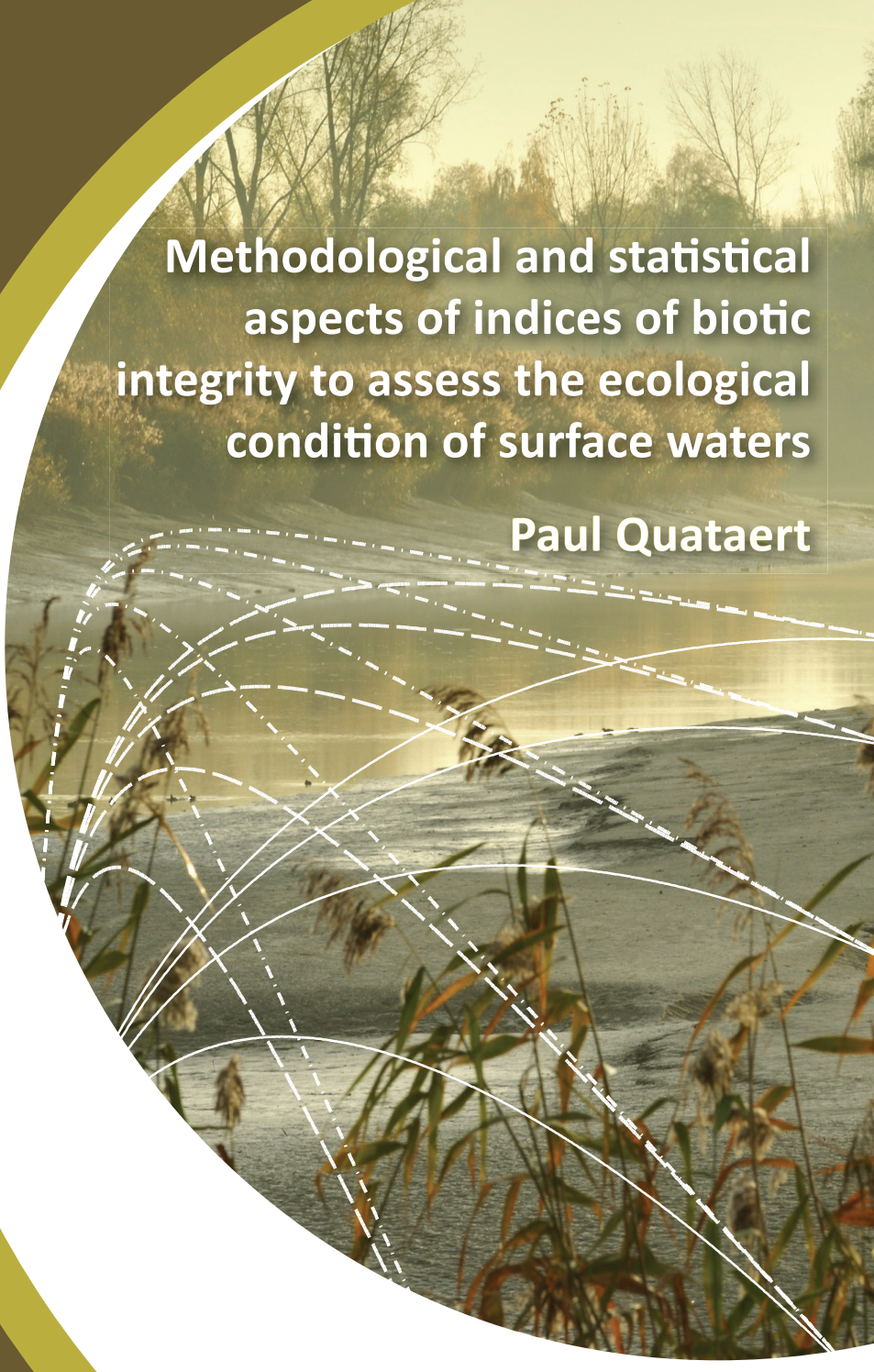


Methodological and statistical aspects of indices of biotic integrity to assess the ecological condition of surface waters

Paul Quataert

inbo

Text:
Paul Quataert

Promotors:
Prof. dr. em. Frans Ollevier
Afdeling Dierenecologie en -systematiek, K.U.Leuven
Departement Biologie
Prof. dr. Geert Verbeke
Leuvens Biostatistiek en Statistisch Bioinformatica Centrum, K.U.Leuven
Departement Maatschappelijke Gezondheidszorg

Research Institute for Nature and Forest
The Research Institute for Nature and Forest (INBO) is the Flemish research and knowledge centre for nature and its sustainable management and use. INBO conducts research and supplies knowledge to all those who prepare or make policies or are interested in them.

Contact:
INBO Brussel
Kliniekstraat 25, 1070 Brussel
www.inbo.be

e-mail:
paul.quataert@inbo.be

Wijze van citeren:
Quataert P. 2011. Een methodologische en statistische benadering van biotische integriteitsindices voor oppervlaktewateren. Doctoraten van het Instituut voor Natuur- en Bosonderzoek 2011 (INBO.T.2011.1). Instituut voor Natuur- en Bosonderzoek, Brussel.

To refer to this thesis:
Quataert P. 2011. Methodological and statistical aspects of indices of biotic integrity to assess the ecological condition of surface waters. PhD theses of the Research Institute for Nature and Forest 2011 (INBO.T.2011.1). Research Institute for Nature and Forest, Brussels.

D/2011/3241/021
INBO.T.2011.1

Responsible editor:
Jurgen Tack

Printing:
Managementondersteunende Diensten van de Vlaamse overheid.

Dit werk is een heruitgave, met toestemming van de auteurs en van de Katholieke Universiteit Leuven, van 'Quataert P. 2011. Methodological and statistical aspects of indices of biotic integrity to assess the ecological condition of surface waters. PhD Thesis. Arenberg Doctoral School of Science, Engineering & Technology. Faculty Science. **Department of Biology. K.U.Leuven. ISBN 978-90-8649-397-5.'**

This volume is a reprint, with permission of the authors and of the **Katholieke Universiteit Leuven, of 'Quataert P. 2011. Methodological and statistical aspects of indices of biotic integrity to assess the ecological condition of surface waters.** Arenberg Doctoral School of Science, Faculty Science. Department of Biology. K.U.Leuven. ISBN **978-90-8649-397-5**.'

Photos:
Yves Adams/Vildaphoto

METHODOLOGICAL AND STATISTICAL ASPECTS OF INDICES OF BIOTIC INTEGRITY TO ASSESS THE ECOLOGICAL CONDITION OF SURFACE WATERS

EEN METHODOLOGISCHE EN STATISTISCHE BENADERING VAN BIOTISCHE INTEGRITEITSINDICES VOOR OPPERVLAKTEWATEREN

Paul QUATAERT

Supervisors:

Prof. dr. em. Frans Ollevier
Prof. dr. Geert Verbeke

Members of the Examination Committee:

Prof. dr. Els Goetghebeur
Prof. dr. Martin Hermy
Prof. dr. Filip Volckaert
Dr. Jurgen Tack
Dr. Joost Vanoverbeke

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Sciences

March 2011

De onderzoeker die voor de samenleving wil werken, zal moeten leren in overleg met die samenleving de vragen scherp te stellen. Dan moet de onderzoeker de problemen operationaliseren, omzetten in een serie meetbare zaken. Etzioni wijst erop dat het heel belangrijk is dat de gebruiker daarbij wordt betrokken. De kans is niet gering dat de onderzoeker toch weer iets gaat meten dat de gebruiker helemaal niet weten wil, alleen maar omdat de onderzoeker bepaalde methoden bij de hand heeft. En als het onderzoek is afgesloten, dan moet de onderzoeker een rapport uitbrengen dat de gebruiker kan lezen en toepassen. Dat lijkt misschien voor de hand te liggen, in de praktijk is dat de grootste moeilijkheid. De onderzoeker is dan toch teveel deel van zijn eigen wereldje, blijkt toch teveel te zijn toegespitst op erkenning in eigen kring om dan nog de belangen van anderen in het oog te houden.

A. De Kool (Natuur & Techniek, 1975/6)

"Given n random cases of some variable" are typical starting words for statistical fables, and each word is misleading for real data. Most data are not "given" – they have to be taken, enticed, captured, or mined. The n is generally not fixed but varies, because of the many imperfections in collection. The selection is not simple "random", but clustered and stratified or otherwise complex. Also one obtains not true "variables", but only observations subject to errors, which the analyst should recognize and control. Such problems are all treated in statistical design.

Kish (Statistical design for research, 1987)

Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. A first, though at first sight, not a very helpful principle, is that all models are wrong; some, though, are more useful than others and we should seek those. At the same time we must recognize that eternal truth is not within our grasp. A second principle (which applies also to artists!) is not to fall in love with one model to the exclusion of alternatives. Data will often point with almost equal emphasis at several possible models and it is important that the statistician recognize and accept this. A third principle recommends more thorough checks on the fit of a model to the data, for example by using residuals and other statistics derived from the fit to look for outlying observations and so on. Such diagnostic procedures are not yet fully formalized, and perhaps never will be. Some imagination or introspection is required here in order to determine the aspects of the model that are most important and most suspect.

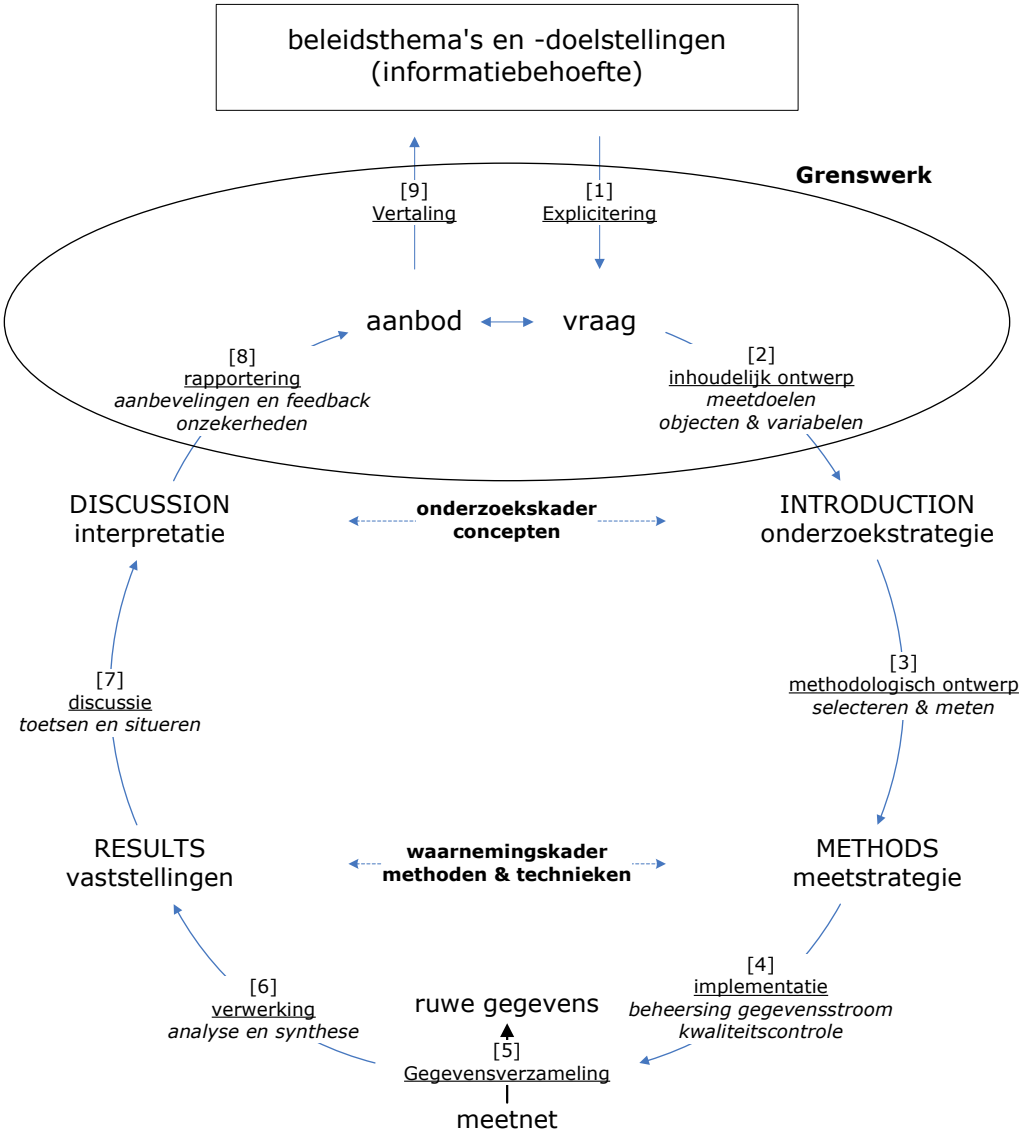
McCullagh P & Nelder J.A. (Generalized linear models, 1989)

Time has now run out. There is nowhere left for the author to go but to discuss just what constitutes a family for simultaneous inference. This is the hardest part of the entire book because it is where statistics takes leave of mathematics and must be guided by subjective judgment. The winner (when the mathematical smoke clears) is frequently an old friend of the author, whose merit could have been established on intuitive, distributional, or simplicity grounds. The experimentor is far more familiar with the data, its virtues and its vagaries, than the reader, so it is his prime responsibility to draw the main conclusions.

Miller (Simultaneous Inference, 1966)

De gegevenscirkel

empirische onderbouwing van het beleid
afstemming vraag naar en aanbod van gegevens



Organisatie/Samenwerking - Beheer/Supervisie - Evaluatie/Bijsturing
Infrastructuur - Competenties - Planning - Kosten

Dankwoord

Het schrijven van een doctoraat is een niet-lineair proces: weinig output in het begin, doodlopende straatjes en steile leercurven, lang uitgesponnen stukken om achteraf vast te stellen dat alles veel eenvoudiger kan. Ik heb hierbij op heel veel begrip kunnen rekenen van mijn promotoren, de professoren Frans Ollevier en Geert Verbeke. Hartelijk dank hiervoor. Frans, zelden werden mijn teksten zo grondig nagelezen en diepgaand becommentarieerd. Geert, je feedback was een goede hulp om het ecologisch model scherp te stellen en tot de essentie te herleiden. Veel dank aan de juryleden, Prof. Els Goetghebeur, Prof. Filip Volckaert, Prof. Martin Hermy, Dr. Jurgen Tack en Dr. Joost Vanoverbeke voor hun bijdrage tot een beter geheel. Els, bedankt om mij mee op te volgen. Je opmerkingen inspireerden me om naar diepere verbanden te zoeken.

Dank zij een sabbatical kon ik dit werk na bijna zeven jaar afronden. Ik wens hiervoor Jurgen Tack, Administrateur Generaal van het Instituut voor Natuur- en Bosonderzoek (INBO) en de afdelingshoofden Maurice Hoffmann, Janine Van Vessem en Dick Van Straaten te bedanken. In het bijzonder Dick. Ik kan me voorstellen dat het als kersvers afdelingshoofd niet evident was meteen een diensthoofd te zien vertrekken. Gelukkig kon ik rekenen op Dirk Bauwens en Gerrit Genouw om mij te vervangen. Dirk en Gerrit, erg bedankt om mijn taken over te nemen.

Mijn team gaf het voorbije jaar blijk van een grote autonomie zodat ik goed kon doorwerken. Pieter, dank voor je bijdrage aan hoofdstuk 5. Ook het rapport over de diatomeeënindex was puik werk (à propos, bedankt Luc, Gaby, Henk, Lidia, Frank en buurman Wim voor de aangename samenwerking). Thierry en Hans, ik was opgetogen over jullie uitwerking van het schaduwmeetnet voor de Vlaamse bosinventarisatie. Toon, ik had maar een maand tijd om je op weg te zetten, maar je hebt je erdoor geslagen (met de gewaardeerde hulp van Desiré, Patrick, Gerald, Dries en Jeroen). Ivy, ik hoor alleen maar lovende woorden over je aanpak van de lessenreeks statistiek.

Mijn voormalige directeurs Jos Van Slycken en Prof. Eckhart Kuijken gaven me krediet om deel te nemen aan het fusieproces. Met hun steun kon ik de basis leggen voor het BMK-team (biometrie en kwaliteitszorg).

Een thesis kan niet zonder onderwerp, geen triviaal probleem voor iemand die statistisch advies verleent zonder "eigendomsrechten". Daarom wil ik van harte Jan Breine bedanken die zonder aarzelen toestemming gaf om op zijn thema te doctoreren. Hierbij wil ik ook graag Claude Belpaire, Gerlinde Van Thuyne, Hugo Verreycken en de vele "vissers" vermelden. Op hun werk kon ik verder bouwen. Een bijzonder woord van dank aan Ilse Simoens. Ik had geen tijd om mee te gaan op cursus in Amsterdam, maar het was goed dat jij en Jan me met vereende krachten weghaalden van achter mijn bureau. Het een bracht het ander met zich mee, en zo zaten we een jaar later samen op een vliegtuig naar Portugal voor FAME, waar een wereld voor mij openging. Dit was de aanpak waar ik van de droomde, op de grens tussen beleid en wetenschap.

Especially, I would like to thank the project leader Stephan Schmutz for the many interesting and thought provoking discussions about the deeper background of FAME and science. Gertrud Haidvogl and Andreas Melcher, I remember as the people taking care of me my first FAME hours in Sesimbra and thank you for arranging an extra meeting in Groenendaal when I asked for help. Didier Pont

invited me to Lyon for a further discussion. Teresa Ferreira, thanks for informing whether I could work on my PhD in Portugal. Thanks to all other FAMERS for their openness and opening my eyes.

Grenswerk, onderzoek op de grens tussen wetenschap en beleid, ik verwees er al kort naar, is voor mij een belangrijke rode draad. Vele mensen begeleidden me hierbij en/of gaven me ruimte. Op het Wetenschappelijk Instituut voor Volksgezondheid, kreeg ik een ruime bagage mee van Herman Van Oyen, Viviane Van Casteren en Patrick Leurquin. Bij de Afdeling Natuur leerde ik van Jeroen Cockx het belang van langetermijnplanning en kon ik als coördinator bij het milieubeleidsplan de eerste bouwstenen ontwikkelen om meetnetten beter op het beleid af te stemmen. Hieraan droegen veel mensen bij: Geert De Blust en Mira Van Olmen (de "snuffelplaatsen"), Dirk Maes en Hans Van Dyck (het multisoortenconcept), Maurice Hoffmann en Else Demeulenaere (meetnet beheer natuurreservaten), Ria Gielis en Hilde Heyrman (monitoring natuurontwikkeling). Dankzij de inzet van Ludo Vanongeval (leidraad ontwerp beleidsgerichte meetnetten), Martine Waterinckx (tweede Vlaamse bosinventarisatie en alles wat erop volgde) en Desiré Paelinckx (Natura 2000) kon ik hieraan verder bouwen. Ook Koen Martens, Bart Roelandt, Philippe Van Haver, Philip Van Avermaet, Ward De Cooman en Bernard Van Eleghem waren kritische en bijgevolg uitstekende gesprekspartners om vorderingen te maken. Dank aan Thierry Onkelinx voor zijn pionierswerk bij de audit van de meetnetten en aan Jasper Wouters, gerugsteund door Dirk Bauwens, om mijn ideeëncocktail te stroomlijnen tot een toegankelijke leidraad.

Enkele mensen gaven me input vanaf de zijlijn. Dank aan Bart Goossens, Bart De Pauw en Herwig Borremans van de bib voor hun speurwerk naar papers die ik niet kon vinden. In Leuven hielpen Conny Coeckelberghs en Indra Hautekiet me altijd heel vriendelijk en efficiënt verder. Van Dirk Maes mocht ik per email regelmatig een leestip ontvangen. Mails van Bruno De Vos, wroetend op zijn thesis, vroegen iets meer engagement, maar zijn vragen hielden mijn kennis statistiek op peil in tijden van fusieperikelen. Caroline Geeraerts verzette veel werk om de Kaderrichtlijn Water mee te helpen doorgronden. Dank je voor je enthousiaste inzet. Ook Bart Vandecasteele wil ik danken. Van heel dichtbij kon ik de groei van je doctoraat meemaken en bijleren hoe kleine, goed gerichte onderzoeksstapjes bijdragen tot iets nieuws. Mijn advies beloonde je steevast met coauteurschap. Morele steun kwam er van Jan, Ilse en Sandra en hun partners Erika, Knut en Ward. Bedankt voor de aangename babbels (en me op tijd en stond te herinneren aan een deadline). De welgemeende interesse van Luc & Leentje en Luc & Lieve mag ik hierbij zeker niet vergeten te vermelden.

Ook mocht ik veel steun ervaren op het thuisfront. Volgens de sinoloog Ulrich Libbrecht is het geen toeval dat in het Engelse "to believe" liefde vervat zit en we in geloven "to love" herkennen. Deze taalverwarring duidt er volgens hem op dat "houden van" en "geloven in" nagenoeg hetzelfde betekenen. Deze gedachte is heel toepasselijk op mijn vrouw. Cecilia, bedankt voor je goede zorgen en luisterend oor. Hierbij denk ik ook aan mijn ouders en grootouders die me heel veel waardevols hebben meegegeven. Bedankt papa en mama voor alle steun. Mama, spijtig dat je er op mijn verdediging niet bij kan zijn. Lieve, An, Piet, Jan, Moeke (bedankt voor het kaartje) en alle schoonbroers en schoonzussen, hartelijk dank voor jullie sympathie. Katrien en Elisabeth, mij ontvoeren voor het omakamp is niet meer nodig. Renee, in mei gaan we naar de zoo. Beloofd!

Paul, zondag 6 februari 2011

Table of contents

Summary

The general aim of an index of biotic integrity (IBI) is to provide policy makers, managers and stakeholders an overall appreciation of the ecosystem condition of a site in one synthetic measure. This is achieved by evaluating the species composition at the community level. The ecological rationale consists in the fact that anthropogenic changes of the environmental conditions and ecosystem resources ultimately result in a shift of the species composition. By quantifying selected attributes of the species composition in a test variable, human impacts can be followed up. An IBI is conceived according to the principles of the Reference Condition Approach (RCA): the index assesses the ecosystem condition of a test site by evaluating the composition of its biological community compared to the expected configuration under reference conditions. If the difference is substantial in comparison to the intrinsic natural variability, it is concluded that the test site is impacted by an anthropogenic source. Ideally, the reference sites are pristine, or nearly so. Yet, it is possible to develop IBIs with respect to any well motivated baseline condition, accepted as a societal goal.

This thesis discusses the IBI concept from a statistical and methodological perspective. A first focus was to obtain a better understanding of the underlying rationale. Several papers describe how to construct IBIs. However, to our knowledge, none of them makes the underlying statistical model explicit. They just give a narrative description of the calculation steps. Yet, we were able to derive a simple but flexible format, phrasing the four transformation steps of IBIs in a formalised statistical language. (i) The model equations start with deriving metrics from the community data. The metrics reflect ecological relevant features of the species distribution and are sensitive to anthropogenic alterations of the ecosystem. (ii) Subsequently, the metrics are scored, expressing how (dis)similar the observed metric values are in comparison to type-specific or site-specific reference conditions taking into account the site typology and the intrinsic natural variability. (iii) In the third step, the individual scores are aggregated (traditionally by simply summing or averaging) into the ecological quality measure (EQM), meant as an overall impact assessment. As such, EQM is hard to interpret because its value does not tell directly how impaired the ecosystem is and whether to decide about restoration. (iv) Hence, the fourth and final step compares EQM with decision thresholds resulting in the ecological quality class (EQC), an ordinal class variable appreciating the degree of biotic integrity.

When developing an IBI, the selection of an optimal metric set from a candidate list remains one of the main challenges. In many instances, a coherent and transparent strategy is lacking leaving much room for subjective decisions and/or personal preferences. Important unresolved questions include how many metrics to select (the model dimension) and how to choose properly between them. In our opinion, two important factors contribute to suboptimal metric choice. First of all, optimisation criteria are seldom made explicit and are not always appropriate. Only occasionally, the crucial distinction is made between false positive (FP) and false negative (FN) errors. A high false negative fraction (FNF) implies that many degraded sites are not restored, not realising the full potential recovery of ecosystem functions, goods and services. Conversely, a high false positive

fraction (FPF) results in unnecessary restoration of many unimpaired sites, detracting resources and/or risking to harm pristine sites. In analogy to diagnostic models in medicine, we propose the Receiver Operating Characteristic (ROC) curve to optimise the diagnostic accuracy of the index. ROC curves plot the true positive fraction (TPF) as a function of the false positive fraction (FPF). To gain a deeper understanding of the impact of the shape of ROC curves, we introduced utility curves plotting the cost implications of index-guided decisions as a function of FPF or TPF of the index. Utility curves link the strength of the index as characterised by the height of the ROC curve with its practical usefulness. We inferred that the main factor determining the usefulness of the index is its capacity to realise a high TPF keeping FPF small. More specifically, we demonstrate how a strong index is capable to realise a high true restoration fraction (TRF) and a high overall restoration benefit (ORB) at a low average restoration cost (ARC).

The second factor leading to a suboptimal metric choice is the insufficient recognition that an IBI is in essence a regression model. Traditionally, an IBI is simply an average (or sum) of scored metrics. This average score model (AVG) is an ordinal logistic regression model (OLR) in disguise because it is not necessary to estimate the regression coefficients which are fixed. Yet, we can borrow concepts, strategies and techniques from statistical model building to search for the optimal suite of metrics. In this context, an important issue is overfitting, i.e. selecting too many variables in comparison to the data available, resulting in a lower diagnostic accuracy than a simpler, more parsimonious model. Another point to consider is that the optimisation criterion is a random variable. Hence, the optimal model is not necessarily the best one. To cope with these and other problems, we propose a modelling strategy which first ascertains the optimal number of metrics and then explores competing models in the vicinity of that optimum.

We illustrate our approach by revising the Estuarine Biotic Index (EBI) for the mesohaline part of the Zeeschelde Estuary in Flanders, Belgium. Statistical modelling techniques, such as best subset regression and bootstrapping, combined with optimisation criteria derived from the ROC curve, are forged together into a powerful and transparent strategy to select the optimal set of metrics. We also demonstrate that the proportional odds logistic regression model results in a very similar model as AVG. This extension to generalised linear models (GLM) opens a perspective to formulate more flexible models better adapted to the sampling design and able to incorporate background variables, adjusting for differences between sites.

A second illustration is the evaluation of the European Fish Index (EFI) as developed by the EU funded FAME project (Development, evaluation and implementation of a standardised fish-based assessment method for the ecological status of European rivers). An important requirement for meeting obligations under the European Water Framework Directive (WFD) is the development of a fish-based index that is able to predict the ecological status of surface waters and particularly to distinguish between (nearly) pristine and disturbed conditions. For the EFI, the overall FPF was 22 % and the FNF 19 %. Comparison of EFI with existing national or regional fish-based assessment methods revealed major discrepancies making intercalibration between them unfeasible.

In our opinion, a representative sampling scheme covering the full spectrum of human impacts in a region, is a prerequisite to retrieve a responsive set of metrics. Therefore, to conclude, we present suggestions to improve data collection for the calibration of IBIs.

Samenvatting

Een biotische integriteitsindex (BI) is een ecologische maatlat die de algemene toestand van een ecosysteem samenvat en herleidt tot een eenduidige categorie ten behoeve van beleidsmakers, terreinbeheerders en belanghebbenden. Dergelijke maatlatten worden vooral toegepast voor aquatische milieus, ofschoon ze ook voor terrestrische ecosystemen bruikbaar zijn. De categorieën geven aan in hoeverre een bepaalde locatie beantwoordt aan een referentiebeeld. Zo werkt de Kaderrichtlijn Water (KRW) met vijf categorieën(kleuren): 1 = hoog (blauw), 2 = goed (groen), 3 = matig (geel), 4 = ontoereikend (oranje) en 5 = slecht (rood). Hierbij corresponderen categorie 1 en 2 respectievelijk met de natuurlijke en licht afwijkende toestand. Vanaf categorie 3 gaat het om duidelijke antropogene veranderingen van het ecosysteem. De KRW beoogt om tegen 2015 alle waterlopen minstens in een goede toestand (categorie 2) te brengen.

Biologische integriteitsindices maken een evaluatie van de soortensamenstelling op het niveau van de levensgemeenschappen. Antropogene veranderingen van milieu en ecosysteem vertalen zich uiteindelijk in verschuivingen binnen de soortensamenstelling. Bijgevolg kan de menselijke invloed op een ecosysteem opvolgd worden door deze verschuivingen te kwantificeren. Om rekening te houden met de natuurlijke variabiliteit, wordt een BI geijkt conform de principes van "Reference Condition Approach" (RCA). Hierbij wordt de index berekend door de soortensamenstelling van een testgebied te vergelijken met een referentiebeeld, zoals afgeleid uit een representatieve steekproef van vergelijkbare referentiegebieden. Alleen wanneer het verschil substantieel is in vergelijking met de natuurlijke variabiliteit, wordt besloten dat het geteste gebied verstoord is. In het ideale geval corresponderen de referentiebeelden met een (nagenoeg) ongerepte toestand.

Deze thesis bespreekt het BI concept vanuit een statistisch en methodologisch perspectief. Een eerste doelstelling was daarom de onderliggende ecologische achtergronden van een BI duidelijk te stellen. Vrij veel auteurs hebben het proces om een index te ijken al grondig beschreven, maar zover we konden nagaan zonder het onderliggende statistische model te expliciteren. Nochtans is een eenvoudig en toch algemeen schema mogelijk dat de vier opeenvolgende transformatiestappen om een BI te berekenen in een statistische taal beschrijft. (i) Het betreft vooreerst de afleiding van de zogenaamde metrieken uit de waargenomen soortensamenstelling. De metrieken worden zo gekozen dat ze de meest relevante kenmerken van de soortensamenstelling voorstellen die gevoelig zijn voor antropogene beïnvloeding van het ecosysteem. (ii) Vervolgens worden aan de metrieken scores toegekend die uitdrukken hoe verschillend de waargenomen metrieken zijn vergeleken met de referentietoestand. Hierbij wordt rekening gehouden met het type gebied en de natuurlijke variabiliteit. (iii) In een derde stap worden de individuele scores samengevoegd in een globale ecologische kwaliteitsmaat (EKM), door bijvoorbeeld het rekenkundig gemiddelde van de scores te berekenen. Deze maat is niet eenvoudig te interpreteren. (iv) Daarom leidt de vierde en laatste stap de ecologische kwaliteitsklasse (EKC) af door de EKM in te passen binnen drempelwaarden. Deze EKC is een objectieve weergave van de graad van biotische integriteit.

De optimale keuze van metrieken voor een BI blijft een grote uitdaging. Veelal ontbreekt een doordachte strategie waardoor veel ruimte wordt gelaten voor persoonlijke voorkeur. Twee

belangrijke factoren dragen hiertoe bij. Een eerste factor is het ontbreken van precieze selectiecriteria. Lang niet altijd wordt immers voldoende het onderscheid gemaakt tussen foutpositieven en foutnegatieven hoewel beide soorten fouten verschillende implicaties hebben. Een hoge foutnegatieve fractie of lage sensitiviteit leidt ertoe dat veel verstoorde locaties niet aangepakt worden. Omgekeerd resulteert een hoge foutpositieve fractie of lage specificiteit in veel onnodige ingrepen. Naar analogie met diagnostische modellen in de geneeskunde, wenden we de ROC (Receiver Operating Characteristic) curve aan om de diagnostische kwaliteit van een index te bestuderen. Dergelijke curve beschrijft hoe de echtpositieve fractie of sensitiviteit toeneemt als een functie van de foutpositieve fractie. Om ROC curven beter te begrijpen, hebben we nutsfuncties (utility functions) afgeleid die de kosten kwantificeren van beslissingen genomen op basis van de index. Deze nutsfuncties leggen het verband tussen de sterkte van een index, zoals aangeven door de hoogte van een ROC curve, en de praktische bruikbaarheid van de index. Meer in het bijzonder tonen we aan dat een index met een steile ROC curve toelaat een groot aandeel van de verstoorde gebieden te herstellen voor een gemiddeld lage restauratiekost.

Een tweede factor die resulteert in een suboptimale keuze van metrieken is dat niet onderkend wordt dat een index in feite een regressiemodel is. Dat komt omdat traditioneel een biotische index eenvoudigweg een rekenkundig gemiddelde is van de metriekscores. Dit gemiddelde scoremodel is vrij verwant aan een ordinaal logistisch regressiemodel, maar de regressiecoëfficiënten liggen op voorhand vast. Bij het ijken van een index moeten we alleen nog de metrieken kiezen. We tonen aan dat dit kan door gebruik te maken van concepten, strategieën en technieken uit het domein van de statistische modelbouw. Hierbij moeten we vermijden dat meer metrieken in de index worden opgenomen dan noodzakelijk, waardoor de diagnostische accuraatheid van het model lager komt te liggen dan bij een model met minder variabelen ("overfitting"). Het criterium om de kwaliteit te beoordelen is een toevalsveranderlijke zodat het optimale model volgens dat criterium soms toevallig als beste wordt aangeduid. Om ook dit knelpunt op te lossen, vergelijken we in de studie meerdere modellen die in de buurt liggen van het optimale.

We illustreren onze aanpak vooreerst door de estuariene visindex voor de Zeeschelde te herzien. Uitgaande van zestien metrieken die belangrijke functies van het estuariene ecosysteem weergeven, selecteerden we een korf van vier metrieken: het percentage migrerende soorten die het estuarium in hun juveniele stadium benutten, het aantal soorten gebonden aan brakke wateren, het percentage piscivoren en het percentage Spiering. Ten tweede evalueren we de Europese visindex die door het FAME project werd ontwikkeld. Dit project had als doel een visindex te ontwikkelen op pan-Europese schaal. Bij gebruik kwam naar voor dat zowel de foutpositieve als de foutnegatieve fractie van deze index ongeveer 20 % bedroeg. Beide fouten zijn dus goed in balans, wat helemaal niet geval was voor de meeste nationale visindices. De verschillen tussen de Europese en de andere indexen zijn zelfs zo groot, dat een intercalibratie niet haalbaar is.

Tot slot brengen we aanbevelingen voor een betere steekproefschema dat in de toekomst de verantwoorde aanmaak van biotische integriteitsindices sterker kan onderbouwen.

List of abbreviations, acronyms and symbols

A.	Cost symbols expressing the benefit of index guided decisions in comparison to a full restoration (reference cost = all sites restored)
A_{Opt}	Maximal value for A_{ROC} (at b.ΔTPF = ΔFPF)
A_{ROC}	Benefit in excess to full restoration; $A_{ROC} = B_{ROC} - (T_G - T_H) = T_H.a_{ROC}$
a_{ROC}	Kernel function of A_{ROC}; $a_{ROC} = Sp - b.FNF$
ABC	Activity-based Costing
AFM	Assessment fraction of the total monetary or management cost
AQEM (EU project)	Development and Testing of an Integrated Assessment System for the Ecological Quality of Streams and Rivers throughout Europe using Benthic Macroinvertebrates
ARC	Average restoration cost
AUC	Area under the curve
aucF	(Full) area under the curve for the full range = aucF(0,1)
aucP(f_1,f_2)	Partial area under the curve for $f_1 \leq FPF \leq f_2$
aucP	Default-value = aucP(0.1,0.3)
AVG	Average score model
b	Benefit ratio at the level of the region; $b = T_G/T_H = b_R.odds(\pi^+)$
b_R	Intrinsic benefit ratio because of restoration; $b_R = R_G/R_H$
B.	Cost symbols expressing the benefit of using an index
B_M	Total benefit after correction for monetary costs; $B_M = B_{ROC} - C_A = E_{ROC} - C_M$
B_{Opt}	Maximal value for B_{ROC} (at b.ΔTPF = ΔFPF)
B_{ROC} (= ORB)	Benefit along ROC curve in comparison with no restoration at all (reference = no sites restored); $B_{ROC} = T_G.TPF - T_H.FPF = T_G.(TPF - T_H/b) = T_G.b_{ROC} = b.T_H.b_{ROC}$
b_{ROC}	Benefit function: kernel of B_{ROC}; $b_{ROC} = TPF - FPF/b$
BAU	Business as usual
BCa	Bias-corrected, accelerated percentile method (bootstrapping)
C	Community data (species data collected according to a standardised protocol)
C.	Generic cost symbol
$C_{FN} \leftrightarrow C_{TP}$	Degraded sites: cost of not restoring (FN) / restoring (TP)
$C_{TN} \leftrightarrow C_{FP}$	Pristine sites: cost of not restoring (TN) / restoring (FP)
C_0	Ecological & societal cost status before restoration (cost of no intervention); $C_0 = \pi^+.E_{FN} + (1-\pi^+).E_{TN} = \pi^+.E_{FN}$ (if $E_{TN} = 0$)
C_A	Assessment cost
C_D	Decision costs (monetary evaluation of cost consequences)
C_E	Entire cost ($C_A + C_D$)
C_{GS}	Cost associated with (mis)classification in comparison to gold standard (GS)
C_M	Total monetary cost or total management cost

C_R	Restoration cost
C_T	Total cost
C_Z	Zero or reference point for cost
CBA	Cost benefit analysis
CEA	Cost effectiveness analysis
CL	Confidence limits
CMA	Cost minimisation analysis
CWA	Clean Water Act (US)
D	Degraded or disturbed condition (see EQC)
DPSIR	Causal chain: Driving forces (human activities) → Pressures → State → Impact and Response of the society to these problems.
$E_\bullet$	Ecological cost
$E_{FN} \leftrightarrow E_{TP}$	Degraded sites: ecological cost of not restoring (FN) / restoring (TP)
$E_{TN} \leftrightarrow E_{FP}$	Reference sites: ecological cost of not restoring (TN) / restoring (FP)
$E_G \leftrightarrow E_H$	Ecological gain/harm of restoring degraded/pristine sites; $E_G = E_{FN} - E_{TP}$, $E_H = E_{FP} - E_{TN}$
E_{ROC}	Ecological restoration benefit. E_{ROC} = ERB = B_{ROC} – ARC
E[X]	Expected value of X
EBI	Estuarine Biotic Index
EDF	Empirical distribution function
EFI	European Fish Index (developed by the FAME consortium)
EFI+	Updated European Fish Index
EI	Ecological integrity
EDA	Exploratory Data Analysis
EQC	Ecological Quality Class (ordinal variable) Binary (bi-state): EQC = R (reference), D (degraded) Ternary (tri-state): EQC = R (reference), M (moderate), D (degraded) Five-tiered (WFD + colour system): EQC = high (1 = blue), good (2 = green), moderate (3 = yellow), poor (4 = orange) and bad (5 = red)
EQM	Ecological Quality Measure (continuous test variable / yardstick)
EQR	Ecological Quality Ratio (WFD: continuous test variable with range 0 – 1)
EQS	Ecological quality status
ERB	Ecological restoration benefit (E_{ROC})
EU	European Union
f (f_R, f_M, f_D)	Density function of the test variable (R = reference, M = moderate, D = degraded)
F (F_R, F_M, F_D)	Cumulative distribution corresponding with the density functions f
F^{-1}	Inverse distribution function = quantile function
FAF	False alarm fraction (1 – PPV)
FAME (EU project)	Development, Evaluation & Implementation of a Standardized Fish-based

	Assessment Method for the Ecological Status of European Rivers
FC, FCF	False classification (misclassification), false classification fraction
FI	Fish Index
FIDES	Fish Database of European Streams (the common database of the FAME project)
FN, FNF	False negative, false negative fraction
FP, FPF	False positive, false positive fraction
FRF	False reassurance fraction (1 – NPV)
FV	Future value
GES	Good ecological status
GIS	Geographic Information System
GLM	Generalised linear model
GS	Gold Standard
h_{ij}	Characteristics of the species i with respect to metric j. e.g. h_{ij} = 1/0 specifies whether species i belongs to group j of the feeding guild
HIC	Human Impact Class (discrete)
HIS	Human Impact Score (continuous)
HQC	Human Quality Class
IBI	Index of Biotic Integrity
logit(x)	logit(x) = log(odds(x)) = log (x /(1 - x))
M	Moderated condition (see EQC)
M_j	Metric j (j = 1, 2, …, J)
$m_j(X\|R)$	Expected value of metric j (or any other location measure)
MEA	Millennium Ecosystem Assessment (2005)
MM	Mixed model
MMI	Multimetric index of biotic integrity
MuLFA	Multi-level concept for fish-based assessment (IBI of Austria)
NPV	Negative predictive value
n_B	Number of bootstrap (re)samples (default: 1000)
<ns>	Prefix: metrics based on the number (#) of species
nsBen	# of benthic species [0 – 10]
nsBra	# of all but freshwater species [0 – 17]
nsDia	# of diadromous individuals [0 – 5]
nsErs	# of estuarine resident individuals [0 – 5]
O/E	Ratio of observed and expected value = original definition and meaning of EQR (ecological quality ratio).
odds(x)	odds(x) = x / (1 - x); $0 \le x \le 1$
OLR	Ordinal logistic regression (model)
ORB	Overall restoration benefit

p_i	Proportion of individuals in the sample of species i
P(.)	Probability
P(W|U)	probability of test outcome W given the unknown status U
P(U|W)	(a posteriori) probability disease status U given test outcome W
<pi>	Prefix: metrics based on the percentage (%) of individuals
piDia	% of diadromous individuals
piErs	% of estuarine resident individuals
piExo	% of invasive individuals
piFlo	% of flounder individuals
piMjm	% of marine juvenile migrating individuals
piOmn	% of omnivore individuals
piPis	% of piscivorous individuals
piSme	% of smelt individuals
PP	Precautionary Principle
PS	Pressure status
PPV	Positive predictive value
PV	Present value
QC/QA	Quality Control / Quality Assurance
R	Reference condition (see EQC)
R**.**	Cost-corrected restoration gain or harm
R_G	Cost-corrected ecological gain of restoring degraded sites; $R_G = E_G - C_R$
R_H	Cost-corrected ecological harm of restoring pristine sites; $R_H = E_H + C_R$
REBECCA (EU project)	Relationships between Ecological and Chemical status of surface waters
RIVPACS	River Invertebrate Prediction And Classification System (software package of the Institute of Freshwater Ecology (IFE) to assess the river quality in the UK)
ROC curve	Receiver Operator Characteristic curve
RCA	Reference Condition Approach
S_j	Scored metric j (often z-score)
$s_j(X\|R)$	Standard deviation of metric j (or any other spread measure)
Se	Sensitivity (= TPF)
Sp	Specificity (= TNF)
STAR (EU project)	Standardisation of River Classifications
Stdev	Standard deviation
T	Decision or detection threshold
T**.**	Total gain or harm on a regional level
T_G	Expected total gain of restoring all degraded sites; $T_G = R_G \, n^+$
T_H	Expected total harm of restoring all pristine sites; $T_H = R_H \, (1 - n^+)$

TC, TCF	True (correct) classification (correct), true classification fraction
TEEB	The Economics of Ecosystems and Biodiversity (2010)
TMC	Total management cost
TMF	Total misclassification fraction
TN	True negative
TNF	True negative fraction (= specificity)
TP	True positive
TPF	True positive fraction (= sensitivity)
TRF	True restoration fraction (= PPV)
$U^{\bullet}$	Real (but unknown) status of the patient, ecosystem, …
U^{+}	Event of interest present: e.g. diseased, degraded (moderately or severely impaired, most impacted), …
U^{-}	Event of interest absent: e.g. not diseased, not degraded (healthy, pristine, reference, baseline, least impacted), …
U_C	Unknown status variable (continuous)
U_D	Unknown status class (discrete: binary or bi-state, ordinal)
<va>	Prefix: metrics averaging indicator values of species
vaTol	Total tolerance value
<vd>	Prefix: metrics based on diversity measures
vdDiv	Simpson diversity index
vdSha	Shannon diversity index
vdSim	Simpson dominance index
VLINA	Vlaams Impulsprogramma Natuurontwikkeling
w_i	General symbol to express the weight given to a variable
$W^{\bullet}$	Test outcome to work with for decision making
W^{+}	Test outcome positive (signal)
W^{-}	Test outcome negative (no signal)
W_C	Test variable (continuous)
W_D	Test class (discrete: binary or bi-state, ordinal)
WFD	Water Framework Directive
WTP	Willing to Pay
X	Predictors: characterisation of the site (environmental characteristics, type, …) & data collection conditions (weather, period, sampling method, …)

List of Greek symbols

α	Level of significance (risk of a type I error)
β	Risk of a type II error (1 – power)
β_j	Regression coefficients (β_0 = intercept; $\beta_{0;c}$ = intercept for the c^{th} cumulative contrast of a proportional odds model)
$\delta_i = 0/1$	Indicator whether species i is present in the sample ($\delta_i = p_i^0$)
$\delta_j = \pm 1$	Multiplier for metric j: +1 for positive metrics (positively associated with ecological quality) / -1 for negative metrics
Δ	Difference (e.g. ΔARC = difference of Average Restoration Cost of two indices)
$\Delta(C,\Phi(X\|R))$	Multivariate distance function calculating the "distance to target" of a vector (e.g. species data) with respect to its reference distribution.
$\eta_\bullet$	Efficacy
η_E	Ecological efficacy of restoration; $\eta_E = E_G/E_{FN} = 1 - C_{TP}/E_{FN}$
η_C	Cost-corrected efficacy of restoration; $\eta_C = R_G/E_{FN} = 1 - (C_{TP} + C_R) / E_{FN} = T_G/C_0$. Note that $\eta_C = T_G/C_0$ if $E_{TN} = 0$
$\pi^\bullet$	Prevalence (proportion of statistical population with property of interest)
$\pi^+ = P(D^+)$	Prevalence of disease / degradation
$\pi^- = P(D^-) = 1 - \pi^+$	Prevalence of health / high ecological status
$\Phi(X\|R)$	Multivariate reference distribution (R) as a function of the predictors X
Φ_C	Reference distribution for community data
Φ_M	Reference distribution for metrics

1 General introduction

1.1 Presentation of the subject

The subject of this thesis is the index of biotic integrity (IBI). The general aim is to provide policy makers, managers and stakeholders with an overall appreciation of the ecosystem condition in one single synthetic measure (Coates *et al.*, 2007). This is achieved by making an evaluation of the species composition at the community level (Attrill and Depledge, 1997). Instead of looking at the response of individual species, IBIs consider the species composition. The ecological rationale is that anthropogenic alterations of the environmental conditions and ecological resources at a site will be ultimately reflected in a shift of the species distribution, as argued by Poff (1997) for aquatic ecosystems. IBIs assess the ecosystem condition of a test site by evaluating the intactness of its biological community in comparison to the species composition under reference conditions. If the difference is large in comparison to the intrinsic natural variability (Bailey *et al.*, 2004), it is concluded the test site is degraded. This is the essence of the Reference Condition Approach (RCA).

The outcome of an IBI is the Ecological Quality Class (EQC). In its simplest form, the classification is binary, distinguishing reference (R) and degraded (D) sites. A straightforward extension is to distinguish also sites which are moderately impacted (M) resulting in a ternary classification (R, M, D). The European Union (EU) Water Framework Directive (WFD) goes one step further and uses a five-tiered classification system ranging from 1 (high) to 5 (bad). Classes 1 (high) and 2 (good) are a further refinement of the reference class (R), class 3 (moderate) is equivalent with the moderate class before (M) and class 4 (poor) and 5 (bad) subdivide the class indicating degradation (D).

We mainly discuss multimetric indices (MMI) of biotic integrity. Metrics are well-chosen attributes of the ecological community, sensitive to human impacts on the ecosystem. By comparing the metrics to their expected value under reference conditions and taking into account the natural variability, the metrics are scored. The average (or another function) of these scores – which we call the ecological quality measure (EQM) – is a test variable quantifying the overall "distance to target" in comparison to the reference situation. From this yardstick, the EQC is derived by setting thresholds. Quite often, metrics are derived from the community data by pooling species according to their ecological strategy or another characteristic sensitive to anthropogenic stress (Jørgensen *et al.*, 2005). For instance, typical species disappear if their ecological niche is disturbed. By evaluating their (relative) abundance in comparison to the reference distribution under natural conditions, we can score how much the ecosystem is impacted by anthropogenic activities as a "distance to target". By combining several of these scores, MMIs are sensitive to a broad spectrum of human impacts and can give a global appreciation of the ecosystem.

MMIs have a long research tradition. They are mainly used in aquatic ecosystems and are less common in terrestrial ecosystems (Andreasen *et al.*, 2001). They became popular in the eighties and nineties of the previous century after a seminal paper of Karr (1981) proposing a strategy to make operational the concept of ecological integrity as launched by the Water Act, to overcome the failure of an emission-based policy to protect aquatic ecosystems (Karr and Dudley, 1981).

