standard, the first step was to compare the indices with respect to each other to get insight in their relative position.

The fish indices developed within FAME

The difference between classes 1 & 2 and 4 & 5 is unreliable. With respect to these classes, the classification by the different indices is not consistent and/or the discriminative capacity is low. As there is no golden standard available, it is hard to say which approach is the best. Regrouping into three classes (1-2, 3 and 4-5) improves the comparability very much. For this redefined indices class 3 takes in most instances an intermediate position and the cumulative curves of the reference situation (1-2) and the poor to bad status (4-5) are well separated. This fulfils the requirement of the WFD to have a good distinction between reference and disturbed sites (the contrast between 1-2 and 3-5).

The performance of the indices developed within FAME (EFI, SBM and SEM) does not vary much. The only method which can be ruled out is SBM-ER in favour of SBM-EU. The quality of the former is uniformly lower than of the latter. The strategy of pooling the fish typologies on a European level as a basis for stratification has been successful.

Evaluation in terms of sensitivity and specificity to detect the pressure status (PS) results in following ranking: SMB-EU ≈ SEM > EFI. The sensitivity and specificity ranges from 90 % (SMB-EU ≈ SEM) to 80 % (EFI). This difference of 10 % in quality is not very important

because the pre-classification of the pressure status (PS) is not without error. Further, a property of EFI is that it does not optimise directly the distinction between disturbed and undisturbed, but first models the reference situation without reference to disturbance. Only in the next step the well behaving metrics are selected in function of the contrast between reference and disturbed sites. Although this indirect optimisation is a handicap in terms of sensitivity and specificity, it is a protection against over-fitting, i.e. adjusting the parameters too much to the available data at the expense of extrapolation power.

In spite of this handicap in the modelling process, the prediction quality of EFI is comparable to the other approaches. This is remarkable and it is an indication that the fundaments of the model are biologically sound and practicable and go beyond a purely empirical fitting. For these two reasons the FAME-project selected the reference condition approach (RCA) as the European Fish Index (EFI).

The relation to the existing national or regional methods (ExM)

That the three basic philosophies lead to very similar results is not unexpected. All models are calibrated with respect to the same variable: the pre-classification of human impact PS. This is not true for the existing methods which stem from a multitude of methods and calibration external to FAME. At the main contrast 1-2/3-5 the correspondence with PS is as high as for the FAME indices. With increasing detail the correspondence goes down quickly. However, also for the FAME methods itself the distinction between 1 & 2 and 4 & 5 is questionable.

B. A FIRST EVALUATION of the EUROPEAN FISH INDEX (EFI)

Behind EFI there is a continuous variable lying between 0 and 1 (after normalization). This property allows a more detailed investigation of the characteristics of EFI.

Performance of EFI with respect to the pre-classification of the Pressure Status (PS)

As required by the WFD, the discriminative capacity between reference and disturbed sites is good. There is a clear separation with respect to PS between 1-2 and 4-5, class 3 having an intermediate position. At the main contrast 1-2/3-5 the sensitivity to detect a poor or bad status (4 or 5) is close to 100 %, for the moderate status (3) it is still 70 % and this with a specificity of 80 %. This gives a good balance between sensitivity and specificity and results in an IPV of about 80 %. At the next threshold 1-3/4-5 the specificity is about 100 % for class 1 and 2 (nearly no undisturbed sites are classified as poor or bad) and 70 % for class 3. Still the sensitivity for class 4 and 5 is more than 80%.

Screening by river group did not lead to important contradictions. If data were available over the total range of the index PS, the contrast between 1-2, 3 and 4-5 is good in general. For Meuse / North Sea the specificity is quite low: 40 % of the reference sites are classified as disturbed at the main contrast. For Sweden class 3 cannot distinguished well from 1-2. In the United Kingdom the specificity is low for class 2 (60 %). Similar results were found by Eco-region.

Position of EFI with respect to the existing national or regional methods (ExM)

The relation of EFI with the existing methods is seldom good. For Austria, EFI cannot reproduce any of the two methods. The sites range from 2 to 4 by the index of Flanders (in

Belgium). EFI classifies them from 3 to 5 and does not discriminate between level 2 or 3. The difference in quality seen from a Flemish perspective vanishes at a European level. In France the correspondence is good at the main contrast; however a lot of discriminating capacity disappears at the next threshold (1-3/4-5). The differentiation made by the index of Lithuania disappears for EFI and a large part of the cases classified as 2 receive a worse score. In Sweden 97 % of the sites have a very good status. About 20 % of them are scored 3 by EFI. Of the 2.6 % in class 2 about 50 % get a worse score. The differentiation made by the salmon index in UK from 1-4 disappears. Nearly always EFI gives a score of 2. For class 5 EFI ranges from 2 to 4.

The poor correspondence is only a first indication. An evaluation split by country and region turned out to be very hard if not impossible: or the number of cases was too limited (for one of the methods of Austria and for Wallonia in Belgium), or the distribution over the different levels of impact was too uneven. For instance, in Sweden most cases are reference sites, whereas in Flanders (Belgium) most cases are disturbed or very disturbed. In fact only for France a good comparison was possible. But there the local method was nearly equal to the EFI, so the very good correspondence with EFI was no surprise.

Performance of the components of EFI with respect to PS

The (normalized) continuous EFI_n is the average of 10 different metric scores which can be grouped two by two into 5 dimensions: trophic structure, reproduction guilds, type of habitat, migration and connectivity and sensitivity to general disturbance. For none of them a deviant behaviour was observed. The metrics related with trophic structure (insectivorous and omnivorous guilds) and disturbance in general (tolerant species and intolerant species) seem

to contribute most to the discrimination capacity. The migration and connectivity metrics seem to have the smallest contribution.