According to Karr (1981), the ability of an ecosystem to maintain an intact biotic community is a strong indicator for the ecological integrity of the waterbody and its capacity to provide the goods and services beneficial for the society (Millennium Ecosystems Assessment, 2005). Thus, Karr's intention was to use IBIs as a proxy for "ecosystem health" (Figure 1.1) comprising both an assessment of the intrinsic ecological values and human and societal values (Boulton, 1999). This combination has been debated strongly (Fairweather, 1999; Dufour and Piegay, 2009; Bunn *et al.*, 1999; Maddock, 1999). By now, similar to biodiversity, the concept of ecosystem health is appreciated as beneficial for nature protection as it links a technical and scientific approach with a political and societal perspective (Boulton, 1999; Meyer, 1997). In this sense, we consider an IBI as a boundary object at the interface between science and society (Turnhout, 2003; Turnhout *et al.*, 2006; Turnhout, 2009). With the introduction of the Water Framework Directive (WFD) in 2000, and its subsequent implementation in the legislation of the member states, the concept of ecosystem health and the use of IBIs has become an important standard in Europe (Hering *et al.*, 2010). Therefore, before discussing the research questions, we first give a brief presentation of this innovative directive (in spite of its shortcomings and growing pains (Nõges *et al.*, 2009a; Hering *et al.*, 2010)) as the general context of our thesis.

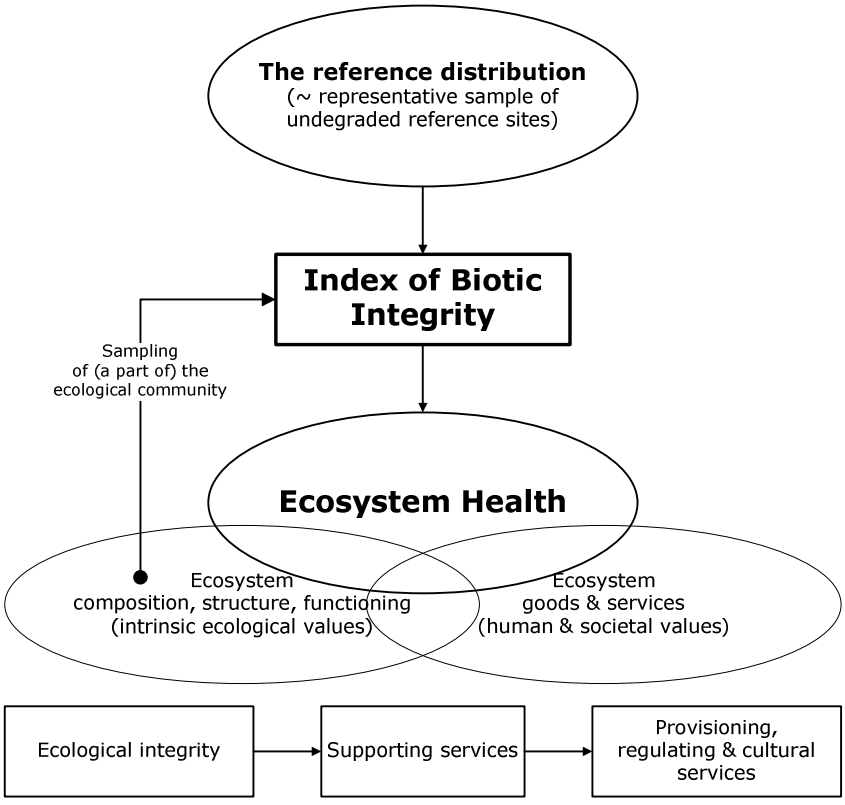


Figure 1.1 **The IBI as a proxy.** The link between ecosystem health (variable of interest) and an index of biotic integrity (test variable). Ecosystem health concept comprises both intrinsic and human and societal values of the ecosystem, while an IBI only assesses the species composition of a part of the ecological community in comparison to the species distribution under reference conditions.

1.2 The Water Framework Directive

1.2.1 The WFD goals

Voted in 2000, after a 12-year-long policy process of negotiation, the EU Water Framework Directive (WFD) has initiated a concerted action of the member states to develop a consistent policy towards (i) a more sustainable use of water resources and (ii) restoration and/or rehabilitation of impaired aquatic ecosystem (Kallis and Butler, 2001; Kaika, 2003). The WFD covers both groundwater and surface water resources. For natural surface waters, four types are distinguished: two freshwater systems (rivers and lakes), transitional waters (estuaries) and coasts. Artificial and modified waters are matched as closely as possible to the natural systems.

The overall objective is to achieve at least a "Good Ecological Status" (GES) for all waters by 2015. The GES objective corresponds with class 2 of the WFD classification comprising high (1 = blue), good (2 = green), moderate (3 = yellow), poor (4 = orange) and bad (5 = red) quality (Figure 1.2). To take into account the changed environmental boundary conditions and societal use, for "Heavily modified waterbodies" (HMWB), a different approach is followed (Borja and Elliott, 2007) and the objective is to achieve a "Good Ecological Potential" (GEP). With respect to the overall policy goal, the classification is a bi-state binary variable (Legendre and Legendre, 1998) contrasting classes 1-2 with 3-5 (Quataert *et al.*, 2007). Still, subdivision in five classes is important. For instance, according to the "stand still principle" embedded in many environmental laws (Macrory, 2004), the amelioration of the general environmental quality may not be at the expense of the best sites. In this respect, it is crucial to follow-up whether there is no degradation from class 1 to class 2. Similarly, it is important to follow-up how the distribution changes in the range from class 5 to class 3 to ascertain the trend towards the target.

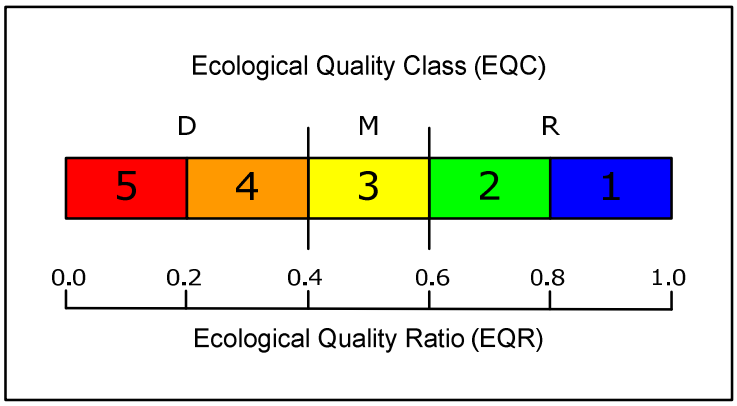


Figure 1.2 **The WFD classification.** The ordinal five-tiered classification of the Water Framework Directive with corresponding colours. Preferably, also a linear continuous measure, the ecological quality ratio (EQR), should be provided ranging from 0 to 1 with cut points as shown in the figure. In a simplified three-tiered classification, classes 1 and 2 are considered as the reference condition (R), class 3 as moderate (M) and class 4 and 5 as degraded or disturbed (D).

1.2.2 The WFD management cycle

In response to the fragmented character of water policy in Europe, the European Union (EU) set up a common management framework organised around river basins as the logical unit to enable an integrated and ecosystem based approach. To guarantee concerted action, the WFD obliges member states to adhere to a prescribed and strict time path. The first period from 2000 to 2015 is meant to set up the appropriate organisational structure including (i) the transposition of the directive in national legislation (2003), (ii) the organisation of river basin districts (2003) and their characterisation with respect to pressures, impacts and economical value (2004), (iii) the establishment of monitoring network (2006) and (iv) the development of river basin management plans (2009) to reach, by 2015, the environmental objectives unless there are "objective" reasons making this impossible, allowing to extend the period till 2027. In fact, many scientists argue that a (much) longer timeframe will be required (Hering *et al.*, 2010; Kail and Hering, 2005; Jones and Schmitz, 2009; Jeppesen *et al.*, 2005; Duarte *et al.*, 2009)). From 2015 on, there is a six-year management cycle (2021, 2027, …) to gradually refine the plans for further improving the water quality in Europe. In this cycle, the WFD has attributed a decisive role to the ecological monitoring and IBIs. By introducing clear and measurable objectives, the WFD introduces the basic principles of business process improvement (Geeraerts and Quataert, 2010), such as the PDCA (plan–do–check–act) cycle of Deming (Deming, 1982) and/or the DMAIC methodology (Snee, 2007): Define the problem and the goals, Measure key aspects of the process, Analyse the process to determine the root causes, Improve the current process, and Control the improved process.

1.2.3 The Common Implementation Strategy (CIS)

To facilitate collaboration, the Common Implementation Strategy (CIS) was installed as a forum to share general problems and challenges and to negotiate issues crossing boundaries (as rivers and other waterbodies do not stop at the frontier). The CIS develops guidance and other technical documents to assist the practical implementation of the Directive. An important issue is the intercalibration of the IBIs and other assessment methods (Van de Bund, 2008; Furse *et al.*, 2006a; Buffagni *et al.*, 2005). The WFD does not prescribe the member states which monitoring methods to use, but requires to intercalibrate in order to achieve a common standard. Therefore, regional working groups were installed to converge to a similar approach. In addition, scientific projects were set up to tackle methodological issues; the most relevant for IBIs are:

- AQEM (Development and Testing of an Integrated Assessment System for the Ecological Quality of Streams and Rivers throughout Europe using Benthic Macroinvertebrates),
- STAR (Standardisation of River Classifications),
- FAME (Development, Evaluation & Implementation of a Standardized Fish-based Assessment Method for the Ecological Status of European Rivers), and,
- REBECCA (Relationships between Ecological and Chemical Status of Surface Waters).

Many of the results of these projects are freely available on the world wide web and are reported in special issues of journals (Hering *et al.*, 2004; Furse *et al.*, 2006b; Schmutz *et al.*, 2007b; Solheim and Gulati, 2009).

1.2.4 The role of monitoring in the WFD

Monitoring has a central role in the management cycle of the WFD (Hering *et al.*, 2010). Vos *et al.* (2000) attribute two fundamental roles for monitoring: the signal and control function. Figure 1.3 adds a third and in our opinion very critical role in a policy context: the report function to communicate the results to other (international) organisations, to other (non environmental) sectors, the stakeholders and the society as a whole (Moffat *et al.*, 2008). The report arrow is two-way to promote active partnership with the stakeholders (Muñoz-Erickson *et al.*, 2007). The report function serves to account for the investments and actions of the government to the international community and the general public (Mulgan, 2000). This external focus implies the monitoring should be well standardised and transparent.

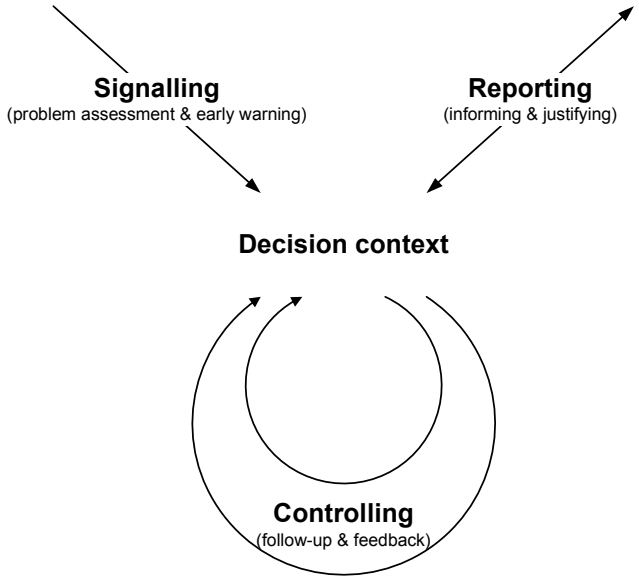


Figure 1.3 **The three main functions of monitoring in a decision context.**

The signal function is to assess the condition of the waterbodies and to detect trends timely (early warning). The control function is to follow-up the management plans and to assess whether they are effective and/or the target is reached. The double feedback arrow in Figure 1.3 refers to the different temporal, spatial and decision levels at which the follow-up of actions should be organised. A short feedback loop is necessary to follow-up in a short term whether the implementation of the decisions is on scheme, the long feedback loop to give information about the longterm success of the management actions and plans. This is especially relevant in a nature conservation context where it can take many years before the results become visible (Jeppesen *et al.*, 2005). For a large concerted action program as the WFD, it is also important to distinguish the program and project decision level (Wholey *et al.*, 2004; Wouters *et al.*, 2008a; Kaczmarek and Ottitsch, 2004; Bouckaert *et al.*, 2009; Mickwitz, 2003). The program level on a large spatial scale is necessary to enable an evaluation of all efforts as a whole and to assess whether there is an

effect beyond the advances at a project level. It requires a representative sample at a regional level. Project monitoring focuses at the success of local and/or specific plans.

1.2.5 Three types of monitoring

To cover information needs at the different decision levels, the WFD provides three complementary types of monitoring: surveillance, operational and investigative monitoring (Common Implementation Strategy (CIS), 2003). Surveillance monitoring is (i) to study long term developments in reference sites to obtain more insight in the functioning in natural conditions (Callahan, 1984; Franklin *et al.*, 1990; Burt *et al.*, 2008) and (ii) to follow-up changes resulting from widespread anthropogenic activity in contrast to the reference sites. With this dual monitoring policy, the WFD aims to judge whether the water policy as a whole achieves its goals. Every six year, in correspondence with the management cycle, new data should be available. This longterm monitoring is at a program level and requires a representative sample on a water basin or higher level to assure an unbiased picture of the condition of the waterbodies as a whole. To enable a coherent and comprehensive analysis of the data, the WFD imposes to follow-up the same quality elements everywhere (Beard *et al.*, 1999). This is not true for the two other types of monitoring which are focused on the project level.

Operational monitoring (Ferreira *et al.*, 2007) serves to assess whether waterbodies locally achieve the required quality and to evaluate success of local measures. It concentrates on waterbodies identified as being at risk or failing to meet environmental objectives. The WFD imposes a three year cycle (or higher if warranted) to serve as an early-warning system. As the monitoring is to evaluate the effect of specific and local measures, it is not necessary to measure all the prescribed quality elements of the surveillance monitoring. It is allowed to focus on a few elements relevant for the local context. Investigative monitoring (Borja *et al.*, 2008) is when the causes of a bad ecological condition are not known or there is no progress without a clear reason. This more intensive type of monitoring can involve a research project to fill in knowledge gaps. Investigative monitoring is also meant at ascertaining the magnitude and impacts of accidental pollution.

1.2.6 The biological quality elements

The ambition of the WFD makers was to go beyond the classical chemical and physical monitoring of the waterbodies and therefore they promoted ecological monitoring, i.e. monitoring capable to asses the condition of the ecosystem and to give information about the underlying causes. Basically, there are two branches of monitoring:

- "Chemical monitoring" refers to the follow-up of toxic substances in the water (including a list comprising 33 priority substances for which ecological quality standards are set to be differentiated for the waterbody type). It does not include the measurement of general chemical characteristics of the waterbodies which is part of the ecological monitoring.
- "Ecological monitoring" is composed of "biological monitoring" of the species composition of specific taxonomical groups and "associated monitoring" of ecologically relevant hydromorphological and physico-chemical variables. The latter group includes chemical

variables considered to be supportive for the interpretation of the biological results (the causes) and should help with decision making.

The biological monitoring comprises a combination of several taxa as fish, macrophytes, macro-invertebrates and phytobenthos, the so-called biological quality elements (BQE). The combination of taxa to follow-up is prescribed for each waterbody type (river, lake, transitional and coastal waters). As advocated by Hering *et al.* (2006b; 2006c), the taxa should be combined to get the maximal resolution on different temporal and spatial scales. For instance, diatoms and macro-invertebrates are more sensitive to local changes of the environment and are reactive on a short time interval. In contrast, fish and macrophytes are influenced by human impacts on a larger temporal and spatial scale. Yet, for IBIs, little progress is made with respect to the complementary value of the taxa. Ironically, all indices are developed to cover all possible pressures instead of striving to the optimal combination (Hering *et al.*, 2010). Currently, information of the different taxa is combined according to the one-out-all-out principle, meaning that the worst conditions of one of the elements determines the final outcome. Although this principle is defensible from a ecological point of view, this approach does not allow to correct for assessment errors. If one element is mistakenly scored too low, this error is propagated in the final conclusion. Many authors signal this problem (Borja *et al.*, 2004; Hering *et al.*, 2010), but as this rule is embedded in the WFD legislation itself, it is not evident to change this practice (Nõges *et al.*, 2009a).

1.2.7 The scientific challenge

The WFD text only provides narrative definitions (Nõges *et al.*, 2009b; Nõges *et al.*, 2009a) of the ecological quality classes, but leaves the ecological interpretation and technical implementation open to the scientists. The WFD covers a broad range of aquatic environments and a universally accepted ecological yardstick does not exist yet. The scientific challenge is to operationalize the rather vague class definitions, giving them an unambiguous and ecologically relevant meaning in a specific context (Borja *et al.*, 2009c). Contributing to a better index development is a central theme of this thesis.

1.3 The original research questions

This thesis originates from two major projects. The first project was the evaluation of the European Fish Index (EFI) developed by the EU funded FAME consortium (Schmutz *et al.*, 2007b; Schmutz *et al.*, 2007a). The second project was the development of fish-based indices in the Flemish context (Breine *et al.*, 2004; Breine *et al.*, 2007; Breine *et al.*, 2010). Recently, we also cooperated for an evaluation of the Flemish phytobenthos index (Denys, 2006a; Verschelde *et al.*, 2010). We briefly explain the context of each project and the associated research questions important for this thesis.

1.3.1 The EU FAME project

1.3.1.1. The European Fish Index (EFI)

Unmistakably, at its tenth anniversary, the WFD has been an important impetus for the development of IBIs (Hering *et al.*, 2010) Yet, we are far from a common generally accepted yardstick measuring the ecosystem condition. There has been a real proliferation of IBIs, most of them constrained locally instead of aiming at broader geographical scale (Borja *et al.*, 2009c). In contrast, the ambition of the FAME project was to develop a fish index (FI) at a pan-European level resulting in the European Fish Index (EFI) developed by Pont *et al.* (2006; 2007). The EFI was possible because fish catch methods are well standardised over Europe and a common database could be composed (Beier *et al.*, 2007). Another reason for the success was the availability of a strong concept on how to develop a fish index on a large geographical scale as tested in France (Oberdorff *et al.*, 2002). An update of EFI already exists (EFI+) extending the geographical range to the new Member States of the European Union (Bady and Pont, 2008; Bady *et al.*, 2009).

1.3.1.2. The diagnostic accuracy of ecological indicators

A work package of FAME was to evaluate EFI and to compare it with existing national or regional fish indices (Quataert *et al.*, 2004). In 2003, very little information was available about the evaluation of IBIs (and ecological indicators in general). Most IBI developers report some form of validation, but without a coherent framework (Roset *et al.*, 2007). To our knowledge, only quite recent papers tackle the issue fundamentally (Hale *et al.*, 2007; Quataert *et al.*, 2007; Hale and Heltshe, 2008; Dos Santos *et al.*, 2011). Yet, already fifteen year ago, Murtaugh (1996) proposed a method to evaluate ecological indicators. His paper is cited in a conceptual paper of Vos *et al.* (2000) discussing how to match environmental monitoring programs better to the needs of policy makers, but the method is seldom or not applied in practice. Interestingly, in another research domain (diagnostic medicine), Pepe (2003) reports in a recent handbook that methods to evaluate medical tests are not part of a statistical training.

1.3.1.3. The total misclassification fraction

The total misclassification fraction depends on the prevalence of the event we aim to detect (Zhou *et al.*, 2002). Thus, the same diagnostic test applied in two regions with a different proportion of degraded sites will result in a different overall misclassification rate. When applying the index in a pristine site, the index should score high (assuming there is a positive correlation between the index and the ecological condition). If not, we have a false alarm or a false positive (FP). The proportion of false positives is called the false positive fraction (FPF). Conversely, in a degraded situation, the index should score low. If not, we have a false reassurance or a false negative (FN). The proportion of false negatives is called the false negative fraction (FNF). If the prevalence of degradation in a region is π^+, then the total misclassification fraction (TMF) is:

$$TMF = \pi^+ FNF + (1-\pi^+)FPF$$

This formula demonstrates the overall misclassification depends on the prevalence of the problem, hampering a comparison between regions and countries. For the FAME project, acknowledging that

diagnostic accuracy is a two-dimensional variable, gave a clue to interpret the bewildering variety of the performance of the existing national or regional indices. Details are given in Chapter 6 (Quataert *et al.*, 2007), but the key finding was that some indices optimised FNF at the expense of FPF and for other indices it was the other way round. Only a few indices controlled both errors.

1.3.1.4. Decision analysis and the usefulness of indices

Moreover, and perhaps more important, the distinction between FPs and FNs is essential in a decision context as the consequences of both errors are different (Quataert *et al.*, 2007). In a restoration project, with a FP, we decide wrongly to restore a pristine site. With a high FPF, resources are spoilt and restoration of pristine sites can be harmful. Conversely, with a FN, a degraded site is not detected and the possible ecological benefit by restoring the site is not realised. When applying an index, the policy maker or field manager should have information of both errors separately to make a cost balance. In fact we should optimise following cost function.

$$C_D = \pi^+(C_{FN}FNF + C_{TP}TPF) + (1-\pi^+)(C_{FP}FPF + C_{TN}TNF)$$

This equation is explored in full depth in Chapter 4. Here we sketch the main idea. C_D represents the (expected) average decision cost if the decisions are based on an index. TP and TN are true positives and true negatives (correct classifications), C_{FN}, C_{TP}, … represent the costs associated with false negatives (FN), true positives (TP), … Benefits are modelled as negative costs. The hard point is not the mathematical equation but the correct assessment of the costs.

1.3.1.5. The cost of monitoring

An application is that we can link the quality of the index with the decision cost. Increasing the quality of the index implies a higher assessment cost (C_A), but a lower decision cost (C_D). Hence, there is a tradeoff between the benefit and the cost of the index and we should minimise the entire cost (C_E) which is the sum of the two:

$$C_E = C_A + C_D$$

The observation that we should distinguish FPs and FNs opened a window on a new research domain enabling to quantify the contribution of ecological indicators to decision making in terms of costs. Quite often, ecological indicators (and, by extension, ecological monitoring programs collecting the data) are questioned about their value in and contribution to decision making and it is asked to motivate the costs. This is more easily said than done as it is not easy to quantify the benefits, while the costs are apparent (Caughlan and Oakley, 2001).

1.3.2 The Flemish indices of biotic integrity

1.3.2.1. The data flow and formalisation of MMIs

The fish-based IBI for upstream brooks comprises the grayling and trout zone (Breine *et al.*, 2004) and completed previous work on the bream and barbel zone (Belpaire *et al.*, 2000). An important challenge at the time was to obtain more insight in "data flow" of an IBI. At first sight, the calculation of an IBI is rather an involved procedure with many rules and data transformation steps

requiring a lot of data and additional information about species and environmental characteristics. Although a series of guideline papers appeared to standardise and consolidate the construction of biotic indices (Karr and Chu, 1997; Hughes *et al.*, 1998; Bailey *et al.*, 2004; Hering *et al.*, 2006a; Roset *et al.*, 2007; Southerland *et al.*, 2007; Stoddard *et al.*, 2008), to our knowledge none of these and other papers, explicitly formulates the statistical model. They just provided a narrative description of all steps. The challenge at the time was to structure the calculation steps of the index to make the data and knowledge transparent required to calibrate and validate an index (Breine *et al.*, 2004). Later, this exercise made clear that many IBIs are essentially an average of scores, each score being a "distance to target" measure reflecting how much a well selected attribute of the sampled community, called metric, deviates from a reference situation. This method is known in the literature (Bailey *et al.*, 2004) as the reference condition approach (RCA).

1.3.2.2. Selection of the optimal basket of metrics from a set of candidates

As mentioned above, an IBI is an average (or another function) of scored attributes or metrics derived from the community data. The bottleneck when constructing an index, is to find out which combination of metrics is most appropriate. In many situations, this is not clear in advance and, therefore, it is common practice to propose a series of candidate metrics from which the optimal set should be selected. In many papers, the selection procedure is rather "ad hoc" and highly subjective (Roset *et al.*, 2007). The optimisation procedure is not transparent and clear criteria are lacking. To develop a fish index for the Zeeschelde Estuary, the idea was to use a procedure similar to stepwise regression to select the best metric basket from the candidate set, minimising the misclassification error (FPs and FNs) as the optimisation criterion (Breine *et al.*, 2007). From this idea, it was a small step to realise that an IBI is in fact a logistic regression model. For this reason, regression building techniques as best subset regression can be used to select to optimal combination of metrics. This link of index construction and model building is well known in diagnostic medicine (Pepe, 1997; 1998; Pepe and Thompson, 2000).

1.4 The objectives and outline of the thesis

1.4.1 Objectives

As explained in the previous section, this thesis originates from two key problems: (i) the evaluation of the European Fish Index (EFI) and how to compare it with existing national or regional fish indices (Quataert *et al.*, 2007) and (ii) the selection of the optimal basket of metrics to include for a fish-based estuarine biotic index (EBI) for the Zeeschelde Estuary (Breine *et al.*, 2007). This thesis builds further on these issues. Chapter 6 presents the results for FAME and Chapter 5 recalibrates the EBI. The key objective is to develop a coherent framework to construct (track 1 in Table 1.1) and evaluate (track 2) IBIs and ecological indicators in general. Chapter 2 translates the ecological concept in a statistical model and Chapter 3 presents a statistical

framework to evaluate ecological indicators. Chapter 4 is an intermezzo exploring the cost benefit tradeoff of ecological indicators giving an onset for a better motivation of monitoring costs.

The five main objectives are (the numbers within brackets refer to the chapters):

(i) to translate the multimetric index concept in a statistical framework;

(ii) to explore the possibilities of ROC curves to evaluate ecological indicators;

(iii) to quantify the benefits of ecological indicators in decision making;

(iv) to develop a strategy to select the optimal basket of metrics for an MMI;

(v) to develop of a strategy to evaluate IBIs.

In the background, three more fundamental questions motivate this thesis. A first intriguing question was to find out how IBIs can cover a broad spectrum of anthropogenic pressures. Besides an intelligent choice of the metrics and their scoring, representative sampling at the calibration stage is an important factor. Yet, because of budget constraints, IBIs are mostly calibrated with existing datasets. A second driving question was to obtain a better understanding why the ROC concept is so pivotal. For this objective, we introduce utility curves quantifying the usefulness in terms of costs and/or benefits. A third impetus was to make more transparent how ecological indicators can improve decision making and fulfil their role as boundary objects at the interface between science, policy making and society.

1.4.2 Outline

In Chapter 2, we explain what is effectively measured by the test variable of an IBI and why this test variable is possibly a good proxy for the ecological integrity of a site. Then, we give insight how the RCA philosophy allows to score attributes of the ecosystem in comparison to a reference situation. We derive a generic four-step calculation scheme transforming the community data in an overall appreciation of the ecological condition, the ecological quality class (EQM). We make clear this IBI model is in fact a categorical regression model.

In Chapter 3, we make a synthesis of the theory of ROC curves. A specific purpose was to understand better why ROC curves are so crucial to evaluate the strength of indicators. Therefore, we introduced utility curves linking the usefulness of an index with its strength as characterised by the ROC curve.

Chapter 4 is an intermezzo in line with the utility curves of Chapter II. Very often the value of ecological indicators is challenged. Although they are cost-effective in comparison to gold standard measures, they are altogether not that cheap. Therefore, based on a cost-effectiveness analysis (CEA) and a cost benefit analysis (CBA), we attempt to obtain more insight in the benefits of an analysis. By demonstrating the benefits, the choice of the optimal indicator can be better motivated.

Chapter 5 represents the first application. We recalibrate a fish-based estuarine biotic index (EBI) for the Zeeschelde estuary. We demonstrate how we can use statistical model building techniques to find the optimal combination of metrics to compose an IBI. Chapter 6 gives the second

application. We evaluate the European Fish Index (EFI) as developed by the FAME project based on the error curve, an equivalent concept as the ROC curve.

Table 1.1 **Structure of the thesis.**

	Track 1 **How to construct an IBI?**	**Track 2** **How to evaluate an IBI?**
	Chapter 1 General introduction	
Theoretical foundation of the concept / tool	**Chapter 2** Understanding the ecological rationale of indices of biotic integrity and the link with statistical modelling	**Chapter 3** Understanding diagnostic accuracy of ecological indicators and the link with decision making
Application to ecological indicators	**Chapter 4** How to make the tradeoff between diagnostic accuracy and costs of the index? How to determine the optimal budget to be allocated to monitoring?	
Application to (fish-based) IBIs	**Chapter 5** Construction of a fish-based estuarine index for the Zeeschelde Estuary (EBI)	**Chapter 6** Evaluation of the fish-based European Fish Index (EFI) of the EU FAME project
	Chapter 7 General discussion and conclusions	

2 The Reference Condition Approach (RCA). A one-line format for multimetric indices of biotic integrity merging the ecological and statistical rationale

In this chapter, we further elaborate the four-step calculation scheme of an IBI as we proposed for the construction of a fish-based index of biotic integrity for the upstream brooks (grayling and trout zone) in Flanders (Breine et al., 2004).

Abstract

In this chapter, we translate the ecological rationale underlying an index of biotic integrity (IBI) in a statistical framework. An IBI is a bioassessment tool to give an integrative measure of the overall ecosystem condition. The outcome is the ecological quality class (EQC) expressing how much the ecosystem is impacted by anthropogenic activities in comparison to a reference condition. Ideally, this reference condition refers to a (nearly) pristine situation, but it may be also the least impacted sites or any other baseline condition in a certain region. In its simplest form, EQC discriminates between reference (R) and degraded (D) sites, but it is possible to adapt the classification to a finer gradation with intermediate states. For instance, the Water Framework Directive (WFD) requires to make a distinction between five quality classes, ranging from 1 (high quality) to 5 (bad quality).

Several methodological papers describe step by step how to construct IBIs based on the reference condition approach (RCA). However, as far as we know, none of these papers make the underlying statistical model explicit. They just give a narrative description of the calculation steps. Yet, we were able to derive a very simple but flexible one-line format, modelling the four transformation steps of an IBI. (i) The model equations start with deriving, from the community data augmented with species properties, metrics reflecting ecologically relevant features of the species distribution that are sensitive to anthropogenic alterations of the environment. (ii) Subsequently, the metrics are scored, expressing how (dis)similar the metric observations are in comparison to type-specific or site-specific reference conditions by taking into account the site typology or by correcting for differences in environmental conditions. (iii) In the third step, the individual scores are aggregated (traditionally by simply summing or averaging) into the ecological quality measure (EQM), meant as an overall impact assessment. As such, EQM is hard to interpret because its value does not tell directly how impaired the ecosystem is and whether restoration is necessary. (iv) Hence, the fourth and final step compares EQM with decision thresholds resulting in the ecological quality class (EQC), an ordinal class variable appreciating the degree of ecosystem degradation (or, expressed positively, the level of biotic integrity).

This simple one-line format is very flexible and hides a lot of complexity. Each of the four steps can be described by one or more mathematical functions. At the calibration stage, the appropriate functions should be derived and the unknown)parameters estimated by matching EQC as closely as possible with the human quality class (HQC), an independent gold standard assessment of the true state of the ecosystem. This mathematical description links the construction of IBIs with statistical model building. For example, traditionally, most IBIs are simply an average of scored metrics. We call this the average score model (AVG). We demonstrate that AVG is closely linked to an ordinal logistic regression (OLR) model, but with fixed regression parameters. In this model, EQC is the ordinal response variable and the metrics are the explanatory variables or predictors. This link with regression models embeds the construction of IBIs in the framework of statistical model building and we can use regression techniques to fit an IBI model.

Keywords

multimetric index (MMI), index of biotic integrity (IBI), bioassessment, ecological model, biological indicators, ecological community, the reference condition approach (RCA), ecological indicators

2.1 Introduction

In this chapter, we translate the ecological rationale underlying an index of biotic integrity (IBI) in a statistical framework. It was Karr's idea (1981; 1986) to use multimetric IBIs (MMI) judging the intactness of the species composition in comparison to a reference as an overall indicator for ecological integrity. Changes in the environmental or ecological conditions by anthropogenic activities are ultimately reflected in shifts of the species composition. By making an analysis at the community level and considering attributes of the ecological community, we can circumvent the necessity for detailed information about individual species (Attrill, 2002). Karr's proposal on how to compose and construct MMIs inspired a rich research tradition of researchers (Laudan, 1977) working at the interface between science and policy making.

These boundary workers (Turnhout *et al.*, 2006) gradually refined and extended the original concept to a broad range of situations (Hughes and Oberdorff, 1999). Theoretical work (Attrill and Depledge, 1997; Yoccoz *et al.*, 2001; Baattrup-Pedersen *et al.*, 2008) improved the conceptual and ecological foundations. A series of guideline papers standardised and consolidated the construction of biotic indices (Karr and Chu, 1997; Hughes *et al.*, 1998; Bailey *et al.*, 2004; Hering *et al.*, 2006a; Roset *et al.*, 2007; Southerland *et al.*, 2007; Stoddard *et al.*, 2008). However, to our knowledge, none of these papers make the statistical model explicit. They give a narrative description of all steps without specifying the corresponding equations. Yet, a very simple but flexible one-line equation is possible, linking IBIs with regression models. This link allows to embed the construction of IBIs in the framework of statistical model building. Construction of an index is not any more an isolated technique with its own rules, but can be motivated from a broader perspective.

In this section, we assume a well standardised sample of the ecological community is available, we call the community data C. This sample should not cover the complete ecological community which is impossible anyhow. It is sufficient that an ecologically relevant assemblage is sampled covering a broad range of ecosystem functions. For instance, in aquatic ecosystems, common choices of assemblages are fish, macroinvertebrates, macrophytes and/or phytobenthos. Although our examples come mainly from an aquatic environment, the results presented also hold for terrestrial ecosystems.

2.2 The one-line format

Although there exist several related variants, many multimetric indices (MMI) of biotic integrity essentially have following four-step model format (see Table 2.1 for definitions):

$$C \overset{(i)}{\rightarrow} M_j \overset{(ii)}{\rightarrow} S_j \overset{(iii)}{\rightarrow} EQM \overset{(iv)}{\rightarrow} \boxed{EQC \overset{(v/vi)}{\leftrightarrow} HIC}$$

The scheme consists of four transformation steps (i) – (iv) transforming the community data C into the ecological condition class (EQC). For the derivation of an index, we distinguish two stages. The first stage, comprising steps (i) & (ii), is aimed at the calculation and scoring of subindices, the so-

called metrics, well-chosen attributes of the community data sensitive to anthropogenic stress. The second stage, steps (iii) & (iv), combines the scored metrics in one overall assessment.

This index model is empirically calibrated (v) and validated (vi) with respect to a "gold standard", we call the human impact class (HIC) which is an alternative assessment of the ecosystem status, independent of the ecological community to avoid circularity (Stoddard *et al.*, 2006). HIC is also called the preclassification as it classifies the sites a priori to tune the index model. The calibration process can be subdivided in two parts. First, a list of candidate metrics is tested for their response to anthropogenic stress to filter the most appropriate ones. Next, in a second step, from this core set the best combination is sought for to reproduce as closely as possible the preclassification. Especially, the second part of the calibration (the choice of an optimal combination of metrics and the required number) remains a hard problem to tackle (Roset *et al.*, 2007). We return to this question in Chapter 5.

Table 2.1 **The one-line format.** The transformation steps (i – iv) and calibration and calibration (v/vi).

	Transformation	Variables
Community data (C) (sample of the ecological community) = original data		
(i) Metrics	**Extract** the ecological information from the community data augmented with species information (guilds, traits, sensitivity to disturbance, …)	► **metrics** (M_j) = attributes of the ecosystem (composition, structure or function) sensitive to human alteration of the environment.
(ii) Scores	**Evaluate** the metric values in comparison to the type-specific and/or site-specific reference distribution (if necessary, also correct for sampling method and conditions)	► **scores** (S_j) = standardised "distance to targets" measures in comparison to the reference condition.
Scores (S) (distances to target) = explanatory variables of the index model		
(iii) EQM	**Combine** the scores into a single test variable (e.g. a weighted average) assessing the global impact of anthropogenic stress	► **ecological quality measure** (EQM) = test variable of the index, the ecological yardstick
(iv) EQC	**Interpret** EQM in relation to a binary or ordinal classification system based on preset classification or decision thresholds	► **ecological quality class** (EQC) = final appreciation reflecting the human impact on the ecosystem
Index of Biotic Integrity (IBI) = response variable of the index model		
(v/vi) HIC	**Inventory** of anthropogenic use and pressures to preclassify the calibration sites	► **human impact class** (HIC) = independent assessment of the human impact on the ecosystem (gold standard)

2.2.1 Calculating and scoring the metrics

Together, steps (i) and (ii) define a series of ecological indicators at the community level (Attrill, 2002) sensitive to anthropogenic alterations of the ecosystem.

(i) The model equations start with extracting ecological indicators called metrics M_j from the community data. The metrics express key attributes or features of the ecological community, selected to be sensitive to human impacts or anthropogenic alterations of the environment. For instance, a possible metric is the abundance of typical species which are expected to decrease under human pressure. To calculate the metrics, the community data is "augmented" with knowledge about the ecological strategy or other relevant ecological properties of the species and the characteristics of the site (Jørgensen *et al.*, 2005). In above example, to calculate the abundance of the typical species from the community data, the type of the site and its typical species should be known.

(ii) Subsequently, the metrics are scored and standardised to make the different metrics unitless and more comparable to each other, facilitating interpretation. This is a very crucial step as explained in a separate section of this chapter. A score S_j is a "distance to target" measure expressing the (dis)similarity of a metric in comparison to its expected value *under reference conditions* in *similar type- or site-specific conditions* (matching), *taking into account the natural variability*. None of the phrases in italics is easily done, but all are essential elements enhancing the discriminatory power of the metrics. For instance, a simple scoring method is to calculate the standardised residuals (z-scores) of the observed values in comparison to a statistical model predicting the reference value of a metric from the environmental characteristics of the site (e.g. river width and stream flow) and the sampling conditions (season, sampling method of the community, …). This matching of the test site to similar reference sites of an equal type and/or of similar environmental conditions considerably reduces the environmental variability (Oberdorff *et al.*, 2001; Oberdorff *et al.*, 2002; Pont *et al.*, 2006; Pont *et al.*, 2007).

2.2.2 Integrating the scored metrics in a single measure

The next two steps are about how to combine the separate test variables in one single measure or test variable making an overall assessment of the ecosystem.

(iii) In a third step, the individual scores are summarized into one single continuous test variable. We call this ecological yardstick the ecological quality measure (EQM). EQM can be any summary function of the scored metrics. In the IBI tradition (Karr, 1981), EQM is often simply an average (or sum) of scores, the average score model (AVG). The EQM values rank the sites according to the human impact and can be considered as a global distance to target, but as such, the numerical value has no clear ecological meaning and, for instance, does not tell how much the ecosystem is impaired and/or whether restoration is necessary.

(iv) Hence, the fourth and final step compares EQM with decision thresholds or cut-off points resulting in the ecological quality class (EQC), an ordinal class variable expressing the degree of degradation (or, expressed positively, the level of biotic integrity). For instance, the WFD distinguishes five quality classes ranging from 1

(high) to 5 (bad). Although the classification is meant to facilitate communication and decision making, the meaning of the classes is not always very clear nor homogeneous. They should be tuned with respect to an ecologically relevant interpretation model appreciating the level of degradation (Bailey *et al.*, 2004; Davies and Jackson, 2006; Fellows *et al.*, 2006; Southerland *et al.*, 2007). A possible interpretation is to link EQC with the effort necessary to restore/rehabilitate the site or the total anthropogenic pressure on a site. This measure we call the human impact class (HIC). At the calibration stage, an index is tuned such that EQC matches HIC as closely as possible.

2.2.3 Construction (validation and calibration) of the index

The construction of the index comprises a calibration and a validation step. Among other information, it is essential to have a consistent preclassification system assessing the human impact independently from the community data used for the index. The index model is calibrated and validated with respect to this human impact classification (HIC) which is the response variable of the model. The scored metrics are the explanatory variables.

(v) Above model calculations involve several variables and parameters unknown at the onset. When calibrating the model, we should search for an optimal functional form of the model, select the appropriate metrics from a set of candidates and estimate the unknown parameters in the model. One of the main problems is to identify the best combination of metrics. Currently a good strategy is lacking to select the metrics on objective grounds (Roset *et al.*, 2007). In Chapter 5, we present a more systematic approach building further on the stepwise regression approach of Breine *et al.* (2007).

(vi) Internal validation based on the same dataset as used to construct the index cannot correct for study flaws and risks to underestimate the true misclassification fraction (Hosmer and Lemeshow, 2000). External validation is necessary to determine whether the IBI is responsive to anthropogenic alterations of the environment. Or, it can be interesting to test the index to other definitions of the human impact or for specific or new emerging problems. In addition, index calibration is a cross-sectional observational study and not an experiment (Cochran, 1983; Rosenbaum, 2002). Therefore, it should be investigated whether an IBI will be appropriate to follow-up restoration projects (Hering *et al.*, 2010; Yallop *et al.*, 2009; Johnson and Hering, 2009; Kelly *et al.*, 2009).

To calibrate and validate the index model, it is necessary to have a dataset of sites for which the true ecological condition is (perfectly) known. As already stated, this preclassification or a priori ranking should be independent of the species composition of the sites to avoid circular reasoning (Stoddard *et al.*, 2006). If preconceived notions about the range of biotic assemblages at a "typical" reference site interfere, we risk to underestimate the natural variability. There is however no gold standard yardstick to measure ecological integrity directly. An often used solution is to make an inventory of the anthropogenic activities and pressures to rank the sites of the calibration dataset with respect to a human impact gradient. For instance, Aubry and Elliott (2006) standardised a scheme to score the anthropogenic activities and pressures in estuaries based on the DPSIR-framework (Elliott, 2002; Borja *et al.*, 2006; Svarstad *et al.*, 2008), which is developed

to analyse systematically the impact of human activities in a causal scheme (D = driving forces → P = pressures → S = state → I = impact). In this approach, reference sites are operationalized as sites with no or nearly no measurable pressures. The reference sites should be carefully selected as they are the anchor of the index calibration (Clarke *et al.*, 2003). A good example is RIVPACS (River InVertebrate Prediction And Classification System) developed in the UK to assess the ecosystem condition based on macroinvertebrates (Wright *et al.*, 2000). The method was tuned over a ten year long period (it started in October 1977) and considerable effort was spent on the selection of reference sites (Wright, 2000).

2.2.4 An example based on the ecological guild concept

In the subsequent sections, we work out the scheme mentioned above in more detail. We first present an example based on ecological guilds. An important guiding principle to find appropriate candidate metrics responsive to anthropogenic stress is the functional guild concept. Functional guilds classify species with respect to their ecological resources and/or environmental conditions required to survive and reproduce (Wilson, 1999). For instance, the feeding guild classifies species in omnivores, piscivores, detritivores, and so on. Comparison of the proportions in each feeding guild class with its expected proportion reference sites gives an indication of the human impact on the ecosystem. Figure 2.1 presents the principle graphically in an aquatic environment. Typically, under the influence of anthropogenic disturbance, the proportion of generalists such as omnivores and detritivores tends to increase, and specialist species with a narrow ecological amplitude such as piscivores tend to decrease. Based on this simple division, we propose a chi-square type statistic EQM to test whether the environment corresponds to a reference (R) or degraded (D) situation by comparing the observed proportions (M_j) with their expected values (E_j):

$$M_j = \sum_{i=1}^{I} p_i \delta_{ij} \rightarrow S_j = \frac{\left(M_j - E_j\right)^2}{E_j(1-E_j)} \rightarrow EQM = -\sum_{j=1}^{J} S_j \rightarrow EQC = \begin{cases} D & if \quad EQM \leq T \\ R & if \quad EQM > T \end{cases}$$

In the equation, we sum over the species i. The δ_{ij} are binary coefficients (1/0) specifying whether species i belongs to group j of the feeding guild. By summing the relative abundance p_i multiplied with δ_{ij}, we obtain the relative proportions (which are the metrics M_j) of omnivores, piscivores, detritivores, … in the sample. Comparison with their expected value under reference values E_j taking into account the natural variability (assuming a binormal distribution) results in the scores S_j. Summing over these scores gives EQM. We use a minus sign to guarantee that EQM is positively associated with the ecological quality. This is not necessary, but it is a general assumption in this thesis, facilitating notation and explanation. As small values are indicative for a low quality, a logical decision rule is to classify a site as degraded if its score is smaller than a certain threshold. As we explain later, to determine the threshold T, it is not necessary to hypothesize a chi-square distribution as the decision thresholds are determine empirically.

Another possibility is to use the Shannon's entropy to combine the metrics in one single measure and to compare it with the reference value as follows:

$$EQM = \sum_{j=1}^{J} M_j \log(M_j) - \sum_{j=1}^{J} E_j \log(E_j)$$

Trousselier and Legendre (1981) used a similar measure as an index of functional evenness for studying bacterial assemblages. In such assemblages, the species level is often poorly defined. The index bypasses the step of species identification, using the proportions M_j of positive responses to a microbiological test j (Legendre and Legendre, 1998). This example is interesting as it illustrates the fundamental idea of how metrics are chosen to reflect key functional properties of an ecological assemblage. Combining this information results in an indicator responsive to ecological change.

To conclude, we briefly mention that we can combine the EQM_k for several functional guilds k as defined above (e.g. reproductive guild, habitat guild, …), by simply adding them together (or any other function):

$$EQM = \sum_{k=1}^{K} EQM_k$$

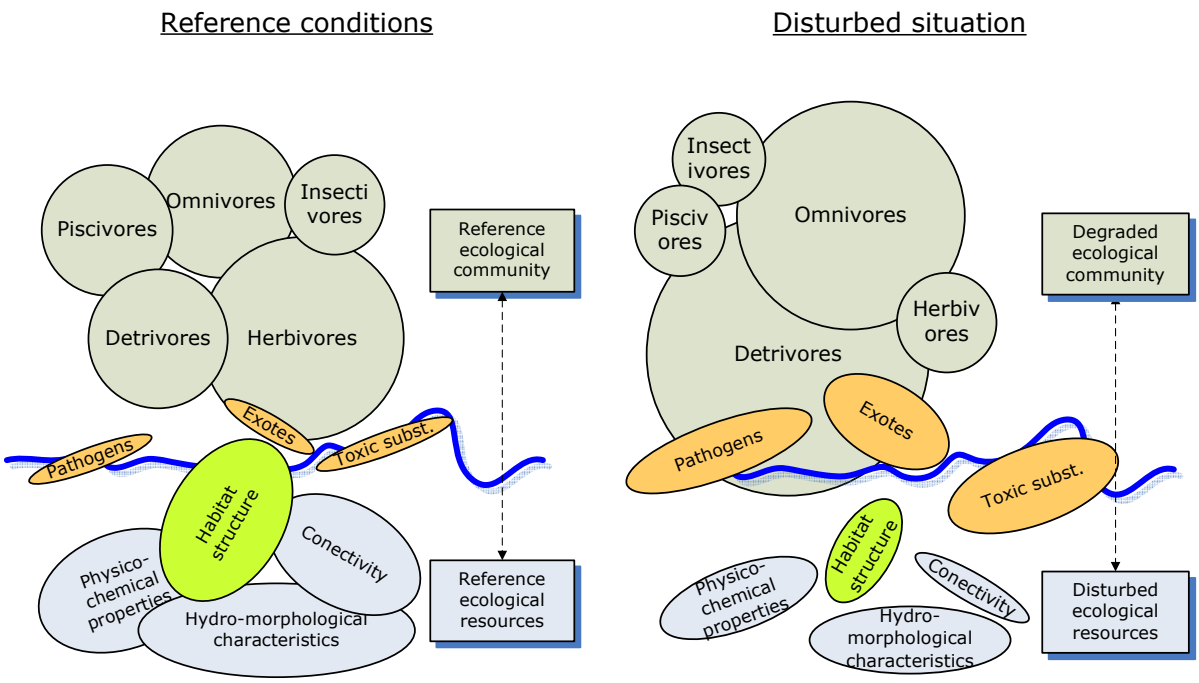


Figure 2.1 **The underlying ecological rationale of a multimetric index of biotic integrity.** The circles (grey) represent how the proportion of the feeding guilds in the community data change (from left to right) because of anthropogenic disturbance of the environmental and ecological resources (the change of the blue and green ellipses) and introduction of harmful organisms or substances (orange). By comparing the ecological community of a test site (right) with its reference (left), it can be derived whether the ecosystem is disturbed. The scheme does not represent the natural variability of the reference condition which has to be taken into account to assess whether the shift of the ecological community significantly differs from the reference situation.

2.3 The ecological quality class

In this section, we provide a more precise meaning of EQC. As an example, we take the five-tiered WFD classification. The WFD text only gives a narrative definition of the EQCs, but leaves the precise ecological interpretation and technical implementation open to scientists. This is the only feasible approach as the directive covers a broad range of aquatic environments and no universal ecological yardstick exists. The scientific challenge is to operationalize the rather vague class definitions in a consistent way (Borja *et al.*, 2009c).

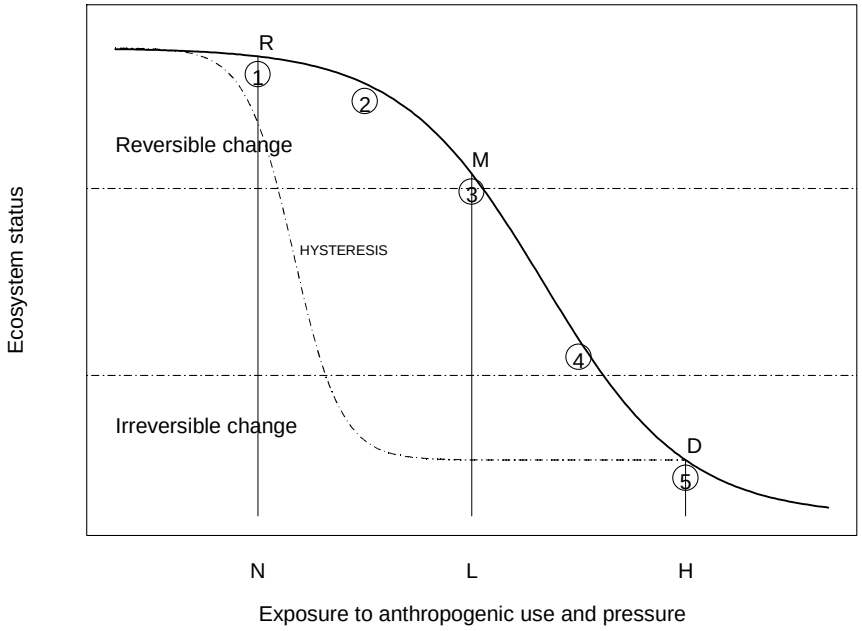


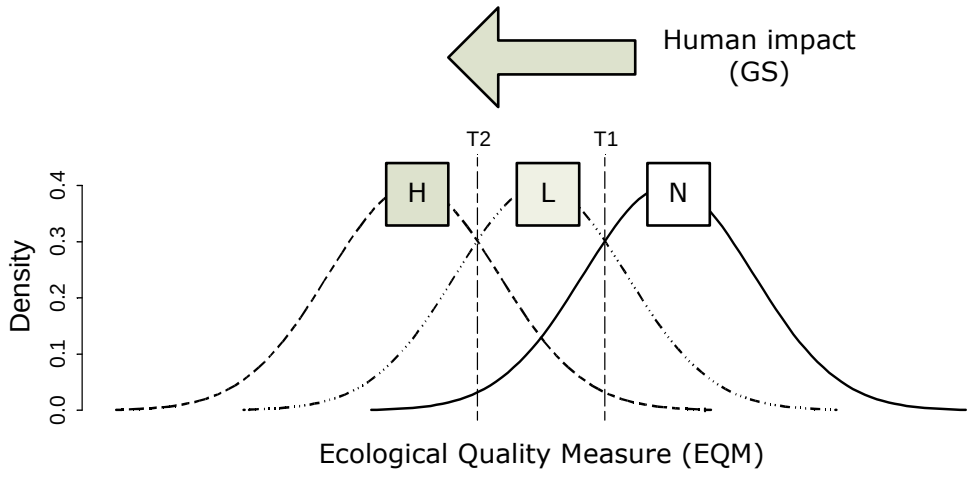
Figure 2.2 **Link of the hypothesized effect of exposure to anthropogenic use and pressure on the ecosystem with the five-tiered WFD classification** (numbers in circles). Exposure to anthropogenic stress: N = (nearly) no, L = low, H = high exposure. Ecosystem status: R = reference, M = moderate, D = degraded. The solid line shows how the ecosystem condition decreases with anthropogenic stress (Cairns *et al.*, 1993; Davies and Jackson, 2006). The dashed line symbolises the hysteresis because of irreversible changes of the ecosystem: eliminating exposure does not immediately result in a recovery of the ecosystem and requires a long time unless specific rehabilitation measures are taken (Brown and Ulgiati, 2005).

2.3.1 A conceptual scheme to interpret the WFD narrative definitions

Figure 2.2 connects the narrative definitions with increasing anthropogenic pressure and associates them with fundamental changes of the ecosystem. The scheme shows how anthropogenic use and pressures deteriorate the ecosystem by surpassing critical pressure thresholds. Level 1 and 2 correspond to (close to) pristine reference conditions (R). At level 2, human impact is present but remains minimal. Level 3 is equivalent to a moderate but definite change (M). There is clearly an

impairment of the ecosystem integrity (function, composition, structure) and/or a reduced delivery of the ecosystem services and goods (societal values). Level 4 and 5 correspond to a degraded (D) ecosystem condition.

The general idea behind this scheme (Cairns *et al.*, 1993; Rapport and Whitford, 1999; Davies and Jackson, 2006) is that up to low exposure (level 2), human impact on the ecosystem is reversible and the ecosystem is not fundamentally altered. If the exposure is diminished, the ecosystem status returns to the original status in a reasonably short time scale without additional restoration measures. From moderate impairment on (level 3), changes become more and more fundamental, precluding a (fast) return to the original status if pressure is diminished (Reynolds, 2002; Rapport and Whitford, 1999). To restore or rehabilitate the ecosystem, more and more, active restoration is necessary in complement to the reduction of anthropogenic exposure to overcome hysteresis (Brown and Ulgiati, 2005), i.e. delayed recovery of the ecosystem when the cause of degradation is removed (Reynolds, 2002).



$$EQC = \begin{cases} R & if \quad T_1 < EQM \\ M & if \quad T_2 < EQM \leq T_1 \\ D & if \quad EQM \leq T_2 \end{cases}$$

Figure 2.3 **Hypothetical response of the ecological quality measure (EQM) to an increasing human impact (HIC) and categorisation to the ecological quality class (EQC).** HIC is measured by a gold standard (GS) tri-state assessment: N = (nearly) no impact, L = low impact, H = high impact. The decision rule allows to distinguish between three different states: R = reference, M = moderate, D = degraded based on the decision thresholds T_1 and T_2. Because of natural variability and model imperfections, the densities of EQM overlap resulting in false positive and false negative errors (for a further elaboration, see Figure 2.5 and Figure 2.6).

2.3.2 Intactness of the community as a yardstick to measure ecosystem health

The issue is to operationalize the ecosystem status represented conceptually in Figure 2.2 by developing a measurable yardstick quantifying the ecosystem degradation. Figure 2.3 is a realistic model describing the response of an ecological quality measure (EQM) to increasing human impact. Because of the natural variability, the distributions overlap and no perfect distinction is possible. The challenge of the index calibration is to make the overlap of the distribution as small as possible and to make an optimal choice of the thresholds in relation to the human impact class (HIC) to interpret EQR in term of the ecological quality class (EQC) distinguishing between different states of the ecosystem. Quantifying changes in the ecological community in comparison to a reference situation provides a powerful tool to measure ecosystem degradation.

Exposure to human activities alters the environmental conditions and/or ecological resources which ultimately affect the community structure through changes in the food web altering competition, predation, etc. Because of this causal chain, changes in the community composition reflect human impact on the ecosystem (Poff, 1997; Attrill, 2002). In this respect, the functional guild concept is a powerful tool. Guilds group species according to their role in the ecosystem and/or their required environmental conditions (Wilson, 1999); e.g. the feeding guild classifies species according to their diet. If the environmental conditions change, the distribution over the feeding group will change too. By observing how the distribution over the guilds changes in comparison to a reference situation, we can track functional changes at the community level (Attrill and Depledge, 1997). This reasoning offers a rationale to measure quantitatively the ecosystem quality. An ecological yardstick can be composed by sampling the ecological community and evaluating the distribution of the functional groups in comparison to the reference.

2.3.3 The RCA concept

The Reference Condition Approach (RCA) is a general method to score biological communities for bioassessment purposes in comparison to reference sites. Ideally, the reference sites represent a pristine situation, but for the method this is not a prerequisite. It can be any baseline to compare with on the condition that it is well defined. Sometimes the ideal situation does not exist any more and the reference should be derived from the best available least impact sites. In the following, to simplify the discussion, we assume that the reference sites are pristine unless it is relevant to say otherwise.

The RCA concept compares species composition at a test site with its reference or null distribution analogous to statistical testing (Bailey *et al.*, 2004) or diagnostic testing in medicine (Quataert *et al.*, 2007) as expressed by the following symbolic and very generic basic RCA equation:

$$EQM = -\Delta(C, \Phi_C(X \mid R)) \quad \& \quad EQC = \begin{cases} D & if \quad EQM \le T \\ R & if \quad EQM > T \end{cases} \qquad (the\ RCA\ equation)$$

In this equation, C represents the observed community collected according to a standardised and documented procedure in the same way as the reference sites are sampled. $\Phi_C(X|R)$ is a statistical multivariate distribution which models the expected species composition of a site as predicted from the site characteristics X under reference conditions (R). The species composition is not fixed but

varies also under reference conditions as characterised by the multivariate distribution. X stands for the type and/or environmental factors of the site. Also, X can contain information to correct for the sampling conditions, for instance, the season and the sampling method to collect the species community. The function Δ is a well-chosen distance (or dissimilarity) measure calculating the "distance to target" from the reference distribution. The negative sign is to make EQM positively associated with the ecological quality as we generally assume here. Hence, a logical decision rule is to conclude that a site is degraded (D), if EQM is beyond a certain decision threshold T; otherwise there is no evidence for degradation and we consider the site to be reference (R). The ecological quality class (EQC) is a binary class variable making a distinction between reference and degraded sites. To give additional information about the degree of the impact, a refinement is possible by defining more thresholds (e.g., Figure 2.3).

2.3.4 The empirical basis

To calibrate the RCA equation crucial practical challenges include (1) to operationalize which combination of site characteristics constitutes a reference situation (define R) and to work out how to match similar sites such that like with like can be compared (define X), (2) to determine how to take an appropriate sample of reference sites and how to sample the ecological community within a site to determine the null distribution (estimate Φ_C), and (3) to preclassify the sites on a gradient of human impact to assess how the ecological community changes to determine the alternative distribution (define deviations from R such that the appropriate Δ can be developed). A well documented example of these steps can be found in Wright *et al.* (2000) amply discussing the practical approach followed to develop RIVPACS (River Invertebrate Prediction and Classification System), a software package developed by the Institute of Freshwater Ecology (IFE) to assess the river quality in the UK based on macroinvertebrates. For its development, a comprehensive database covering all sizes and types of river sites in the UK was compiled (Wright *et al.*, 1984).

2.3.5 The null or reference distribution and the alternative distribution

Similarly to statistical testing, the appropriate decision threshold T is chosen with respect to the null distribution of EQM under reference conditions. As the null is theoretically not normally known, the RCA solution is to collect a representative sample of reference sites with similar characteristics as the test site from which the reference distribution of EQM is empirically derived. Figure 2.4 gives a hypothetical example for a given ecotype. The distribution represents both the intrinsic natural variability and the sampling error to select the reference sites and the sampling of the ecological community within the site (Clarke *et al.*, 2002). With respect to this observed reference distribution, a percentile determines the decision threshold T. If a site is an outlier with respect to this intrinsic natural variation, it is considered as degraded.

This approach allows to control for the "type I error", declaring wrongly a site as degraded while it is not, but does not give information about the sensitivity of the EQM yardstick to anthropogenic impact. Controlling "the type II error" (the complement of the sensitivity) is important as well (Brosi and Biber, 2009; Lemons *et al.*, 1997; Fidler *et al.*, 2006). Therefore, the RCA approach prescribes to sample degraded sites impacted by anthropogenic pressure and to investigate the sensitivity of the index to human alterations of the ecosystem. Figure 2.5 illustrates the principle

for a hypothesized situation with two different indices A and B. For simplicity, we assume that the reference distributions of both indices are equal (which is not necessarily true, as both indices are different variables). A second simplification is that we assume the effect of degradation is only a shift of the reference distribution. In reality, most often also the spread and the shape of the distribution will be changed because of degradation. Under these simplified conditions, it is readily seen that index B has a greater discriminatory power than index A. More generally, we can quantify the diagnostic accuracy of indices by assessing the misclassification.

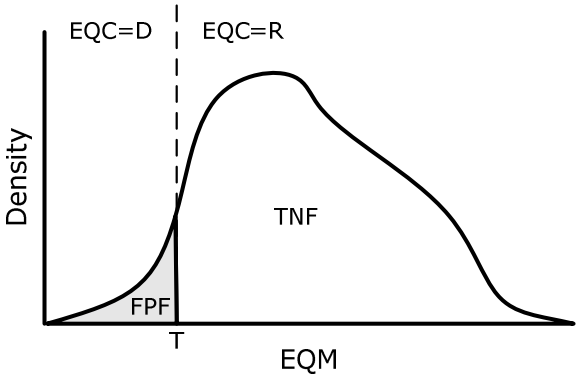


Figure 2.4 **The false positive fraction (FPF).** Decision rule with threshold fixed to a percentile of the reference (null) distribution of the ecological quality measure (EQM) under undisturbed or least impacted conditions. T = decision threshold. EQC = ecological quality class as determined by T. FPF = false positive fraction = proportion of density function below T representing sites wrongly classified as degraded, TNF = 1 – FPF = true negative fraction = specificity = proportion of the sites above T which are correctly classified.

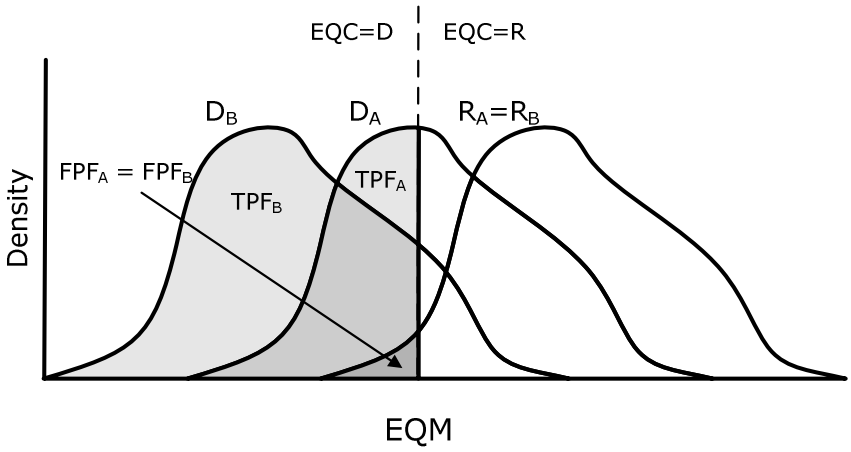


Figure 2.5 **True positive fraction (TPF) or sensitivity.** Alternative distributions D_A & D_B for two different indices A & B. For simplicity, the distribution under reference conditions is assumed to be equal for both indices (R_A = R_B) and degradation does not alter the shape of the distributions. As with Figure 2.4, the false positive fraction is fixed to a preset value (FPF_A = FPF_B). The fraction of the alternative density functions below T equals the true positive fraction (TPF) = correctly classified.

2.3.6 False positives and false negatives

The classification of Figure 2.5 is not perfect. A fraction of the sites has an EQM value below the decision threshold and will be falsely classified as degraded. In a statistical framework, they are called type I errors. To distinguish from truly statistical testing, we prefer an alternative framework coming from signal theory (Murtaugh, 1996) currently applied in diagnostic medicine to evaluate the diagnostic accuracy of biomarkers to detect a disease. Our key idea is that IBIs are very similar to diagnostic clinical tests assessing the health condition of persons (Quataert *et al.*, 2007) and we work out this analogy in detail in the chapters to come.

In signal theory, a test is called positive if it signals the event of interest (for instance, a ship on a radar). It is called negative if there is no signal. In this binary framework, four combinations are possible (Table 2.2). If the event of interest is absent (the true situation), we can have a true negative (TN) or a false positive (FP). In the latter case, the test is positive although in reality the event is not present. If the event of interest is present (the true situation), we can have a true positive (TP) or a false negative (FN). In the latter case the event of interest is not detected. In this language, a type I error is FP and a type II error is FN. In Figure 2.4, the area under the density curves left from the threshold represents the false positive fraction (FPF), i.e. the fraction of reference sites concluded to be degraded. In Figure 2.5 for index A and B, the area under the density curves left from the threshold represents the true positive fraction (TPF), as now a classification as degraded is correct.

Table 2.2 **Measures of diagnostic accuracy for a binary classification.** TPF = true positive fraction, FNF = false negative fraction, PFP = false positive fraction, TNF = true negative fraction. Y = yes or positive (index gives signal), N = no or negative (index does not give signal).

Diagnostic accuracy		Test variable or yardstick (index of biotic integrity)	
		Y (positive)	N (negative)
Gold standard (perfect knowledge)	Y (degraded)	TPF (sensitivity)	FNF (1 – sensitivity)
	N (reference)	FPF (1 – specificity)	TNF (specificity)

2.3.7 Basic assumptions

In essence, the RCA method works as follows. With respect to the expected community, a "distance to target" measure of the species composition found at a test site is defined. If this ecological quality measure (EQM) is significantly different from its expected value after taking into account the natural variability, it is concluded the test site is disturbed. A key assumption is that there exists a relatively stable natural ecological community such that models can be developed predicting the expected species composition or its attributes from the environmental characteristics

of a site. This is not always true (Bunn and Davies, 2000). For instance, the RCA principle can be hard to apply in transient and/or in highly dynamic and stressed ecosystems as estuaries (Elliott and Quintino, 2007). Yet, in spite of these restrictions, the concept offers a powerful and versatile method to create an ecological yardstick that can be empirically calibrated and validated (Bailey *et al.*, 2004).

Also, as already stated, it is not strictly necessary to choose pristine sites to apply RCA. It is also possible to calibrate an index in comparison to any other favourable baseline. In a densely populated regions as Flanders, it can be referred to the best available, least impacted sites unless similar reference sites abroad can be found. A problem with indices based on a non pristine reference is that we cannot judge recovery beyond this baseline. For instance, the estuarine biotic index for the Zeeschelde as developed by Breine *et al.* (2007), was calibrated in with respect to a baseline corresponding to class 3 of the WFD (moderate impact). If in future, this ecosystem further recovers a recalibration of the index will be necessary, for instance, by including reference situations from similar estuaries mouthing in the North Sea.

2.4 Metrics

2.4.1 Two schools

Community data consists of species variables indicating whether a species is present in the sample (presence/absence data) and/or specifying the abundance of the species. Because most species are rare and/or not easily observed, the community data is highly multivariate, variable and sparse, i.e. containing many zeros. As a consequence, it is not easy to work out the RCA equation and to estimate the multivariate reference distribution. To cope with these problems, generally speaking, there exist two different schools of researchers (Reynoldson *et al.*, 1997; Fore, 2003): the multivariate and multimetric school. The former school uses multivariate techniques to reduce the dimensionality of the data and to extract the main features of the species composition. Ordination techniques (McCune *et al.*, 2002; Legendre and Legendre, 1998) are used as principal components analysis (PCA), correspondence analysis (CA) (Jongman *et al.*, 1995), multidimensional scaling (MDS) (Borg and Groenen, 2005), and their modern more robust variants. These techniques allow to extract the strongest pattern in species composition of the reference sites and use this information to evaluate the species composition of a test site. For recent examples, we refer to Romero *et al.* (Romero *et al.*, 2007), Muxika *et al.* (Muxika *et al.*, 2007), Lücke and Johnson (Lücke and Johnson, 2009) and Primpas *et al.* (Primpas *et al.*, 2010).

In contrast, the latter school first derives well-chosen characteristics from the community data, the so-called metrics, hypothesized to be sensitive to human impact. As a result, it is possible to work with a small set of variables which is more easy to handle. More specifically, applying the RCA principle results in following equation:

$$EQM = -\Delta(M, \Phi_M(X \mid R)) \quad \& \quad EQC = \begin{cases} D & if \quad EQM \le T \\ R & if \quad EQM > T \end{cases} \qquad (the\ MMI\ equation)$$

In comparison to the basic RCA equation, the species variables of the community data are replaced with the metric variables M reducing the dimensionality of the data considerably. Moreover, as the metrics represent global characteristics of the community, the variability of the individual and especially rare species is dampened. However, this approach crucially depends on the choice of the appropriate set of metrics and interesting features of the ecological community can be filtered away. For instance, the presence of some rare species can be highly informative. In contrast, because the multivariate approach works directly on the species composition matrix, it avoids to make subjective choices of the multimetric approach which can misguide the analysis. However, metrics have a more direct meaning than the synthetic variables extracted from a multivariate approach and are derived from a priori formulated ecological hypotheses tested at the calibration stage. In addition, if it is known some rare species are highly informative indeed, an appropriate metric can be composed to cover this signal.

In general, both approaches should be considered as complementary and there is no single best method. Members of both schools sometimes report which method is to be preferred (Jordan and Vaas, 2000; Fore, 2003). In this text, we chose for multimetric models as the variables (metrics) have a more direct meaning and – as we will demonstrate – we can link them to regression models that can be optimised transparently. In this section, we further introduce the metric concept. In the next section (2.5), the scoring of the metrics is discussed. In a final section (2.6), we link MMIs to statistical regression models.

2.4.2 The metric concept

Metrics are attributes of the ecological community hypothesized to be sensitive to anthropogenic stress. To enhance the sensitivity of the bioassessment, knowledge about the ecological strategy of species (Jørgensen *et al.*, 2005) is integrated in the community data. By incorporating ecological knowledge in the study variables, we augment the power by focusing the scientific hypothesis which is generally favoured by many authors from different backgrounds (Carver, 1993; Gill, 1999; Agresti, 2002; Desbiens, 2004; Ford, 2009; Underwood, 2009). However, focussing the ecological hypothesis is not without risk. We can look in the wrong direction and miss the important event, or the species information used can be wrong or not precise (enough). Also the metric can depend too much on rare species, such that the sampling variability of the metric blurs the relation with human impact. Therefore, at the calibration stage, the response of a series candidate metrics is tested and only the best combination is retained (as elaborated further in Chapter 5). The challenge is to find a suite of metrics that in combination are sensitive to a broad spectrum of pressures (Noble *et al.*, 2007). In the following subsections, we describe some strategies to compose a set of candidate metrics with a high potential as input for the calibration of an IBI. We start with an example to illustrate the idea.

2.4.3 The number of type-specific species

An often used metric with high potential is the number of type-specific species. With increasing disturbance of the ecosystem, species requiring the specific ecological resources and environmental context of the type gradually disappear and are replaced by other, impact associated and/or less

typical species. We can quantify this process by looking at the number of typical species present in the community data. A possible ecological quality measure (EQM) can be following "z-score":

$$EQM = S_{typ} = \frac{O - E(X \mid R)}{Stdev(X \mid R)}$$

The measure compares the observed number of typical species present in the community data (O) with its expected value (E) under reference conditions (R) but otherwise similar environmental characteristics (X). To take into account the natural variability, the difference (O – E) is divided by the corresponding standard deviation (Stdev). It is important to realise this score is a theoretical construct to be checked at the calibration stage of the index. Then, we have to ascertain its discriminatory capacity and compare it with probably more powerful candidate indices (Figure 2.5).

A good focus of the candidate metrics is critical. Suppose, instead, we propose the total number of species in the sample. At low disturbance levels, quite often, first the number of species tend to increase as other less typical species invade the ecosystem increasing the species number before the original species disappear. As a consequence, this metric is expected to be less sensitive to anthropogenic stress. By focusing on the typical species, we improve the discriminatory power of the metric.

Also the numerical format of the metric influences its resolution (Oberdorff *et al.*, 2002). In our example, perhaps a better choice is to take into account the abundance of the species and look at the proportion of typical species in the community sample. It can take a long time before a typical species totally disappears from the ecosystem. In contrast, a proportion based metric immediately drops when the relative abundances of the species change or when new species enter the disturbed ecosystem. Working with proportions requires to assess abundance data consistently, otherwise the information gained can be destroyed because of the sampling variability.

2.4.4 Incorporating ecological properties of species in the metrics

The general idea of metrics is to replace the community data C with metrics M by incorporating ecological properties about the species. Symbolically, we write:

$$M = C \otimes species\ characteristics$$

In many instances, the metrics are simply linear combinations of the species properties weighted by information about the presence or abundance of the species. We distinguish two situations depending on whether information about presence/absence or abundance of species is used:

$$M_j = \sum_{i=1}^{I} w_i h_{ij} = \overline{h_j} \quad \begin{cases} (i)\ w_i = \delta_i = p_i^{\,0} & \sum_{i=1}^{I} \delta_i = N_S \\ (ii)\ w_i = p_i & \sum_{i=1}^{I} p_i = 1 \end{cases}$$

The p_i equal the proportions of individuals of species i in the sample. The sum over the p_i is one. The δ_i equal 0/1 indicating absence/presence of the species (note: $p_i^0 = 0$ for $p_i = 0$). We obtain the number of species present in the sample (N_S) by summing over the δ_i. The h_{ij} express the characteristics j of the species i. This can be simply a binary indicator (0/1) indicating whether

species i has a certain property j or belongs to group j (group membership). Alternatively, they can quantify the strength of a certain property on an ordinal or continuous scale.

We work out some often-used examples in more detail and comment on the underlying hypotheses about how the metrics M_j will respond to ecosystem impairment (Bailey *et al.*, 2004). The challenge is to find characteristics of the ecological community that are in combination sensitive to a broad spectrum of pressures (Noble *et al.*, 2007).

2.4.4.1. Type-specific species (revisited)

As already stated, we expect that the type-specific species will decrease because of anthropogenic pressure. To test for this hypothesis, the h_{ij} indicate whether a species i is characteristic for a type j yes (1) or no (0). Then, with absence/presence data, M_j equals the number of type-specific species associated with type j. With proportions, M_j is equal to the proportion of type-specific species.

A step further is to define a continuous indicator value h_{ij} expressing the level of type-specificity. For instance, the IndVal (indicator value) procedure of Dufrêne and Legendre (1997) takes into account the frequency of occurrence of the species to determine their indicative value. The average of IndVal of the species weighted by their abundance in the sample can be used as a metric.

2.4.4.2. Sensitive and tolerant species

The proportion of species known to be sensitive to anthropogenic stress is expected to decrease under the influence of anthropogenic stress, while the proportion of stress tolerant species will increase. With increasing anthropogenic disturbance, species with a narrow habitat or biotic requirements will become less abundant, and, conversely, generalists or species with adaptations to harsh circumstances, will become more dominant. To pick up this tendency, we can define binary h_{ij} coefficients specifying whether a species i is sensitive or tolerant. Based on this principle, Denys (2006a; 2006b) classified diatoms in impact sensitive (intolerant) and impact associated (tolerant) species to construct a phytobenthos index (Verschelde *et al.*, 2010).

A step further is to express the degree of tolerance or sensitiveness for degradation of the ecosystem with the h_{ij}. Then M_j equals the average tolerance or sensitivity weighted by abundance. For instance, Ellenberg scores of plants express their dependence of plants on environmental key variables as light, water and nutrient availability, acidity of the soil (Ellenberg *et al.*, 2001; Hill *et al.*, 1999; Hill *et al.*, 2009). A weighted average of the Ellenberg scores gives an indication of the ecological condition and we can compare this measure with a reference value.

2.4.4.3. The guild concept

Species are typical because they depend on specific ecological resources associated with the type. A decline of typical species is an indication the associated resources are disturbed. This principle is further refined by the functional guild concept. Ecological guilds categorise species according to their functional role in the ecosystem. Wilson (1999) distinguished two types: alpha guilds group species according to their ecological resources required for the species to survive and beta guilds refer to the necessary environmental conditions. The guild approach offers a consistent framework to construct metrics expressing functional and structural symptoms of the ecosystem (Aarts and

Nienhuis, 2003). A suite of metrics can be composed to reflect the most essential properties of the ecosystem. Another related approach to construct metrics is to group species based on their basic biological traits adapted to their (harsh) environment. For instance, Cummins (1988) advocates to use mouthpart morphology and behaviour to create functional groups of food acquisition instead of using feeding guilds.

2.4.5 Considering general characteristics of the species distribution: diversity metrics

A separate class of metrics are the diversity measures assessing characteristics of the abundance distribution without looking at specific species characteristics. Although heavily contested (Hurlbert, 1971) because a clear ecological meaning is often lacking, they are useful as indicators. Diversity measures characterise how the total number of individuals is scattered over the different species. A state where very few species dominate is totally different from a state with the same number of species, but more evenly distributed. For these metrics, one should realise that also under natural circumstances, the species distribution is uneven. However, in general, under human disturbance, a few (tolerant) species dominate more and more resulting in a lower evenness. Evenness is also often assumed to be an important characteristic on its own. As illustrated by a recent experiment with denitrifying bacterial communities (Wittebolle *et al.*, 2009), with a decreased evenness, the ecosystem functioning becomes less resistant to environmental stress.

There exist many numerical formats expressing diversity. It can be shown that most of them are function of both species richness (the number of species) and species evenness (Legendre and Legendre, 1998). This is no problem as, because of disturbance, both richness and evenness tends to decrease. Interestingly, as mentioned in Legendre and Legendre (1998) and Magurran (1998; 2004), three often used diversity indices are specific cases of the generalised entropy formula of Rényi (1961) which helps to understand the relationship between them.

$$M_q = \left(\sum_{i=1}^{I} p_i^{\,q} \right)^{\frac{1}{1-q}} = \begin{cases} q=0: \quad M_0 = \sum_{i=1}^{I} p_i^{\,0} = N_S \quad (number\ of\ species) \\ q=1: \quad H = \log(M_1) = -\sum_{i=1}^{I} p_i \log p_i \quad (Shannon's\ entropy) \\ q=2: \quad M_2 = 1/\sum_{i=1}^{I} p_i^{\,2} \quad (Simpson's\ diversity\ index) \end{cases}$$

$$Note\,1: \quad Simpson's\ dominance\ index = \sum_{i=1}^{I} p_i^{\,2} = 1/M_2$$

$$Note\,2: \quad \max(M_q) = N_S \le I \quad if\ \ p_i = \frac{1}{N_S}$$

In above formula, q represents the order of generalised entropy. Increasing the order q decreases the impact of rare species in comparison to abundant species (Hill, 1973). It is easy to understand this because p_i^q decreases faster for small p_i than for large p_i with increasing q. In fact, for q = 0, we obtain the number species N_S present in the community sample because all – thus also rare – species receive the same weight (p_i^0 = 1 for p_i > 0). For q = 1, the log of the metric equals Shannon's entropy (for the interested reader, this can be proven by applying l'Hôpital's rule). The case with q = 2 is known as Simpson's diversity measure. In the original publication, Simpson

(1949) derived the inverse of a diversity index, the dominance index (note 1 in above formula), which equals the probability that two individuals, sampled randomly, belong to the same species.

It can also be easily demonstrated that the maximal value of M_q equals N_S for any order (note 2). The maximum is obtained if all species have the same abundance or the species distribution is perfectly even. For decreasing evenness of the species distribution, M_q decreases (except for q = 0 which does not take the abundance into account). Hence, M_q can pick up the increasing skewness of the species distribution because of degradation. To be explicit, the optimal reference value is not equal to N_S. As already stated, under natural conditions, the species distribution can be extremely uneven (Magurran, 1998), but under anthropogenic pressure, we expect the evenness will further decrease and ultimately also the number of species will go down. There exist pristine ecosystems with an extremely low evenness. In this respect, evenness as such has no meaning, but only in relation to the reference value. This is true for many metrics, and therefore scoring is necessary as elaborated in the next section.

2.5 Scoring the metrics

2.5.1 The trisection method of Karr

By incorporating knowledge about the ecological strategy of species into the community data, we create powerful indicators to detect anthropogenic disturbance of the ecosystem (Jørgensen *et al.*, 2005). However, environmental conditions may strongly interfere blurring the relationship of the metric with anthropogenic stress. Therefore, to correct for this environmental variability, metrics are scored, considerably improving their discriminatory power (Pont *et al.*, 2006). In the original concept, Karr (1981) scored each metric very roughly in three possible categories: 1 (low quality), 3 (moderate quality) or 5 (high quality). The scoring thresholds for the metrics were derived by the so-called trisection method subdividing the total range of metric values of a sample of reference as well as degraded sites (excluding outliers) in three equal parts. Strictly speaking, the trisection method is not truly RCA compliant as the scoring was in comparison to a mixture of good and bad sites (Appelberg *et al.*, 2000). However, the major step forwards was Karr's idea to link the scoring with environmental background variables as illustrated by Figure 2.6. The figure shows how the relative abundance of typical species in an aquatic environment increase with stream flow (m/s) (Karr *et al.*, 1986; Bailey *et al.*, 2004). The terciles (originally fitted by eye) as a function of the stream flow define the boundaries for the scoring. Without this correction, the indicative value of the typical species metric would vanish. By simply subdividing the range in terciles irrespective of the stream flow, the sites with a slow stream flow (on the left) would receive mostly a small score and the sites with a rapid stream flow (at the right) a high score.

2.5.2 Criticisms

The trisection method is still used (Breine *et al.*, 2004; Blocksom, 2003). However, Karr's proposal is not RCA proof. It is better to derive the regression equation based on a sample of reference sites alone as disturbance can also change the relation with environmental variables. Also, instead of discrete scoring, continuous scoring is to be preferred because more powerful (Blocksom, 2003).

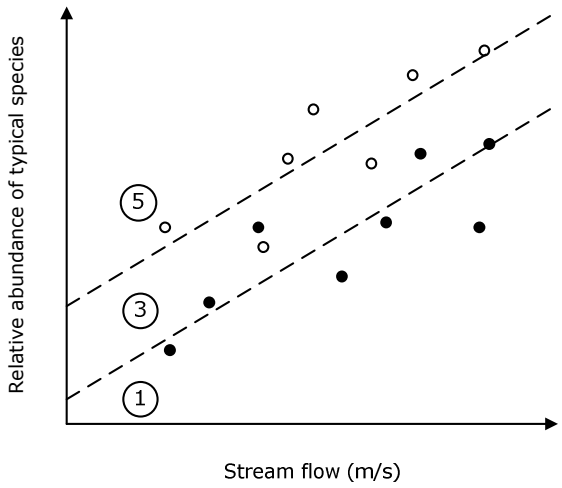


Figure 2.6 **The original trisection scoring method of Karr (1981)** (○ = Reference sites, ● = Degraded sites). Two lines are drawn (originally by trial and error) subdividing the original observations in three equal parts. For positive metrics, the points in the upper tercile receive a score of 5 (as is the case here); for negative metrics, the points in the lower tercile receive a score of 5.

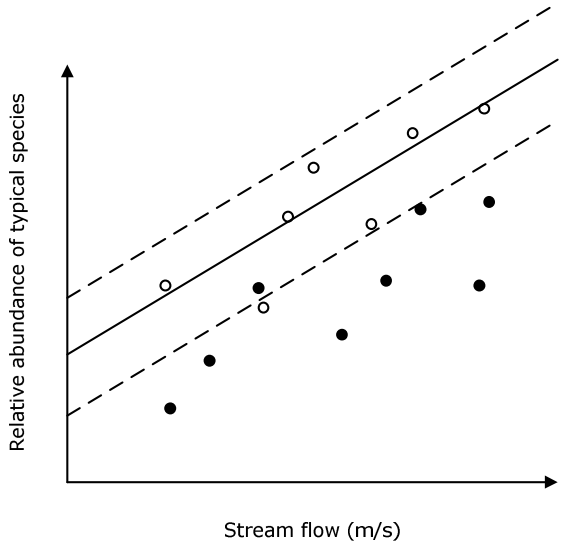


Figure 2.7 **Standardised residual approach based on z-scores** = "distance to target" measure in comparison to the reference situation (○ = Reference sites, ● = Degraded sites). A regression line (solid line) is fitted to the reference data correcting for the environmental variability because of stream flow. The dashed lines give an indication of the natural variability under reference conditions (±2 times the standard deviation). In comparison to Figure 2.6 the points stemming from degraded sites consistently receive a low score.

Figure 2.7 gives an improved version of the scoring. The dashed lines represent the prediction limits taking into account the natural and sampling variability under reference condition (this conceptual representation does not comprise the model uncertainty on the regression line). Now all disturbed sites consistently receive a negative score. As demonstrated by Oberdorff *et al.* (2002) for an IBI in France, this improvement substantially increases the discriminatory power. To cover the broad geographical scale, scoring was based on complex regression models containing several environmental variables. A similar method was subsequently applied to develop the European Fish Index (EFI), the pan-European fish index (Pont *et al.*, 2007) we evaluate in Chapter 6.

2.5.3 Standardised residuals

We generalise and formalise the concept of Figure 2.7 as follows:

$$S_j = \delta_j \frac{M_j - m_j(X \mid R)}{s_j(X \mid R)} \qquad with\ \delta_j = \pm 1 \quad \sim \quad positive / negative\ metric$$

The z-scores S_j are standardised residuals of predictive statistical models describing the reference distribution of the metrics as a function of the environmental characteristics of a site X (which can be a vector of values describing multiple components of the ecosystem). These z-scores express the difference between the observed metric value M_j and its expected value $m_j(X|R) = E[M_j|X;R]$ divided by the standard deviation $s_j(X|R)$ as predicted from the environmental variables X under reference conditions R. The multiplication factor δ_j guarantees each score is positively associated with the ecological quality. A metric is positive when positively associated with ecological quality (e.g., the number of typical species present at a site), and negative when negatively associated (e.g., the proportion of impact-associated species of the diatom index). Scoring also standardises the metrics making them more comparable and facilitating interpretation. The scores are unitless and have a similar scale. Yet, the main reason is to correct for the environmental variability improving the resolution of the indicator as illustrated by Figure 2.6.

2.5.4 Alternative scoring of the metrics

Although scoring is very important, with the exception of Blocksom (2003), we did not find papers systematically exploring the issue. As used in Box 2.1, she proposed a two-fold categorisation: (i) continuous or discrete (q specifies the number of classes; q = 0 denotes continuous scoring), and (ii) with respect to the reference (R) or full (F) distribution. Even if the metrics are not normally distributed, z-scores give a flavour of the "distance to target" to the reference condition by taking the natural variability into account. A possible drawback is that deviations are judged in a linear fashion. A z-score of -4 is not necessarily twice as bad as a z-score of -2. For instance, if a critical threshold is surpassed at -3, a scoring of z = -4 can worse than indicated. To improve scoring in this respect, exploration of the full distribution may be helpful, to derive thresholds linked with critical changes in the environment.

Discrete scoring was originally used by Karr based on terciles (Karr, 1981) giving the scores 1, 3 and 5 to observations in the lower, middle and upper tercile respectively. In Box 2.1, we generalise this scoring to any number of subdivisions, e.g. quintiles for q = 5. To see the link with Karr, it is important to recognize that scoring with (0, ½, 1) is equivalent to using (1,3,5). It is just a linear

transformation. The advantage of discrete scoring is that we take into account the shape of the distribution. However, in the aforementioned study of Blocksom (2003), it is recommended against discrete scoring as it decreases the sensitivity as we will confirm in Chapter 5.

Finally, as demonstrated in Box 2.1, continuous scores with respect to the entire distribution are linear transformations of continuous scores with respect to the reference distribution. So, implicitly, we give the metrics another weight.

Box 2.1 **Scoring typology.** R/F = with respect to reference or full (degraded sites included) distribution. 0/q = continuous/discrete with q classes (e.g., q = 3 or 5 refers to scoring with respect to terciles or quintiles). (1) R0: Reference scoring (continuous): E_R & $Stdev_R$ = expected value & standard deviation of the reference distribution of the metric. (2) F0: Full range scoring (continuous): E_F & $Stdev_F$ = expected value & standard deviation of the entire mixture distribution of the metric in reference and degraded sites. (3 & 4) Discrete scoring with respect to reference (Rq) or full distribution (Fq): e.g. for q = 3, we obtain (0, ½, 1) as possible scores, and, for q = 5, (0, ¼, ½, ¾, 1); $Q^{Rq}(Y)$ = class number of Y in comparison to the percentiles (e.g., for q = 5, if Y is larger than the third quintile, but not larger than the fourth, $Q^{R5}(Y) = 4$; and $S^{R5} = ¾$). Note: to guarantee the scores are positively associated with the ecosystem quality, we multiply negative metrics j by $\delta_j = -1$. The X allow to correct for the environmental conditions and site typology.

(1) $R0$: *Continuous scoring* $(q = 0)$, *with respect to the reference distribution* (R)

$$S_j^{\ R} = \delta_j \frac{M_j - E_R[M_j \mid X]}{Stdev_R[M_j \mid X]}$$

(2) $F0$: *Continuous scoring* $(q = 0)$, *with respect to the full* (*mixture*) *distribution* (F)

$$S_j^{\ F} = \delta_j \frac{M_j - E_F[M_j \mid X]}{Stdev_F[M_j \mid X]} = \frac{Stdev_R[M_j \mid X]}{Stdev_F[M_j \mid X]} S_j^{\ R} + \delta_j \frac{E_R[M_j \mid X] - E_F[M_j \mid X]}{Stdev_F[M_j]}$$

$$\Rightarrow \quad \boxed{S_j^{\ F} = a_j S_j^{\ R} + b_j} \qquad \boxed{a_j = \frac{Stdev_R[M_j \mid X]}{Stdev_F[M_j \mid X]}}$$

(3) Rq : *Discrete scoring*, *based on percentiles of the reference distribution*

$$S_j^{\ Rq} = \frac{Q_{Rq}(\delta_j M_j) - 1}{q - 1} \qquad e.g.\ S_j^{\ R5} = \frac{Q_{R5}(\delta_j M_j) - 1}{4} \in \{0, ¼, ½, ¾, 1\}$$

(4) Fq : *Discrete scoring*, *based on percentiles of the full* (*mixture*) *distribution*

$$S_j^{\ Fq} = \frac{Q_{Fq}(\delta_j M_j) - 1}{q - 1}$$

2.6 The link with regression models

2.6.1 Multimetric indices of biotic integrity

MMIs combine several metrics into one single index for an overall assessment the ecosystem condition. At the calibration stage, the candidate metrics are chosen to cover a broad spectrum of possible degradations of the ecosystem and an optimal basket is selected from this list. As already stated in section 2.4.2, we can define EQM as follows:

$$EQM = -\Delta(M, \Phi_M(X \mid R))$$

In this formula, EQM is the distance between the observed metrics M and their multivariate statistical distribution $\Phi_M(X|R)$. In contrast to the more general formula in section 2.3.3, the community data C is replaced with metric variables M. Apart from augmenting the community data with species knowledge, this replacement reduces the dimensionality of the data and stabilizes its statistical properties. Community data is highly multivariate, variable and sparse, i.e. containing many zeros as most species are rare and/or are not easily observed. By considering attributes of the community, the metrics are less dependent on rare and/or hard to determine species. As a consequence, some of the sensitivity may be lost, but in general, this drawback is compensated by far because of the better statistical properties of the metrics. If monitoring of rare species is an important objective, a more specific monitoring program is necessary. An IBI is not the appropriate instrument to monitor specific species but tests for the ecological condition.

2.6.2 The average score model (AVG) and the link with regression models

A specific format of above formula is a simple linear combination of the metric scores. Instead of considering the multivariate metric distribution, each metric is individually scored and a weighted average is calculated:

$$EQM = \sum_{j=1}^{J} w_j S_j \quad \overset{w_j=\frac{1}{J}}{\rightarrow} \quad EQM = \frac{1}{J}\sum_{j=1}^{J} S_j$$

If the weights are equal, EQM is simply an average of scores most commonly found in the literature. We refer to this model as the average score model (AVG). In fact, AVG is a special case of a regression model with fixed coefficients. This link becomes more apparent by replacing the weights by regressions coefficients:

$$EQM = [\beta_0] + \sum_{j=1}^{J} \beta_j S_j \qquad \Rightarrow \qquad HIC \sim S_j$$

For this regression model, the ecological status is the (binary or ordinal) response variable and the metric scores are the explanatory variables. The weights are the regression parameters. The intercept is immaterial for the discriminatory power as it just implies a shift of the location of the distribution of EQM as whole. Because of the link with regression models, we can use model building strategies and techniques to select the optimal combination of metrics (Pepe *et al.*, 2006; Pepe and Thompson, 2000). This is main topic of Chapter 5.

2.6.3 The proportional odds model

At the calibration stage, we should seek for the optimal set of metrics out of a set of candidates to optimise the match with the human impact class (HIC), an independent a priori classification of the ecosystem condition. As HIC is binary or ordinal, an evident choice is a logistic regression model, e.g. the proportional odds model (Agresti, 2002; McCullagh and Nelder, 1989). This model is instructive because of its latent variable interpretation (Anderson and Philips, 1981) as represented by Figure 2.8. The proportional odds model assumes there exists an underlying continuous latent variable expressing the ecological quality status (EQS). We cannot observe the latent variable, but only assess the human impact class (HIC). By fitting the proportional odds model to HIC, we estimate this underlying variable.

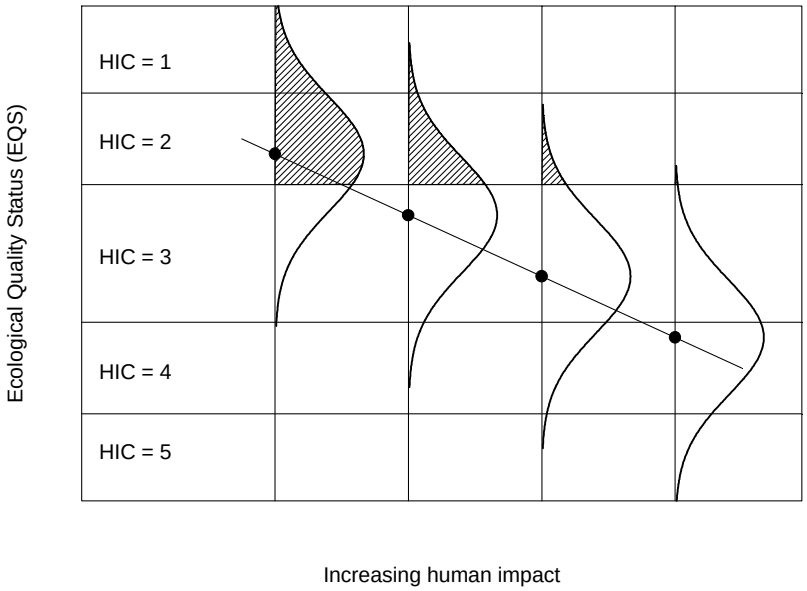


Figure 2.8 **Observation model underlying an ordinal logistic regression.** Measurements on an ordinal scale of an underlying continuous (latent) variable representing the ecological quality status (EQS) by the human impact class (HIC). The regression line models how an increasing human impact decrease the average EQS. The density curves represent the natural variability around the regression line. The horizontal lines show how the measurement method categorizes EQS. The figure shows how the proportions in each category are expected to change for an increasing human impact. For instance, with respect to the contrast 1-2 / 3-5, the expected proportion of observations in the upper category will change as shown by the shaded region.

2.7 In summary

2.7.1 The data flow

To measure the ecosystem condition, the WFD prescribes to use the reference condition approach (RCA) for the development of ecological indices. An RCA index measures the "distance to target" of the observed community C from the community as expected in reference conditions. Figure 2.9 assembles the main formula of this chapter (top panel) and visualises the data flow starting from the raw data: the community data C and the site descriptors X enabling to match the test site with similar reference sites. The X specify the site type, the environmental conditions, and also, if relevant, the sampling conditions and methods.

(i) As it is quite hard to compare the highly multi-dimensional and sparse species data directly, the community data C are first translated into metrics M. These metrics are calculated from the species data augmented with indicative values of the species reflecting their niche in the ecosystem. By incorporating this functional information, metrics express properties and patterns sensitive to human exposure on the ecosystem level. Because species properties are used instead of the species themselves, the original crude species data is translated in "classical" variables amenable for statistical (regression) analysis. In this respect, the metric approach has the potential to work at larger geographical scales. It is possible that the species change along large distances, but that other species fulfil the same ecological functions such that the same metrics can be used.

(ii) An important next step is the metric scoring. Although the metrics have a conceptual meaning, as such, their values do not directly express the level of human impact. Therefore, it is necessary to score them against a reference which is the pivotal step of the RCA. The metric scores S are conceived as "distance to target" measures from a reference value taking into account the natural and sampling variability under reference conditions, a simple example being the z-scores. Many other scoring algorithms exist, but the essential point is that a reference distribution is known to compare with. This can be a quite involved predictive model that should be estimated from a sufficiently broad sample of the reference sites in the region in which IBI is to be used.