C. SYNTHESIS

All approaches developed within FAME give very similar results. Seen from the general performance the choice between the different methods is not that critical. Only the SBM-ER can be ruled out in favour of SBM-EU, which is the same methodology on a more global European scale.

The choice for EFI is based on the fact that it is based on a sounder biological concept. Only metrics with a good distribution in reference conditions are allowed for the calibration, and the quality of the metric is judged with respect to its distribution in reference conditions.

The discriminative capacity of EFI between reference and disturbed sites is good: a sensitivity and specificity of 80 %. Splitting by country, eco-region or river group or by component did not reveal major inconsistencies.

With the exception of France, the relation with the existing national or regional methods is poor. There are no real inconsistencies present (the same trends on a global level), but EFI is incapable to predict the individual local scores at an acceptable level.

However, one should be aware this is not an external data validation. Also, the data are often too weak (small subgroups or imbalanced distribution over the impact classes) to arrive at definite conclusions. In the future the performance of the EFI should be monitored with new data.

References

Agresti, A., 2002. Categorical Data Analysis. Wiley-Interscience. 710 pp.

Beier, U., Degerman, E., Melcher, A. & C. Rogers, 2005. Fides founding FAME – standardisation and mergence of existing data. This issue.

van Belle, G., 2002, Statistical Rules of Thumb. Wiley Series in Probability and Statistics, Wiley & Sons. 281 pp.

Belpaire, C., R. Smolders, I. Vanden Auweele, D. Ercken, J. Breine, G. Van Thuyne, & F. Ollevier, 2000. An Index of Biotic Integrity characterizing fish populations and the ecological quality of Flandrian waterbodies. Hydrobiologia 434: 17-33

Böhmer, J., 2005. Standardised European Model. This issue.

Chambers, J.M., W.S. Cleveland, B. Kleiner & A. Tukey, 1983. Graphical methods for data analysis. Wadsworth, Belmont, California.

Degerman E., Beier U., Breine J., Melcher A., Quataert P., Rogers C., Roset N. & I. Simoens, 2005. Assessment of human impact in European running waters. This issue.

van Dijk, G.M., van Liere, L., Admiraal, W., Bannink, B.A.& J.J. Cappon, 1994. Present state of the water quality of European rivers and implications for management. Sci. Total Environ. Vol. 145(1-2):187-195.

EU Water Framework Directive, 2000. Directive of the European parliament and of the council 2000/60/EC establishing a framework for community action in the field of water policy. Official Journal of the European Communities 22.12.2000 L 327/1

Fausch, K.D., J. Lyons, J.R. Karr & P.L. Angermeier, 1990. Fish communities as indicators of environmental degradation. American Fisheries Society Symposium 8: 123-144.

Fore, L. S., J.R. Karr & L.L. Conquest, 1993. Statistical properties of an index of biological integrity used to evaluate water resources. Can. J. Fish. Aquat. Sci. 51: 1077-1087.

Goudie, A. 1993. The human impact on the natural environment. (4th edition). Blackwell Publ., 454 p.

Hughes, R.M., P.R. Kaufmann, A.T. Herlihy, T.M. Kincaid, L. Reynolds & D.P. Larsen, 1998. A process for developing and evaluating indices of fish assemblage integrity. Can. J. Fish. aquat. Sci. 55: 1618-1631.

Hughes, R.M. & T. Oberdorff, 1999. Applications of IBI Concepts and Metrics to Water Outside the United States and Canada. In Simon, T.P. (ed.), Assessing the Sustainability and Biological Integrity of Water Resources Using Fish Communities. CRC Press LLC, Washington, D.C.: 62-74.

Karr, J.R., 1981. Assessment of biotic integrity using fish communities. Fisheries 6: 21-27.

Karr, J.R., K.D. Faush, P.R. Angermeier, P.R. Yant & I.J. Schlosser, 1986. Assessing Biological Integrity in Running Waters: A Method and its Rationale. Illinois Nat. Hist. Surv. Spec. Publ. 5, 28 pp.

Lyons, J., L. Wang & T.D. Simonson, 1996. Development and validation of an Index of Biotic Integrity for coldwater Streams in Wisconsin. N. am. J. Fish. Manag. 16: 241-256.

McCullagh, P. & J.A. Nelder, 1983. Generalized Linear Models. Chapman and Hall, London. 511 pp.

Melcher, A., 2005. The Spatially Based Method at a European level. This issue.

Motulsky, H., 1995, Intuitive Biostatistics. Oxford University Press, New York, Oxford. 386 pp.

Oberdorff, T. & R.M. Hughes, 1992. Modification of an index of biotic integrity based on fish assemblages to characterize rivers of the Seine Basin, France. Hydrobiologia 228: 117-130.

Pont, D., 2005. The European Fish Index. This issue.

Quataert, P., Breine J. & I. Simoens, 2005. Comparison of the European Fish Index with the Standardised European Model, the Spatially Based Models (eco-regional and European), and Existing Methods. Report of WP 10 of the FAME project.

Schmutz, S., 2005. The Spatially Based Methods on a Eco-regional level. This issue.

Shoukri, M.M., 2004, Measures of Interobserver Agreement. CRC Press. 152 pp.

Sokal R.R. & F.J. Rohlf, 1995, Biometry. W.H. Freeman and Company. 887 pp.

Strandberg, C. 1971. Water pollution. In: G.H. Smith (ed.), Conservation of natural resources (4^{th} edition). Wileys, New York:189-219.

Whittier, T.R. & R.M. Hughes, 2001. Comment: Test of an Index of Biotic Integrity. Trans. Am. Fish. Soc. 130: 169-172.