(iii) The third step integrates the different scores into one single yardstick, the ecological quality measure (EQM). Quite often, this is simply the average of the scores, but it can be any function. This ecological yardstick is the test variable underlying the IBI of which we can study its potential to discriminate between the different states of the ecosystem.

(iv) The fourth step of RCA is an appreciation of EQM on a degradation scale. The categorisation can be binary (D or R) or ordinal, e.g. ternary (D, M, R); or five-tiered as for the WFD (1,2,3,4,5). This ecological quality classes (EQC) result from comparing EQM with decision thresholds reflecting fundamental ecosystem changes.

(v) To calibrate and validate the IBI model, a preclassification of the sites is necessary expressing the human impact class (HIC) independent of the ecological community to avoid circular reasoning. The common approach is to score the human activities at a site by an anthropogenic impact assessment and to integrate this information in an overall measure. Although seldom made explicit, this involves some "impact model" describing how different pressures result in changes of

the ecosystem. To make a systematic inventory of anthropogenic impacts, the DPSIR model is useful, expressing a causal path starting from the anthropogenic activities (Aubry and Elliott, 2006) = the driving (human) forces ("D"), resulting in pressures ("P") changing the state of the environment ("S") which causes an impact ("I") on the ecosystem.

$$principle: IBI \overset{RCA}{=} -\Delta(C, \Phi_C(X)) \overset{MMI}{=} -\Delta(M, \Phi_M(X))$$

$$\triangleright\ Index\ calculation: \quad C \xrightarrow{m_j} M_j \xrightarrow{s_j(X)} S_j \xrightarrow{p} EQM \xrightarrow{d} EQC$$

$$\triangleright\ Index\ construction: \quad HIC \overset{q(ROC)}{\sim} EQC = i(C, X)$$

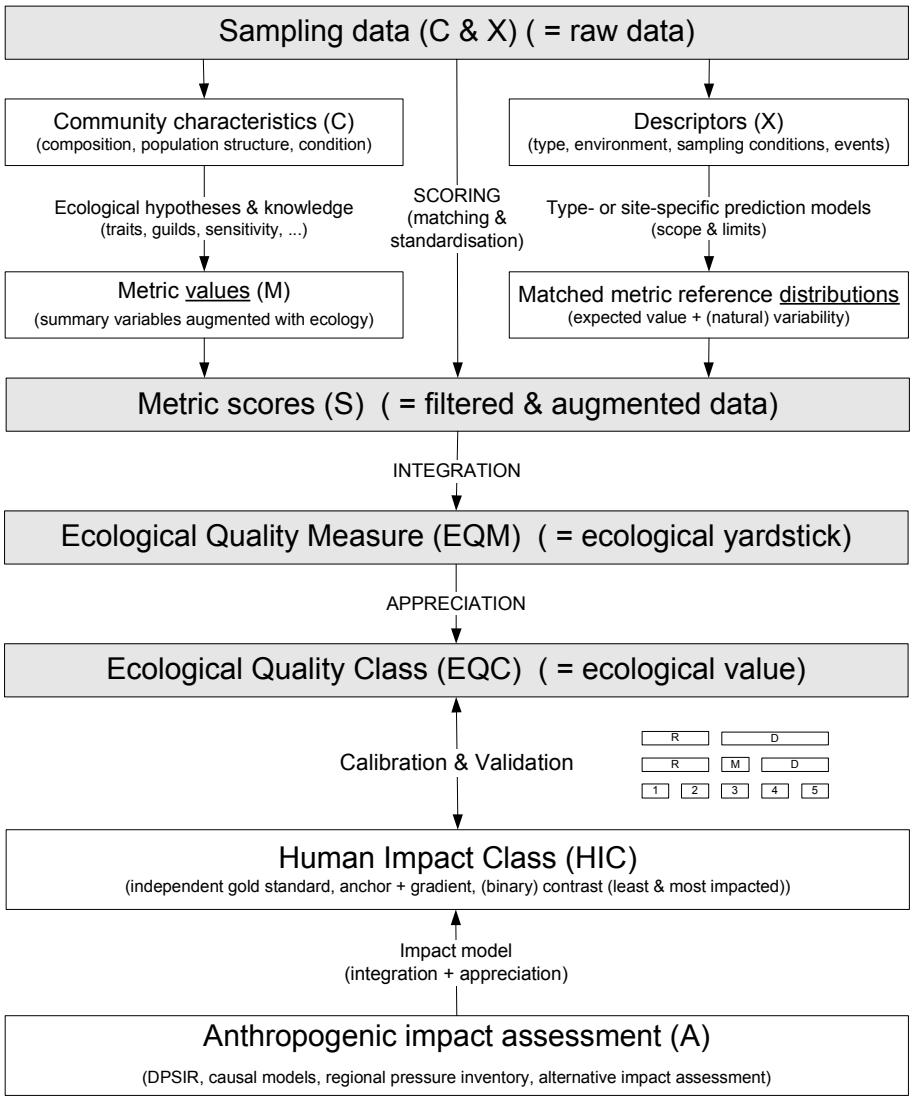


Figure 2.9 **The measurement model**. Summary of the main formula of the chapter and generic data flow of an index of biotic integrity (IBI) as based on the reference condition approach (RCA).

2.7.2 The index model

In this chapter, we aimed for a deeper understanding of the ecological rationale of an IBI in order to derive the appropriate statistical model.

- **The model format.** Traditionally, an IBI is an average of 'distance to target' scores which is seldom recognised as a statistical model. Yet, this average score model (AVG) is a model closely linked to an ordinal logistic regression model but with fixed regression coefficients. Because of this link, we can use regression building techniques to construct IBIs more objectively and coherently, for instance, to select an optimal set of metrics from a list of candidates. This is the subject of Chapter 5 where we recalibrate the fish-based estuarine biotic index (EBI) for the Zeeschelde (Breine *et al.*, 2007). In this chapter, we also investigate the relation of AVG with a proportional odds model.

- **The explanatory variables.** The basic building blocks of an IBI and the explanatory variables of the index model are the scored metrics which are 'distance to target' measures. Metrics are attributes of the community data covering some essential features of the ecosystem. The metrics incorporate background information about the ecological strategy of species to increase the specificity and sensitivity to anthropogenic stress. Scoring of the metrics is essential to match the sites to similar reference sites based on type and site specific criteria and to take into account the natural variability of the metrics under reference conditions. This is the essence of the Reference Condition Approach (RCA).

- **The response variable.** The IBI model is calibrated and validated with respect to an independent preferably gold standard classification of the human impact (HIC). To avoid circularity and preconceived ideas about what constitutes an intact ecosystem, we should determine HIC independently from the biological community (Stoddard *et al.*, 2006). The common approach is to score the human activities at a site. Sites without any (measurable) human activity are reference or anchor sites. The other sites are ranked with respect to these anchor sites in a gradient of human impact.

- **The optimisation criterion.** When constructing an IBI, the purpose is to match EQC as closely as possible to HIC. We observed that IBIs are very analogous to diagnostic tests in medicine for which the Receiver Operating Characteristic (ROC) curve is the central tool (Zhou *et al.*, 2002; Pepe, 2003). Our evaluation of the European Fish Index (EFI) and the comparison with other existing national indices was based on this idea (Quataert *et al.*, 2007) as presented in Chapter 5.

3 The Receiver Operator Characteristic (ROC) curve. An indispensible concept and tool to understand and assess the validity and usefulness of ecological indicators

In this chapter, we further justify our idea to evaluate IBIs in a way similar as diagnostic tests in medicine as we worked out for the evaluation of the European Fish Index (EFI) of the FAME project (Quataert et al., 2004; Quataert et al., 2007). We summarize the vast literature about ROC curves and translate it to the context of ecological indicators. A novelty is the introduction of utility functions visualizing and quantifying the cost implications of index-guided decisions and linking the diagnostic accuracy of the index with its usefulness in a decision context.

Abstract

According to statistical decision theory, the Receiver Operating Characteristic Curve (ROC) concept is the appropriate tool to evaluate the diagnostic accuracy of continuous test variables. ROC curves plot the true positive fraction (TPF = sensitivity) as a function of the false positive fraction (FPF = 1 – specificity) for all possible decision thresholds discriminating between degraded and pristine sites. To obtain a deeper understanding of ROC curves, we introduce utility curves as tools to evaluate index performance. Similarly to ROC curves, utility curves plot the cost implications of index-guided decisions as a function of FPF or TPF for all possible decision thresholds. By doing so, they link the diagnostic accuracy of an index with its practical usefulness. As a major insight we infer that the main factor determining the usefulness is the capacity to realise a high TPF controlling FPF at a low level. We refer to this property as the strength of the index as visualised by the steepness of the ROC curve and quantified by summary statistics as the full or more focused partial area under the ROC curve (aucF and aucP). More specifically, in a river restoration context, we demonstrate how a strong index is better capable to realise a high true restoration fraction (TRF) and a high overall restoration benefit (ORB) keeping the average restoration cost (ARC) low.

Keywords

ROC curves, utility curves, sensitivity, specificity, (partial) area under the curve, false positives and false negatives

3.1 Introduction

Indices of biotic integrity (IBIs) are designed to assess the condition of the ecosystem in a cost-effective way in comparison to a more full ecological investigation (the gold standard). To be useful in a decision context, the diagnostic accuracy of an IBI should be sufficiently high. Otherwise the cost because of wrong decisions will be larger than the savings realised by using an index instead of a gold standard. For the evaluation of an IBI, we should differentiate between false positive (FP) and false negative (FN) errors as they imply different ecological and societal costs. An FN error occurs when the index does not detect that the ecosystem is degraded. An FP error is when the ecosystem is signalled to be degraded while it is pristine. It is necessary to keep both errors small. When deciding whether a site should be restored, a high false negative fraction (FNF) or low sensitivity (TPF = true positive fraction = 1 – FNF) results in many undetected degraded sites, not properly treated implying a high ecological loss. Conversely, with a high false positive fraction (FPF) or low specificity (TNF = true negative fraction = 1 – FPF), many pristine sites will be unnecessarily treated depleting resources and implying a risk for the ecosystem.

An IBI is similar to a diagnostic medical test used to assess the health condition from biomarkers (Quataert *et al.*, 2007). In diagnostic medicine, Receiver Operating Characteristic (ROC) curves are routinely used to evaluate the diagnostic accuracy of medical tests. According to statistical decision theory (Zhou *et al.*, 2002; Pepe, 2003), the ROC curve is the appropriate tool to characterise the diagnostic accuracy in a binary decision framework. Originally, ROC curves were developed in the context of communication technology to examine the detection quality of radars (Swets and Pickett, 1982; Swets, 1988; Swets *et al.*, 2000). Gradually, the tool found its way in many other areas where detection of an event is central. Already in 1996, Murtaugh argued to employ ROC curves founded in signal theory to evaluate ecological indicators in comparison to a gold standard, but to date, the technique is not applied unless recently (Breine *et al.*, 2007; Quataert *et al.*, 2007; Hale *et al.*, 2007; 2008; Dos Santos *et al.*, 2011).

The Receiver Operating Characteristic (ROC) curve is central in this thesis. The aim of this chapter is to demonstrate the value of ROC curves as a tool for calibrating (Chapter 5) and validating (Chapter 6) indices. We start with a profound recapitulation of the concept. To achieve a deeper understanding, we investigate how the usefulness of an index is linked to ROC curves, and how this relationship is modified by parameters describing the operational context. For this summary, we borrowed extensively from two recent handbooks about statistical methods in diagnostic medicine (Pepe, 2003; Zhou *et al.*, 2002), but we make a consistent transcription and extension of the terminology and concepts to ecological indicators and IBIs. A novelty is the introduction of utility curves which link the strength of the index as characterised by the ROC curve with its contribution to decision making. In Chapter 4, we further investigate the potential of ecological indicators to improve decision making.

3.2 Material & Methods

3.2.1 Diagnostic accuracy in a binary decision framework

3.2.1.1. The binary decision framework

First, we imagine a simple binary decision framework in the context of river management for which an IBI is used to decide about restoration of a site or waterbody. We assume that there are only two ecosystem conditions: degraded "D" (restoration necessary) or pristine "R" (restoration not necessary and possibly harmful). The relative frequency or prevalence of degraded sites in the target region of the index is π^+, the prevalence of pristine sites is $1-\pi^+$ (the complement as only two conditions are possible).

The test variable of an IBI is the ecological quality measure (EQM) which can be a single continuous variable (or metric) or a synthesis of many metrics (a multi-metric index). We assume EQM is positively correlated with the quality for the ecosystem. High values correspond with a good ecosystem status, low values are indicative for degradation (this is no restriction, as we can always change the sign of EQM).

An IBI resembles a diagnostic model in medicine assessing the health or disease status of a patient from a combination of biomarkers. Typically, the outcome of a medical diagnostic model is a continuous test variable, yardstick or metric of which, say, low values are indicative for the disease and high values for absence of the disease. As such the biomarker does not make clear whether a person is diseased and whether treatment is necessary. To discriminate, it is necessary to determine a decision threshold T below which a patient is considered diseased and is consequently treated.

Typically, no perfect distinction will be possible with this classification. Because of intrinsic natural variation in the population, model limitations and/or measurement error, some healthy persons will have a score below the decision threshold and, conversely, there will be some persons with the disease score above the threshold. The same holds for IBIs. Because of the natural variation and sampling variation, the EQM can be unexpectedly large for degraded sites or very low for pristine sites. For decision making, we should ascertain the accurateness of the index.

3.2.1.2. False positive and false negative errors

In a binary decision framework (yes or no, diseases or healthy, pristine or degraded), two mistakes are possible: false positive (FP) errors when the event of interest is not detected (e.g. a healthy person has a test value below the decision boundary), and, conversely, false negative (FN) errors when the event of interest is absent, but is signalled (e.g. a diseased person has a test value above the decision boundary). It is crucial to differentiate between both errors, as they have different implications. With a high false positive fraction (FPF), many healthy persons are unnecessarily treated (detracting resources from where they are required) with sometimes invasive techniques (possibly harmful). Conversely, with a high false negative fraction (FNF), many diseased persons do not receive the appropriate treatment.

Box 3.1 **Definition of the main measures of diagnostic accuracy.** The intrinsic diagnostic accuracy is defined from the perspective of the true but unknown status U. The operational diagnostic accuracy or informativeness makes an evaluation with respect to the observed outcome W of the index. For definition of the symbols, see Table 3.1.

$$decision\ rule: EQC = \begin{cases} W^+ & if \quad EQM \le T \\ W^- & if \quad EQM > T \end{cases}$$

▷ *Confusion matrix & prevalence of degradation*

	W^+	W^-
U^+	TP	FN
U^-	FP	TN

$\pi^+ = P(U^+)$

$\pi^- = P(U^-) \quad (=1-\pi^+)$

▷ *Intrinsic diagnostic accuracy* : $P(W^\bullet \mid U^\bullet)$

	W^+	W^-
U^+	$Se = TFP = P(W^+ \mid U^+)$	$FNF = P(W^- \mid U^+)$
U^-	$FPF = P(W^+ \mid U^-)$	$Sp = TNF = P(W^- \mid U^-)$

▷ *Operational diagnostic accuracy* : $P(U^\bullet \mid W^\bullet)$

	W^+	W^-
U^+	$PPV = P(U^+ \mid W^+)$	$FRF = P(U^+ \mid W^-)$
U^-	$FAF = P(U^- \mid W^+)$	$NPV = P(U^- \mid W^-)$

3.2.1.3. The confusion matrix with respect to the gold standard

To estimate FPF and FNF, a gold standard method should be available assessing the true health condition without error, based on more extensive, more invasive, and/or more sophisticated techniques, or by repeating observations over a long time period and engaging several experts. Although in reality perfect knowledge does not exist, we can imagine that, given the current state of the art, there exists (or we can develop) a method which is at least one order of magnitude more accurate than the test, but of which the high costs (or other constraints such as the required time to detect the event) preclude a practical application. The whole point of diagnostic testing is that it is possible to approximate the gold standard at a lower cost (in terms of the monetary as well as societal, ecological or health consequences) and/or in a shorter time frame, with a sufficient level of diagnostic accuracy to be useful in practice. To realise this idea and to guarantee its validity, calibration and validation studies are necessary. With calibration studies, we seek for the optimal test variable (or combination of them) compromising cost and diagnostic accuracy. Validation studies assess the diagnostic accuracy preferably on an independent sample of the sites in the target region (the statistical population).

The assessment of the diagnostic accuracy starts from the binary confusion matrix (Box 3.1) which is a contingency table of the test outcomes and the true status (of the ecosystem or health) as determined by the gold standard. For degraded sites (U^+), there are true positives (TP: W^+) and false negatives (FN: W^-); for pristine sites without degradation (U^-), there are true negatives (TN: W^-) and false positives (TP: W^+). TN and TP are correct classifications, FN and FP are misclassifications. From the confusion matrix, it is possible to derive various measures of diagnostic accuracy. A first group of statistics is defined from the perspective of the true (but unknown) status (the rows of the confusion matrix), a second from the perspective of the test outcome (the columns). As the statistics in the former group are calculated within each ecosystem condition, they do not depend on the relative occurrence or prevalence of the condition. Therefore, they are referred to as measures of intrinsic diagnostic accuracy. They do not depend on the decision context. In contrast, statistics of diagnostic accuracy of the latter group are defined over the mixture of different ecosystem conditions. As a consequence, they depend on the relative occurrence of the conditions in a specific operational context. We call these statistics measures of operational diagnostic accuracy as they change as a function of the decision context.

3.2.1.4. Intrinsic diagnostic accuracy: sensitivity (TPF) and specificity (TNF)

A first series of measures of diagnostic accuracy evaluates the diagnostic test from the perspective of the true (ecosystem) condition. When the ecosystem is pristine (true condition), the specificity (Sp) or true negative fraction (TNF) expresses which fraction of the sites is recognised as a pristine site. The complement is the false positive fraction (FPF), the fraction of pristine sites misclassified as degraded. Conversely, when the ecosystem is degraded (the true condition), the sensitivity (Se) or true positive fraction (TPF) equals the faction of the sites detected by the index. The complement is the false negative fraction (FNF), the fraction of degraded sites not detected and misclassified as pristine. Mathematically, these measures are conditional probabilities of the test outcome $W^\bullet$ given the true status $U^\bullet$: $P(W^\bullet|U^\bullet)$. For instance, $FPF = P(W^+|U^-)$, or in words: the false positive fraction is the conditional probability that a site will be (mis)classified as degraded given it is pristine. Within each condition the sum of the probabilities equals one, e.g. $P(W^+|U^+) + P(W^-|U^+) = 1$ as there are only two events possible in each group: a correct classification or a misclassification. Thus, Se = TPF = 1 – FNF and Sp = TNF = 1 – FPF.

To simplify notation, in most cases, we will use TPF and TNF instead of Se and Sp (Pepe, 2003). To characterise the diagnostic accuracy, it is sufficient to pick only one element from each couple as the other is just the complement. In most situations, we will choose for the combination (TPF, FPF) because of its relation with the ROC curve plotting TPF as a function of FPF (to be introduced in next section). Another instructive measure of the intrinsic diagnostic accuracy is the ratio TPF/FPF, expressing the signal to noise ratio, i.e. how well an index is capable to detect degraded sites without confounding with pristine sites. In a Bayesian framework, TPF/FPF is the likelihood ratio of the probability of obtaining a positive test result in a degraded site divided by the probability of obtaining a positive test result in a pristine site (Motulsky, 1995). As will be derived further on, the ratio is closely linked to the positive predictive value (PPV) of the index, a measure of the operational diagnostic accuracy.

Table 3.1 **Measures of diagnostic accuracy: symbols and definitions.**

ASSESSMENT of the STATE of the ECOSYSTEM		
"Gold standard"	$U^{\bullet}$	Gold standard outcome assessing the unknown situation correctly
	U^{+}	Degraded (moderately or severely impaired)
	U^{-}	Not degraded (healthy, pristine, reference, baseline)
Outcome diagnostic test (index)	$W^{\bullet}$	Outcome diagnostic test
	W^{+}	Test/index positive, indicating/signalling degradation
	W^{-}	Test/index negative, not indicating/signalling degradation
Index variables	EQM (= W_C)	Ecological quality measure (= continuous test variable)
	EQC (= W_D)	Ecological quality class (= discrete: binary / ordinal)
		+ bi-state (binary): R (reference) / D (degraded) or U^{-} (event of interest absent)/ U^{+} (event of interest present)
		+ tri-state (ordinal): EQC = R (reference), M (moderate), D (degraded);
		+ five-tiered (WFD): 1 (good = blue), 2 (fair = green), 3 (moderate = yellow), 4 (poor = orange), 5 (bad = red)
Classification	T	Decision threshold to distinguish degraded from non-degraded sites.
	TC & FC	True (correct) & false (incorrect) classification
	TP & FP	True & false positive
	TN & FN	True & false negative
MEASURES / STATISTICS of DIAGNOSTIC ACCURACY		
From perspective of the true but unknown status	TPF & FNF	True positive fraction and false negative fraction (sum = 1)
	Se = TPF	Sensitivity (capacity to detect disease / disturbance)
	TNF & FPF	True positive fraction and false negative fraction (sum = 1)
	Sp = TNF	Specificity (capacity to recognise healthy / pristine status)
From perspective of index outcome (observed value)	PPV	Positive predictive value (fraction of positive outcomes of index corresponding with presence degradation)
	FAF	False alarm fraction (FAF = 1 – PPV)
	NPV	Negative predictive value (fraction of negative outcomes of index corresponding with absence degradation)
	FRF	False reassurance fraction (FRF = 1 – NPV)
Area under the ROC curve (AUC)	aucF	(Full) area under the curve for the full range = aucF(0,1)
	aucP(f_1,f_2)	Partial area under the curve for $f_1 \leq FPF \leq f_2$
	aucP	Default-value = aucP(0.1,0.3)

Table 3.1 **Measures of diagnostic accuracy: symbols and definitions.** (Continued)

PROBABILITIES		
Prevalence	π^+	prevalence (frequency) of disease/degradation in population
	$P(U^+) = \pi^+$	(a priori) probability a person/site is diseased/degraded
	π^-	prevalence of health/high ecological status in the population
	$P(U^-) = \pi^-$	(a priori) probability that a person/site is healthy
Conditional probabilities	$P(W^\bullet \mid U^\bullet)$	probability of test/index outcome $W^\bullet$ given the ecosystem or health status $U^\bullet$ → sensitivity, specificity
	$P(U^\bullet \mid W^\bullet)$	(a posteriori) probability of health status $U^\bullet$ given test/index outcome $W^\bullet$ → predictive values (informativeness of index)
Odds	odds(π)	$= \pi / (1- \pi)$; $0 \leq \pi \leq 1$
	logit(π)	$= \log(\text{odds}(\pi)) = \log (\pi /(1- \pi))$
UTILITIY CURVES		
CEA: Cost effectiveness analysis	ARC	Average restoration cost
	TRF	True restoration fraction (= PPV)
	C_R	Restoration costs
	C_A	Assessment costs
CBA: Cost benefit analysis	ORB	Overall restoration benefit
	ERB	Ecological restoration benefit (ERB = ORB – ARC)
	B_{ROC} = ORB	Expected benefit of the index; $B_{ROC} = T_G.TPF – T_H.FPF$
	b_{ROC}	Benefit function (kernel of ORB); $b_{ROC} = TPF – FPF/b$
	b	Benefit ratio (regional scale) = $T_G/T_H = b_R.\text{odds}(\pi^+)$
	b_R	Intrinsic benefit ratio (local scale) = R_G/R_H
Cost tradeoffs	$T_G \leftrightarrow T_H$	Expected gain/harm on a regional level; $T_G = R_G.\pi^+$ & $T_H = R_H.(1-\pi^+)$
	$R_G \leftrightarrow R_H$	Cost-corrected gain/harm on a site level: $R_G = E_G – C_R$; $R_H = E_H + C_R$
	$E_G \leftrightarrow E_H$	Expected gain/harm on a site level
STATISTICAL DISTRIBUTIONS		
Distribution of the test variable	$F (F_R, F_M, F_D)$	Cumulative distribution function of the test variable (in reference, moderate, degraded situation)
	$f (f_R, f_M, f_D)$	Corresponding density functions
	F^{-1}	Inverse distribution function = quantile function
Bootstrapping	n_B	Number of bootstrap (re)samples (default: 1000)
	BCa	Bias-corrected, accelerated percentile method to estimate confidence limits (Efron, 1987)
	EDF	Empirical distribution function (= bootstrap population)

3.2.1.5. Operational diagnostic accuracy: positive and negative predictive value

When using the index, the true ecosystem status is unknown. We only have the index results. The question is how sure we can be about the health condition given the outcome of the index. This leads to a second series of measures of diagnostic accuracy assessing the informativeness or predictive values of a diagnostic test from an end user perspective (Box 3.1). Mathematically, predictive values are conditional probabilities of the ecosystem status given the index result: $P(U^{\bullet}|W^{\bullet})$. The positive predictive value (PPV) expresses which fraction of the sites with a positive test result is indeed degraded, and, it is the conditional probability that a site is degraded given a positive outcome of the test: $P(U^{+}|W^{+})$. Similarly, the negative predictive value (NPV) specifies which fraction of sites with a negative test result is pristine. It is the conditional probability a site is pristine given a negative test result: $P(U^{-}|W^{-})$. Although not often used, but instructive for a better interpretation of the concepts, we can also define the complements of these statistics in Box 3.1 (Mayer, 2004). The false alarm fraction (FAF) equals the complement of PPV and specifies which proportion of the sites with a positive outcome is in fact not degraded: $P(U^{-}|W^{+})$. Similarly, the false reassurance fraction (FRF) is the complement of NPV and quantifies the relative proportion that a negative test outcome refers to a degraded site: $P(U^{+}|W^{-})$.

In contrast to the sensitivity and specificity, PPV and NPV are influenced by the relative proportion of degraded and pristine sites in the region. For instance, PPV is about all sites with a positive test outcome which contains degraded as well as pristine sites. The expected number of truly degraded sites with a positive test result is proportional to π^{+}.TPF, i.e. the prevalence multiplied by the proportion of the degraded sites detected which is the sensitivity. Similarly, the expected number of pristine sites but with a positive test result is proportional to $(1-\pi^{+})$.FPF, i.e. the prevalence of no disturbance $(1-\pi^{+})$ multiplied by the fraction misclassified as positive which is FPF. As the predictive value (PPV) equals the proportion of correct positive test results, $PPV = \pi^{+}.TPF / (\pi^{+}.TPF + (1-\pi^{+}).FPF)$. The important message is that the prevalence has an important impact. For instance, if the prevalence is very small, $1-\pi^{+} \approx 1$ and $\pi^{+}.TPF << (1-\pi^{+}).FPF$, $PPV \approx \pi^{+}.(TPF/FPF)$ or PPV is proportional to π^{+}. Thus PPV is the result of both the intrinsic diagnostic quality (TPF/FPF) and the prevalence. Therefore PPV (and NPV) are measures of the operational diagnostic accuracy; assessing how useful the index is in practical circumstances.

3.2.1.6. The relation between intrinsic and operational diagnostic accuracy

As illustrated for the PPV in the previous paragraph, both groups of diagnostic accuracy measures can be derived from each other. Box 3.2 gives a derivation inspired on Motulsky (1995) leading to a Bayesian interpretation of a test. The odds of the positive predictive value (PPV/(1-PPV)) equals the odds of the prevalence $(\pi^{+}/(1-\pi^{+}))$ multiplied with the ratio TPF/FPF. Interpreting the prevalence as the *a priori* probability of a site to be degraded before testing, and PPV as the *a posteriori* probability to be degraded after the test, the ratio TPF/FPF quantifies the gain in information by testing.

Box 3.2 **Relation between intrinsic and operational measures of diagnostic accuracy and Bayesian interpretation.** The odds of predictive values (PPV and NPV) are equal to the odds of the prevalence multiplied by the ratio of the intrinsic measures of diagnostic accuracy. Interestingly, the true restoration fraction (TRF) is equivalent to the positive predictive value (PPV) linking the quality of allocation of the budget with informativeness of the index.

▷▷▷ *Positive predictive value* (PPV) $\boxed{PPV \triangleq P(U^+ \mid W^+)}$

$$\boxed{odds(PPV) = \frac{TPF}{FPF} odds(\pi^+)} \quad \Rightarrow \quad \boxed{logit(PPV) = logit(\pi^+) + \log(\frac{TPF}{FPF})}$$

$$\triangleq \frac{PPV}{1-PPV} = \frac{P(U^+ \mid W^+)}{1-P(U^+ \mid W^+)} = \frac{P(U^+ \mid W^+)}{P(U^- \mid W^+)} = \frac{P(U^+ \mid W^+)P(W^+)}{P(U^- \mid W^+)P(W^+)} = \frac{P(U^+, W^+)}{P(U^-, W^+)}$$

$$= \frac{P(W^+ \mid U^+)P(U^+)}{P(W^+ \mid U^-)P(U^-)} = \frac{TPF}{FPF}\frac{\pi^+}{1-\pi^+} = \frac{TPF}{FPF} odds(\pi^+)$$

▷▷▷ *Negative predictive value* (NPV) $\boxed{NPV \triangleq P(U^- \mid W^-)}$

$$\boxed{odds(NPV) = \frac{TNF}{FNF} odds(\pi^-)} \quad \Rightarrow \quad \boxed{logit(NPV) = logit(\pi^-) + \log(\frac{TNF}{FNF})}$$

$$\triangleq \frac{NPV}{1-NPV} = \frac{P(U^- \mid W^-)}{1-P(U^- \mid W^-)} = \frac{P(U^- \mid W^-)}{P(U^+ \mid W^-)} = \frac{P(U^- \mid W^-)P(W^-)}{P(U^+ \mid W^-)P(W^-)} = \frac{P(U^-, W^-)}{P(U^+, W^-)}$$

$$= \frac{P(W^- \mid U^-)P(U^-)}{P(W^- \mid U^+)P(U^+)} = \frac{TNF}{FNF}\frac{\pi^-}{1-\pi^-} = \frac{TNF}{FNF} odds(\pi^-)$$

▷▷▷ *Bayesian interpretation*

$$\boxed{Post\text{-}test\ odds = pretest\ odds \cdot likelihood\ ratio} \qquad \boxed{likelihood\ ratio = \frac{TPF}{FPF}}$$

▷▷▷ *True restoration fraction* (TRF) = *positive predictive value*

$$\boxed{TRF = \frac{\cancel{C_R}\,\pi^+ TPF}{\cancel{C_R}\,\pi^+ TPF + \cancel{C_R}\,(1-\pi^+)FPF} = PPV}$$

It is important that the contrast between the sensitivity and the false positive fraction is as large as possible. If TPF = FPF, the ratio is one or the index does not contribute to a prediction of the real status of the ecosystem. We have an indifferent index, not discriminating between the two states of the ecosystem condition. Still worse, if TPF < FPF, the information about the true status decreases. The *a posteriori* probability of the disease will be smaller than the *a priori* probability. The index gives the wrong information. In this case, reversing the decision rule (i.e. interchange the conclusions: decide pristine instead of degraded and vice versa) corrects the situation. For this reason, we can restrict our discussion, to ROC curves above the TPF = FPF.

The factorisation of the (odds of the) predictive values as a product of the (odds of the) prevalence and the likelihood ratio highlights how both the intrinsic quality of the index and the operational context determine the predictive value. When applying the index, we should pay attention to the operational context in which the index is applied. If the prevalence in the target region is small, it will be very hard to achieve a high PPV requiring a high intrinsic diagnostic accuracy which can be very expensive. An alternative strategy is to think about an approach to make the prevalence higher. One possibility is to focus at a population at risk. This is possible if we can define strata in the population with different risk levels. Another strategy is two work in two steps. First using a cheap index to screen, and then using a more expensive index for a confirmative test.

3.2.2 ROC curves

3.2.2.1. The binormal model with equal variance

As already stated, because of intrinsic natural variation, model imperfections and/or measurement errors, some sites will be misclassified. To explore the impact of the variability of the test variable, we assume EQM is normally distributed with a homoskedastic error, meaning that the variance of the test variable under pristine and degraded conditions is equal. This idealisation facilitates to clarify the principles, without altering the main conclusions. It can be shown that most of the relations hold for many other even irregularly shaped distributions, but the actual numerical results can be quite different.

Figure 3.1 shows the density function and (cumulative) distribution function of the test variable for two hypothetical IBIs (A and B). For both indices, as assumed above, the distribution curve is shifted to the left because of disturbance. The test variable of degraded sites is stochastically lower than the test variable of pristine sites. For each index, the curve on the right represents the reference distribution modelling the variability of the EQM under pristine or reference conditions (R), while the curve on the left is the distribution under disturbance (D). For index A, the distributions strongly overlap and the discrimination is poor. Index B is more powerful because of two reasons: the distributions are better separated and the variability of EQM is smaller.

3.2.2.2. The impact of the threshold

For both indices, we set the decision threshold T such that FPF is 20 %. For index A, this corresponds with a sensitivity of 36.2 % (likelihood ratio = 1.81), while for index B, the sensitivity amounts to 95.8 % (likelihood ratio = 91.2). We can derive FPF and TPF graphically. The area under the density curve left from T (top panels of Figure 3.1) is proportional to the probability of having a value below the decision threshold. This probability equals FPF for the reference distribution and TPF for the degradation distribution. Because the cumulative distribution is the integral of the density distribution, we can directly read these two values from the cumulative curves. When working with sample data, the cumulative curves can be easily estimated with the empirical distribution function (EDF) which is smoother than the density function (integration makes data smoother).

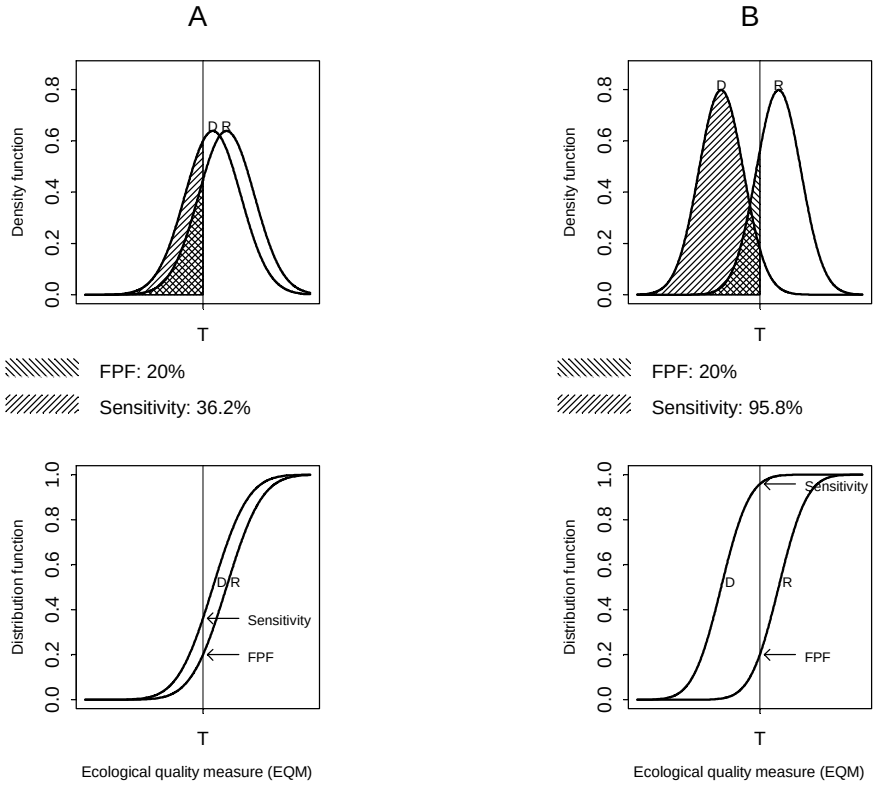


Figure 3.1 **Binormal equal variance model of a test variable to distinguish between two conditions.** The normal distributions represent the natural variation of the ecological quality measure (EQM) for two different indices or biomarkers (A & B) under reference (R) and degraded (D) conditions. The shape of the distribution does not change by degradation, there is only a shift. For index B, the separation between the two distributions is larger than for A. The top panels contain the density function, the bottom panels the corresponding cumulative distributions. T represents the decision threshold. If EQM is smaller than T, it is decided that the site is degraded. The area under the density curves left from T represents the false positive fraction (FPF) for the reference sites (a wrong conclusion) and the true positive fraction (TPF) for degraded sites (a correct decision). As the cumulative curves equal the area under the curve of the density function, FPF and TPF can be read directly from the cumulative distributions. T is set such that the FPF = 20 %.

The sensitivity for index A is small; only about one site in three will be detected. We can try to improve TPF by moving the decision threshold upwards. As is immediately clear graphically (Figure 3.1), this is not possible without increasing FPF. When moving the threshold to the right, the area under the density curves left from the threshold increases for both distributions and, by definition, both cumulative curves climb to higher values. For a given index, we cannot simultaneously improve sensitivity and specificity. Increasing T improves the sensitivity, but the specificity deteriorates, and, conversely, decreasing T results in a higher specificity but a lower sensitivity.

This inverse relationship becomes also clear from a more abstract reasoning. Lowering the decision threshold implies a more conservative approach: EQM values should be smaller before it is decided a site is disturbed. This is advantageous for the reference situation as less sites will be classified as degraded (= specificity higher). However, as the same rule applies for every site investigated, also less really degraded sites will be detected (= sensitivity smaller). The converse holds for increasing T.

In summary, there is an intrinsic tradeoff between sensitivity and specificity. Given the index, setting the decision thresholds for one of the parameters (say, FPF), immediately fixes the value of the other parameter (TPF), and vice versa. We have only one degree of freedom and we should seek for a compromise. If none of the choices of the thresholds results in an acceptable balance, we should look for a better alternative and/or develop another index with better characteristics. A tool visualizing this fundamental tradeoff for all possible thresholds and giving insight in the relative performance of indices, is the ROC curve to be introduced in the next section.

3.2.2.3. Definition of the ROC curve

The dependence of sensitivity and specificity on the decision threshold implies we are faced with an infinite number of combinations. The Receiver Operator Characteristic (ROC) curve synthesises this information in a single line. It plots the sensitivity of a diagnostic test (TPF) against its false positive fraction (FPF = the complement of the specificity). Figure 3.2 presents the ROC curves for two indices of Figure 3.1. At one glance, it is apparent that the diagnostic capacity of index B is (much) higher than index A. Over the full range of FPF, the ROC curve of index B is above that of index A. For any value of FPF (vertical line), the TPF of index B is superior, or for any value of the TPF (horizontal line), the FPF of index B is smaller.

The ROC curve characterises the diagnostic accuracy by showing all and only possible combinations of sensitivity and specificity. The lower left side corresponds a stringent and conservative policy with low decision thresholds of T keeping FPF small often at the expense of TPF. The upper right side corresponds with a liberal policy with high thresholds T increasing TPF but allowing for a high FPF. A powerful or strong index has the capacity to realise a high TPF at small FPF values. In the limit, the ideal gold standard index realises a sensitivity of one at FPF = 0. In contrast, for a weak index, one has to allow for a high FPF to realise a high sensitivity. In the limit, we have an indifferent index for which TPF = FPF over the full range (ROC curve = the positive diagonal) which is of no use at all.

The ROC curves allows to choose the decision threshold. For index B, it may be meaningful to set the decision boundary at a FPF of 10 %, but not for index A for which the sensitivity is too small. It is necessary to accept a higher FPF to have a useful index, for instance, by setting the sensitivity and specificity equal to each other. This results in a sensitivity of 56 % for index A, at the expense of a high FPF (44 % = 100 % – 56 %). In some situations, this can be acceptable if the benefit of detection and subsequent treatment is very high in comparison to no detection (and no treatment).

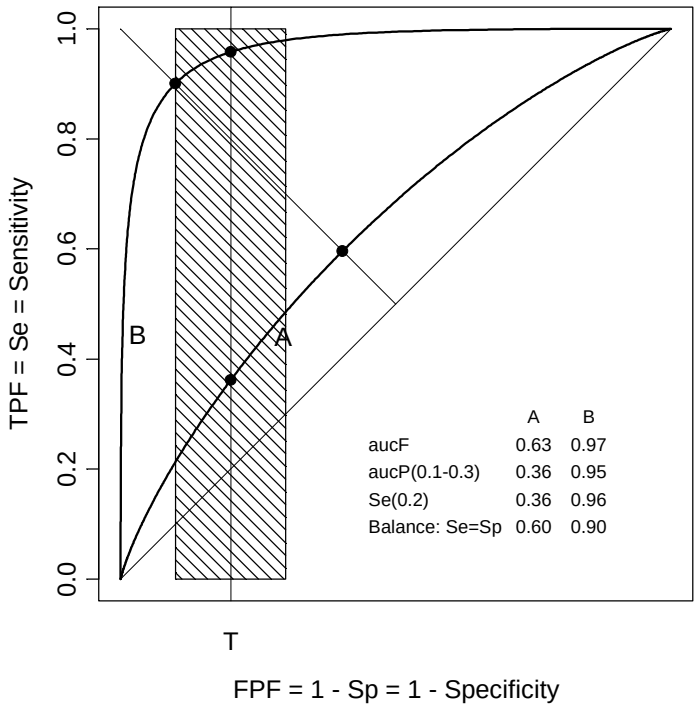


Figure 3.2 **The ROC-curve and measures of diagnostic accuracy** for the two hypothetical binormal equal variance indices A and B of Figure 3.1: (i) aucF = full AUC; (ii) aucP = partial AUC for $0.1 \leq FPF \leq 0.3$ (shaded area); (iii) Se(0.2) = TPF(0.2) = sensitivity at FPF of 20 % and (iv) sensitivity at balance with equal sensitivity and specificity (the diagonal line where TPF = TNF = 1 - FPF).

3.2.3 Some additional considerations

3.2.3.1. Mathematical representation of the ROC curve

The decision threshold T uniquely defines a "decision point" on the ROC curve with coordinates (FPF,TPF) = ($F_R(T)$, $F_D(T)$). F_R is the (cumulative) distribution function under reference conditions and F_D is the (cumulative) distribution function for degraded sites as depicted graphically in Figure 3.1. Because of the one-to-one relation of T and a point on the ROC curve, we can use any element of the triplet (T, FPF, TPF) to define a decision point. Elimination of T directly expresses TPF as a function of FPF: $TPF = F_D(F_R^{-1}(FPF))$. Box 3.3 also calculates the first derivative of the ROC curve by the chain rule. For continuous ROC curves, it is simply the ratio of the density functions under degradation and reference conditions. For equal variance models, the curve is concave: at the origin (0,0), $f_D >> f_R$ because f_R is located to the right f_D (Figure 3.1). Close to the origin, the ROC

curve departs sharply and then levels off. If the separation between the distributions is large (index B), the curve will be steeper than for a smaller separation (index A).

Box 3.3 **Mathematical description of the ROC curve.** The parametric representation expresses (FPF,TPF) as a function of the decision threshold T. FPF and TPF are equal to the cumulative value of the test variable under reference and degraded conditions (F_R and F_D) as illustrated in Figure 3.1 (lower panels). By eliminating the common parameter T, the ROC function is obtained. By the chain rule, we can calculate the first derivative. It is the ratio of the density function of the test variable under degraded and reference conditions (f_D and f_R). For the equal variance model (only a shift), the derivative at the origin is infinity and then gradually levels off to 0 in the end point.

▷▷▷ *Parametric representation of ROC curve*

$$\boxed{ROC \;\leftrightarrow\; \begin{cases} FPF = F_R(T) \\ TPF = F_D(T) \end{cases} \quad -\infty < T < +\infty} \;\Rightarrow\; \boxed{T = F_R^{-1}(FPF) = F_D^{-1}(TPF)}$$

$$(FPF, TPF) = \begin{cases} T = -\infty \;\Rightarrow\; (0,0) & (\textit{starting point}) \\ T = +\infty \;\Rightarrow\; (1,1) & (\textit{finishing point}) \end{cases}$$

▷▷▷ *The ROC function (elimination of the threshold T)*

$$\boxed{ROC \;\leftrightarrow\; TPF = F_D(F_R^{-1}(FPF))} \quad \Leftrightarrow \quad \boxed{FPF = F_R(F_D^{-1}(TPF))}$$

▷▷▷ *First derivative of the ROC function (chain rule)*

$$\boxed{\frac{dTPF}{dFPF} = \frac{dTPF}{dT}\frac{dT}{dFPF} = \frac{dTPF}{dT} \Big/ \frac{dFPF}{dT} = \frac{f_D(T)}{f_R(T)}}$$

$$\frac{dTPF}{dFPF} = \begin{cases} T = -\infty \;\Rightarrow\; +\infty & (\textit{starting point}) \\ T = +\infty \;\Rightarrow\; 0 & (\textit{finishing point}) \end{cases}$$

for binormal equal variance model

3.2.3.2. Differential response

Figure 3.1 also represents the same index, but applied in two different situations A and B. There are three important oft occurring situations. Firstly, for different levels of degradation, B depicts the resolution of the index with a severely degraded ecosystem, while A holds when the ecosystem is only moderately degraded. Similarly to power calculations in statistics, at the design stage, we should specify in advance the minimum effect to detect, to judge whether the index suffices. If moderate degradation is considered as important to detect, then the index is not acceptable. Secondly, A and B can represent different types of degradation. Ideally, the index should have a uniform quality for all possible degradation types of the region. This is unfeasible. Therefore, it is important to specify in advance the types of degradation (broad classes) for which the index should be sensitive in the target region and strata (ecoregions). Thirdly, an index is always developed in a

specific context. When extended to other situations or regions, recalibration, or at least a validation study is advisable to be sure about the performance in the new region.

3.2.3.3. Spectrum bias and representative sampling

A further application of Figure 3.1 is spectrum bias (Begg and Greenes, 1983; Ransohoff and Feinstein, 1978; Zhou *et al.*, 2002). Spectrum bias occurs when the sample used for calibration does not cover the total variability of the target region (the statistical population of the index). As a consequence, the performance assessed at the calibration stage of the index will be higher (case B) than in real circumstances (case A). To control for spectrum bias, representative sampling of the target region is necessary. Representative sampling is a difficult point as its realisation can be quite hard and expensive. Hence, it is difficult to convince decision makers but also many scientists about the added value of a truly representative sample, although many authors demonstrate that in the long run the investment pays back (Dauer and Llansó, 2003; Hughes *et al.*, 2000; Llansó *et al.*, 2003; Overton, 1993; Paul *et al.*, 2008; Southerland *et al.*, 2009). As a specific example, an evaluation of IBIs (Fore, 2003) showed that a probability sample (a random selection procedure for which the probability of each possible sample is known and larger than zero (Overton, 1993)) was superior to cover a broader spectrum of pressures in contrast to the intuition of many researchers involved in the project who argued for a judgement based sample. Equally important for the development of the index, a probability sample is also instrumental to assess the relations between variables correctly (Kish, 1987). An IBI has the potential and the ambition to be reactive to a broad range of impacts on the ecosystem. If, however, the sample available for the index construction is not representative for the impacts in the region and does not cover the full gradient of pressures, there is little hope the biotic index will fulfil its aims.

3.2.4 Measures of (intrinsic) diagnostic accuracy derived from the ROC curve

An ROC curve condenses in one single line the intrinsic diagnostic accuracy for all possible decision thresholds. This integration of information is a strong point. Yet, comparing full curves as a whole is not evident, as the curves are estimated from sample data and not smoothed. Therefore, a common solution is to further synthesize the information in summary statistics of the curve.

3.2.4.1. Sensitivity at specific points of the ROC curve

When comparing the diagnostic accuracy of several indices, it is crucial to base the judgements on an equal footing. The sensitivity of one index can be larger than another index, just because the false positive fraction of the former index is larger. Therefore, comparison of indices should be at similar decision points, for instance by fixing FPF in the operational range of the index, i.e. the decision points which are considered to be relevant in practice. For instance, rather seldom we will choose to set FPF as small as 0.05, because in this point the sensitivity will be too low for a practical application. Similarly, FPF = 0.5 is seldom relevant (but it can be) because in this case the burden of FPs is very large (unless the corresponding benefits are higher, a point we work out further). In general, for an IBI, we are interested in values in the range 0.1 – 0.3. Within this range, we can check the sensitivity for a series of FPF values (e.g. 0.1, 0.2 and 0.3). With this solution, we still have to judge several points and, with sample ROC curves, the uncertainty of the

point estimates or the sample variability can be very large. A solution is to take more global "integrated" statistics as presented below.

3.2.4.2. The (full) area under the curve (AUC): aucF

An often used numerical criterion to characterise the diagnostic accuracy in its entirety, is the (full) area under the ROC curve (AUC). To make a distinction with the third alternative, the partial AUC (aucP), to be introduced in the next section, we denote this statistic by aucF (the full AUC). This statistic has several interpretations (Pepe, 2003; Zhou *et al.*, 2002). First, it can be interpreted as the average sensitivity (the y-axis of the ROC curve) when varying the specificity (the x-axis of the ROC curve) from 0 to 1. In addition, it can be shown aucF is the average value of the specificity when ranging the sensitivity from 0 to 1 (Metz, 1989). Hence, a high aucF indicates that the index has a great diagnostic capacity over a broad range of decision points. Because integrals smooth numbers, the statistical variability is lower than the point estimates of the sensitivity for fixed FPF values or of the specificity for fixed TPF values (Zhou *et al.*, 2002). The interpretation of the statistic is not always unequivocal, as ROC curves can cross. Yet, as a measure of the overall performance of a diagnostic test, it is well accepted. However, as AUC is an integrated measure over the full range of the sensitivity or specificity, also values useless for decision making influence aucF. As argued in 3.2.4.1, for IBIs, the relevant range is 0.1 – 0.3. In this respect, the partial AUC – to be defined in next section – has a much better indicative value (see also Figure 3.3).

An alternative interpretation links aucF more directly to the distance of the reference to the alternative distribution. It can be shown that aucF equals the probability that the ecological quality measure (EQM) from a randomly selected pair of pristine and degraded sites are correctly ordered, i.e. $P(EQM_R > EQM_D)$ = aucF (Bamber, 1975). Thus if we have two EQM values, one coming from a reference site and the other from a degraded site, aucF gives the probability we correctly classify them. However, in practice, it is hard to imagine a situation where we know in advance the two observations come from a pristine and degraded site. The statistic aucF is also equal to the p-value of the two-sample Wilcoxon rank sum statistic (Hanley and McNeil, 1982). In the same vein, it can be shown there is a close relation of the distance of the ROC curve and the first diagonal with the Kolmogorov-Smirnov two-sample statistic for comparing two groups (Campbell, 1994). Hence, we can conclude the ROC curve essentially provides a distribution-free description of the separation of the distribution of EQM_R and EQM_D (Pepe, 2003).

The statistic aucF ranges from 0.5 (minimal quality, the indifferent index) to 1 (maximum quality, the gold standard). Values below 0.5 are possible because of sampling variability, but values higher than 1 are never possible. A gold standard index discriminates perfectly and has a sensitivity of 1 over the full range of FPF, or, aucF = 1. On the contrary, an indifferent index does not differentiate at all. There is no difference with a random classification. In this situation, FPF = TPF. The ROC curve is the diagonal line (of indifference) from (0,0) to (1,1) and aucF = 0.5. Indices with aucF < 0.5 exist if the ROC curve is lying below the diagonal line. In this situation, the decision rule is set in the wrong direction. TPF < FPF implies that pristine sites are judged to be degraded in more instances than degraded sits. Reverting the decision rule solves the problem.

3.2.4.3. The partial area under the curve (aucP)

As already suggested, the strength of aucF is also its weakness. It synthesizes in one single number the diagnostic accuracy for all possible decision thresholds, but quite different curves can have a similar aucF. When ROC curves cross, the diagnostic accuracy of one index is not uniformly better over the total range. In many situations, the total ROC curve is not of practical value since FPF is too high or TPF is too low, but still these parts of the curve considerably contribute to aucF. An index can have the best overall performance, but does not function optimally in the operational region, for instance, from FPF = 0.1 to 0.3.

Box 3.4 **The partial area under the ROC curve (aucP).** The partial AUC (aucP) is the general formula comprising the full AUC (aucF) and the sensitivity (TPF) at a fixed point of FPF = f_m. We defined f_m as the midpoint of the interval $[f_1,f_2]$. Hence, in the limit for $\Delta f = f_2 - f_1 \rightarrow 0$, aucP $\rightarrow$ TPF(f_m).

$$aucP(f_1, f_2) = \frac{1}{f_2 - f_1} \int_{f_1}^{f_2} TPF(FPF)\,dFPF$$

$$\rightarrow \quad aucP(f_1, f_2) = \frac{1}{\Delta f} \int_{f_m - \Delta f/2}^{f_m + \Delta f/2} TPF(FPF)\,dFPF = aucP(f_m - \frac{\Delta f}{2}, f_m + \frac{\Delta f}{2})$$

$$with: f_m = \frac{f_1 + f_2}{2} \quad \& \quad \Delta f = f_2 - f_1$$

$$\rightarrow \quad aucF = aucP(0,1) = \int_0^1 TPF(FPF)\,dFPF$$

$$\rightarrow \quad Se(f_m) = TPF(f_m) = aucP(f_m, f_m) = \lim_{\Delta f \to 0} aucP(f_m - \frac{\Delta f}{2}, f_m + \frac{\Delta f}{2})$$

This reasoning results in the partial AUC, averaging TPF from FPF = f_1 to f_2 (Dodd and Pepe, 2003). The ROC curve of Figure 3.2 gives aucP for an FPF interval from 0.1 to 0.3. A full notation is aucP(f_1,f_2). By default, aucP equals aucP(0.1, 0.3). The lower limit 0.1 is chosen to guarantee a sufficiently high sensitivity. Very close to the origin TFP approaches zero. The upper limit controls FPF to less than one third of the pristine sites misclassified as degraded. We chose above values on our experience with the well-developed European Fish Index (Quataert *et al.*, 2007) for which the balanced misclassification error (TPF and FPF) was around 20% which is reasonable for ecological data. It should be clear that for each application this choice should be reconsidered taking into account the relative costs as we will elaborate further on.

The aucP statistic has good statistical properties. It is an integrated measure less variable than the point-estimate TPF at a certain FPF (Zhou *et al.*, 2002). Its interpretation is better than aucF. As illustrated by Figure 3.3, averaging TPF over a narrow range of FPF keeps its relation with the TPF values at fixed points of FPF. Instead for aucF, integrating over the full range, this connection is

totally lost. However, the focus can be felt as subjective, because it allows to gerrymander, i.e. to choose the interval optimising the diagnostic accuracy in a specific case. A correct way is to specify the range in advance, without looking at the data. Also, choosing a range can make a comparison between published results more difficult because of the different ranges (Zhou *et al.*, 2002). In addition, particularities outside the range can escape attention. For all these reasons, we believe it is best to use aucP in complement to aucF when model building (see Chapter 5).

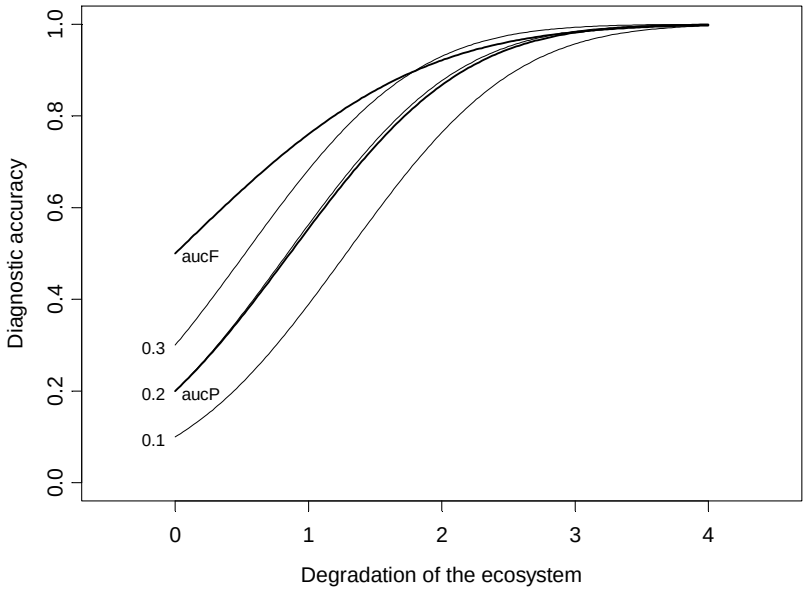


Figure 3.3 **Relation between AUC measures as a function of the degradation.** The lines in bold represent the partial AUC (aucP(0.1,0.3)) and the full AUC (aucF); the other finer lines the sensitivity for fixed values of the FPF in the range of aucP. These lines are similar to power curves for an increasing difference between the null and alternative hypothesis.

3.2.4.4. Relations between the diagnostic measures

Box 3.4 defines aucP mathematically and makes a link with the other summary statistics of the ROC curve. In fact, we choose a variant of aucP normalising the statistic by dividing the integral over its range, sometimes referred to as the partial area index (McClish, 1989; Jiang *et al.*, 1996). Interestingly, in its normalised format, aucP encompasses the two other summary statistics of the ROC curve. In the limit of shrinking the interval to zero, aucP = TPF(FPF), i.e. the sensitivity at a specific false positive fraction. Conversely, expanding the integration interval to the full FPF range gives aucP = aucF. Figure 3.3 compares the measures of diagnostic accuracy for an increasing degradation of the ecosystem assuming a homoskedastic binormal model. Evidently, these curves resemble the sigmoid power curves of statistical tests. Without degradation (at the origin), the

reference and alternative distribution coincide such that TPF = FPF, aucF = 0.5 and aucP(0.1,0.3) = 0.2. Interestingly, with increasing separation between degraded and pristine sites, the curve for aucP does not differ much from the curve with FPF = 0.2. This is partly due to the choice of the underlying model (homoskedastic binormal). Yet, with alternative skewed distributions based on the beta distribution, a similar close relationship was found (results not shown).

Box 3.5 **Extension of measures of diagnostic accuracy to an ordinal tri-state variable**: FN and FP can be easily generalised. FP: misclassification for which the outcome of the index is worse than in reality, FN: misclassification for which the outcome of the index is better than in reality. U represents the unknown status, W = outcome of the index.

▷▷▷ $3\text{-}level\ ordinal\ classification: reference\ (R) / moderate\ (M) / degraded\ (D)$

$$\boxed{EQC = \begin{cases} "D" & \Leftrightarrow & EQM \le T_M \\ "M" & \Leftrightarrow & T_M < EQM \le T_R \\ "R" & \Leftrightarrow & T_R < EQM \end{cases}} \quad \leftrightarrow \quad WFD \quad \begin{cases} 4-5\ (low\ quality) \\ 3\ (moderate\ quality) \\ 1-2\ (high\ quality) \end{cases}$$

▷▷▷ *Confusion matrix*

$\boxed{R,M,D}$	$W=R$	$W=M$	$W=D$
$U=R$	TC	FP	FP
$U=M$	FN	TC	FP
$U=D$	FN	FN	TC

$\Rightarrow$

$\boxed{R,M,D}$	$W=R$	$W=M$	$W=D$
$U=R$	$TCF_{R,R}$	$FPF_{R,M}$	$FPF_{R,D}$
$U=M$	$FNF_{M,R}$	$TCF_{M,M}$	$FPF_{M,D}$
$U=D$	$FNF_{D,R}$	$FNF_{D,M}$	$TCF_{D,D}$

▷▷▷ $Separate\ binary\ contrasts \Rightarrow aucF_{R/M} / aucP_{R/M}, ...$

$\boxed{R/M}$	$W=R$	$W=M$
$U=R$	$TNF_{R/M}$	$FPF_{R/M}$
$U=M$	$FNF_{R/M}$	$TPF_{R/M}$

$\boxed{R/D}$	$W=R$	$W=D$
$U=R$	$TNF_{R/D}$	$FPF_{R/D}$
$U=D$	$FNF_{R/D}$	$TPF_{R/D}$

$\boxed{M/D}$	$W=M$	$W=D$
$U=M$	$TNF_{M/D}$	$FPF_{M/D}$
$U=D$	$FNF_{M/D}$	$TPF_{M/D}$

▷▷▷ $Successive\ binary\ constrasts \Rightarrow aucF_{R/M-D} / aucP_{R/M-D}, ...$

$\boxed{R/M-D}$	$W=R$	$W=M-D$
$U=R$	$TNF_{R/M-D}$	$FPF_{R/M-D}$
$U=M-D$	$FNF_{R/M-D}$	$TPF_{R/M-D}$

$\boxed{R-M/D}$	$W=R-M$	$W=D$
$U=R-M$	$TNF_{R-M/D}$	$FPF_{R-M/D}$
$U=D$	$FNF_{R-M/D}$	$TPF_{R-M/D}$

3.2.4.5. Extensions to ordinal variables

ROC curves are primarily developed for binary classification schemes. However, this methodology can be extended to ordinal classifiers. This is necessary when the IBI is aimed at discriminating several degradation stages. Here, Box 3.5 presents some possibilities for a three level ordinal

variable. The approach can easily generalised to more levels. The first table retains the full detail, the two other tables transform the ordinal variable to a series of binary contrasts. For latter two tables, we can use the original terminology, however for the former, some care is necessary. We still can interpret the off-diagonal elements (misclassifications) as FPs if the classification of a site is worse than in reality, and as FNs if the classification of a site is better than in reality. However, for the diagonal elements (correct classification), it is not directly clear how to talk about TPs and TNs. Therefore we use true classification (TC) and true classification fraction (TCF). For the successive binary contrast, we pool classes. For instance, R/M-D is the contrast between R and the combination of M & D.

3.2.5 ROC curve estimation

3.2.5.1. Synoptic diagrams

The Mann-Whitney U-test (or equivalently the Wilcoxon two-sample rank sum test) is a non-parametric rank-based test sensitive to whether one distribution is stochastically higher than the other, i.e. shifted upwards without crossing (Lehmann, 1975). If both the null and alternative distribution have the same shape, the test is sensitive for a shift in location and a significant result can be interpreted as a shift in the median. An index with non-crossing cumulative distributions as a function of the ecosystem degradation is exactly what we are aiming for when developing an IBI. The ecological quality measure (EQM) should be a yardstick of which the distribution is consistently shifted downwards (or upwards) such that its reference distribution is stochastically higher than the distribution under degradation.

Above interpretation of the Mann-Whitney U-test makes plausible that the test is equivalent with aucF as first noticed by Bamber (1975; Hanley and McNeil, 1982). In fact, boxplots, cumulative distributions, and ROC curves are three complementary graphical tools essentially investigating the same underlying statistical hypothesis, namely exploring whether there is a consistent response of the index to human impact. Because of this close relationship, we can use aucF (or aucP) to synthesize the response of the ecological indicators or indices to anthropogenic stress. In the results (Figure 3.11), we will illustrate this relationship to screen the potential of candidate metrics with a synoptic diagram comparing at one glance the response to an anthropogenic stress gradient of a series of candidate metrics. Surely, boxplots and cumulative distributions remain useful for a more profound exploratory data analysis (EDA), for instance, to pick up outliers or other anomalies in the distribution, or to investigate the shape of the distributions.

3.2.5.2. The empirical ROC curve

The cumulative distribution function F can be estimated from a sample of n observations by the empirical distribution function (EDF) defined as follows:

$$EDF \leftrightarrow y = \hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x) = \frac{number\ of\ \ items\ in\ sample \le x}{n}$$

The EDF is simply equal to the proportions of items in the sample smaller or equal than x. I() is the indicator function which equals 1 if the test within brackets is true, and 0 otherwise:

$$I(x_i \le x) = \begin{cases} 1 & if\ x_i \le x \\ 0 & if\ x_i > x \end{cases}$$

If the test variable EQM is positively associated with ecological quality, as assumed by default, then the cumulative distributions of degraded (F_D) and reference sites (F_R) represent TPF and FPF as a function of the decision threshold T (lower panels of Figure 3.1). Hence, we can estimate the ROC curve empirically from the EDF of pristine and degraded sites (Shapiro, 1999). By varying T, we obtain the empirical ROC with in the x-axis the estimated FPF and in the y-axis the estimated TPF.

$$\widehat{ROC} \leftrightarrow \begin{cases} x = \widehat{FPF} = \hat{F}_{R;n_R}(T) = \dfrac{1}{n_R}\sum_{i=1}^{n_R} I(EQM_{R;i} \le T) & \{EQM_{R;i}\} = sample\ of\ \ reference\ sites\ (size\ n_R) \\ y = \widehat{TPF} = \hat{F}_{D;n_D}(T) = \dfrac{1}{n_D}\sum_{i=1}^{n_D} I(EQM_{D;i} \le T) & \{EQM_{D;i}\} = sample\ of\ \ degraded\ \ sites\ (size\ n_D) \end{cases}$$

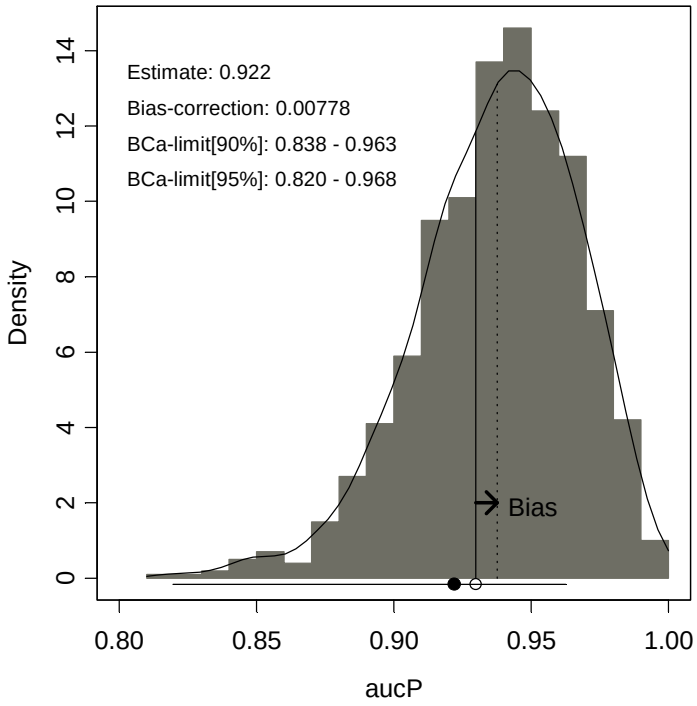


Figure 3.4 **Principle of the accelerated bias-corrected percentile method** (BCa method (Efron, 1987)). The white point is the estimate of aucP derived from the original sample. The histogram gives the resampled distribution (n_B = 1000) of aucP of which the mean is represented by the dotted line. The shift with respect to the original estimate assesses the bias. Subtracting the estimated bias from the original estimate, results in the bias-corrected estimate for aucP (the black point). The confidence limits are based on the percentiles of the sampling distribution, but a correction is necessary to improve statistical properties as described by Efron (1987).

3.2.5.3. Estimation of aucF and aucP and bootstrapping

A nonparametric estimate of aucF is the trapezoidal area under the empirical ROC curve, which equals the Mann-Whitney form of the two-sample Wilcoxon rank-sum statistic (Hanley and McNeil, 1982; Shapiro, 1999). Similarly, we apply the trapezoidal rule to estimate the partial area under the ROC curve as illustrated in Figure 3.9. The statistical sampling distributions of aucF and aucP are not well known and complex (Zhou *et al.*, 2002). They critically depend on the shape of the distribution. In this situation, resampling methods as bootstrapping are ideal tools to estimate the standard error and confidence limits (Efron and Tibshirani, 1993; Shao and Tu, 1995; Lunneborg, 1999; Davison and Hinkley, 2006). Measures of diagnostic accuracy are bias prone. The estimate is often higher than the real value. Because of the fitting, the model is partly bend towards the data risking to give a wrong impression of the diagnostic accuracy. Again, bootstrapping can be used to correct for bias. We shortly sketch the principle visualised in Figure 3.4. For bootstrapping, the sample is considered as the population and the any parameter calculated from this sample is considered as the true value. Resampling from this population allows to evaluate the sampling distribution of the parameter. For each resample, the parameter is estimated and by repeating the procedure the sampling distribution is reconstructed. The difference between the mean of this sampling distribution and the parameter estimated from the original sample, is an estimate of the bias. Subtracting the estimated bias from the observed statistic, allows to correct for bias. From the percentiles of this bootstrapped sampling distribution also the confidence limits can be derived as we also have information about the variability of the estimate. Based on this principle, Efron (1987) developed the accelerated bias-corrected percentile method considerably improving the coverage of the confidence intervals for small samples.

3.2.6 Utility curves derived from the ROC curve

ROC curves serve as a tool for a utility analysis, by quantifying the consequences (monetary, but also ecological, economical and societal implications) of index-based decisions. To investigate the link between the ROC curve and its utility, a crucial assumption is that the index is effectively integrated in decision making to choose between different options. Otherwise, data is merely accumulated with at best a vague link to management decisions. In turn, the utility curves give a deeper understanding of ROC curves by showing how the position (the height and the shape) of the curve characterises the index strength, i.e. its capacity to realise a high sensitivity keeping the false positive fraction low, and determines its utility. For a particular index, we cannot "escape" the ROC curve. It describes all possible combinations of sensitivity and specificity. The sensitivity represents the gain realised by using the index, the false positive fraction represents the harm the index cannot avoid (in this sense, the specificity expresses the harm avoided by using the index).

Similarly to ROC curves, the utility curves evaluate the usefulness of the index as a whole and plot measures of the utility as a function of the decision threshold. Because of the one-to-one relation of (T, FPF, TPF) for a given index (Box 3.3), we can choose any element of this triplet as the argument of utility functions, and, if convenient, we can use a mixture (Box 3.6). For a particular utility measure, following formats are possible: $U = u_1(T) = u_2(FPF) = u_3(TPF) = u_4(FPF, TPF)$. The first function ($u_1$) is not used here, but is relevant in a practical setting, when choosing the

appropriate decision threshold. For a graphical representation of the curves, u_2(FPF) or u_3(TPF) are most convenient. To tighten the relationship with the underlying ROC curve, we will choose FPF as the x-axis for u_2(FPF), and TPF as the y-axis for u_3(TPF). The fourth format is most instructive to link the utility with the decision point on the ROC curve.

To assess the utility of an index, we adopt two complementary perspectives (Gold *et al.*, 1996). The cost effectiveness analysis (CEA) evaluates the costs in relation to ecological (and/or societal) targets, for instance, the proportion of degraded sites we want to restore, without expressing the targets in monetary units. If the effects are (assumed to be) equal, a CEA resolves to a special case: a cost minimisation analysis (CMA). A step further is a cost benefit analysis (CBA) which expresses the (ecological and societal) targets in monetary terms. The aim of CBA is to maximise the overall benefit. Although in principle, CBA is the more comprehensive method, we should consider both approaches as complementary as it is not evident to value ecological and societal values in monetary terms. If the outcomes of both approaches are in concordance with each other, the decision is easy. However, disagreement is not necessarily bad, as it can initiate a discussion about fundamental options to choose between.

The next theoretical paragraphs define utility curves and give a concise mathematical discussion assuming concave ROC curves. In the results section, we discuss the utility curves as a function of the quality of the index. We assume the ROC curves are well shaped concave curves lying above the positive diagonal. In reality, this is not always true and we can have "improper" ROC curves having a "hook" at the bottom left (Zhou *et al.*, 2002; Pan and Metz, 1997). As our focus is to develop a global insight in the relation between the strength of the index and its utility, we make abstraction of this and similar problems. In particular cases, verification is always necessary.

3.2.6.1. Utility measures for a cost effectiveness (CEA) (Box 3.6)

3.2.6.1.1. The average restoration cost (ARC)

A first utility function is the average restoration cost (ARC). It is the expected restoration cost averaged over all sites covered by the restoration program. With perfect knowledge (the gold standard index), ARC equals $C_R.\pi^+$, i.e. the restoration cost (per site) C_R multiplied with the prevalence of degraded sites π^+. In reality, only a part of the degraded sites (TPF) is detected and resources get lost because some pristine sites (FPF) will be misclassified. The expected expenditure is $C_R.\pi^+.TPF$ for degraded sites and $C_R.(1-\pi^+).FPF$ for pristine sites, resulting in ARC = $C_R.\pi^+.TPF$ + $C_R.(1-\pi^+).FPF$. As C_R is a common parameter, we can drop it to study the impact of the index and analyse the function $\pi^+.TPF + (1-\pi^+).FPF$ to which we refer to as the ARC curve or function.

From the format of this function (type u_4), we can easily derive the impact of the strength of the index. All ARC curves start and end in the same point: ARC(0,0) = 0 (no sites are restored) & ARC(1,1) = C_R (all sites restored). In between, ARC strongly depends on the strength of the index. When fixing the sensitivity (TPF) at a target value, for the weaker index A, we have to allow for a higher FPF than for index B ($FPF_A > FPF_B$) to realise the same TPF (see Figure 3.2). Hence, $ARC_A > ARC_B$. With a strong index B, the target sensitivity is reached at a lower restoration cost. However, for a full cost picture, we should take into account the assessment cost C_A which will be higher for

index B. As will be elaborated in the cost-effectiveness analysis (CEA) in Chapter 4, we have to minimise the total monetary cost $C_M = C_A$ + ARC. Note that in this case, we should know C_R, or at least its relation with C_A.

Box 3.6 **Utility functions based on a cost effectiveness analysis (CEA).** The average restoration cost (ARC), true restoration fraction (TRF) and total management cost (C_M). For ARC and TRF, the arrows give the evolution of the utility curve from origin and endpoint on ROC curve. For an explanation of the derivations, see text.

▷ *Average restoration cost* (ARC)

$$\boxed{AVC = C_R(TP) + C_R(FP) = C_R\pi^+ TPF + C_R(1-\pi^+)FPF} \qquad (0,0) = 0 \;\nearrow\; (1,1) = C_R$$

▷ *True restoration fraction* (TRF) = *positive predictive value* (PPV)

$$\boxed{TRF = \frac{C_R\pi^+ TPF}{ARC} = \frac{\cancel{C_R}\pi^+ TPF}{\cancel{C_R}\pi^+ TPF + \cancel{C_R}(1-\pi^+)FPF} = PPV} \qquad (0,0) = 1 \;\searrow\; (1,1) = \pi^+$$

▷ *Minimise total monetary cost* (C_M) $(C_A = \text{assessment costs})$

$$\boxed{C_M = ARC + C_A} \quad \downarrow\downarrow\downarrow$$

3.2.6.1.2. The true restoration fraction (TRF)

Because of misclassification, ARC is larger than when only degraded sites are restored. The true restoration fraction (TRF) evaluates which proportion of the budget is correctly allocated to degraded sites. As derived in Box 3.6, TRF is equivalent to the positive predictive value (PPV) as the restoration cost C_R cancels out from the ratio. The equivalence between PPV and TRF links the informativeness of the index with its utility to have a high quality allocation of the resources.

The TRF curves start and end in the same points: TRF(0,0) = 1 and TRF(1,1) = π^+. Very close to the origin (in the limit), TPF climbs fast as a function of FPF even for weak indices. Hence, TPF >> FPF, such that TPF.π^+ dominates the denominator: TRF = TPF.π^+/(TPF.π^+ + FPF.(1-π^+)) ≈ 1. At the end, TRF(1,1) = π^+. Without any discrimination, all sites are restored implying that the fraction of correctly restored sites equals the prevalence. In between the end points, the quality of an index is important. A similar reasoning as for ARC, reveals $TRF_A < TRF_B$ to realise the same sensitivity because $FPF_A > FPF_B$. The more powerful index B results in a better allocation of the budget.

3.2.6.2. Utility measures for a cost benefit analysis (CBA) (Box 3.7)

3.2.6.2.1. The overall restoration benefit (ORB)

We now express the possible (cost) consequences of the restoration in monetary units (or some other common yardstick). We give a derivation in a nutshell. Chapter 4 will explore the information necessary for the calculations. We define E_G = the ecological gain by restoring a degraded site, $R_G = E_G - C_R$ = the (cost-corrected) gain after correcting for restoration costs (subtraction!), $T_G = \pi^+.(E_G - C_R)$ = the total gain by restoring all degraded sites in the region, and T_G.TPF = the gain realised by basing the decisions on the index. Similarly for pristine sites, we define E_H = the ecological harm caused by restoring a pristine site, $R_H = E_H + C_R$ = the harm augmented with the restoration cost (addition!), $T_H = (1-\pi^+).(E_H + C_R)$ = the total harm by restoring all pristine sites, and T_H.FPF = the harm caused by using the index. Subtracting harm from benefit results in the overall restoration benefit: $ORB = T_G.TPF - T_H.FPF = T_G.b_{ROC}$ with b_{ROC} = TPF – FPF/b = the benefit function and $b = T_G/T_H$ = the benefit ratio. To analyse the benefit function, we should not know T_G and T_H. A good estimate of b is sufficient. This situation changes if we also intend to incorporate the assessment costs C_A (Chapter 4).

ORB changes along the ROC curve proportionally to TPF – FPF/b ≤ TPF ≤ 1. Irrespective of the index strength, the benefit curve starts and ends in the same point: $b_{ROC}(0,0) = 0$ and $b_{ROC}(1,1) = 1 - 1/b$. A maximum is reached where the first derivative of the ROC curve equals 1/b. ORB appears to be proportional to a function which is composed of the ROC curve itself (TPF) minus a penalty term (FPF/b) which directly depends on the false positives, but divided by the benefit ratio b. As b expresses the balance between gain and harm, this is logical. If b is high, the negative effect of FPs is limited. The converse is true if b is small (especially if b < 1). As for the TRF curve, the benefit curve makes clear that the key factor determining the optimum is the capacity of the index to realise a high sensitivity keeping FPF low.

It is instructive to notice that b can be factorised as the odds of the prevalence odds(π^+) and what we call the intrinsic benefit ratio $b_R = R_G/R_H$ which expresses at the scale of the site the contrast between gain and harm of restoration. To judge the potential success of the restoration, we should also consider the prevalence of degradation. If the prevalence is small, a strong but expensive index will be necessary to keep the penalty small. In this situation, it can be more attractive to target the restoration program at a stratum at risk, i.e. a subpopulation of which it is known of having a higher prevalence.

3.2.6.2.2. The ecological restoration benefit (ERB)

The ecological restoration benefit function (ERB) is totally analogous to ORB, except that it is not corrected for the costs of restoration. Instructively, both approaches are mathematically linked: ORB = ERB – ARC. As it should be! The formula makes transparent how the restoration costs diminish ORB. Minimising costs will contribute to a higher ORB, but not necessarily minimal costs will correspond to an optimal benefit. If ecological benefits dominate, the optimal solution based on purely ecological basis can be quite different from a monetary analysis.

In the same vein, it can be shown that if we take the assessment costs into account, the ecological benefit corrected for all monetary costs, B_M = ORB – C_A, equals ERB – C_M, i.e. the ecological benefit minus the monetary costs defined above as the sum of the assessment and the restoration costs. Optimising for C_M is equivalent to neglecting ERB and setting it zero.

Box 3.7 **Utility functions based on a cost benefit analysis (CBA)**. Overall restoration benefit (ORB), Ecological restoration benefit (ERB) and ecological benefit corrected for all monetary costs (B_M). (0,0) & (1,1) = origin and endpoint of ROC curve

▷ *Overall restoration benefit* (ORB) & *benefit function* (b_{ROC})

$$\boxed{ORB = B_{ROC} = T_G TPF - T_H FPF = T_G b_{ROC}} \qquad \begin{cases} T_G = (E_G - C_R)\pi^+ = R_G\pi^+ \\ T_H = (E_H + C_R)(1-\pi^+) = R_H(1-\pi^+) \end{cases}$$

$$\boxed{b_{ROC} = TPF - FPF/b} \quad \boxed{b = \frac{T_G}{T_H} = odds(\pi^+)b_R} \quad \boxed{b_R = \frac{R_G}{R_H}} \quad (0,0) = 0 \;\nearrow\; \searrow\; (1,1) = 1 - 1/b$$

$$Maximum \;\Leftrightarrow\; \frac{\partial b_{ROC}}{\partial FPF} = 0 \;\Leftrightarrow\; \frac{\partial TPF}{\partial FPF} = \frac{1}{b} \qquad (note: \frac{\partial^2 b_{ROC}}{\partial^2 FPF} = \frac{\partial^2 TPF}{\partial^2 FPF} < 0 \; as \; ROC \; is \; concave)$$

▷ *Ecological restoration benefit* (ERB) & *ecological benefit function* (e_{ROC})

$$\boxed{ERB = E_{ROC} = U_G TPF - U_H FPF = T_G e_{ROC}} \qquad \begin{cases} U_G = E_G\pi^+ \\ U_H = E_H(1-\pi^+) \end{cases} \qquad note: \boxed{ORB = ERB - ARC}$$

$$\boxed{e_{ROC} = TPF - FPF/e} \quad \boxed{e = \frac{U_G}{U_H} = odds(\pi^+)e_R} \quad \boxed{e_R = \frac{E_G}{E_H}} \quad (0,0) = 0 \;\nearrow\; \searrow\; (1,1) = 1 - 1/e$$

▷ *Maximise ecological benefit corrected for all monetary costs* (B_M) $(C_A = assessment\ costs)$

$$\boxed{B_M = ORB - C_A} = ERB - ARC - C_A = \boxed{ERB - C_M} \uparrow\uparrow\uparrow$$

3.3 Results

3.3.1 Utility analysis with ROC curves

The purpose of this theoretical exercise, is to clarify how the relative magnitude of the sensitivity and false positive fraction as described by the ROC curve determines the utility of the index and to explore how the operational context influences this relationship. We will interpret and comment the

utility functions for indices A & B as defined in Figure 3.1 and Figure 3.2. We start with the benefit function b_{ROC} = TPF – FPF/b which gives insight on how to tune the index, i.e. choose the optimal decision point maximising the overall restoration benefit (ORB) as a function of the decision context as characterised by the benefit ratio b (at a regional level). As will be motivated in the next chapter, we vary b from ½ to 8 in multiples of 2 which covers most practical circumstances.

3.3.1.1. The benefit function (Figure 3.5)

The benefit function strongly depends on the strength of the index. For any value of b, the benefit curve for B is (much) above the curve for A. The optima are higher and found in a relatively narrow range of FPF compared to A. Index B can realise a higher benefit at a lower FPF in contrast to A, with reduced the costs (as derived for TRF and ARC in the Methods section). In addition, index B appears to be more robust for uncertainties in b. Robustness is a useful property to cope with the uncertainty. For large b values, the maxima are relatively "broad". In this case, a safe strategy is to set the decision point somewhat below the optimum.

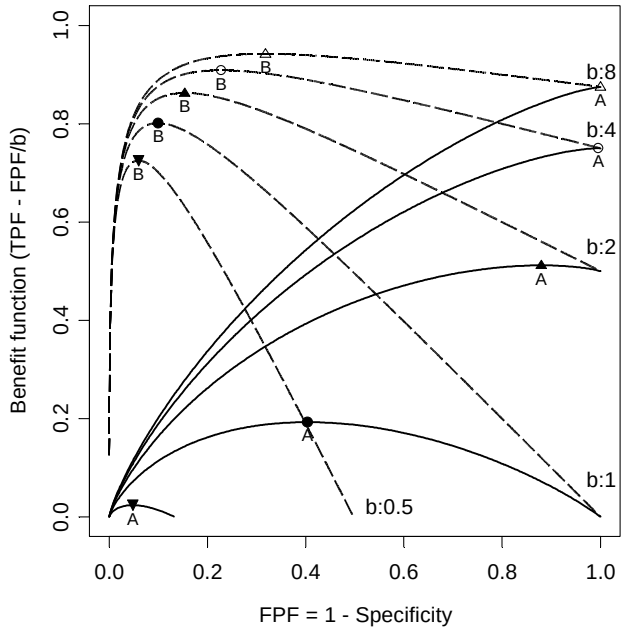


Figure 3.5 **The benefit function (b_{ROC} = TPF – FPF/b)** as a function of the quality of the index (A < B) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8).

To understand the underlying mechanism, an instructive case is b = ½ implying that the expected harm is twice as large as the expected benefit. In spite of this low value, the strong index B can realise a fairly high benefit, because the index discriminates well between degraded and pristine

sites. In contrast, index A can at most realise a breakeven at b = ½ (and the same holds for any b < 1). Overoptimistic b values result in a negative overall restoration benefit.

As already stated in the methods section, the benefit function (TPF – FPF/b) as a function of FPF is the ROC curve (TPF versus FPF) minus a penalty (FPF/b). The ROC part represents the gain because of restoring degraded sites, the penalty part the negative effect of restoring pristine sites. We may expect that the benefit curves resemble the ROC curves: steep for strong indices and gentle for weak indices. As the penalty term decreases with b, its impact is reduced resulting in higher maxima and flatter curves, resembling more and more the ROC curve (for b → ∞, the benefit curves become equivalent to ROC curves).

3.3.1.2. The true restoration fraction (TRF) as a function of FPF

The true restoration fraction is defined as the fraction of the management resources correctly allocated to true positive cases. From a management point of view, it is sensible to require that misallocation is as small as possible, for instance by requiring that TRF ≥ ¾ or ½ (Figure 3.6). This turns out to be a severe restriction for index A as its TRF curve drops sharply; nearly immediately TRF < 0.5 and none of the optimal decision points has a TRF above 0.4. For a strong index B, the decrease is more gradual and unless b = 8, we can realise the optimal benefit with TRF > ½.

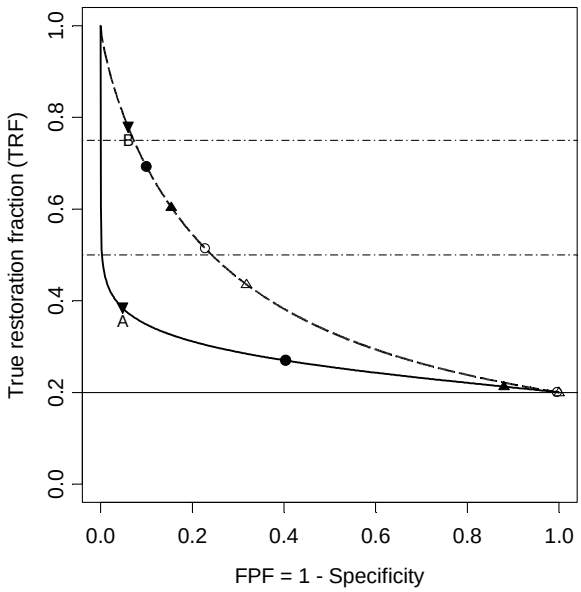


Figure 3.6 **The true restoration fraction (TRF = PPV) utility function** as influenced by the index quality (A < B). The decision points are optimally tuned with respect to the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, Δ = 8). The prevalence $\pi^+ = 0.2$ = under limit of TRF.

The pattern observed can be readily explained from the ROC curve. For the weak index A, the ROC curve resembles that of an indifferent index for which TPF = FPF. Then TRF = TPF.π^+ / (TPF.π^+ + TPF.(1-π^+)) = π^+ / (π^+ + (1-π^+)) = π^+. Except very close to the origin, FPF is of the same order of magnitude as TPF (TPF ≈ FPF), and as a consequence, PPV drops fast to π^+. In contrast, for a strong index B, the ROC curve climbs fast to TPF = 1 and over a much longer range TPF > FPF.

3.3.1.3. The true restoration fraction (TRF) as a function of the sensitivity

Figure 3.7 gives an alternative view by plotting TRF as a function of the sensitivity (y-axis). This is interesting to explore the effect of the index quality when fixing the sensitivity at a target value. For index A, TRF is below ½ for nearly the total sensitivity range.

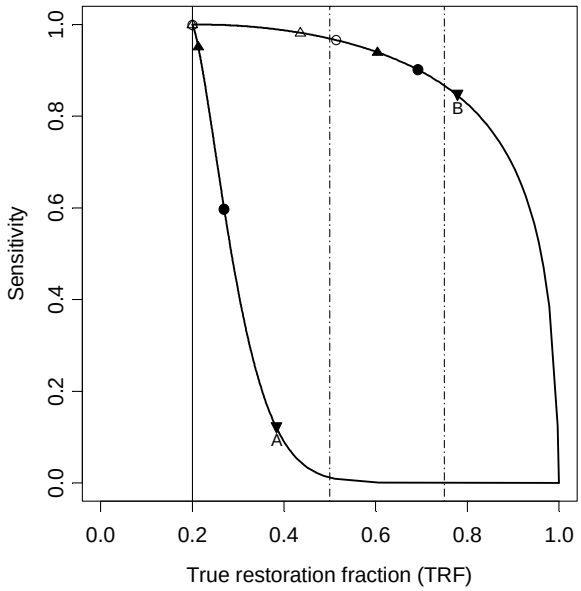


Figure 3.7 **The true restoration fraction (TRF) as a function of the sensitivity (TPF)** as influenced by the index quality (A < B). The decision points are optimally tuned with respect to the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). The prevalence π^+ = 0.2 = under limit of TRF.

3.3.1.4. The average restoration cost (ARC) as a function of the sensitivity

Because of the unfavourable budget allocation for weak indices, we proved that $ARC_A > ARC_B$ for a fixed TPF. Figure 3.8 visualizes this difference. We interpret $\Delta ARC = ARC_A - ARC_B$ as the resources we recuperate by choosing for the better index. We also added the optimal decision points as a function of b. For index A as well as index B, ARC increases as a function of b. This is because a larger b implies we can accept a higher FPF to realise the optimal overall benefit. For a weak index

as A, this effect is very strong and we should be prepared to pay a high price when b large. For a strong index, we have a better control of the costs.

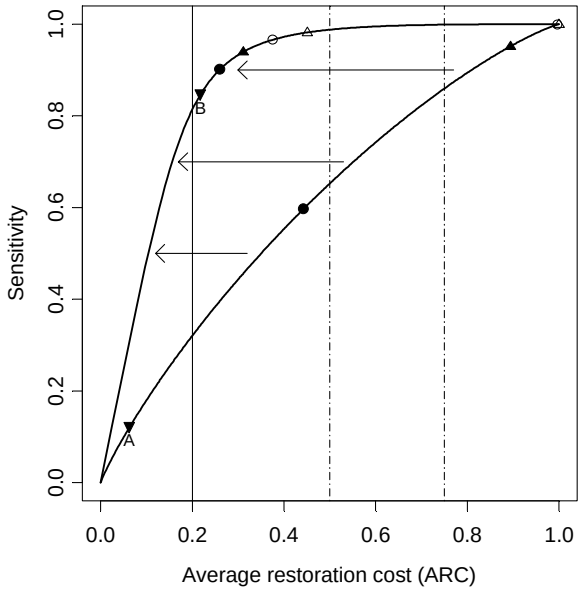


Figure 3.8 **The average restoration cost (ARC) function** as influenced by the quality of the index (A < B). The prevalence is $\pi^+ = 0.2$. The decision points are optimised with respect to the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8).

3.3.2 Exploratory Data Analysis (EDA) with ROC curves

ROC curves are a powerful exploratory data analysis (EDA) tool to screen the potential of candidate metrics to discriminate between different conditions of the ecosystem. As argued in the methods section, there is a close relation between boxplots, cumulative distributions and ROC curves and we illustrate how to synthesize the information of these graphs in synoptic plots based on the area under the ROC curve which is a measure of the distance between distribution functions. We illustrate the idea with data borrowed from a study to develop an estuarine biotic index for the Zeeschelde (Breine *et al.*, 2007) presented in Chapter 5.

Here, we consider three metrics in detail representative for respectively a high, moderate and low response to anthropogenic stress: piMjm = the percentage of marine juvenile species in the sample, piPis = the percentage of piscivores in the sample and vdSha = the Shannon diversity index. The gradient of anthropogenic stress ranges from 3 (moderate quality) to 5 (very low quality). Class 3 is least impacted and is considered as the baseline situation. We first explore the binary contrast 3/4-5 (coded as 1/0), then we evaluate the full gradient.

3.3.2.1. Relation between boxplots, EDF and ROC curves (Figure 3.9)

The first metric (piMjm) is a highly responsive metric which is actually the best metric of the candidate list. The boxplots show a clear separation between degraded (0) and pristine (1). The EFD curve of the degraded sites is shifted to the right in comparison to the baseline (bold). The ROC curve is fairly steep: aucP = 0.777 (90% CI: 0.606 – 0.875), meaning that the sensitivity in the range 0.1 ≤ FPF ≤ 0.3 is nearly 80 % on average, which is quite high for a single metric. Conversely, vdSha has no potential at all. Its ROC curve does not differ from the line of indifference (aucF = 0.5). In between, piPis has a moderate discriminatory power (aucF = 0.692; aucP = 0.5).

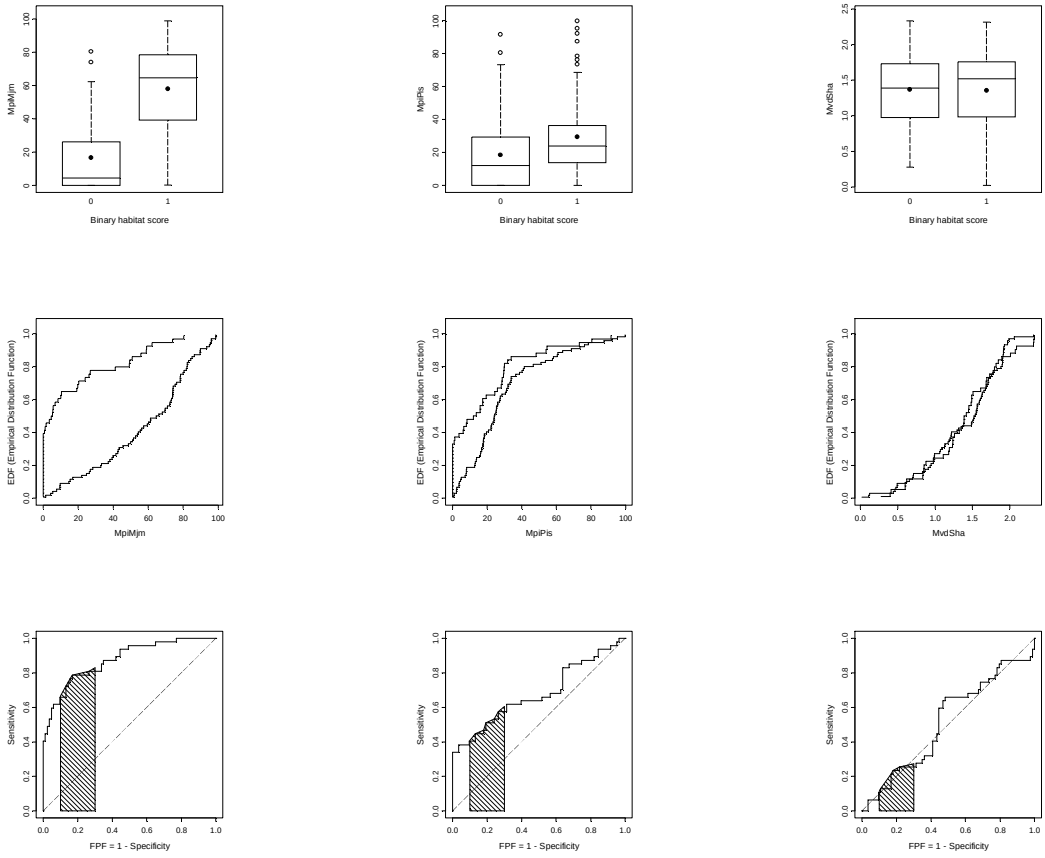
piMjm	**piPis**	**vdSha**
		
aucF[3/4-5]: 0.885 (0.802 – 0.929)	aucF[3/4-5]: 0.692 (0.576 – 0.774)	aucF[3/4-5]: 0.531 (0.414 – 0.621)
aucP[3/4-5]: 0.777 (0.606 – 0.875)	aucP[3/4-5]: 0.489 (0.384 – 0.631)	aucP[3/4-5]: 0.224 (0.114 – 0.363)

Figure 3.9 **Boxplots, EDF and ROC curves for the first binary contrast 3/4-5.** Essentially, all curves provide the same information but provide different details. In the boxplots, class 3 (baseline) is coded as 1, and class 4-5 as 0. The EDF curves in bold represent the baseline class 3. The shaded area of the ROC curves is the partial area under the curve (aucP). The values between brackets are BCa confidence limits (90%) for aucF and aucP with respect to 3/4-5.

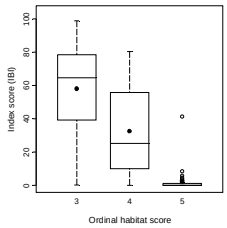
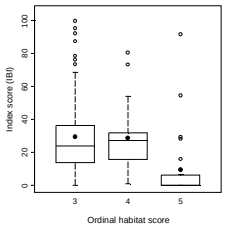
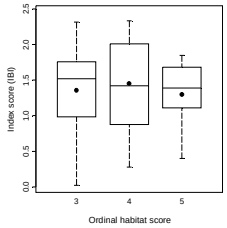
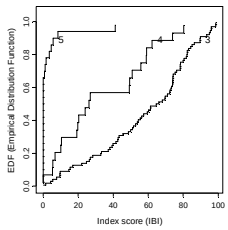
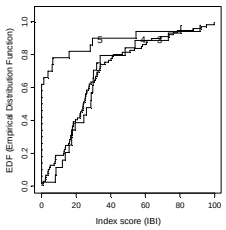
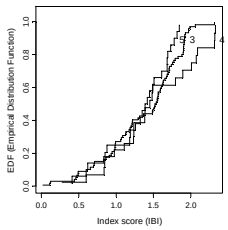
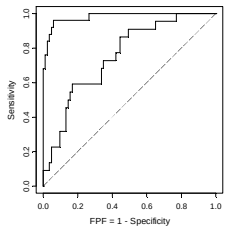
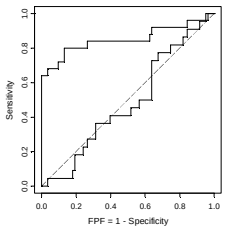
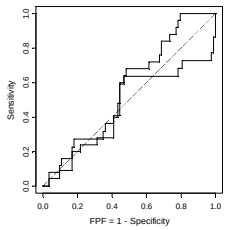
piMjm	piPis	vdSha
		
		
		
aucF(3/4): 0.777	aucP(3/4): 0.570	aucF(3/4): 0.513	aucP(3/4): 0.173	aucF(3/4): 0.496	aucP(3/4): 0.243
aucF(3/5): 0.948	aucP(3/5): 0.947	aucF(3/5): 0.817	aucP(3/5): 0.815	aucF(3/5): 0.577	aucP(3/5): 0.220

Figure 3.10 **Boxplots, EDF and ROC curves for the full gradient.** The ROC curves are drawn for two binary contrasts 3/4 and 3/5 investigating the response to an increasing anthropogenic pressure. The higher curve corresponds to contrast 3/5 except for vdSha with the ROC curves intermixed. The statistics aucP and aucF are calculated for the two ROC curves. For piMjm and piPis, aucF(3/4) < aucF(3/5) reflecting the response trend (a similar relation holds for aucP). aucF ≈ 0.5 refers to an indifferent contrast, e.g. for piPis aucF(3/4) = 0.513, visible in the boxplots, EDF and ROC curves. The synoptic diagram (Figure 3.11) is based on the two aucF statistics and summarises all graphs in one single picture for 11 additional metrics.

3.3.2.2. Extension to ordinal variables

For the full gradient (Figure 3.10), we observe piMjm has a consistent response to an increasing anthropogenic stress. The boxplots and EDF curves are well and logically separated for the three levels of human impact. The ROC curve for second contrast 3/5 is higher than the first contrast 3/4 revealing an increasing response. The same message is given by the summary statistics for the two contrasts: aucF increases from 0.777 to 0.948, and aucP from 0.570 to 0.947. To detect the

difference between class 3 and 5, the average sensitivity in the range 0.1 ≤ FPF ≤ 0.3 is nearly 95 %.

The metric piPis is not sensitive to low pressure (aucF(3/4) = 0.513), but the discrimination is high with respect to class 5 (aucF(3/5) = 0.817). The type of response is different than for piMjm, but, as confirmed in Chapter 5 when building the index model, piPis offers complementary information which results in a selection for the final index. Not necessarily, the best metrics are included in the multi-metric index, but those that offer the highest predictive power in combination. Finally, we can conclude the Shannon index does not have any discriminatory power at all.

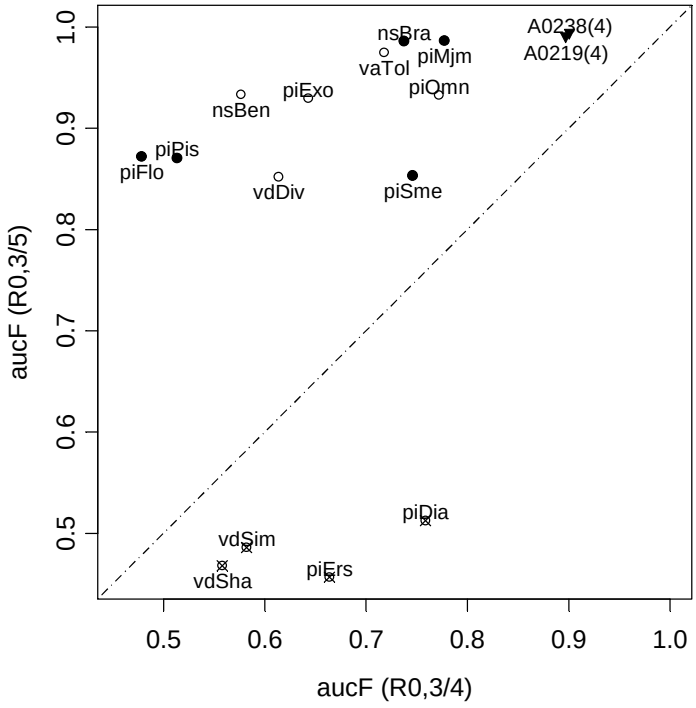


Figure 3.11 **Synoptic diagram of the diagnostic accuracy to screen the candidate metrics.** (aucF for the binary contrasts 3/4 and 3/5). Black points: metrics included in the two best models with four metrics (note that piFlo & piPis are interchangeable, see text). Triangles: best indices with four metrics (R0.A0238 and R0.A0219). Crossed points: metrics excluded because not informative.

3.3.2.3. Synoptic plot to evaluate metrics and indices

We can summarise this detailed information in a single synoptic plot (Figure 3.11). The x-axis represents the diagnostic accuracy for contrast 3/4 and the second axis for contrast 3/5. With this synoptic plot, we see at one glance that four metrics have no potential at all (see Chapter 5 for a

further discussion, here we only explain the principle). We can safely drop these four metrics for the subsequent analyses reducing the workload considerably (a reduction from 14 to 10 metrics, implies 2^{10} - 1 instead of 2^{14} - 1 models to investigate). All other metrics we keep in the model as variables with a low response can be very informative once other variables are included in the model. In this respect, it is instructive to notice that not all of the metrics in the final four-metric index (labelled as black points) are the best ones.

3.4 Discussion

3.4.1 A summary of the essential ideas

The first objective of this chapter was to give a coherent overview of the main ideas and formulas about diagnostic accuracy and ROC curves and to make a transcription to the context of ecological indicators.

The decision context and decision rule

To fix the ideas, we imagined that an IBI is applied to decide about restoration of a waterbody in the context of river management. For this case, the event of interest is degradation. By default, low values of the test variable of the index, the ecological quality measure or EQM, are indicative for degradation. If EQM is smaller than a preset decision threshold T, it is inferred the site is degraded and it is decided to restore the site. We call the outcome of the index "positive", when it signals degradation; it is "negative", when there is no signal.

False positives and false negatives, sensitivity and specificity

A first important insight is that we should distinguish between false positive (FP) and false negative (FN) errors as their ecological and societal consequences are of a different nature. A high false positive fraction (FPF) results in many unnecessary and possibly harmful restorations depleting resources and risking to cause harm to the pristine ecosystem. With a high false negative fraction (FNF), the necessary action is not taken, leaving many sites impaired with a reduced ecological functioning less capable to sustain the ecosystem services for the society. The true positive fraction (TPF) is more commonly known as the sensitivity expressing the capacity of the index to detect the event of interest. The false positive fraction (FPF) is the complement of the specificity or true negative fraction (TNF), which is the capacity of the index to recognise that the event of interest is not present.

The role of the decision threshold

For a continuous test variable, the relative magnitude of FPF and FNF depends on the decision threshold. For a given index, it is impossible to improve both errors simultaneously by changing T: decreasing FPF increases FNF and vice versa. This tradeoff is an inherent characteristic of any decision procedure. A restrictive policy diminishes FPs but increases FNs, and the opposite holds for a liberal policy. There is no way out, unless a better index can be developed.

When evaluating the diagnostic accuracy, we should consider both criteria simultaneously to guarantee a fair comparison of indices as it is always possible to optimise one criterion at the

expense of the other. To maximising the utility of the index in a given context, we can tune the index, i.e. seeking the optimal decision point on the ROC curve taking into account the parameters characterising the decision environment.

If the cost is unacceptable in the optimum, we have to seek for a better index. In this respect, we define the strength of the index as the capacity to realise a high sensitivity while keeping the false positive fraction low. A stronger index results in a greater potential to find an acceptable compromise. Although the optimal decision point depends on the relative costs of FPs and FNs, the general rule is that we should keep both errors small.

The ROC curve

The Receiver Operator Characteristic (ROC) curve visualises the intrinsic strength of the index by graphing the true positive fraction (TPF = the sensitivity) as a function of the false positive fraction (FPF = 1 – the specificity) for all possible decision thresholds T. The shape of the ROC curve characterises the potential of the index. A steep climbing high ROC curve implies that the index can detect an event with a high TPF keeping FPF small. In contrast, a flat low ROC curve means that we have to allow a high FPF to detect an event with a high TPF.

As an overall summary measure of diagnostic strength, we can use the area under the ROC curve (aucF). As aucF also depends on parts of the ROC curve irrelevant for a practical application, an appropriate alternative is to use the partial area under the curve (aucP) focusing on a relevant region. As a default, we average on $0.1 \leq FPF \leq 0.3$. With IBIs, $FPF < 0.1$ results in a too small sensitivity, while with $FPF > 0.3$ the FP burden is too high.

The informativeness or predictive value of the index

The positive predictive value (PPV) of an index is the fraction of positive outcomes of the index corresponding with degradation. PPV expresses how confident we can be about a positive test result. If it is high, we can be pretty sure the event is present in reality. Otherwise we cannot trust the result very much. The important factor determining PPV is the ratio of TPF and FPF. It can be shown that $odds(PPV) = odds(\pi^+).(TPF/FPF)$. Interpreting the prevalence of degradation as the *a priori* probability of a degraded site, and PPV as the *a posteriori* probability after using the index, then TPF/FPF is the contribution of the index to improve the predictive value of the test. It is important to recognise that PPV is strongly determined by the prevalence. If π^+ is small, the ratio TPF/FPF should be very large to achieve a high PPV.

3.4.2 The usefulness of an index

A second purpose was to get better insight in the usefulness of an index for decision making. If an index is effectively used and integrated in decision making, we can link its informative value to its contribution to decision making.

Cost effectiveness analysis (CEA) and cost benefit analysis (CBA)

A CEA considers the costs implications without valuing benefits in monetary terms. For instance, it is decided to keep FPF beyond a certain limit as a general management option (evidence-based or not) and the cost implications are investigated as a function of the strength of an index. Or, the

management goal is set to restore a certain proportion of degraded sites. This fixes the sensitivity and for this choice the cost consequences of different indices are evaluated. Conversely, with a CBA, the benefits are valued in monetary terms and a global evaluation is performed balancing costs and benefits. As CBA integrates the same monetary costs as CEA, the results of both approaches can be very similar. The CBA philosophy is more ambitious, but at the same time more risky as it not evident to value non-monetary costs.

The true restoration fraction (TRF) and average restoration cost (ARC)

A first series of utility functions is CEA-based. The true restoration fraction (TRF) quantifies which portion of the budget is correctly allocated to degraded sites. TRF is equivalent to PPV. This was an unexpected but retrospectively logical result. As the total restoration budget spent is proportional to the total number of positive signals, the fraction of the total budget correctly allocated (TRF) is equal to the proportion of correct positive signals (PPV). To control the loss of resources and/or the ecological risk by restoring pristine sites below a certain limit, it is sensible to require TRF ≥ ¾ or ½. With weak indices, this requirement results in a very low sensitivity as their TRF curves drop sharply. Similarly, to realise a high sensitivity with weak indices, we have to accept a high FPF resulting in a high average restoration cost (ARC). In contrast, the evolution of TRF is much more gradual for strong indices, such that we can realise a high TPF and simultaneously keep ARC low.

The overall restoration benefit

A second series of utility functions is based on a CBA philosophy. The overall restoration benefit (ORB) is defined as the difference of the benefit realised by restoring degraded sites and the harm because some pristine sites are unnecessarily restored, ORB appears to be proportional to the benefit function TPF – FPF/b which is the ROC curve (TPF) from which a penalty term is subtracted (FPF/b). The first term reflects the benefit realised and the penalty term is because FPs cannot totally be avoided.

The penalty is inversely proportional to the benefit ratio b defined as the quotient of the expected gain and harm at a regional level. This is logical. If the expected gain is large in comparison to the risk, FPs are less influential than TPs and we can afford a higher FPF level. The converse holds if b is small. As for TRF, the CBA analysis makes clear it is impossible to realise a high TPF with a weak index without accepting a high FPF. In addition, if b is low, a strong index is necessary to realise any benefit. In this case, FPF should be kept very small which is only possible with strong indices.

3.4.3 Towards a deeper understanding of the usefulness of ROC curves

The third objective was to get a deeper insight why the ROC concept is so fundamental to assess the validity and usefulness of indices. The utility functions reveal the close connection between strength, informativeness and usefulness. They are simple, monotonous transformations of ROC curves reflecting transparently how parameters characterising the decision context reshape the intrinsic strength as characterised by ROC curves in the operational strength.

A first example of this link between the different diagnostic measures is that odds(PPV) equals odds(π^+) multiplied with TPF/FPF. The latter term in this multiplication is a measure of the capacity of the index to realise a high TPF at a low FPF. The former term represents the impact of the

decision context. Under "harsh" circumstances, with a low prevalence, it will be very hard to realise a high PPV unless TPF/FPF is very high.

A second example is the mathematical equivalence between PPV and TRF. The former expresses which proportion of positive outcomes corresponds with degraded sites, the latter specifies the budget fraction correctly attributed to degraded sites. This equivalence makes a direct connection between the average restoration cost (ARC) and the predictive value of the index. With a small PPV, we have to accept a high ARC.

Thirdly, the overall restoration benefit (ORB) is proportional to the benefit function TPF – FPF/b. The first term equals the ROC curve (TPF as a function of FPF) reflecting the gain realised by the index. The second term is a penalty (FPF/b) because of the FPs that could not be avoided by the index. The benefit ratio b characterises the operational context and can be factorised as the $\text{odds}(\pi^+).b_R$. The latter term expresses the local intrinsic benefit ratio, but we should take into account the prevalence of degradation to fully characterise the operational quality of the index.

3.5 Conclusions

Inspired on a discussion in Zhou (2002) and the references herein (Fryback and Thornbury, 1991), we organised the quality measures hierarchically in Table 3.2 and added two other quality criteria: on top, we placed the technological aspects of the index, on the bottom the societal implications. The scheme is hierarchical, because without a sufficient quality at a lower level (the first lines in the table), quality at a higher level may not be expected.

Table 3.2 Hierarchical relationship between strength, informativeness and utility of the index

Concept	Related concepts	Criteria ~ Operational context
Reproducibility	Data quality	Reproducibility and repeatability measures
↓		Interobserver variability
Strength	Intrinsic diagnostic accuracy	FNF | TPF, FPF | TNF
↓	Potential of the index	ROC curve (aucF, aucP)
Informativeness	Operational diagnostic accuracy	PPV | FAF, NPF | FRF ~ π^+
↓	Predictive value	
Utility	Efficacy	CEA: ARC, TRF (= PPV) ~ π^+
↓		CBA: ORB, ERB ~ b = $\text{odds}(\pi^+).b_R$
Societal value	Decision value	CEA: C_M = ARC + C_A ~ π^+, C_R, C_A
(Chapter 4)	Pragmatic value	CBA: B_S = ORB – C_A = ERB – C_M ~ T_G, T_H, C_A

The first requirement is to have a high reproducibility and data quality which includes issues as sampling, data storage and data availability. This item gives still another interpretation of Figure 3.1. A and B represent the same index, but in situation A, data quality and reproducibility is low, while in situation B, they are high. Lack of a quality policy (QA/QC) can destroy an intrinsically good index.

At the bottom of the table, the pragmatic value takes into account the assessment costs C_A. In some cases C_A can be so large that it is more beneficial to take a lower quality index. In Chapter 4, we will tackle this tradeoff more in detail. It is important to realise that the pragmatic value represents the potential of the index to realise a benefit in a certain operational context. Its real impact however depends on whether the index is effectively used for decisions. Evaluation of the impact requires a sociological approach (Turnhout *et al.*, 2008).

Because of the close link between the intrinsic quality of an index and its usefulness, it is sensible to use the ROC curve and/or its summary statistics as optimisation criteria to calibrate and validate IBIs. Chapter 5 presents a methodology based on the area under the ROC curve to retrieve the optimal basket of metrics for a fish-based estuarine biotic index for the Zeeschelde estuary. In Chapter 6, we demonstrate how an evaluation based on the error curve, an equivalent of the ROC curve, allows to interpret differences in diagnostic accuracy between the European Fish Index (EFI) and existing local indices.

A step further could be to directly optimise the utility function. In fact, our proposal to directly optimise the more focused aucP is a step in this direction. For the optimisation, we concentrate on the part of the ROC curve which is relevant for a practical optimisation. In the same vein, we can chose any other summary statistic of the ROC curve incorporating parameters of the decision context. To take into account the uncertainty of the context parameters sensitivity analyses or still better a Bayesian approach is recommended involving a lot of flexibility to cope with uncertainty (Box, 1980; Brosi and Biber, 2009; Clark and Bjørnstad, 2004; Ellison, 1996; Engel *et al.*, 2009; Fox, 2001; Wintle *et al.*, 2003).

4 Do we pay too much for monitoring? A cost analysis of ecological indicators based on ROC curves.

The aim of this original theoretical exercise is to demonstrate how we can use the ROC framework to analyse the cost consequences of decisions based on ecological indicators in the context of river restoration. We fully parameterize the decision context and link basic "field parameters" describing features as the restoration efficacy and the prevalence of degradation with a few "cost parameters" governing the cost implications of index-based decision making.

Abstract

By using relatively cheap ecological indicators instead of expensive gold standard measurements, we can save resources if the diagnostic accuracy remains sufficiently high for decision making. Improving index quality decreases the fraction of wrong decisions, but assessment costs increase. To get insight in the cost tradeoff and the factors governing an optimal allocation of the resources to monitoring, we elaborate a hypothetical but realistic example in the context of river restoration and/or rehabilitation where an index is effectively used to decide about treatment of a waterbody. We imagine a series of indices is available ranging from nearly no diagnostic accuracy to the gold standard and we model the relation between the assessment cost and diagnostic accuracy with a quadratic relation representing the (fast) increase in cost with the index quality. By a full parameterisation of this (simple) binary decision example, we link the costs of the diagnostic accuracy of an index with the corresponding costs and benefits of management decisions. In this framework, the original question is rephrased as an enquiry for the best index, maximising the total benefit (a cost benefit analysis – CBA) or minimising the total management costs (a cost effectiveness analysis – CEA). The latter criterion adopts a strictly monetary perspective, the former criterion is more comprehensive and takes into account all possible benefits and costs of management decisions. As low costs imply high benefits, in most cases, both approaches result in similar outcomes, but they diverge if non-monetary benefits are an important component.

Keywords

ROC curve, utility curves, informativeness, predictive value, true restoration fraction, cost benefit analysis, cost-effectiveness analysis

4.1 Introduction

Monitoring can contribute to better decisions if the output is relevant, sufficiently accurate, cost-effective and well integrated in the decision cycle. Cost-effectiveness is a neglected, but crucial and decisive element in the success of monitoring (Caughlan and Oakley, 2001). Although monitoring is generally appreciated, in particular cases, it is felt as an overhead with a low perceived benefit. For this reason, the usefulness of monitoring programs should be carefully motivated. To reduce costs, ecological indicators are often advocated (Vos *et al.*, 2000) as proxies assessing a complex reality at a relatively low cost in comparison to gold standard measurements (Murtaugh, 1996). This is only true if the indicators are well calibrated and validated against the gold standard. The cost reduction by using a proxy should be larger than the negative consequences of the approximation, for instance, as measured by the cost of making wrong decisions.

We will apply this basic economical principle to a simple situation where we use an IBI to decide about the restoration and/or rehabilitation of a waterbody (Dufour and Piegay, 2009). If the index signals the waterbody is degraded, it is straightforwardly concluded without further investigations to restore or rehabilitate the site. In this simplified decision framework, it is possible to calculate the expected cost associated with correct and wrong decisions. Benefits are modelled as negative costs. With an increasing diagnostic accuracy, the assessment costs generally increase but the decision costs decrease because the misclassification becomes smaller. The optimum is where the sum of these two costs is minimised.

The purpose of this theoretical exercise is to get more insight in the key factors governing the optimal choice of an index. The mathematical discussion is sometimes tedious but in essence very simple. Step by step, we derive and interpret the cost equations from the parameters describing the restoration context. In the end, few parameters and functions determine the optimum, but, for a practical application, it is important to link these parameters with more basic information. Finally, we extrapolate the results to more global principles and insights helping to orient the design and choice of ecological indicators.

4.2 Material & Methods

4.2.1 Parameterisation of an index-based binary decision framework

An index can only contribute to better decisions if it is integrated and really used in a decision process. Otherwise, we merely accumulate data with at best a vague link to decision making. To quantify the impact of diagnostic accuracy, we first parameterise the decision context describing the link between the index outcomes and the cost consequences. Table 4.1 and Box 4.1 give an overview of the main symbols and definitions.

4.2.1.1. The decision context

We imagine a simple binary decision framework in the context of river management for which an IBI is used to decide about restoration. We suppose there are only degraded sites (to be restored) and pristine sites (restoration depleting resources and possibly harmful for the ecosystem). The

prevalence or relative frequency of degraded sites in the region is π^+, the prevalence of pristine sites is $1-\pi^+$ (the complement as there are only two states). Decisions are straightforwardly based on an index. If the test variable of the index a value below the decision threshold, it is concluded to restore without further investigations. If the index value is above the threshold, no action is taken. With this binary decision rule, we can distinguish four different decision types depending on the true (but unknown) condition of the site. If the site is really degraded, we have a true positive (TP) decision if degradation is detected, and a false negative (FN) if it is not. For pristine sites, a true negative (TN) denotes a correct decision, and a false negative (FN) corresponds with a misclassification as a degraded site.

Box 4.1 **Elementary information required for the cost calculations (analytical parameters).** See Table 4.1 for a definition of the symbols. For an explanation, see text.

$$\boxed{C_T = C_A + C_{TP}P_{TP} + C_{FN}P_{FN} + C_{TN}P_{TN} + C_{FP}P_{FP}} \qquad with: P_{FN} + P_{TN} + P_{TP} + P_{FP} = 1$$

▷▷▷ *Monetary costs*

$$\boxed{\begin{array}{l} C_A : assessment\ cost \\ C_R : restoration\ cost \end{array}}$$

▷▷▷ *Probability matrix*

$$\boxed{\begin{array}{l} \pi^+ : prevalence\ of\ degraded\ sites \\ (FPF, TPF) = coordinates\ ROC\ curve\ (at\ least\ one\ decision\ point) \end{array}}$$

$$\begin{bmatrix} P_{FN} & P_{TP} \\ P_{TN} & P_{FP} \end{bmatrix} = \begin{bmatrix} P(U^+)P(W^-|U^+) & P(U^+)P(W^+|U^+) \\ P(U^-)P(W^-|U^-) & P(U^-)P(W^+|U^-) \end{bmatrix} = \begin{bmatrix} \pi^+(1-TPF) & \pi^+TPF \\ (1-\pi^+)(1-FPF) & (1-\pi^+)FPF \end{bmatrix}$$

▷▷▷ *Cost matrix : monetary + ecological costs associated with decision types*

$$\begin{bmatrix} C_{FN} & C_{TP} \\ C_{TN} & C_{FP} \end{bmatrix} = \begin{bmatrix} 0 & C_R \\ 0 & C_R \end{bmatrix} + \begin{bmatrix} E_{FN} & E_{TP} \\ E_{TN} & E_{FP} \end{bmatrix} = \begin{bmatrix} E_{FN} & C_R + E_{TP} \\ E_{TN} & C_R + E_{FP} \end{bmatrix}$$

▷▷▷ *Influence of the baseline of zero cost C_Z*

$$C'_T = C_A + (C_{TP} - C_Z)P_{TP} + (C_{FN} - C_Z)P_{FN} + (C_{TN} - C_Z)P_{TN} + (C_{FP} - C_Z)P_{FP}$$
$$C'_T = C_A + (C_{TP}P_{TP} + C_{FN}P_{FN} + C_{TN}P_{TN} + C_{FP}P_{FP}) - C_Z(P_{TP} + P_{FN} + P_{TN} + P_{FP})$$

$$\Rightarrow \quad \boxed{C'_T = C_T - C_Z} \quad \Leftrightarrow \quad \boxed{C_T = C'_T + C_Z}$$

Table 4.1 **Overview of the main symbols and definitions for the cost calculations.**

TOTAL COST FUNCTION		
Cost decomposition	C_T	Total cost function; $C_T = C_0 + C_A - B_{ROC} = C_0 - B_M$; C_0 = ecological & societal cost status before restoration; C_A = assessment cost; B_{ROC} = Overall restoration benefit
	B_M	Total benefit after correction for monetary costs; $B_M = B_{ROC} - C_A = E_{ROC} - ARC$
Additional definitions	C_Z	Zero or reference point for cost calculations (all costs should be consistently expressed with respect to C_Z; e.g. $E_{TN} = 0$)
DECISION COSTS		
Basic cost types	$C_{FN} \leftrightarrow C_{TP}$	Degraded sites: cost of not restoring (FN) / restoring (TP)
	$C_{TN} \leftrightarrow C_{FP}$	Pristine sites: cost of not restoring (TN) / restoring (FP)
Monetary cost of restoration	C_R	Restoration costs
	C_A	Assessment cost (development & maintenance and design & implementation monitoring program = activation cost)
Ecological (and societal) costs	$E_{FN} \leftrightarrow E_{TP}$	Degraded sites: ecological cost of FN & TP
	$E_{TN} \leftrightarrow E_{FP}$	Reference sites: ecological cost of TN & FP
TRADEOFF BENEFIT & RISK (restoring degraded ↔ pristine sites)		
Costs at a site level (local)	$E_G \leftrightarrow E_H$	Ecological gain/harm of restoring degraded/pristine sites; $E_G = E_{FN} - E_{TP}$, $E_H = E_{FP} - E_{TN}$
	η_E	Ecological efficacy of restoration; $\eta_E = E_G/E_{FN} = 1 - E_{TP}/E_{FN}$
	$R_G \leftrightarrow R_H$	Cost-corrected ecological gain or harm of restoring degraded or pristine sites; $R_G = E_G - C_R$, $R_H = E_H + C_R$
	η_C	Cost-corrected efficacy of restoration; $\eta_C = R_G/E_{FN} = 1 - (E_{TP} + C_R) / E_{FN}$. Note that $\eta_C = T_G/C_0$ if $E_{TN} = 0$
	b_R	Intrinsic benefit ratio because of restoration; $b_R = R_G/R_H$
Tradeoff at a regional level taking into account π^+	C_0	Societal and ecological cost situation before restoration (no intervention yet); $C_0 = \pi^+.E_{FN} + (1\text{-}\pi^+).E_{TN} = \pi^+.E_{FN}$ ($E_{TN} = 0$)
	$T_G \leftrightarrow T_H$	Expected total gain / harm of restoring <u>all</u> degraded / pristine sites; $T_G = (E_G - C_R).\pi^+$ & $T_H = (E_H + C_R).(1 - \pi^+)$
	b	Benefit ratio (on a regional level); $b = T_G/T_H = b_R.odds(\pi^+)$
TRADEOFF ALONG ROC CURVE		
Reference: no restoration at all	B_{ROC} (B_{Opt})	Benefit along ROC curve in comparison with no restoration at all; reference = no sites restored; $B_{ROC} = T_G.TPF - T_H.FPF = T_G.b_{ROC} = b.T_H.b_{ROC}$ (Optimal value for B_{ROC} at $b.\Delta TPF = \Delta FPF$)
	b_{ROC}	Benefit function; kernel of B_{ROC}; $b_{ROC} = TPF - FPF/b$
	B_M	B_{ROC} corrected for the assessment costs ($B_{ROC} - C_A$)
Reference: with full restoration	A_{ROC} (A_{Opt})	Benefit along ROC curve by avoiding FPs in comparison to full restoration; reference = all sites restored; $A_{ROC} = T_H.TNF - T_G.FNF = T_H.a_{ROC}$ (Optimal value for A_{ROC} at $b.\Delta TPF = \Delta FPF$)
	a_{ROC}	Avoidance function; kernel of A_{ROC}; $a_{ROC} = TNF - b.FNF$

4.2.1.2. The expected total cost C_T

Although the outcome of individual decisions is unknown, we can calculate the expected total cost C_T associated with the use of the index based on the probabilities and the costs of each decision type (the equation on top of Box 4.1). C_T is the sum of the assessment cost (or monitoring cost) necessary to calculate the index (C_A = assessment cost) and the cost consequences associated with the four possible decision types (C_{FN}, C_{TP}, C_{TN}, C_{FP}) multiplied by their probability (P_{FN}, P_{TP}, P_{TN}, P_{FP}). As all possible outcomes are covered, the sum of the decision probabilities equals one. We express all costs as an average per management unit (waterbodies or sites). To estimate the total budget for a region, multiplication by the number of waterbodies is necessary. We get rid of this extra parameter by working consistently with average costs simplifying equations. Benefits are modelled as negative costs. The purpose of the cost optimisation is to minimise C_T.

4.2.1.3. The probability matrix associated with the decision types

From the sensitivity and specificity of the index (which involves knowledge of the ROC curve or at least of one decision point on the curve) and the prevalence of degraded sites (π^+), we can derive the probability of each decision type as summarized in the probability matrix in Box 4.1. Each element of the matrix is a product of the probability of a the ecological condition (its prevalence) and the (conditional) probability the condition will be detected by the index. As an example, for pristine sites misclassified by the index as degraded, $P_{FP} = (1-\pi^+)$.FPF which is the product of the prevalence of pristine sites ($1-\pi^+$) and the conditional probability a pristine site will be misclassified as degraded (FPF). As the matrix contains all possible outcomes (four in total), the sum of the probabilities equals one. The probabilities themselves vary as a function of the decision threshold of the index. Hence, an index is characterized by an infinite number of probability matrices, each matrix corresponding with a decision point on the ROC curve. Variation of the decision point will change the relative contribution of the costs in the equation of C_T and hence will influence C_T itself. For a given index, we can look for the point where C_T is minimal.

4.2.1.4. The cost matrix associated with the decision types

The cost types corresponding with the four decision types are: false negative costs (C_{FN}) and true positive costs (C_{TP}) for degraded sites (restoration necessary), true negative costs (C_{TN}) and false positive costs (C_{FP}) for pristine sites (restoration not necessary). We split these costs in a monetary and ecological component (the ecological component also incorporates societal aspects, if present). A positive decision (i.e. the decision to restore, can be a TP or a FP), always comprises the (monetary) restoration cost (C_R). To assess the total cost, we add the ecological (and societal) cost consequences of the decision: $C_{TP} = C_R + E_{TP}$ and $C_{FP} = C_R + E_{FP}$. With a negative decision, there are no restoration costs, and hence, the decision costs are purely ecological: $C_{TN} = E_{TN}$ and $C_{FN} = E_{FN}$. In total, including the restoration cost and assessment cost, there are six elementary cost parameters characterising the decision context. Box 4.1 puts this basic information in a cost matrix.

4.2.1.5. The baseline of zero cost

To avoid double counting, it is important to understand the costs consistently as state variables and to estimate changes as differences of state variables. We illustrate this point by calculating the ecological gain of restoring a degraded site. The change of the ecological cost is the difference between E_{TP} (end status after restoration = status of a TP) and E_{FN} (status before restoration = status of a FN): $E_{TP} - E_{FN}$. As a benefit is a negative cost, the gain because of restoration (E_G) is $(-E_{TP}) - (-E_{FN}) = E_{FN} - E_{TP}$. If the restoration of a degraded site is totally successful without any (harmful or beneficial) side-effects, the cost will become zero ($E_{TP} = 0$), resulting in $E_G = E_{FN}$. However, setting (incorrectly) $E_{TP} = -E_{FN}$ to describe the new state, results in a double counting: $E_G = 2.E_{FN}$.

Another point of attention is the choice of the baseline of zero cost (C_Z) which is not always evident. In above example, if restoration increases safety for the people living close to the river, one possibility is to specify an E_{TP} value smaller than zero expressing the societal benefit after restoration. An alternative is to include the original lack of safety as a cost in E_{FN}. This degree of freedom can introduce ambiguity and subjectivity. Fortunately, for most (but not all) cost calculations, only the differences between costs are relevant on which C_Z has no impact. For instance, the ecological benefit of restoration remains equal for any value of C_Z: $E_G = E_{FN} - E_{TP} = (E_{FN} - C_Z) - (E_{TP} - C_Z)$.

More generally, as demonstrated on the bottom of Box 4.1, changing the baseline by subtracting C_Z from all basic cost terms (e.g. $C_{TP} - C_Z$) diminishes C_T with C_Z because $P_{FN}+P_{TP}+P_{TN}+P_{FP} = 1$ without impact on the fundamental structure of the formula. By moving the term to the left site of the equation, we obtain an identical formula, but now for $C_T - C_Z$. The equation makes clear that all costs have to be expressed and interpreted in comparison to the (same) baseline value. This implies, we only have to look at the marginal costs, i.e. the costs that change under alternative scenarios. This considerably lowers the burden of the cost calculations.

Yet, for some calculations, a careful choice of the "cost anchor" is necessary. For instance, the efficacy of restoration depends on the baseline. If a "natural reference" exists in a given context, it is recommended to set this cost equal to zero. For the example elaborated in this chapter, a logical choice is to consider a pristine site as the optimal situation without any costs. We will refer all other costs in comparison to this reference. As a pristine site corresponds with a true negative, we set $E_{TN} = 0$. Then E_{TP} (the cost status of a restored, originally degraded site) represents the residual cost after restoration in comparison to a pristine site and the ecological efficacy of restoration $\eta_E = E_G / E_{FN} = 1 - E_{TP}/E_{FN}$. If restoration if fully successful, $E_{TP} = 0$ and $\eta_E = 1$.

4.2.1.6. Present value of costs

In the equations, we assume the costs and benefits are instantaneous. In reality, the restoration benefit is only gradually realised and also the costs are spread over many years. Yet, it is possible to calculate its present value (PV) equivalent with the future gain by taking into account the interest rate (Fisher, 1930). For instance, if the value in the future is FV and this benefit is realised over *y* year, then by the formula of compounded interest (*i* = interest rate):

$$PV = \frac{FV}{(1+i)^y}$$

Thus, for a discount rate of 3 % and a period of 24 year, the PV is about ½ of the FV. For benefits gradually realised over the year, the formulas are more involved, but the principle remains the same: technically it is feasible to integrate future benefits in the cost analysis. The present value technique puts them on equal footing.

4.2.2 The three term decomposition of the overall cost

4.2.2.1. The basic equation and definitions

The parameters defining the decision context in Box 4.1 are rather abstract and hard to determine. Fortunately, it is possible to combine the original quantities in a few key parameters that are more easy to understand (Box 4.2). By substituting these key parameters in the total cost, a simple equation emerges comprising three terms: $C_T = C_0 + C_A - B_{ROC}$ (the derivation at the bottom of Box 4.2 provides a proof). Each term has a specific and easy to understand meaning: (i) C_0 represents the ecological and societal cost before restoration. (ii) C_A comprises the assessment costs to establish the ecological condition with the index. (iii) B_{ROC} equals the benefit of restoration as guided by the index. As suggested by the index, the latter term depends on the decision point of the index on its ROC curve. In the next section, we will show how to tune the index, i.e. choosing the optimal decision threshold such that B_{ROC} is maximal (B_{Opt}).

$B_M = B_{ROC} - C_A$ represents the overall restoration benefit after correcting for the assessment costs of the index. To realise an overall gain, it is necessary that $C_A < B_{ROC}$, otherwise B_M is negative. Both C_A and B_{ROC} increase as a function of the diagnostic accuracy and the question is for which index the maximal benefit is reached. An alternative expression (derivation in Box 4.2) is $B_M = E_{ROC} - C_M$, splitting B_M in a purely ecological term (the ecological gain E_{ROC}) and a monetary term (the overall monetary cost C_M = ARC + C_A; i.e. the sum of the average restoration cost and the assessment cost). This second decomposition clarifies the relation between a cost-effectiveness analysis (CEA) and a cost-benefit analysis (CBA). With a CEA, we fix a certain ecological objective without an explicit valuation, and then aim for a minimisation of the monetary costs C_M. With a CBA, we aim for an overall optimisation of B_M and consider the benefits of alternative ecological objectives. It is possible that a more costly index results in a higher overall benefit if the monetary costs are low in comparison to the ecological benefits.

The mathematical proofs of the equations are given in Box 4.2. To enhance insight, in the following paragraphs we will introduce step by step the terminology and make the cost decomposition intuitively plausible. Figure 4.1 sketches the flow of the calculations.

4.2.2.2. The ecological and societal cost C_0 before restoration

The situation before restoration C_0 (the first term of the cost decomposition) implies a FN cost for degraded sites and a TN cost for pristine sites (first table in Figure 4.1). There are only ecological costs as there is no restoration yet (see cost matrix in Box 4.1). With a prevalence of degradation π^+, the average cost is $E_{FN}.\pi^+$ for degraded sites and $E_{TN}.(1-\pi^+)$ for pristine sites. In total C_0 =

$E_{FN}.\pi^+ + E_{TN}.(1-\pi^+)$. As argued before, $E_{TN} = 0$ is a natural zero cost baseline. Then $C_0 = E_{FN}.\pi^+ =$ cost associated with degraded sites multiplied with the prevalence. Only if both π^+ and the FN cost are sufficiently high, there will be an incentive to set up a restoration program.

▾ **Cost situation before restoration**

Pristine sites	Degraded sites
$1-\pi^+$	π^+
$C_{TN}.(1-\pi^+)$	$C_{FN}.\pi^+$

▾ **Cost situation after restoration**

Pristine (TN)	Degraded (FP)	Pristine (FN)	Degraded (TP)
$(1-\pi^+).TNF$	$(1-\pi^+).FPF$	$\pi^+.FNF$	$TPF.\pi^+$
$C_{TN}.(1-\pi^+).TNF$	$C_{FP}.(1-\pi^+).FPF$	$C_{FN}.\pi^+.FNF$	$C_{TP}.\pi^+.TPF$

▾ **Cost difference**

$C_{TN}.(1-\pi^+).TNF + C_{FP}.(1-\pi^+).FPF - C_{TN}.(1-\pi^+)$	$C_{FN}.\pi^+.FNF + C_{TP}.\pi^+.TPF - C_{FN}.\pi^+$
$(C_{FP} - C_{TN}).(1-\pi^+).FPF$	$-(C_{FN} - C_{TP}).\pi^+.TPF$
$T_H.FPF$	$-T_G.TPF$

▾ **Overall restoration benefit**

$$ORB = B_{ROC} = T_G.TPF - T_H.FPF = T_G.b_{ROC}$$

$$T_H = (E_H + C_R).(1-\pi^+) \quad \& \quad T_G = (E_G - C_R).\pi^+$$

$$b = T_G/T_H = (E_G - C_R)/(E_H + C_R).odds(\pi^+)$$

$$b_{ROC} = TPF - FPF/b$$

▾ **Cost benefit analysis (CBA)**

$$B_M = ORB - C_A = ERB - (ARC + C_A) = ERB - C_M$$

▾ **Cost effectiveness analysis (CEA)**

$$C_M = ARC + C_A$$

Figure 4.1 **Sketch of the calculations.** The first table describes the situation before restoration: all sites are TN for pristine sites and FNs for degraded sites. The second table shows how the sites are classified by the index and assesses the corresponding costs. The third table is obtained by subtracting the first table from the second. The fourth table combines the gain and harm in the overall restoration benefit (ORB). As derived in the text and Box 4.2, the parameters can be expressed as functions of parameters reflecting the basic features of the management context. In the fifth and sixth table, the assessment cost is taken into account.

Box 4.2 **Derivation of the three term cost decomposition and definition of the key parameters.** We link the basic elementary parameters as C_{FN}, π^+ and efficacy of restoration (see Box 4.1) with a few, easy to understand, synthetic parameters defining the restoration context.

▷▷▷ *Cost decomposition*

$$\boxed{C_T = C_0 + C_A - B_{ROC}} \qquad \begin{cases} C_0 = E_{FN}\pi^+ + E_{TN}(1-\pi^+) = E_{FN}\pi^+ \quad (\textit{setting } E_{TN} = 0) \\ B_{ROC} = T_G TPF - T_H FPF = T_G b_{ROC} \quad \textit{with} \quad b_{ROC} = TPF - FPF / b \end{cases}$$

▷▷▷ *The restoration benefit corrected for monetary costs* (B_M)

$$\boxed{B_M = C_0 - C_T = B_{ROC} - C_A = E_{ROC} - C_M} \quad \textit{with} \quad \boxed{C_M = ARC + C_A} = \textit{monetary costs}$$

▷▷▷ *Gain / harm because of restoration*

$$R_G = C_{FN} - C_{TP} = (E_{FN} - E_{TP}) - C_R = E_G - C_R \quad \Rightarrow \quad T_G = R_G\pi^+$$
$$R_H = C_{FP} - C_{TN} = (E_{FP} - E_{TN}) + C_R = E_H + C_R \quad \Rightarrow \quad T_H = R_H(1-\pi^+)$$

▷▷▷ *Benefit ratio*

$$b_R = \frac{R_G}{R_H} = \frac{E_G - C_R}{E_H + C_R} \quad \& \quad b = \frac{T_G}{T_H} = \frac{R_G}{R_H}\frac{\pi^+}{1-\pi^+} = b_R odds(\pi^+)$$

▷▷▷ *Efficacy of restoration*

$$\eta_E = \frac{E_G}{E_{FN}} = 1 - \frac{E_{TP}}{E_{FN}} \quad \& \quad \eta_C = \frac{T_G}{C_0} = \frac{E_G - C_R}{E_{FN}} = \eta_E - \frac{C_R}{E_{FN}} = 1 - \frac{E_{TP} - C_R}{E_{FN}} \qquad note: b = \frac{\eta_C C_0}{T_H}$$

▷▷▷ *Overall restoration benefit* ($B_{ROC} = ORB$) ↔ *Ecological restoration benefit* ($E_{ROC} = ERB$)

$$\boxed{B_{ROC} = E_{ROC} - ARC} \qquad \begin{cases} E_{ROC} = E_G\pi^+ TPF - E_H(1-\pi^+)FPF \\ ARC = C_R(\pi^+ TPF + (1-\pi^+)FPF) \end{cases}$$
$$= T_G TPF - T_H FPF = (E_G - C_R)\pi^+ TPF - (E_H + C_R)(1-\pi^+)FPF$$
$$= \boxed{E_G\pi^+ TPF - E_H(1-\pi^+)FPF} - \boxed{C_R(\pi^+ TPF + (1-\pi^+)FPF)}$$

▷▷▷ *Derivation of cost decomposition*

$$\begin{aligned} C_T &= C_A + C_{TP}P_{TP} + C_{FN}P_{FN} + C_{TN}P_{TN} + C_{FP}P_{FP} \\ &= C_A + C_{TP}\pi^+ TPF + C_{FN}\pi^+(1-TPF) + C_{TN}(1-\pi^+)(1-FPF) + C_{FP}(1-\pi^+)FPF \\ &= C_{FN}\pi^+ + C_{TN}(1-\pi^+) + C_A - (C_{FN} - C_{TP})\pi^+ TPF + (C_{FP} - C_{TN})(1-\pi^+)FPF \\ &= (E_{FN}\pi^+ + E_{TN}(1-\pi^+)) + C_A - (R_G\pi^+ TPF - R_H(1-\pi^+)FPF) \\ &= C_0 + C_A - (T_G TPF - T_H FPF) \\ &= C_0 + C_A - B_{ROC} \end{aligned}$$

4.2.2.3. The gain because of restoring degraded sites

For degraded sites, we expect restoration to be beneficial: $E_{TP} < E_{FN}$. The ecological gain of restoration equals minus the change of the cost: $E_G = -(E_{TP} - E_{FN}) = E_{FN} - E_{TP} > 0$. Correcting for the restoration cost gives the cost-corrected restoration gain for the degraded site: $R_G = E_G - C_R$. If we restore all sites in the region, the expected gain at a regional level is $T_G = (E_G - C_R).\pi^+$. T_G integrates monetary and ecological costs in one single measure and represents the maximal gain that can be realised by restoring all degraded sites in the region. We define the cost-corrected efficacy of restoration as the ratio of T_G and the cost status before restoration: $\eta_C = T_G / C_0 = (E_G - C_R).\pi^+ / E_{FN}.\pi^+ = 1 - (E_{TP}+C_R)/ E_{FN}$. As π^+ cancels out, η_C only depends on the cost parameters at the individual site level. T_G is the first element defining the operational context of the index, i.e. it is a constraint the index cannot alter as it depends on external factors including the restoration cost and the residual ecological cost after restoration. As an index is not perfect, we only realise a benefit of T_G.TPF. The derivation is also sketched in the right part of Figure 4.1 (first three tables).

4.2.2.4. The harm because of restoring pristine sites and the benefit ratio

For pristine sites, the reasoning is analogous (left part of Figure 4.1). Restoration is now harmful and we expect $E_{FP} > E_{TN}$. We define the ecological harm as $E_H = (E_{FP} - E_{TN}) > 0$. Adding the restoration costs gives the total loss because of a wrong restoration: $R_H = E_H + C_R$. The total harm T_H for restoring all pristine sites is $(E_H + C_R).(1 - \pi^+)$. T_H is the second element defining the operational context of the index. It depends on the restoration cost and the harm caused to pristine sites by unnecessary restoration. As an index is not perfect, we cannot avoid FPs totally and have to accept a harm T_H.TPF.

4.2.2.5. The benefit ratio

As it will turn out, the relative magnitude of potential benefit (T_G) and harm (T_H) is a key parameter for the potential benefit of the index and its optimisation. We call the ratio of T_G and T_H the (overall) benefit ratio $b = T_G/T_H$. This ratio incorporates the tradeoff between the benefits and risks of restoration at a regional level. It is strongly influenced by the prevalence of degradation. It is possible to factorise b as a product of the odds of the prevalence $\text{odds}(\pi^+) = \pi^+/(1-\pi^+)$ and the intrinsic benefit ratio $b_R = R_G/R_H = (E_G - C_R) / (E_H + C_R)$. Both b_R and π^+ should be large to have a high benefit ratio. To avoid overly optimistic results, it is crucial to sufficiently consider the ecological risk associated with restoring pristine sites and not to underestimate it. The lower limit of the denominator ($E_H + C_R \geq C_R$) corresponds with the assumption there is no ecological harm ($E_H = 0$). We deplete budget by wrong allocation of the budget. This is the minimal price to pay when restoring pristine sites. Supposing the ecological risk is about equal to the restoration costs ($E_H \approx C_R$), decreases the benefit ratio by a factor two leading to a more prudent approach.

4.2.2.6. The tradeoff along the ROC curve (ORB = B_{ROC})

Combining above definitions, it is possible to derive the overall "index-based" restoration benefit B_{ROC}. The difference between both the benefit and harm results in the overall benefit realised: $B_{ROC} = T_G.TPF - T_H.FPF$. Interestingly, B_{ROC} can be rewritten as $T_G.b_{ROC}$ with the benefit function $b_{ROC} = TPF - FPF/b$ (≤ 1). This kernel only depends on the benefit ratio b and the ROC curve and we

should not know T_G to determine the optimal point. The equation $T_G.b_{ROC}$ integrates the operational context of the index (as summarised by b and T_G or $T_H = T_G/b$) and its intrinsic quality (as characterised by ROC curve) in one single formula. T_G refers to the maximal gain, b_{ROC} represents the fraction that is effectively realised. With a (close to) gold standard index: TPF ≈ 1 keeping FPF ≈ 0, $b_{ROC} \approx 1$ and $B_{ROC} \approx T_G$, or the full potential is realised.

4.2.2.7. Budget available for monitoring by avoiding false positives (A_{ROC})

If we restore all sites without making a distinction (TPF = FPF = 1), $B_{ROC} = T_G - T_H = T_H.(b - 1)$. If $b > 1$, $B_{ROC} > 0$. In this situation, it can be tempting to choose for a "blind restoration" as we save the costs associated with assessment. From a pure cost benefit perspective (neglecting other possible advantages of monitoring), the solution with an index is only more beneficial, if the total benefit realised including the assessment costs is higher than with blind restoration: $B_M = B_{ROC} - C_A > T_G - T_H$ or $A_{ROC} = B_{ROC} - (T_G - T_H) > C_A$. A_{ROC} is an upper limit for the restoration cost. It is the budget available for monitoring in comparison to blind restoration by avoiding false positives. Box 4.4 shows that, similarly to B_{ROC}, A_{ROC} can be rewritten as a product of a constant and a simple kernel function only dependent on b: $A_{ROC} = T_H.a_{ROC}$ with $a_{ROC} = TNF - b.FNF < 1$.

4.2.2.8. Graphical representation of the cost decomposition

Figure 4.2 gives a graphical representation of the three term decomposition for $b > 1$ ($T_G > T_H$). The left side corresponds with the situation before restoration with a baseline cost of C_0. The maximal gain (to realise) and harm (to avoid) is T_G and T_H respectively. Assessment with the index increases the cost to $C_0 + C_A$. This additional cost is recuperated by discriminating between degraded and pristine sites resulting in a benefit (T_G.TPF) keeping the harm (T_H.FPF) as small as possible. As will be shown in next section, we can optimise the index to maximise B_{ROC} (B_{Opt}). Thus after restoration, the cost is $C_T = C_0 + C_A - B_{Opt}$.

4.2.2.9. The maximal assessment budget

The scheme also visualises the maximal assessment budget available. First, the assessment cost should be lower than the overall benefit, otherwise the total benefit is smaller than zero: $C_A < B_{Opt} \leq T_G$. Secondly, $C_A < A_{Opt} \leq T_H$ with $A_{ROC} = B_{ROC} - (T_G - T_H)$, the extra benefit realised by the index in comparison to blind restoration by avoiding FPs. Combining the two inequalities, $C_A < \min(A_{Opt}, B_{Opt}) \leq \min(T_G, T_H)$, defining the maximal budget available for assessment.

If $E_H = 0$ and $b > 1$, $C_A < T_H = (1 - \pi^+).C_R$ which suggests it is not unreasonable to spend a budget on monitoring of the same order of magnitude as on restoration. For instance, for $\pi^+ = 0.2$, an assessment cost half as large as the restoration cost, is more cost-effective than blind restoration as $C_A = 0.5\ C_R < 0.8\ C_R$. If $E_H \approx C_R$ (possible harm about equal to restoration costs), the budget spent on monitoring may be even larger to prevent from harm. Evidently, this may not be understood as an invitation for blind monitoring neither. Monitoring is an overhead and should be kept as small as possible, but in absence of cheap alternatives, it is reasonable to invest in the development of indicators to reduce the costs.

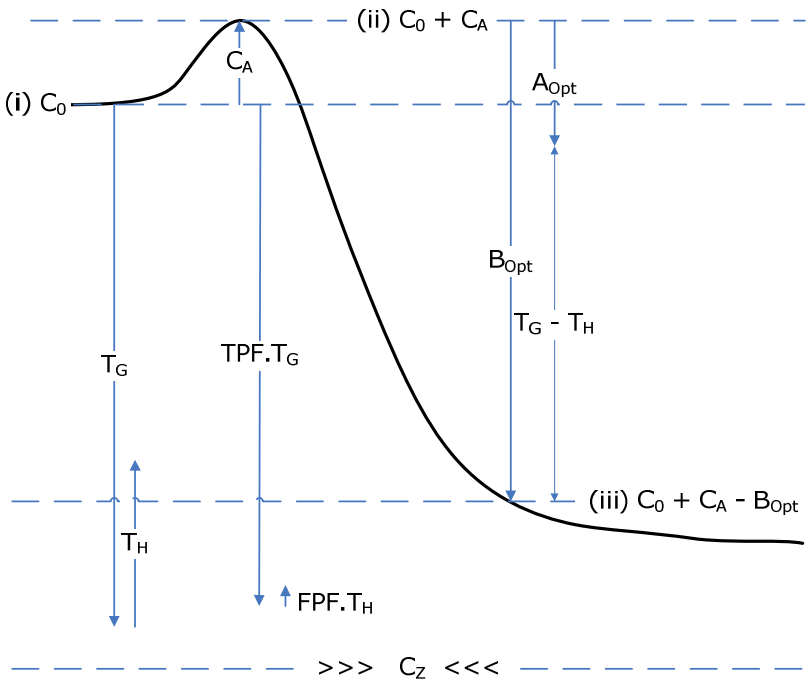


Figure 4.2 **Graphical representation of the basic cost decomposition.** The horizontal dashed lines specify the costs in comparison to C_Z (zero level). (i) On the left, the situation before restoration (C_o) is given. (ii) To restore, we have first to assess the situation adding a cost C_A. (iii) Restoring the sites results in a benefit (B_{Opt}) decreasing the costs. The arrows left characterise the decision context: the maximal possible gain & harm (T_G & T_H). In arrows in the middle represent the gain effectively realised (T_G.TPF) and the harm not avoided (T_H.FPF). The difference between these two terms in the optimal decision point (B_{Opt}) should be larger than the assessment costs (C_A): $B_{Opt} > C_A$. Also $A_{Opt} > C_A$. With blind restoration, the gain is $T_G - T_H$. From a purely cost perspective $B_{Opt} - C_A > T_G - T_H$ or $A_{Opt} = B_{Opt} - (T_G - T_H) > C_A$ which is depicted by the arrows on the right.

4.2.3 Tuning the index (determining the optimal decision point)

Moving the decision point along the ROC curve changes the relative proportion of TPs and FPs and hence the balance B_{ROC} between gain and harm realised by the index. The optimal decision point is where B_{ROC} is maximal. We give a mathematical and economical argumentation.

4.2.3.1. Mathematical optimisation

For concave ROC curves (lying above the first diagonal), the benefit starts at zero: $B_{ROC} = 0$ as in the origin FPF = TPF = 0, corresponding with no restoration at all. First, it increases up to a maximum B_{Opt} and then decreases to the limit value where all sites are restored: $B_{ROC} = T_G - T_H = (b - 1).T_H$ as in the end FPF = TPF = 1, corresponding to a (blind) restoration of all possible sites. The maximum is reached where the first derivative of B_{ROC} is zero. This is where the slope of the ROC curve equals 1/b (Box 4.4).

Box 4.3 **Cost-effectiveness analysis (CEA).** Average restoration cost (ARC), true restoration fraction (TRF) and the overall monetary cost (C_M)

$\triangleright\triangleright\triangleright$ *Gold standard cost* (C_G) $\leftrightarrow$ *Average restoration cost* (ARC)

$$\boxed{C_G = C_R\pi^+} \quad \leftrightarrow \quad \boxed{ARC = C_R(TP) + C_R(FP) = C_R\pi^+ TPF + C_R(1-\pi^+)FPF}$$

$\Rightarrow$ *The overall monetary cost* : $\boxed{C_M = C_A + ARC}$

$\triangleright\triangleright\triangleright$ *The maximal budget available for the gold standard* $(A_{\max})$

$$gold\ standard: ARC_G = ARC(Se = 1, FPF = 0) = C_R\pi^+$$
$$blind\ restoration: ARC_B = ARC(Se = FPF = 1) = C_R$$

$$\Rightarrow \quad \boxed{\frac{ARC_B}{ARC_G} = \frac{1}{\pi^+} \quad \& \quad A_{\max} = ARC_B - ARC_G = (1-\pi^+)C_R}$$

$\triangleright\triangleright\triangleright$ *True Restoration Fraction* (TRF) = *Positive Predictive Value* (PPV)

$$\boxed{TRF \triangleq \frac{C_R(TP)}{ARC} = \frac{C_R(TP)}{C_R(TP) + C_R(FP)} = \frac{\cancel{C_R}\,\pi^+ TPF}{\cancel{C_R}\,\pi^+ TPF + \cancel{C_R}(1-\pi^+)FPF} = PPV}$$

$TRF \geq 0.75$ $(or\ 0.5)$

4.2.3.2. Economical optimisation

From an economical perspective, we can reason as follows. In the vicinity of the origin, with low FPF values, the ROC curve increases fast and the marginal gain of restoring degraded sites is larger than the marginal harm of treating pristine sites ($T_G\ \Delta(TPF) > T_H\ \Delta(FPF)$) implying $\Delta B_{ROC} > 0$. Gradually, the ROC curve levels off, the term $T_H\ \Delta FPF$ becomes more important and ultimately starts to dominate. The benefit along the ROC curve has reached its maximum and starts to decrease. At this point, $T_H\ \Delta(FPF) = T_G\ \Delta(TPF)$, or, more formally the first derivative of the benefit function is zero.

4.2.3.3. The impact of the benefit ratio

A small benefit ratio b results in a steep slope (1/b), implying that the optimum is reached in the left part of the ROC curve. Sensibly, with a low benefit and/or a high risk, it is advisable to follow a restrictive policy, keeping the FPF low at the expense of TPF. Conversely, a large b implies a flat slope (1/b), found in the right liberal part of the curve. Because the expected benefit is high and/or the risk low, a high FPF is recommended to realise a high TPF.

Box 4.4 **Tuning of the index.** The optimal decision point is where the B_{ROC} is maximal.

▷▷▷ *Equations ROC curve*

$$Function: triplet\,(T, FPF, TPF) \quad \leftrightarrow \quad \boxed{\begin{cases} FPF = F_D(T) \\ TPF = F_R(T) \end{cases}} \quad \leftrightarrow \quad TPF = F_R(F_D^{-1}(FPF))$$

$$Coordinate\ on\ ROC\ curve: \quad (FPF, TPF) \quad e.g. \quad \begin{cases} (0,0) = origin \\ (1,1) = endpoint \end{cases}$$

$$First\ derivative: \boxed{\frac{dTPF}{dFPF} = \frac{dTPF/dT}{dFPF/dT} = \frac{f_R(T)}{f_C(T)}} \quad (chain\ rule)$$

▷▷▷ *Overall benefit along ROC curve*

$$\boxed{B_{ROC} = T_G TPF - T_H FPF = T_G.b_{ROC} \quad < \quad T_G = b.T_H} \quad \rightarrow \quad kernel: \quad \boxed{b_{ROC} = TPF - FPF/b} \quad < \quad 1$$

$$b_{ROC}(0,0) = 0 \quad \leftrightarrow \quad b_{Opt} < 1 \quad \leftrightarrow \quad b_{ROC}(1,1) = 1 - \frac{1}{b}$$

$$\boxed{B_{ROC} = B_{Opt} \quad \Leftrightarrow \quad \frac{\partial b_{ROC}}{\partial FPF} = 0 \quad \Leftrightarrow \quad \frac{\partial TPF}{\partial FPF} = \frac{1}{b} \quad \Leftrightarrow \quad T_G \Delta TPF = T_H \Delta FPF}$$

▷▷▷ *Benefit in comparison to blind restoration = assessment budget available by avoiding FPs*

$$\boxed{A_{ROC} \triangleq B_{ROC} - (T_H - T_G) = T_H a_{ROC} \quad < \quad T_H} \quad \rightarrow \quad kernel: \quad \boxed{a_{ROC} = TNF - bFNF} \quad < \quad 1$$

$$A_{ROC} \overset{E_H = 0}{=} (1 - \pi^+) C_R a_{ROC} \quad \Rightarrow \quad \boxed{C_A < A_{Max} = (1 - \pi^+) C_R}$$

4.2.4 The average restoration cost (ARC) and the true restoration fraction (TRF)

To investigate the monetary implications of an index, we calculate the average restoration cost (ARC) and analyse which fraction of the budget is correctly allocated to degraded sites (TRF).

4.2.4.1. The average restoration costs (ARC)

With perfect knowledge, ARC = $C_R.\pi^+$, i.e. a simple multiplication of the restoration cost (per site) C_R and the prevalence of degraded sites. In reality, some degraded sites will not be detected (FN)

and some pristine sites will be wrongly restored (FP). The expected expenditure is $C_R.\pi^+.TPF$ for degraded sites and $C_R.(1-\pi^+).FPF$ for pristine sites: $ARC = C_R.\pi^+.TPF + C_R.(1-\pi^+).FPF$ (Box 4.3).

4.2.4.2. The assessment budget available for a high quality index

For a gold standard index (TPF = 1, FPF = 0), $ARC = C_R.\pi^+$. At the other extreme, blind restoration of all sites (TPF = FPF = 1) results in $ARC = C_R$. In comparison to the gold standard, we loose $(1-\pi^+).C_R$ on restoration of pristine sites and the budget of blind restoration is $1/\pi^+$ times as high. If $C_A < (1-\pi^+).C_R$, we are better of with the gold standard index. For instance, if $\pi^+ = 0.2$ & $C_A = 0.5\ C_R$ for (a close to) gold standard, the total management cost is $C_M \approx 0.5\ C_R + 0.2\ C_R = 0.7\ C_R$.

4.2.4.3. True restoration fraction (TRF) and positive predictive value (PPV)

We define the part of the restoration budget correctly attributed to degraded sites as the true restoration fraction (TRF). $TRF = C_R.\pi^+.TPF\ /\ (C_R.\pi^+.TPF + C_R.(1-\pi^+).FPF) = PPV$ (= the positive predictive value). If the manager has a limited budget, it can be an option to require TRF is larger than ¾ or even ½ to no deplete resources on the restoration of pristine sites.

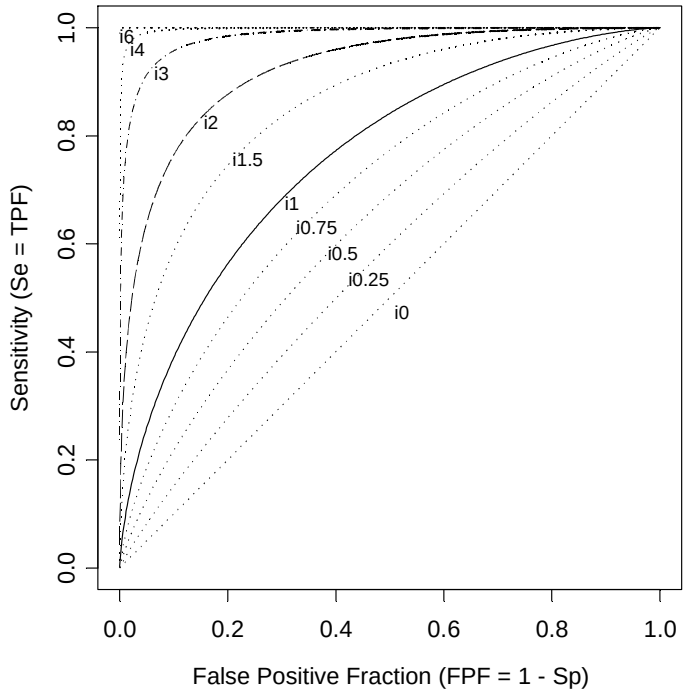


Figure 4.3 **Definition of the indices (i0 → i6)**. Ranking of the indices with respect to the diagnostic accuracy: i0, i0.25, i0.5 = (close to) indifferent indices <<< i1 < i2 < i3 <<< i4, i6 = (close to) gold standard indices. The numbers refer to the separation under reference and degraded conditions (standardised units).

4.2.5 The relation between diagnostic accuracy and the assessment cost

To explore the tradeoff between the diagnostic accuracy and the assessment cost, we introduce hypothetical indices ranging from almost no diagnostic accuracy (indifferent indices) to (nearly) gold standard quality. We assume a quadratic relationship between the index quality and C_A.

4.2.5.1. The hypothetical indices (ranging from indifferent to gold standard)

The ROC curves of the hypothetical indices are presented in Figure 4.3. We assume a homoskedastic binormal distribution. The underlying test variables are normally distributed and the variance is equal for pristine and degraded sites. We varied the standardised distance Δ/σ from 0 to 6 units to cover a broad range of index quality. Index i0 represents the indifferent index with no separation at all, and i6 is the (nearly perfect) gold standard. The ROC curve of i0 is the positive diagonal (TPF = FPF) and, the ROC curve of i6 equals one (TPF = 1) for the total FPF range. Index i1 represents a weak index and i3 is typical for a strong but not prefect index; i2 characterises an intermediate situation.

4.2.5.2. A typology of the indices

To connect the hypothetical indices to real situations, we present a loose typology. The indifferent index i0 is equivalent to no assessment. The indices between i0 and i1 (i0.25, i0.5, i0.75) are slightly better. They model a situation in which sites are inspected superficially at a minimal cost. For the calculations, we pick i0.5 as the representative for these nearly indifferent indices.

The range (i0.5) i1 – i2 (i3) covers experience-based judgement where managers make decisions based on personal experience and expertise, combined with historical information, reports of the fieldwork and other convenient data collected ad hoc. Although strictly speaking, decisions are not index-based, in principle, it is possible to establish sensitivity and specificity of the decision process and the assessment costs, for example as a part of a quality assurance program. If the assessment is integrated in daily practice, the overhead can be small, but it is not unlikely that the decision costs are hidden and are much larger than expected. To guarantee a fair comparison with a "real" index, it is necessary to determine the costs accurately. Although experience-based judgement has its merits, a major disadvantage is the lack of standardisation not guaranteeing a constant quality and the objectivity can be easily contested when many parties are involved.

Next, we consider the range (i1) i2 – i3 (i4) as models for a "real" index. Important characteristics include: the index is calibrated and validated with respect to a gold standard; the measurements for the index are standardised and described in a protocol; the interpretation of the index outcomes is predefined or rules exist for its interpretation; the end users are trained to integrate the information in the decision process and know how to cope with special cases. Thus, the approach does not preclude personal judgement and experience, or the use of additional information sources. Generally, assessment costs will be an order of magnitude larger than the previous category. On the other hand, because of the standardisation, assessment costs can be reasonable. Standardisation alleviates the decision burden and prevents from protracted discussions about the interpretation of the results. The end users trust the signals of the index knowing its overall performance.

At the other end of the spectrum, the range (i3) i4 – i6 are gold standard indices with high assessment costs precluding a practical application except for validation, follow-up studies and scientific research. The ROC curves of i4 and i6 do not differ much (Figure 4.3), yet i6 outperforms i4 for the average restoration cost (see results). We imagine i4 as a cost optimized version of the gold standard: a very high diagnostic accuracy is maintained, but at a much lower price. This cost reduction can be meaningful when a close to gold standard reference is necessary for (routine) quality assurance / quality control programs.

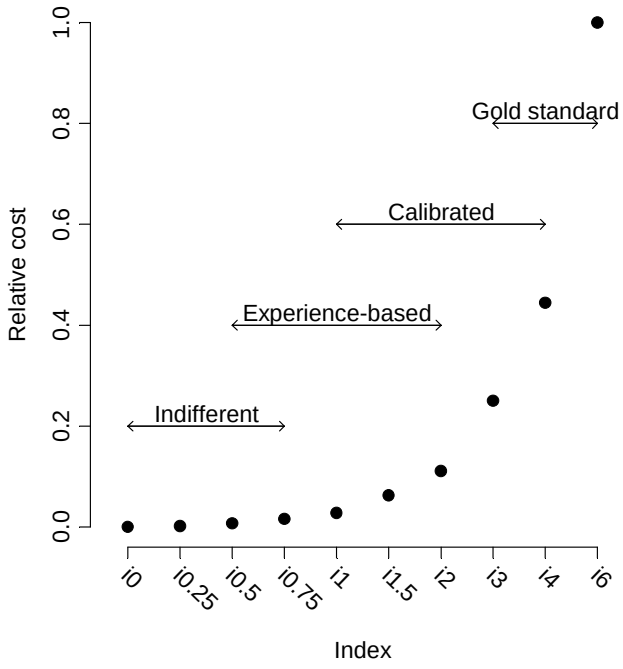


Figure 4.4 **Quadratic cost model.** Indices: i0, i0.25, i0.5 = (close to) indifferent indices <<< i1 < i2 < i3 <<< i4, i6 = (close to) gold standard indices.

4.2.5.3. The cost as a function of the diagnostic accuracy

We assume a quadratic relationship between diagnostic accuracy and assessment costs to model that the assessment cost rapidly increases as a function of the index quality. The equation of Figure 4.4 is $C_A = q^2.(1-\pi^+).C_R$, with $0 \leq q = (\Delta/6\sigma) \leq 1$. Δ/σ represents the standardised difference between the distributions of degraded and pristine sites. The total management cost is $C_M = ARC + C_A = \pi^+.TPF + (1- \pi^+).(FPF + q^2)$. The purpose is to minimise this function. (Box 4.5).

As $q = 1$ for the gold standard index ($\Delta/\sigma = 6$), $C_A = (1-\pi^+)C_R$, i.e. the maximal budget available for assessment in comparison to a blind restoration. For this value, the overall management cost $C_M = \pi^+.C_R + (1-\pi^+).C_R = C_R$, which is equal to the cost of restoring all sites blindly (Box 4.3). By construction (but realistically), we exclude the gold standard index from an application in practice.

The interesting question is whether i4, the index developed as a close, but cheap alternative of the gold standard, will be competitive in comparison to i3, i2 and i1. For i4, q^2 = 4/9 = less than half the cost of the gold standard, while for i3, i2 and i1, q^2 = 1/4, 1/9, 1/36. The assessment costs rapidly decrease by lowering diagnostic accuracy, and we will investigate from which point on the reduction of assessment cost is offset by the loss because of misclassification.

Box 4.5 **Quadratic cost model for assessment**.

▷▷▷ *Total monetary cost* (C_M) & *assessment fraction of the monetary cost* (AFM)

$$\boxed{C_M = ARC + C_A} \quad \& \quad AFM = \frac{C_A}{C_M}$$

▷▷▷ *Quadratic cost model*

$$\boxed{C_A = c_0 + c_m q^m} \qquad 0 \le q = \frac{1}{6}\frac{\Delta}{\sigma} \le 1 \qquad \frac{\Delta}{\sigma} = \textit{standardised distance}\ (0 \to 6)$$

$$\overset{\textit{quadratic model}}{\Rightarrow} \qquad C_A = c_2 q^2 \quad \Rightarrow \quad \begin{cases} C_A(i0) = 0 \\ C_A(i4) = \frac{4}{9} c_2 < C_A(i6)/2 \\ C_A(i6) = c_2 \end{cases}$$

▷▷▷ *Cost effectiveness analysis* (CEA)

$$C_M(i6) = C_M(i0) = C_R$$
$$= AVG(i6) + C_A(i6) = \pi^+ C_R + c_2 \quad \Rightarrow \quad c_2 = (1-\pi^+) C_R$$
$$\Rightarrow \quad \boxed{C_A = (1-\pi^+) q^2 C_R}$$

▷▷▷ *Cost benefit analysis* (CBA)

$$B_T(i6) = B_T(i0) = T_G - T_H$$
$$= T_G - C_A(i6) = T_G - c_2 = T_G - T_H \quad \Rightarrow \quad c_2 = T_H$$
$$\Rightarrow \quad \boxed{C_A = T_H q^2 = T_G q^2 / b} \quad = \quad (1-\pi^+)(1+h) q^2 C_R \quad (\textit{as } T_H = (1-\pi^+)(1+h) C_R \textit{ with } E_H = h C_R)$$
$$\Rightarrow \quad \boxed{B_T = B_{ROC} - C_A = T_G (Se - FPF/b - q^2/b)}$$

▷▷▷ *Assessment fraction* (AFM)

$$\boxed{C_M = (\pi^+ Se + (1-\pi^+)(FPF + q^2(1+h)))\, C_R} \quad (h = 0 \textit{ for } CEA)$$

$$\boxed{AFM = \frac{(1-\pi^+)(1+h) q^2}{\pi^+ Se + (1-\pi^+)(FPF + q^2(1+h))}}$$

4.3 Results

As the benefit ratio b is a crucial parameter governing the maximal benefit that can be achieved, we first discuss its plausible range to cover the full spectrum of possibilities. Next, we investigate how to tune indices, i.e. determine the optimal decision point on the ROC curve. This tuning guarantees a fair comparison at the points where the indices achieve the highest benefit. Then, we evaluate the cost implications of the optimised indices to retrieve the most cost-effective index, i.e. the index minimising the total monetary cost C_M (sum of the restoration and assessment cost) to realise a preset objective. Based on this CEA analysis, we determine which fraction of C_M should be optimally allocated to assessment, contributing to the discussion whether we spend sufficiently on monitoring. In complement, we analyse the cost problem from a CBA perspective integrating monetary, ecological and societal costs and benefits. The best index realises the highest total benefit taking into account ecological and societal benefits.

4.3.1 The range of the benefit ratio

4.3.1.1. The factors determining b

The benefit ratio $b = T_G/T_H = b_R.\text{odds}(\pi^+)$ with $b_R=(E_G-C_R)/(E_H+C_R)$ expresses the balance between the potential benefit and risk of the restoration at a regional level. A benefit ratio of one ($b = 1$) implies that the expected gain and harm are equal to each other. If $b > 1$, the expected gain is larger, if $b < 1$, the expected gain is lower. It is important to notice the strong influence of the prevalence of degradation. The parameter b can be factorised as a product of two factors: (i) the "intrinsic" restoration benefit ratio on a site level, and, (ii) the (odds of the) prevalence of degraded sites. Even if the restoration benefit of a local site is intrinsically high, b can be quite low if the prevalence is low. The fundamental reason is that the tradeoff between benefit and harm depends on their expected values at a regional level. A low prevalence implies there are many pristine sites, increasing the expected number of pristine sites falsely restored which negatively affects the benefit ratio. This relationship also makes clear, in order to make decisions, we should reason at a regional level, i.e. consider the totality of the sites.

4.3.1.2. The range of b

The benefit ratio is a key parameter for both the optimal tuning of the indices and the selection of the best index. However it is not always possible to have an accurate estimate of the benefit ratio as the underlying cost parameters are hard to estimate. By considering different alternative scenarios, we can explore the impact of the uncertainty on the optimum with respect to changes in the benefit ratio. A trial and error exploration of b_{ROC} (Figure 4.7) showed that ranging the benefit ratio b from ½ to 8 in multiples of 2 (½, 1, 2, 4, 8) covers a broad array of possibilities. To make this range more plausible, we consider a hypothetical example in detail and demonstrate at the same time how to assess b if detailed ecological information is lacking. The idea is to express all costs relative to the restoration cost C_R by educated guesses. We also found a real example in literature resulting from a full-blown economical study about the costs to restore a koala population (Field *et al.*, 2004). This example is instructive because, as mentioned by the authors, it is extreme with a very high ecological benefit in comparison to the restoration costs.

4.3.1.3. The hypothetical example

In absence of real data, for the hypothetical example, we express the decision costs of Box 4.1 relative to the restoration cost C_R and make some guesses about the possible gain and harm of the restoration program. We assume $E_{FN} = 12C_R$, $E_{TP} = 2C_R$, $E_{TN} = 0$ and $E_{FP} = 2C_R$. With this cost configuration, the ecological gain of a restoration (see Box 4.2 for the equations) is ten times the restoration cost ($E_G = E_{FN} - E_{TP} = 10C_R$) and the ecological harm is five times as small ($E_H = E_{FP} - E_{TN} = 2C_R$). The (cost-corrected) effectiveness of the restoration $\eta_C = (E_G - C_R) / E_{FN} = 9\ C_R/12\ C_R = 0.75$ and the intrinsic benefit ratio $b_R = (E_G - C_R) / (E_H + C_R) = 9\ C_R/3\ C_R = 3$. Setting the odds ratio of the prevalence equal to 1, 2/3, 1/3 and 1/6 (or $\pi^+ = 0.5, 0.4, 0.25$ and 0.14) results in b = 3, 2, 1 and ½. As an (optimistic) alternative, we assume the ecological harm of restoring pristine sites is zero ($E_H = 0$); then $b_R = 9$ (three times as large), resulting in b = 9, 6, 3 and 1 ½. Varying b from ½ to 8 covers the range of this example well.

4.3.1.4. Monitoring a koala population

The previous example makes clear that the choice of both E_H and π^+ is quite critical as we will demonstrate with the example of Field *et al.* (2004). In this paper, the cost of a koala population decline was estimated to be $ 21 million (E_{FN}) and the restoration cost $ 0.83 million (C_R). Hence, $E_{FN} \approx 25C_R$, which is twice as large as with the hypothetical example. If we set (as in the example above) $E_H = 2C_R$, then $b_R = (25C_R - C_R) / (2C_R + C_R) = 8$. In the paper, the authors assumed π^+ = ½ or odds(π^+) = 1 (in the framework of the paper, π^+ = the a priori probability of the decline of the koala population). As a consequence, $b = b_R = 8$. In contrast, the equations in Field *et al.* (2004) do not take into account the harm of a false restoration implying $E_H = 0$. As a consequence, $b = b_R = (25C_R - C_R)/(0+C_R) = 24$. In spite of this large difference ($b = 8 \leftrightarrow 24$), our calculations turned out to be very close to Field *et al.* (2004) because b = 8 turns out to be already extreme. Yet, the example demonstrates the necessity of a correct estimation of the risk to prevent from an inflation of the benefit ratio. It is important to mention that although our numerical results are very similar to the paper discussed, our conclusions are quite different from the authors. A point we return to at the end of the discussion.

4.3.2 The choice of the optimal decision point

As the decision threshold T uniquely defines the coordinates on the ROC curve, we can use any element of the triplet (T, FPF, TPF) to refer to the decision point. To facilitate the link with ROC curves which plot TPF versus FPF, we consistently choose FPF as the x-axis and TPF as the y-axis. We restrict the analyses to three indices (i1 < i2 < i3) to avoid cluttering of the graphs. Only when discussing the selection of the optimal index, we consider the full range of the indices (i0 → i6).

4.3.2.1. The position of the optimal decision point on the ROC curve

For all three indices, the optimal decision point realising the highest benefit as derived in Box 4.4 shifts to the right as the benefit ratio increases (Figure 3.8). This pattern is most pronounced for the weakest index (i1), where the decision point ranges from FPF = 0.12 (b = ½) to 0.94 (b = 8). In contrast, for the best index (i3) the FPF range is limited from 0.04 (b = ½) to 0.21 (b = 8). A special case is when the expected gain equals the expected harm (b = 1). Then, sensitivity and

specificity are balanced (TPF = TNF). This property is not true anymore if the variances of the degraded sites are not equal anymore (Figure 4.6). Yet, the main pattern remains essentially the same: increasing b shifts the decision points to the right and the effect is most pronounced for weaker indices. In comparison to the homoskedastic model, the impact of b appears to be larger: the optimal decision points are more spread over the FPF range.

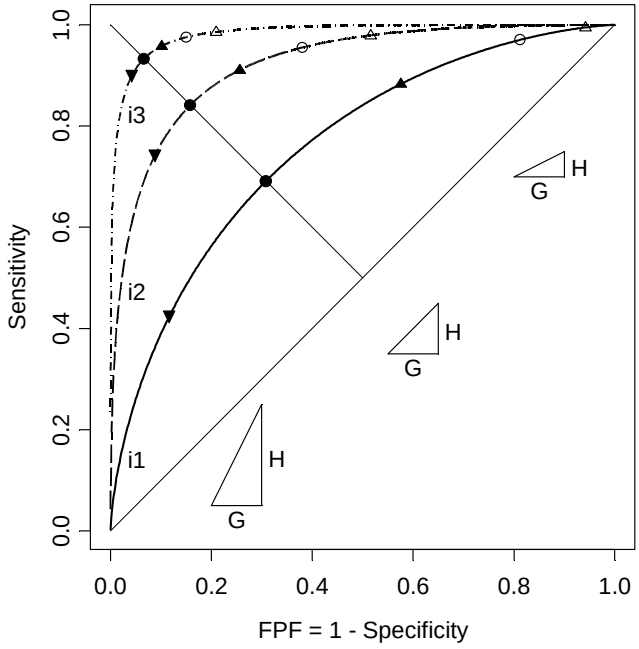


Figure 4.5 **Optimal decision points on the ROC curve** for a binormal equal variance model as function of the index quality (i1 < i2 < i3) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, Δ = 8). Triangles: slope of ROC curve at optimal decision point for b = ½ (left), 1 (middle) and 2 (right).

The observed pattern is easy to understand. A high benefit ratio implies a higher potential benefit T_G in comparison to the risk T_H. We can afford a higher FPF, because extra TPs compensate the loss. The effect of b is smaller on strong indices, because they have a steeper ROC curve: only a small increase of FPF is necessary to realise a higher sensitivity.

From a mathematical perspective, the optimum is where the slope of the curve equals 1/b. For b = 1, the slope is one, which is situated on the negative diagonal for the equal variance binormal model which results in a symmetric ROC curve. A larger b-value (> 1) implies a flat slope (1/b < 1) and the optimal decision point is shifted to the upper liberal part of the ROC curve. Conversely, a small b (< 1) results in a high slope (1/b > 1) found in the lower more conservative part of the index. As strong indices have steep ROC curves with a rapidly changing slope, the optimal decision points as a function of b are situated closely to each other. Conversely, as weak indices have a flat ROC curve, the optimal decisions points are (much) further apart.

Also from an economical reasoning, the location of the optimal decision points is logical. A general principle is that the cost optimum is reached where the marginal profit becomes equal to the marginal loss. For the cost balance T_G.TPF – T_H.FPF, this is where the marginal benefit of increasing TPF (T_G.ΔTPF) equals the marginal harm because of the simultaneous increase of FPF (T_H.ΔFPF). If T_G is small compared to T_H (b small), the balance between benefit and loss is reached early in the conservative region of the ROC curve, whereas, if T_G is large in comparison to T_H, the optimal point is located in the liberal part. The triangles in Figure 3.8 visualize this cost tradeoff for three cost scenarios (b = ½, 1 and 2). The slopes of triangles are equal to the slope of the ROC curve in the optimal decision point.

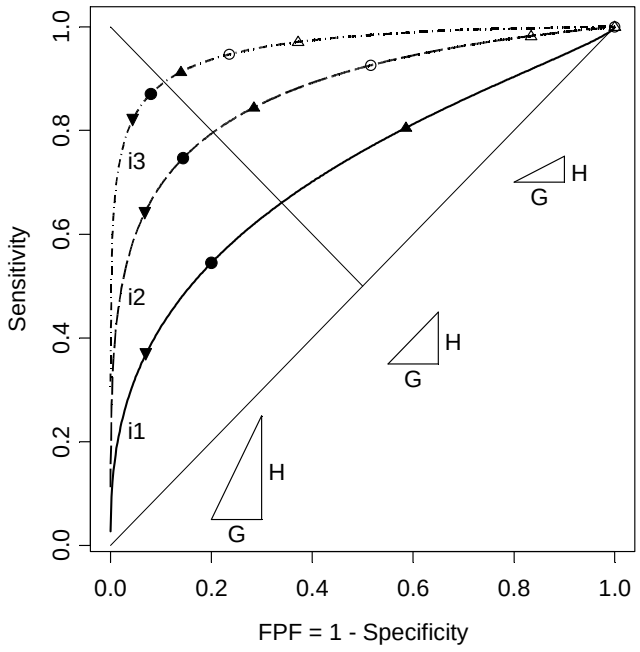


Figure 4.6 **Optimal decision points on the ROC curve** for a binormal model with unequal variance (variance in degraded sites twice as large as in the pristine sites) as function of the index quality (i1 < i2 < i3) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). Triangles: slope of ROC curve at optimal decision point for b = ½ (left), 1 (middle) and 2 (right).

4.3.2.2. The change of the benefit along the ROC curve (sensitivity analysis)

The overall benefit realized by restoration (B_{ROC}) changes along the ROC curve proportionally with the benefit function b_{ROC} = TPF – FPF/b ≤ 1. This kernel of B_{ROC} is 0 at the origin of the ROC curve (FPF=0, TPF=0) and is equal to (1 – 1/b) at the end (FPF = 1, TPF = 1). In between, a maximum is reached (Figure 4.7) as ROC curves are concave. As observed before, there is strong interaction between the index quality and the benefit ratio. The general pattern is that the optima shift to the right and become higher with an increasing benefit ratio. A higher index quality results in a larger

benefit and the optima are achieved sooner. For i3, the benefit curve climbs fast and reaches its maximum early. The maxima are lying in a narrow FPF range. The curves of the weaker indices are flatter. The maxima differ strongly and are spread over a broad FPF range. With b small, the optima are several times smaller than what is maximally possible.

These curves give a better understanding of the position of the optimal decision point on the ROC curve (Figure 4.5) and clarify the robustness of strong indices to variations of b. It is instructive to notice that TPF–FPF/b is an ROC curve (TPF versus FPF) from which a penalty (FPF/b) is subtracted. The ROC part represents the gain because of restoring degraded sites, the penalty part the negative effect of restoring pristine sites. We may expect that the benefit curves resemble the ROC curves: steep for strong indices and slow for weak indices. As the penalty term decreases with b, its impact is down weighted resulting in higher maxima and flatter curves, resembling more and more the ROC curve (for $b \rightarrow \infty$, the benefit curves become equivalent to ROC curves).

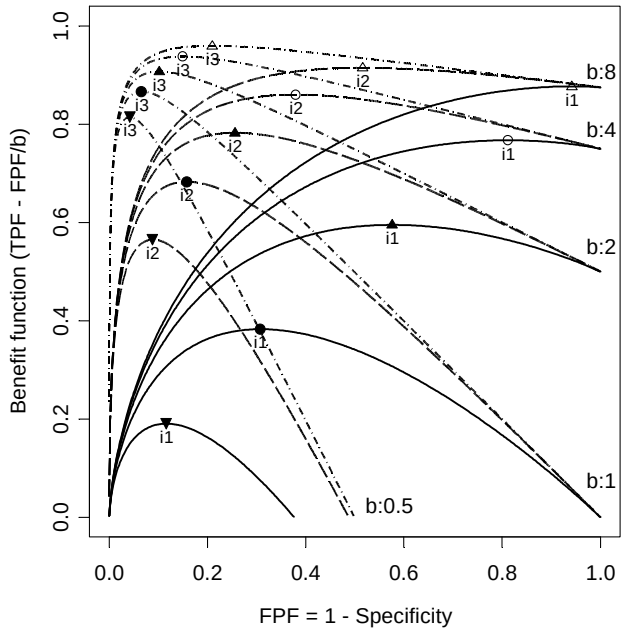


Figure 4.7 **The benefit function** (b_{ROC} = TPF – FPF/b)**.** Evolution along the ROC curve of the overall benefit in units relative to T_G as a function of the quality of the index (i1 < i2 < i3) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8).

In reality, we seldom know the value of b exactly. It is often an educated guess. The same holds for the strength of the index. We only have an estimate of the curve. Also it is not always possible to choose the optimal decision threshold. For instance, for comparability in an international context, the decision thresholds should be kept equal. It is mandatory to make sensitivity analyses

assessing the impact on the uncertainty on some of the parameters. For instance, the robustness of a strong index implies, that small uncertainties in the value of b will have less impact than on weak indices. On the other hand, when b < 1, for all indices the choice of the threshold is very critical. For a weak index, a wrong choice can result in a loss instead of a gain.

4.3.2.3. The true restoration factor as a criterion to choose the optimal decision point

The ecological benefit of restoration does not generally result in a direct monetary benefit for the manager. It is an advantage in the long run for the stakeholders or the society as a whole. Unless the society agrees with the cost benefit analysis and is willing to pay (WTP), the benefit is not available as an extra budget. Within the constraints of the budget, the manager should seek for an optimal solution. One possible way is to control the misallocation of the budget to pristine sites for instance by requiring the true restoration fraction (TRF) not dropping below a certain threshold, say ¾ or ½. For a prevalence of 20 % ($\pi^+ = 0.2$), Figure 4.8 compares the optimal decision points with this threshold. The TRF curves drop fast and, consequently, most points are located below the threshold. Only for index i3, there are two points (b = ½ & 1) for which TRF > ¾. In contrast, for i2, always TRF < ¾, and for i1 TRF < ½.

All curves start at one and decrease to π^+ (= 0.2) in a monotonic way. For i1, the curve drops very fast below ¾ at FPF = 0.01 and below ½ at FPF = 0.18. But also for the much stronger i3, with a steeper ROC curve, TRF drops below ¾ and ½ quite early (FPF = 0.11 & 0.33 respectively). This is a general pattern for any value of the prevalence. At the end of the ROC curve (TPF = FPF = 1), all sites are restored irrespective their status. Hence, by definition, fraction of the budget correctly allocated to degraded sites is equal to the prevalence (TRF = π^+). Conversely, at the origin of the ROC curve, no sites are restored at all. In this limit FPF ≈ 0, or TRF = π^+.TPF / (π^+.TPF + (1-π^+).FPF) ≈ 1. The steep decrease of TRF as a function of FPF is because the weight of FPF in the denominator is relatively high (~ (1- π^+) = 0.8) compared to the other term (TPF).

The analysis highlights that in the optimal decision point a considerable part of the budget is spent on restoring pristine sites (= 1 – TRF) and this part increases fast with b. Apparently, optimising for the overall benefit pushes in the direction of a liberal policy resulting in a high fraction of the budget spent on restoring pristine sites. This effect is very important for weak indices and (much) smaller for strong indices. This observation suggests that, especially if the expected benefit is high (then TRF → 1 for weak indices), it can be advantageous to choose for a better index even if the assessment costs are (considerably) higher.

Another reason to prefer stronger indices, is their higher robustness against uncertainties in the benefit ratio. There is a general tendency to overestimate benefits and underestimate risks. The overestimation of b by insufficiently discounting the possible harm can result in a serious overall loss, especially for weak indices. As an example, suppose we set b = 1 while it is in reality ½. From Figure 4.7, for b = 1, FPF ≈ 0.35 in the optimal decision point of i1. However, FPF ≈ 0.35 for b = ½ results in a benefit that is nearly zero (see the line for b = ½ in Figure 4.7). This effect is less dramatic for i2. For b = 1, FPF ≈ 0.15 and for b = ½ the optimum is only slightly below its optimum.

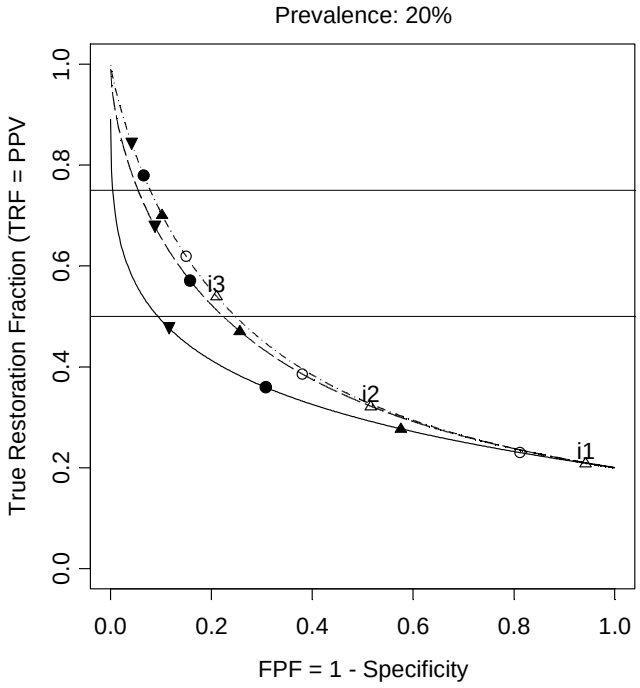


Figure 4.8 **The TRF function.** Evolution along the ROC curve of the true restoration fraction (TRF = PPV) and the position of optimal decision points as a function of the index quality (i1 < i2 < i3) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). The horizontal lines fix TRF at 0.75 and 0.50. Prevalence of degradation = 20 % (n^+ = 0.2).

4.3.3 Cost implications as a function of the index quality and the benefit ratio

The trajectories in Figure 4.9 to Figure 4.12 represent indices after tuning to guarantee a fair comparison (i.e. selecting the optimal decision point taking into account the benefit ratio). They show how the optimal benefit (y-axis) and different cost measures (x-axis) change as a function of the benefit ratio. In this way, we combine a CBA with CEA perspective for the full range of indices.

4.3.3.1. The FPF in the optimal decision point

The trajectories in Figure 4.9 show how both FPF and b_{Opt} (optimal benefit in units relative to T_G) increase as a function of the benefit ratio. The strong indices (i6, i4 and i3) are quite robust: the benefit remains high and FPF is controlled well. In contrast, for the weaker indices (i0.5 – i2), the trajectories are stretched over a broad range. For large b values (b ≥ 4), the overall benefit of weak indices is nearly as high as for strong indices, but the high benefit comes with a large FPF. Choosing for a better index mainly improves the control of the FPs while, relatively, the gain in the overall benefit is small. In contrast, for small values of b (b < 2), the overall benefit of weak

indices is several times smaller than with a more powerful index. The major impact of improving the index, is an increase in the overall budget. This effect is most pronounced for b = ½ (b < 1), where the optimal decision points for i1 and i0.5 are such that nearly no benefit is realised. This pattern can be derived easily from the benefit curves in Figure 4.7 which show that for high b the optima are very similar, while for low b the difference is great.

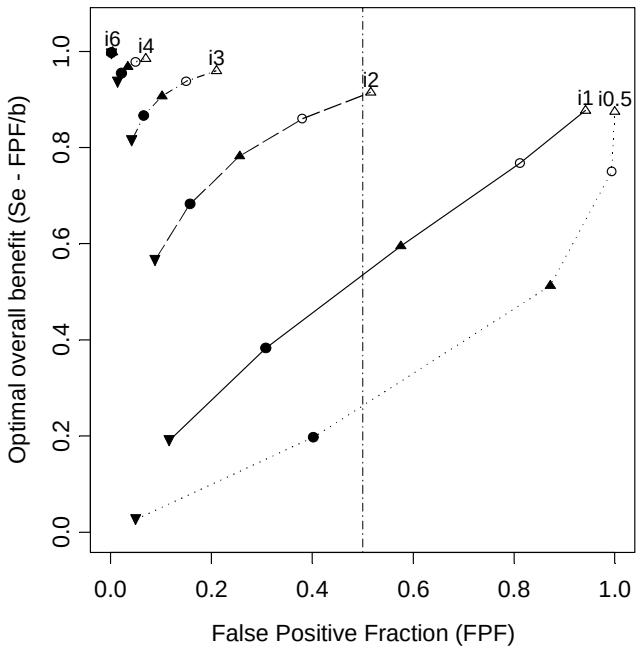


Figure 4.9 **Impact of the index quality on the false positive fraction (FPF) in the optimal decision point.** Trajectories of the optimal overall benefit (B_{Opt} in units relative to T_G) and FPF for indices with an increasing diagnostic accuracy (i0.5 → i6) as a function of the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). The trajectories increase because a higher b implies that it is favourable to allow for more false positives. Because stronger indices can realise a higher benefit at a lower FPF level, their trajectories are lying higher.

4.3.3.2. The average restoration cost (ARC)

With perfect knowledge, the budget required to restore all degraded sites is $\pi^+.C_R$. In reality, because of the imperfectness of the index, we only detect a fraction of the degraded sites (TPF < 1) and we have to accept that some of the pristine sites will be restored unnecessarily (FPF > 0). ARC = $\pi^+.C_R.TPF + (1-\pi^+).C_R.FPF$. The former term is the restoration cost for degraded sites, a (small) reduction in contrast to $\pi^+.C_R$ because TPF < 1. The latter cost component represents the

budget misallocated to pristine sites. For b ≥ 1, this misallocation is larger than the cost reduction and the overall effect is that ARC > $\pi^+.C_R$ (Figure 4.10). For b < 1, ARC is smaller for the two weakest indices. For i0.5 & i1, the optimal strategy is to restore only sites which are extremely degraded. Very few sites are restored, keeping ARC small. Still the quality of allocation is not very good (TRF is low, see Figure 4.11).

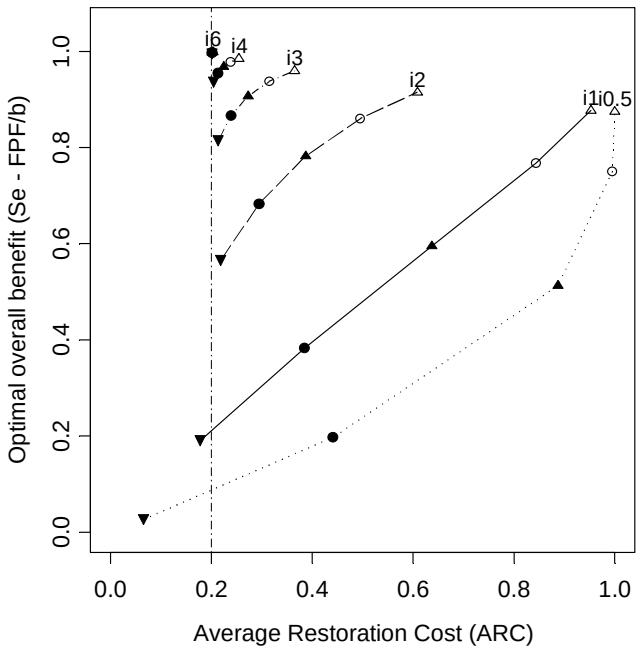


Figure 4.10 **Impact of the index quality on the average restoration cost (ARC) in the optimal decision point.** Trajectories of the optimal overall benefit B_{Opt} (relative to T_G) and ARC (in units relative to C_R) of indices with diagnostic accuracy (i0.5 → i6) as a function of the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). Prevalence of degradation = 20 % (π^+ = 0.2).

The trajectories in Figure 4.10 assess the cost implications of the FPs (Figure 4.9). Both ARC and the optimal overall benefit increase for an increasing benefit ratio. For b ≥ 1, by choosing a stronger index, a higher overall benefit is realized at a lower cost. For a gold standard (i6), the restoration cost remains (about) constant at $\pi^+.C_R$. With decreasing quality of the index, and more pronounced for larger values of b, ARC becomes larger, till for nearly indifferent index i0, ARC approaches C_R. The pattern observed can be directly explained from the evolution of the ROC curves (Figure 3.8). For a strong index, we can realize a high sensitivity at a relatively low FPF, resulting in a good cost balance. For a weak index, the opposite is true. The cost balance is never good and deteriorates further as the benefit ratio increases.

4.3.3.3. The true restoration fraction in the optimal decision point

Figure 4.11 evaluates the quality of the budget allocation with the true restoration fraction, i.e. the proportion of the budget is correctly allocated to degraded sites in case 20 % of the sites is degraded ($\pi^+ = 0.2$). The slope of the trajectories is negative: for an increasing benefit ratio, the overall benefit increases but TRF decreases. Only for the two gold standard indices (i4 and i6), PPV remains above 0.75 for the full range of b-values. Conversely, for the two weakest indices, PPV is always below 0.5, implying more than half of the budget is misallocated to pristine sites.

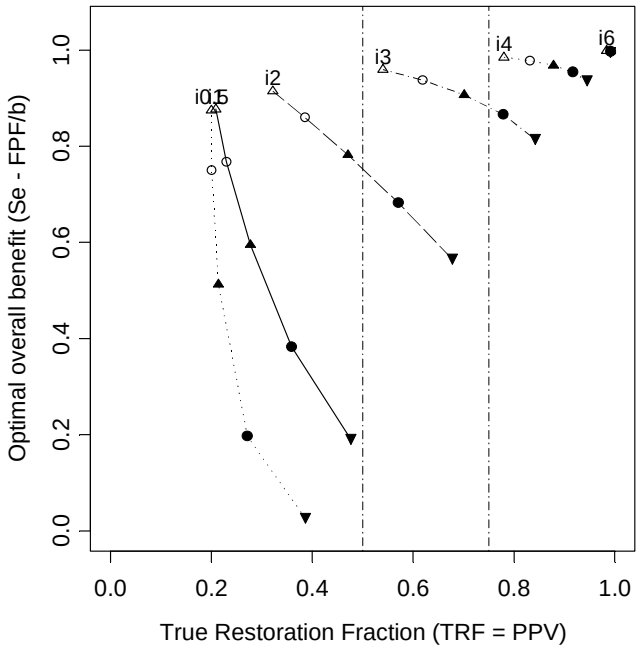


Figure 4.11 **Impact of the index quality on the true restoration fraction (TRF) in the optimal decision point.** Trajectories of the optimal overall benefit B_{Opt} (relative to T_G) and TRF = PPV of indices with an increasing diagnostic accuracy (i0.5 → i6) as a function of the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). The budget allocation of weak indices is consistently worse than for strong indices and in addition B_{Opt} is lower. Over the full range of b, the strong indices succeed in realising a high B_{Opt} keeping TRF high.

4.3.3.4. Increase of the restoration costs because of FP

To have a better picture of the cost consequences of FPs, Figure 4.12 compares ARC with the cost if we restore the same amount of degraded sites without FPs ($\pi^+.C_R.TPF$). The ratio ARC/$\pi^+.C_R.TPF$ = 1/TRF. We now observe how weaker indices consistently increase the budget manifold because of their low discriminatory power especially if b is large.

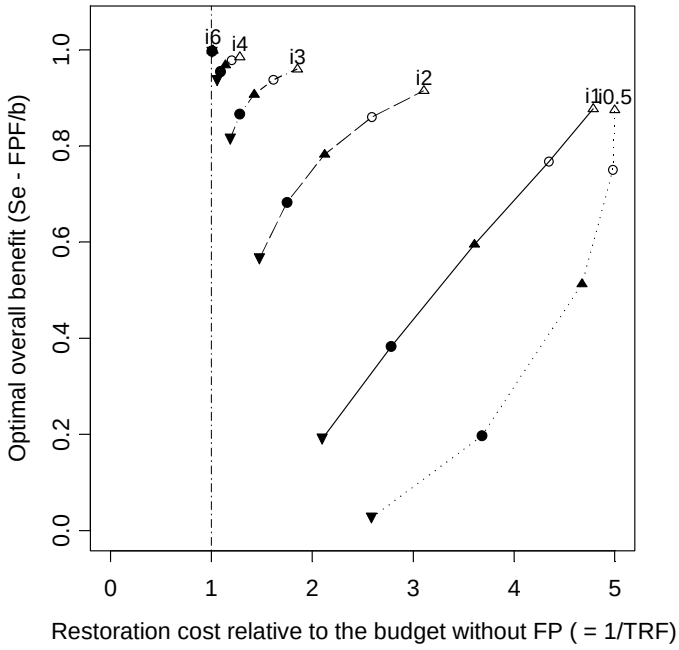


Figure 4.12 **Impact of the index quality on the restoration cost relative to no FPs.** Trajectories of the optimal B_{Opt} (relative to T_G) and the restoration cost expressed relatively to the budget required without FP (=1/TRF) for indices with an increasing diagnostic accuracy (i0.5 → i6) as a function of the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). Results are for a prevalence of degradation = 20 % (π^+ = 0.2). Weak indices consistently require a higher budget, but realise a lower B_{Opt}. The discrepancy of B_{Opt} is maximal for low values of b. Strong indices still succeed to realise the B_{Opt} at a relatively low price. Conversely, for high values of b, also weak indices are capable to attain a high B_{Opt}, but at the expense of a high cost.

4.3.4 Minimising the overall restoration costs (cost-effectiveness analysis)

From a cost perspective, the best index minimises the total management cost, i.e. the sum of the average restoration cost and the assessment cost (C_M = ARC + C_A). A high quality index realises a high sensitivity keeping FPF low. By choosing for a better index, we recuperate the extra assessment cost by avoiding restoration of pristine sites. The best index offers the highest reduction in FP at the lowest price. As discussed before, to analyse the cost tradeoff, we assume a quadratic relationship between costs and quality (Figure 4.4).

4.3.4.1. The choice of the best index

Figure 4.13 shows the evolution of the total management cost as a function of the index quality and the benefit ratio. For b = ½, the index with the lowest C_M is i0.5, for b = 1, i2 (with i1 as a close competitor) is best, for b = 2, i2 and i3 are close, and, finally, for b = 4 and 8, i3 is best. For still higher benefit ratios, the curves suggest i4 will become better, but the gold standard i6 is outcompeted by construction (the cost is prohibitively high).

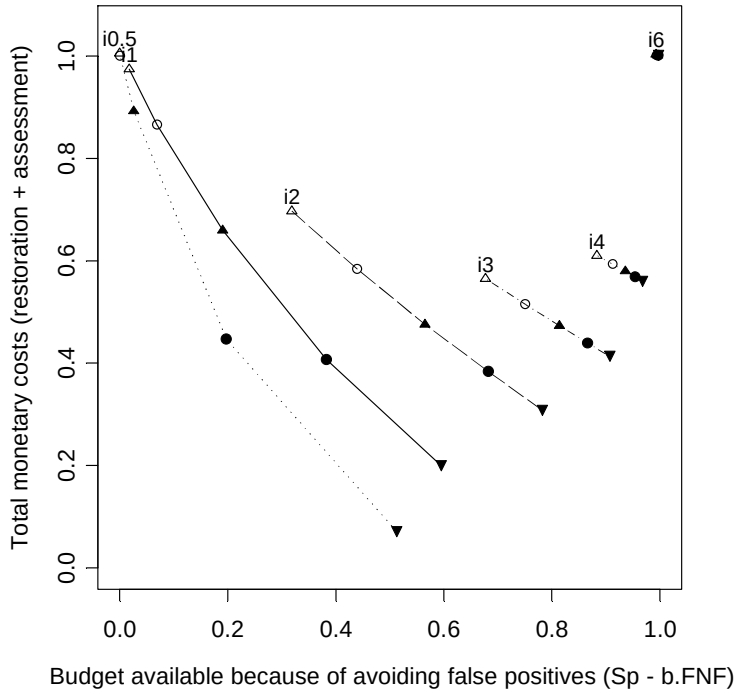


Figure 4.13 **CEA: optimisation of the monetary cost C_M.** As shown by the index trajectories of indices with an increasing strength (i0.5 → i6), strong indices are more apt in avoiding FPs (A_{ROC}) as the benefit ratio increases (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). Hence, when b is high, relatively more budget is available for assessment in comparison to weak indices, shifting the optimum to stronger indices. Prevalence of degradation = 20 % (π^+ = 0.2). C_M is expressed relative to C_R, and A_{ROC} is relative to $T_H = (1-\pi^+).C_R$.

The general pattern is that with increasing benefit, the diagnostic accuracy of the index becomes higher and the monetary cost of the restoration increases. This optimisation is sensible from a cost-benefit perspective. If the expected return is high (as suggested by the benefit ratio), a higher investment results in a higher return. A particularity is that for b = ½ there seems no real minimum (Figure 4.13). The (imaginary) line connecting all points with b = ½ seems to suggest that the minimum cost is zero, implying it is optimal to not restore at all. This is a solution found

by adopting a pure cost perspective. When considering also the benefit in the next section a clear optimum will be found (Figure 4.17).

4.3.4.2. The assessment fraction

As it is often questioned if we do not spend too much on monitoring, Figure 4.14 investigates the assessment fraction in comparison to the total monetary costs (AFM) as a function of the diagnostic accuracy and benefit ratio b. As expected, the trajectories shift to the right for an increasing index quality, but within each index, AFM goes down as a function b. This is because a higher b shifts the optimal decision points to a higher average restoration cost (ARC). As C_A is a constant for the index and does not depend on b, its relative contribution to the total cost decreases.

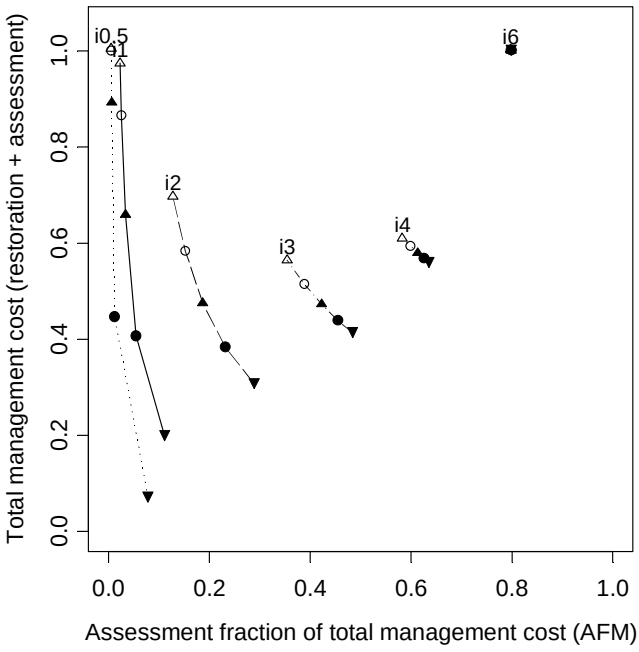


Figure 4.14 **CEA: evolution of the optimal assessment fraction.** Because a higher index quality (i0.5 → i6) is chosen for an increasing benefit (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8), the optimal assessment fraction (AFM), i.e. where the monetary cost C_M (total management cost) is minimal, tends to increase with b when a better index turns out to be optimal, e.g. going from i2 to i3. However, because of the discrete nature of the indices, the same index can remain optimal for more than one b-values (e.g. i3). In this case, the assessment fraction tends to decrease.

However the overall effect of b, is that C_A increases because gradually it becomes more interesting to choose for a better index. For instance, comparing the trajectories of i2 and i3 reveals that for b = ½ & b = 1, the total cost of i2 is lower than for i3. As the average restoration cost for i2 increases faster, from b ≥ 2, i3 realises the lower overall management cost. Changing from i2 and i3 implies a major increase in the assessment cost, but this loss is compensated by keeping ARC

sufficiently low. At b = 8, i4 becomes a close competitor for the same reason. Ultimately, for still higher b-values, i4 will become the best index. The implication of these results is that in some cases, it is not unreasonable to spend about 0.4 C_R (per site) on assessment (for i3).

4.3.4.3. The impact of the gold standard cost

With the cost model, we excluded by construction the gold standard i6. ARC is minimal ($\pi^+.C_R$), but the high assessment cost ($1-\pi^+).C_R$ precludes a practical application. Figure 4.15 shows what happens if the cost of the gold standard is reduced by a factor two: C_A = ½ ($1-\pi^+).C_R$ = 0.4 C_R.

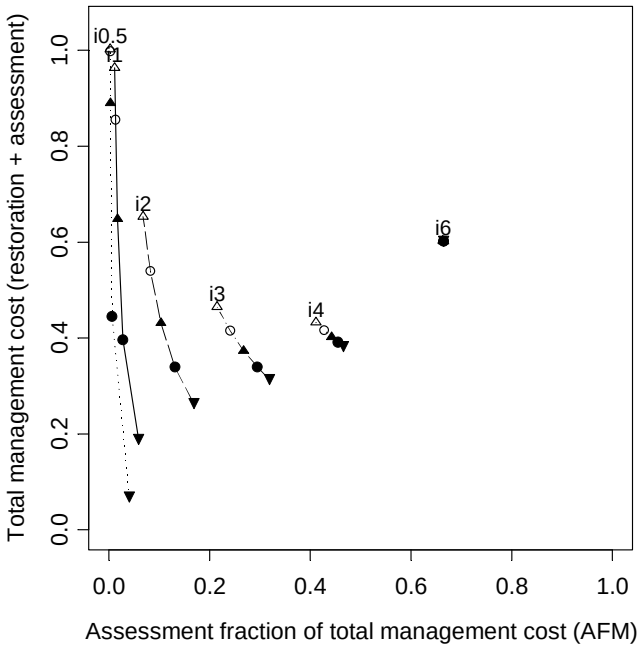


Figure 4.15 **CEA: impact of a cheaper gold standard.** Total management or monetary cost C_M (relative to C_R) and assessment fraction (AFM) as a function of the index quality (i0.5 → i6) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8). Prevalence of degradation = 20 % (π^+ = 0.2). Gold standard cost is 0.4 C_R (twice as small as before) resulting in C_M = 0.6 C_R.

With the same parabolic relation between cost and quality as before, index i4 outcompetes all other indices from b = 4. The index i6 is not excluded in advance as the overall cost is 0.6 C_R and, indeed, i6 is more cost-effective than i2 for b ≥ 4. If only i2 and i6 are available, i6 is the best index. In this case, a better solution would be to develop an index similar to i4, i.e. investigating whether we cannot prune i6 to make it more cost-effective, keeping its optimal qualities. Important to notice that in this new cost configuration, the assessment fraction is again 0.4 C_R (per site) but now for i4.

4.3.5 Maximising the total benefit (cost benefit analysis)

We now adopt a complementary perspective and perform a cost benefit analysis (CBA) integrating the monetary, ecological and societal costs. The optimal index realises the highest total benefit which equals the ecological benefit corrected for the monetary costs (restoration + assessment).

4.3.5.1. The relation between restoration benefit and the risk budget

The trajectories in Figure 4.16 link the overall benefit (B_{Opt}) and the risk budget available for assessment by avoiding false positives (A_{Opt}). For high b values, the benefit realised by weak and strong indices is comparable, but the risk budget differs strongly. Choosing for a better index allows to recuperate the assessment cost by the resources gained by avoiding FPs. Conversely, for low b values, there is high contrast between the indices with respect to the benefit. Choosing a better index increases the benefit many times, but less resources can be recuperated.

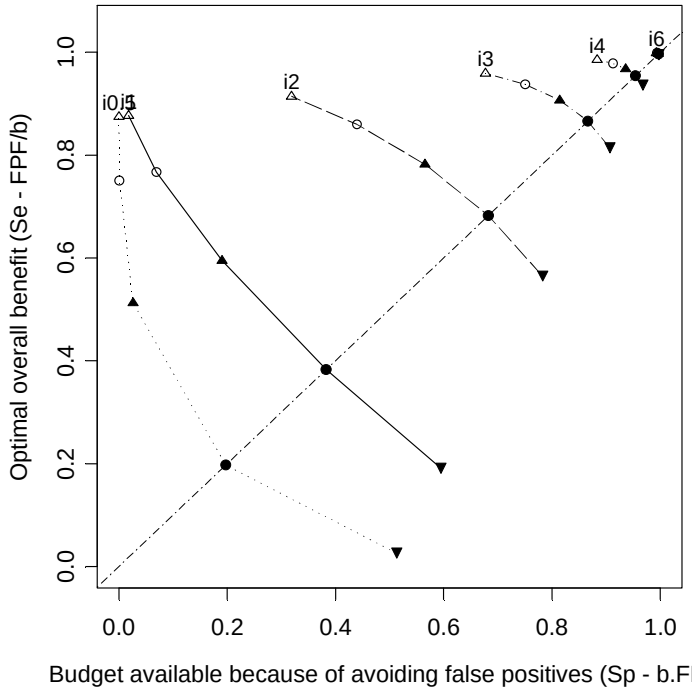


Figure 4.16 **CBA: risk budget and overall restoration benefit.** Relation of the overall benefit B_{Opt} (relative to T_G) and risk budget to avoid false positives A_{Opt} (relative to T_H) as a function of the diagnostic accuracy of the index (i0.5 → i6) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, Δ = 8).

4.3.5.2. The choice of the best index

Figure 4.17 plots the cost-corrected benefit $B_M = B_{ROC} - C_A = T_G\,(TPF - FPF/b - q^2/b)$ as a function of the risk budget available for assessment by avoiding false positives A_{ROC}. Similarly to CEA, with b increasing, the quality of the optimal index increases. As for the CEA, the gold standard index is

ruled out by construction. Index i2 is optimal for b = ½, i3 for b = 1 till 4, and i4 starts to outperform i3 from b = 8. If we are uncertain about the true value of b, i3 is overall the best choice. However, for low b values, in contrast to the CEA, i2 and not i0.5 turns out to be the best index. The reason is that with a CEA only the restoration cost is considered and no benefits.

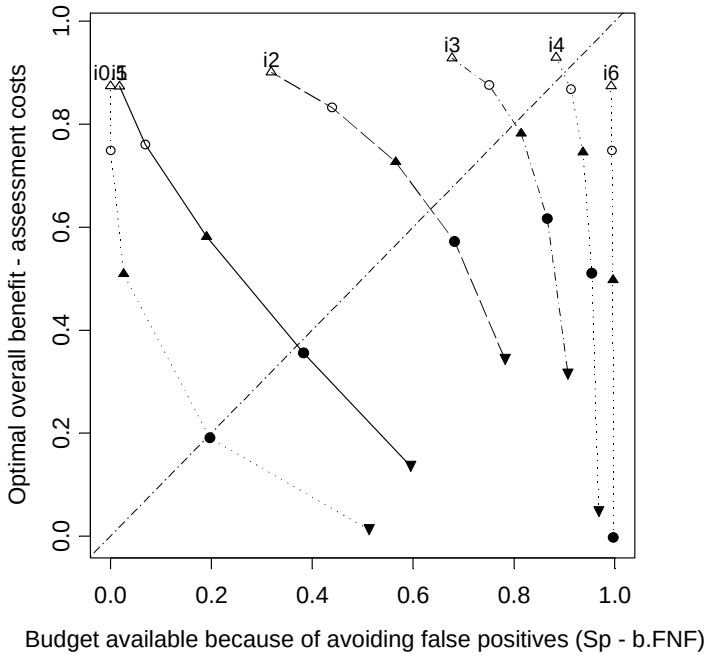


Figure 4.17 **CBA: risk budget and cost-corrected restoration benefit.** Relation of the cost-corrected benefit (= B_{Opt} – C_A; in units relative to T_G) and the risk budget available because avoid false positives A_{ROC} (relative to T_H) as a function of the diagnostic accuracy of the index (i0.5 → i6) and the benefit ratio (b: ▼ = ½, ● = 1, ▲ = 2, ○ = 4, ∆ = 8) (n^+ = 0.2).

4.4 Discussion

4.4.1 The general picture

4.4.1.1. Aims and scope of the study

The main objectives were to get insight in the factors governing the costs and benefits of an index and to develop criteria and a strategy to retrieve the optimal index. To fix ideas, we elaborated on an example in the context of river restoration. As such, the numerical results cannot be used as benchmarks, as they depend on the models chosen. For instance, because of the equal variance

assumption, the ROC curves of all hypothetical indices were concave and symmetric which is generally not true. Also, by construction, we ruled out the gold standard index (i6) by its high assessment cost but favoured a close variant (i4) by assuming a quadratic model decreasing C_A by more than one half. Yet, the main patterns do hold generally and reveal the underlying mechanism of the cost and/or benefit optimisation. For instance, a high benefit ratio shifts the optimal decision point to the more liberal part of the ROC curve and the effect is larger for weak than for strong indices. In the next paragraphs, we will summarise the main findings and extrapolate them on a more abstract level making reference to the example.

4.4.1.2. The parameterisation of the decision context

By parameterisation of the (simplified) binary decision context, we constructed a framework linking the diagnostic accuracy of an index with the expected costs and benefits of decisions. In reality, decisions about restoration are more complex. The signal of an index is not normally decisive, but only one element of consideration. Additional judgement and/or fieldwork are necessary to confirm degradation and to determine the cause. However, we can apply the same principles to this more complex situation and it is straightforward to accommodate the framework by considering the cost implications for each decision step. The essential point of the framework is that the index outcomes are integrated and effectively used in the decision process. Without application, it is not possible to link benefits to the index. In this respect, the design of an index may not focus solely on technical matters but must pay attention to the application context also. It should be investigated which type of information is really relevant and appropriate to support the decisions.

4.4.1.3. Cost effectiveness analysis (CEA) and cost benefit analysis (CBA)

To optimise the benefit of the index, we followed two complementary tracks. The first track was a cost effectiveness analysis (CEA) only taking into consideration the monetary implications of the index-based decisions. From this perspective, the best index minimises the total management cost (C_M), i.e. the sum of the average restoration cost (ARC) and the assessment cost (C_A). By increasing the diagnostic accuracy, ARC goes down because of a better control of the fraction FP decisions (restoring pristine sites). We can use the recuperated resources to compensate the higher assessment cost. As long as $\Delta C_A < \Delta ARC$, it is profitable to choose for an index of higher quality. The alternative track was a more comprehensive cost benefit analysis (CBA) integrating the monetary, ecological and societal benefits in one single measure. The best index maximises the total benefit (B_M), which equals (Box 4.2) the difference of the ecological and societal benefit (E_{ROC}) minus the total management cost (C_M). As could be expected, there exists a close connection between both approaches, but they are not equivalent. Minimising the total management cost contributes to a higher benefit, but in some situations a higher cost can be motivated by a higher non-monetary benefit. This is apparent for b = ½, where CBA choose for i2 (Figure 4.17), while CEA suggested no restoration at all is the best (Figure 4.13).

Although CEA is less comprehensive, a strict monetary approach has practical advantages. Ecological valuation studies are easily contested. It is far from evident to value ecological goods and services and no generally accepted figures exists. Monetary costs are easier to define and to collect, although some care is necessary in defining the costs properly and fully enumerating them

(Caughlan and Oakley, 2001). Quite often, basic information is lacking about the restoration costs. Another advantage of a pure cost optimisation it the immediate gain for the manager who is faced with a fixed budget. If a better index decreases the average restoration cost, it is possible to recuperate the resources saved to compensate the higher assessment cost. In contrast, if the extra assessment cost is motivated to improve the ecological benefit, the benefits are for the stakeholders and/or society as a whole. In addition, ecological benefits are only realised after many years. Unless society is willing to pay (WTP), it will be impossible to realise the extra benefit.

By following a dual track combining a CBA and CEA perspective, we obtain a higher transparency on whether the improvement of the index is motivated by monetary and/or non-monetary profits. In situations where the monetary cost is a dominant cost factor, the results of both approaches will be very similar. Otherwise, if non-monetary benefits dominate, the results diverge. If there are direct monetary profits, we have a strong bottom line to ask for better indices. If the profits are rather indirect, it is important to make the assumptions explicit. Transparency is important for both internal (in the organisation) and external (the policy makers and/or stakeholders) communication and discussion. Internally, the question is how to allocate in an optimal way the available budget over the different priorities. Externally, the point is to motivate the level of the budget.

4.4.1.4. The key results

(a) By parameterisation of the decision context (Box 4.1), it is possible to decompose the expected total cost associated with the decision framework as a sum of three easy to understand terms (Box 4.2): (i) the cost before the decision, (ii) the assessment costs, (iii) the benefit realised by the decision. The latter term is the result of the benefit realised (the sensitivity), the harm avoided and the unnecessary cost prevented (the specificity). Box 4.3 connects the cost benefit approach with the cost effectiveness approach by showing that the total benefit is the difference between the ecological benefit and the total management cost.

(b) To make a fair comparison between the indices, the first step is to tune the index (Box 4.4) taking into account the operational context. The optimal decision point depends on three key parameters: (i) the shape of the ROC curve (the intrinsic diagnostic accuracy), (ii) the intrinsic benefit ratio b_R (ratio of possible benefit and risk on a site level), and (iii) the prevalence (relative frequency) of degraded sites (π^+). The latter two factors define the operational context of the index and can be summarised in one single parameter: the overall benefit ratio (b) at a regional level which is a product of the odds of the prevalence and the intrinsic benefit ratio.

(c) For an increasing benefit ratio, the optimal decision points tend to shift to the upper, liberal part of the ROC curves with a high FPF level (Figure 4.5 & Figure 4.6 for the equal and unequal variance model respectively. This effect is most pronounced for weak indices. In addition, weak indices cannot realise a high benefit, if b is low. In contrast, strong indices are more robust and are capable to realise a high benefit over the total range of b (Figure 4.7).

(d) Figure 4.10 is a key result. For $b \geq 1$ and weak indices, the liberal policy with many FPs increases the average restoration cost (ARC) because resources are misallocated. By choosing for a better index, we can diminish the average restoration cost and recuperate resources by avoiding

restoration of pristine sites. For b = ½ (b < 1), a stronger index mainly improves the benefit. In this situation, the CBA (Figure 4.17) and CEA (Figure 4.13) diverge.

(e) Because stronger indices are more robust, in spite of their higher cost, powerful indices become more and more attractive with an increasing benefit ratio (Figure 4.13) as they recuperate the resources otherwise lost by restoring pristine sites (Figure 4.14).

4.4.1.5. The general mechanism

The general effect is that with an increasing benefit ratio, it becomes more and more attractive to choose for a more accurate and expensive index. This can be explained by a three step reasoning. (1) We observe that for an increasing b, the optimal decision point shifts to the right of the ROC curve, i.e. the more liberal part where the FPF is quite large (Figure 3.8). More and more pristine sites are unnecessarily restored (Figure 4.9), and, as a consequence, the average restoration cost increases (Figure 4.10). (2) This effect is much more pronounced for weak indices than for strong indices which are more robust. The optimal decision points of weak indices vary over a broad range and shift completely to the right for high b values (Figure 3.8). (3) Because of this differential effect, it becomes more and more attractive to invest in a more expensive index as we recuperate the resources gained by lowering the restoration cost for the more expensive assessment. As long as the investment in a better index pays back, there is room to improve the index further.

4.4.2 The cost decomposition as a guidance for the design of a restoration plan

We decomposed the total cost C_T in three meaningful components: (i) the baseline ecological and/or societal cost status before restoration (C_o), (ii) the assessment cost for the index (C_A) to discriminate between pristine and degraded sites and (iii) the restoration benefit as guided by the index (B_{ROC}) which depends on the diagnostic accuracy of the index and changes along the ROC curve: $B_{ROC} = T_G.TPF - T_H.FPF = T_G.(TPF - FPF/b)$. Figure 4.2 makes a graphical representation of this equation, serving as a guidance for three decisions when setting up a restoration plan or any other action plan of which assessment is an essential and integrated part.

4.4.2.1. The cost analysis before restoration

Before starting a restoration project, it is necessary to estimate the potential gain T_G we can realise and harm T_H to avoid. As derived in Box 4.2, both parameters depend on the prevalence π^+ of degradation, the effectiveness of restoration η_R (function of the restoration cost C_R, the ecological residual E_{TP} and the ecological costs of degradation E_{FN}) and the risk associated with treating pristine sites (E_{FP}). The relative magnitude of T_G and T_H or the benefit ratio $b = T_G/T_H$ is a very important parameter determining the maximal potential of the restoration. If $b < 1$, we can only realize a positive benefit in very narrow range and the optimal benefit is only high with a powerful index. If $b \geq 1$, the situation is more favourable, but the optimal management cost for weak indices can become high, requiring a careful motivation.

A first critical factor is π^+, the prevalence of degradation. The prevalence has an important impact on both the potential gain $T_G = \pi^+.R_G$ and the benefit ratio (at a regional level) which can be factorised as a product of the intrinsic benefit ratio b_R and the odds of the prevalence ($odds(\pi^+) =$

$\pi^+/(1-\pi^+)$). If few sites are degraded (π^+ small), the degradation will not be perceived as important at a regional level, even if the (individual) degraded sites are in a very poor status (E_{FN} large). For a proper understanding of the impact of π^+, we should realize no knowledge is available about where the (few) degraded sites are located. First assessment is necessary. With perfect knowledge and no assessment costs (a contradiction in terms), the total management cost C_M is simply $\pi^+.C_R$. To realise a benefit, it is sufficient that $C_R < E_G = (E_{FN} - E_{TP})$. With assessment, the cost picture can change completely. For a (close to) gold standard index with TPF ≈ 1 and FPF ≈ 0, $C_M = \pi^+.C_R + C_A = \pi^+.(C_R + C_A/\pi^+)$. As C_A/π^+ is inversely proportional with π^+, the burden of the assessment increases fast with a decreasing prevalence. For instance, if $\pi^+ = 0.2$, $C_M = \pi^+.(C_R + 5C_A)$. With a low prevalence, we have to assess many pristine sites increasing the relative cost of assessment. In addition, even with a low FPF, there will be many pristine sites misclassified and unnecessarily restored increasing the average restoration cost. We should contemplate to narrow the scope of the restoration program to the sites at risk to increase π^+. For instance, if the risk of degradation is strongly correlated with some easy to measure characteristic of the sites, confining the restoration program to the sites at risk should be considered.

A second parameter of interest is the (cost-corrected) effectiveness of restoration $\eta_C = (E_{FN} - E_{TP} - C_R) / E_{FN}$ (Box 4.2) from which we can derive the maximal restoration gain: $T_G = \eta_C.C_0$. In practice, we only realize the fraction equal to the sensitivity at the decision point of the index: $T_G.TPF = \eta_C.C_0.TPF$. If $\eta_C << 1$, T_G will be too small to start a restoration project. The parameter η_C expresses the fraction of the unfavourable status E_{FN} we can restore. Assessment of this cost involves a monetary evaluation of the loss of ecosystem functions and services because of the degradation of the ecosystem, which is not evident. We also should have information about the residual cost after restoration E_{TP} (the ecological efficacy) and cost of restoration C_R. Although the latter parameter is a monetary term and is easy to determine in practice, very often figures are lacking. It is still more difficult to have good estimates of the ecological efficacy. Ideally, to have an optimal η_C, the restoration cost as well as the ecological residual should be as small as possible.

The third factor setting the scene is the cost associated with the risk of false restoration of pristine sites: $T_H = (1-\pi^+).(E_H + C_R) = (1-\pi^+).(h+1).C_R$, with $E_H = h.C_R$ ($h \geq 0$) expressing the ecological harm relative to the restoration cost. By using the index, the expected harm is reduced to $T_H.FPF$. The lower limit of T_H ($h = 0$) refers to a situation (real of assumed) without ecological side-effect. Interestingly, $T_H = (1-\pi^+).C_R$ equals the resources detracted by restoring all pristine sites. The function of the index is to keep this term as small as possible, by keeping the FPF small. However, from a strict cost perspective, as soon as $C_A \geq (1-\pi^+).C_R$, it can be tempting to restore all sites without monitoring. In general, the assessment cost should be lower than the expected harm of blind restoration: $C_A < T_H$. This limit can be quite high; for instance, for $\pi^+ = 0.2$ and $h = 2$ (ecological damage twice as large as the restoration costs), $C_A/ C_R < (1 - 0.2) (1 + 2) = 2.4$, or it can be defensible to spend twice as much on monitoring than on restoration!

4.4.2.2. The selection of the optimal index

The assessment costs do not only include costs associated with the data collection: field visits, data storage, analysis and interpretation, reporting and decision making. We should also take into

account the design and maintenance of the index and the monitoring program: index development (calibration, validation, follow-up studies), sampling design, quality control and assurance. There are always assessment costs, but sometimes they are hidden and not well known. For instance, assessment based on experience-based judgement also requires data collection and/or time budget for field visits. We suspect the real cost is often larger than admitted. For a fair comparison of personal judgement with an index-based approach, it is necessary to quantify the cost and determine the diagnostic accuracy of the different possible strategies ranging from a minimal assessment to a (close) to gold standard approach.

It can be instructive to compare the cost of the gold standard and the cost of the personal judgement to make a tradeoff between costs and quality for a range of assessment scenarios going from very basic to very involved. The role of the index is to improve the cost-effectiveness of the monitoring by providing a cheaper alternative than gold standard measurements. It will enhance the support for the index, if we can prove that the index is more cost-effective in comparison to the gold standard as well the personal judgement. It is possible that an approach based on personal experience is indeed the best, but in this case, a standardisation should be worked out.

In practice, it is impossible to consider all possible indices for optimisation. We rather compare a few alternatives coming in packages. For instance, to assess the condition of the ecosystem, we can rely on abiotic variables as oxygen and water temperature, habitat characteristics, or species information of one or more taxa (fish, benthos, diatoms, …), or on a combination of them. The outcome of the inventory is a series of indices, ranked by quality and price, to choose from.

4.4.2.3. Tuning the index

Once it is clear which index (package) to choose, we should tune the index by determining the optimal decision point. Implicitly in the previous step, we already used information about the optimum, however the evaluation was rather in global terms to find out the level of diagnostic accuracy required. Circumstances can change, and therefore we should select a robust index under a range of plausible alternatives. In the third stage, the focus is on the optimal decision point itself and it can be necessary to calculate some parameters more precisely or to focus on the most plausible alternative.

The optimal decision point for an index is a function of (i) the intrinsic benefit ratio of the gain and harm associated with restoration (b_R), (ii) the prevalence of the degraded sites (π^+) and (iii) the ROC curve. The ROC curve represents the intrinsic quality of the index, the other two factors reflect the operational context of the index and can be synthesized in the benefit ratio $b = b_R.odds(\pi^+)$ or alternatively $b = E_G/E_H$. The expected gain because of restoration is the benefit of restoring a degraded site multiplied by its prevalence (π^+), the expected harm is the harm of an unnecessary restoration of a pristine site multiplied by its prevalence ($1-\pi^+$). If both terms are equal, $b = 1$ and the optimal decision point (for a binormal model) is where the sensitivity and specificity are equal. Intuitively, this is a logical choice. If the expected gain is larger ($b > 1$), the optimal decision point shifts to the more liberal part of the ROC curve. In contrast, $b < 1$ results in a more conservative policy.

4.4.3 Recommendations for a practical application

4.4.3.1. Knowledge of the ROC curve

To compare the benefit of the indices, we should know their ROC curves. It is not common practice to document the sensitivity and specificity fully. At best, the diagnostic accuracy is known at one single decision threshold. Rather seldom, budget or time is available to construct ROC curves. A way out is to attribute the indices under consideration to a quality class similar to the hypothetical indices in this paper (i1, i2, i3), for instance by searching the literature for quality reports of similar indices. For large scale projects, where there is room to develop new indices, determination of the ROC curve should be an integral part of the design.

It is important to distinguish between internal and external validation (Hosmer and Lemeshow, 2000). Internal validation uses data from the same research context as the index calibration and cannot correct for shortcomings in the study design (e.g. representativeness of the data, gold standard measurements). Hence, internal validation tends to overestimate the true diagnostic accuracy. The design of an external validation study will involve collection of new data. These follow-up studies are rare as their perceived benefit is low for decision-makers. A possible strategy to get funding is to embed "post-marketing" evaluation of the index in a general policy of quality assurance and quality control. Although QA/QC programs are not evident neither, a possible advantage is that funding is structural once there is an agreement about its implementation.

4.4.3.2. The cost information

The monetary valuation of ecological and societal implications of the decisions is far from evident. In general, information required for cost matrices (Box 4.1) is lacking. Even for the most easy part, the monetary restoration costs, consistent documentation is often inexistent. We also lack data about the efficacy and side-effects of many restoration measures. As a consequence, cost benefit studies are rare. We are faced with a typical chicken and egg situation. Because experience with economical studies is lacking, they are avoided. To start a learning process, a possible strategy is to start with a few modest but well-focused economic studies gradually building experience. Uncertainties about parameter values, may not prevent from making cost calculations and drawing conclusions. Instead we should carefully investigate the consequences of the uncertainty with sensitivity analyses. Depending on the circumstances, we can apply the framework at different levels of detail. For a global orientation, the general principles give insight in what is crucial, for instance, they can be helpful to rule out some implausible scenarios. On a semi-quantitative level, as illustrated with the hypothetical indices, we can express the costs relative to the restoration cost to estimate the benefit ratio. Finally, for large projects, a real economical cost benefit analysis is to be recommended, as this will increase the support of decision makers and society for an index-based restoration project. For this type of study, involvement of economists is mandatory.

4.4.3.3. The role of the cost benefit ratio

All analyses confirm the benefit ratio b as the crucial parameter describing the decision context. It is the ratio on a regional level of the potential benefit (to realise) and harm (to avoid). If the risk is high in comparison to the potential gain (b small), it is difficult to realise a high potential. The

tuning of the index should be in a narrow range and if the index quality is low, the maximal benefit remains small. For a high benefit, the optimum is broader, and also with weak indices we can realise a high benefit, however at the expense of a high positive fraction increasing the average cost. The benefit ratio is the crucial parameter to select to optimal index. If the benefit ratio is low, indices of lower quality are preferred. As b increases, indices with a higher accuracy is optimal.

The general pattern is that with an increasing benefit ratio, the overall costs of the restoration increase. A more expensive index with higher diagnostic accuracy is chosen and the optimal decision point is shifted to the more liberal part of the ROC curve resulting a higher average restoration cost (ARC). From a cost benefit perspective, these results sound sensible. With a high potential benefit, it is worthwhile to invest more. However, the benefit ratio is an estimate. If the real benefit is lower in comparison to what is assumed, we risk to invest too much in the index.

4.4.3.4. The koala example

As an interesting final note, Field *et al.* (2004) concluded against monitoring from a pure cost benefit point of view. Although they favour monitoring in general, they arrived at the conclusion that with a high benefit ratio, monitoring is not warranted. Analysis of their calculations shows that they do not take into account the risk of an unnecessary restoration. Although for their specific case, they are possibly correct (do not delay the necessary measures with monitoring), we do not agree with their generalisation. We demonstrate that, even from a "narrow" economical perspective, it is warranted to invest in monitoring if b is high. The fundamental reason being that – on the long run – we better control the costs associated with the risks.

Sometimes, in the context of environmental monitoring, it is argued, referring the precautionary principle (PP) (Lemons *et al.*, 1997; Morris, 2000; Dickson and Cooney, 2005), that in case of a high danger for environmental or human health, we should lower the level of significance to increase the power of the test (Buhl-Mortensen, 1996; Westra, 1997). Our results suggest, that in this case, first of all, a more powerful and hence more expensive test is required to realise a high sensitivity keeping the false positive fraction small. This is in agreement with many discussants reflection on the practical implication of PP (Gray, 1990; Peterman and M'Gonigle, 1992; Gray, 1996; Hansson, 1997; Underwood and Chapman, 2003; Keiding and Budtz-Jørgensen, 2005).

4.5 Conclusions

Many of the derivations are not new and can be found elsewhere as the techniques presented are already routinely applied in other areas where the detection of an event is central, for instance in diagnostic medicine (Metz, 1978; Zhou *et al.*, 2002; Pepe, 2003; Mayer, 2004). The novelty of our approach is the transcription to nature conservation giving each mathematical symbol a clear ecological interpretation.

It is not evident to get the development of IBIs (and other ecological indicators) well funded. Yet, their potential to inform decision makers about the ecosystem in a cost effective way is high. Gold standard measures are perfect but their assessment costs preclude a practical application. The ideal index is a good approximation of the gold standard but much cheaper and minimises the total

management cost (from a cost effectiveness perspective) and/or maximises the total benefit (from a cost benefit angle). As long as the negative consequences of the approximations are less than the cost reduction of assessment, a further simplification is possible. This chapter illustrated how savings can be realised by elaborating a simple example into full detail.

To investigate the factors governing an optimal allocation of the management budget, we fully parameterised a simple decision context in which an index is used to decide about river restoration. For a series of hypothetical indices ranging from very low quality (nearly indifferent indices, having a low discriminatory power) to very high quality up to (nearly) gold standards, we hypothesized a quadratic cost relationship modelling a sharp increase of the assessment costs as a function of the diagnostic accuracy of the index. By construction, the model realistically ruled out the gold standard. Although the numerical outcomes cannot be used as a benchmarks, the key factors are in our opinion quite general.

The optimal decision point from a cost benefit perspective depends on three key parameters: (i) the strength of the index (as characterised by the ROC curve); (ii) the intrinsic benefit ratio (b_R), and (iii) the prevalence (relative frequency) of degraded sites (π^+). The latter two factors can be summarised in one single parameter, the overall benefit ratio (b) which is a product of the odds of the prevalence and the intrinsic benefit ratio. The parameter b is a tradeoff of the expected gain and harm at a regional level. It is crucial to distinguish between the local and the regional level. It is possible that individual sites are severely harmed and that intrinsically the restoration is highly beneficial. However, if the are few impacted sites, the benefit at a regional level may be too low to set up a restoration program. The fundamental reason is we do not know which sites are degraded and assessment is necessary. If the prevalence is low, too many pristine sites will be misclassified as degraded because of the approximation, unless a gold standard is used which is too expensive. A solution for this situation is to focus the restoration on the strata at risk which presumes we have additional information about how to make a preselection.

For an increasing benefit ratio b, the optimal decision point tends to shift to the upper, more liberal part of the ROC curve with a high false positive fraction (FPF). As a higher FPF implies that more pristine sites are treated unnecessarily, the average restoration cost (ARC) increases and the true restoration fraction (TRF) decreases because of the misallocation of management resources. The impact of the benefit ratio is most pronounced for weak indices for which FPF values close to one can be optimal. Strong indices are more robust, as they can realise a high sensitivity at a low FPF. This different behaviour of weak and strong indices, explains why with an increasing benefit ratio, the best indices are more and more competitive and finally outperform weaker indices. By improving the index, we can recuperate part of the resources lost and use them to pay the extra assessment costs because of the improvement.

With an increasing benefit ratio b, the total management cost corresponding with the optimal decision tends to increase. The reason is not that the assessment costs increase. It is because, with a high benefit ratio b, a more liberal policy is optimal, resulting in a higher average restoration cost. As strong indices can simultaneously realise a high sensitivity and keep the FPF small, it is more cost-effective to invest in better indices, recuperating part of the resources lost by keeping

ARC small. This is a logical result. A high benefit ratio implies a high discrepancy between the potential gain and harm. In this situation, investment in a high discriminatory power pays off because of a better distinction between harm and gain. Also, it seems sensible to increase management efforts if the expected benefit is high. However, we should realise that the benefits do no result in a direct monetary return. The benefits are for the stakeholders and/or the society as a whole, in the long run. The manager is faced with a fixed budget unless society is willing to pay. To increase societal support, transparent and standardised procedures are necessary to assess the efficacy of restoration and to value ecosystem functions and services. In this respect, involvement of specialists with an economical background in natural resource valuation is mandatory.

A key result of the cost analysis is that the optimal proportion of the total management budget allotted to assessment becomes higher with an increasing benefit ratio. As the difference between weak and strong indices grows with an increasing benefit, at a certain point, the increase in the assessment cost is smaller than the resources saved by reducing the false positive fraction, and it is more cost-effective to choose for the index with the higher diagnostic accuracy. Although our results cannot be used as real benchmarks, they make plausible that for a high benefit ratio, it is sensible to spend a high fraction of the management budget on assessment because of the better discrimination between degraded and pristine sites. This assessment cost may not be understood as an additional cost on top of the restoration cost. In contrast, by allotting a higher part to the budget to assessment, we can reduce the average restoration cost such that the total management cost is lower than with a less powerful and less expensive index.

5 How to determine the optimal number of metrics in an index of biotic integrity? A coherent strategy based on ROC curves, statistical model building and bootstrapping.

This chapter is a revision of the stepwise regression approach used to construct the Estuarine Biotic Index (EBI) for the Zeeschelde Estuary (Breine et al., 2007).

Abstract

When developing an index of biotic integrity (IBI), the selection of an optimal metric set from a candidate list remains one of the main challenges. In many instances, a coherent and transparent strategy is lacking leaving much room for subjective decisions and/or personal preferences. Important unresolved questions include how many metrics to select (the model dimension) and how to choose properly between metrics. In our opinion, two important factors contribute to this situation.

Firstly, optimisation criteria are seldom made explicit. Rather seldom the important distinction is made between false positive (FP) and false negative (FN) errors. A high false negative fraction (FNF) implies that many degraded sites are restored, not realising the full potential of ecosystem functions, goods and services; a high false positive fraction (FPF) results in unnecessary restoration of many unimpaired sites, detracting resources and possibly harmful for pristine sites. In analogy to diagnostic models in medicine, we propose the Receiver Operating Characteristic (ROC) curve to optimise the diagnostic accuracy of the index model. ROC curves plot the true positive fraction (TPF = 1 - FNF) of an index as a function of the FPF and visualize its potential to limit simultaneously FPs and FNs. To keep both errors small, the index can be optimised with respect to the (full) area under the ROC curve (aucF). As only a part of the ROC curve is relevant for decision making, we suggest to use a more focused complementary variant, the partial area under the curve (aucP).

Secondly, it is insufficiently recognised that an IBI is in essence a regression model. Traditionally, an IBI is simply an average (or sum) of scored metrics. In fact, this average score model (AVG) is an ordinal logistic regression model (OLR) in disguise with prespecified regression coefficients. Because of this connection, we can borrow concepts, strategies and techniques from statistical model building to search for the optimal suite of metrics. In this context, an important issue is overfitting, i.e. selecting too many variables in comparison to the data available, resulting in a lower diagnostic accuracy than a simpler, more parsimonious model. Another point to consider is that the optimisation criterion is a random variable. Hence the optimal model is not necessarily the

best one. To cope with these problems, we propose a modelling strategy to ascertain the optimal number of metrics first and to explore competing models in the vicinity of the optimum.

We illustrate our approach by revising the Estuarine Biotic Index (EBI) we developed for the mesohaline part of the Zeeschelde Estuary in Flanders, Belgium. Specifically, we demonstrate how statistical modelling techniques, such as bootstrapping and best subset regression, combined with optimisation criteria derived from the ROC curve, can be forged together to a powerful and transparent strategy to select the optimal basket of metrics from a candidate list. We also compare the traditional AVG with OLR revealing that the proportional odds model results in a very similar model. This extension to generalised linear models (GLM) opens a perspective to formulate more flexible index models better adapted to the sampling design and able to incorporate background variables adjusting for differences between sites. As a side result, we demonstrate that continuous scoring of the metrics is more powerful than the still often used discrete scoring. We conclude with suggestions to improve data collection for IBI construction. In our opinion, a representative sampling scheme covering the full spectrum of human impacts in a region, is a prerequisite to retrieve a responsive set of metrics.

Keywords

Index of biotic integrity. Diagnostic accuracy. ROC curves. Partial area under the curve. Statistical model building. Metric selection. Overfitting. Bootstrapping. Proportional odds model.

5.1 Introduction

The value of indices of biotic integrity (IBIs) to monitor the ecological status of rivers is well accepted (Hughes and Oberdorff, 1999; Schmutz *et al.*, 2007b). In fact, Karr's original proposal (Karr, 1981; 1986) on how to compose and construct a multi-metric index of biotic integrity (MMI) started a rich research tradition – as advocated by the science philosopher Laudan (1977) to make scientific progress (Hubbard *et al.*, 1998; Kagel and Roth, 1995; Thompson, 1994) – of scientists collaborating at the interface of science and policy. These boundary workers (Turnhout, 2009) gradually refined and extended the original concept to a broad range of situations (Hughes and Oberdorff, 1999). Over the years, a series of guideline papers appeared to standardise and consolidate further the construction of IBIs (Karr and Chu, 1997; Hughes *et al.*, 1998; Hering *et al.*, 2006a; Roset *et al.*, 2007; Southerland *et al.*, 2007; Stoddard *et al.*, 2008). Yet, despite the considerable progress over the previous years (Hering *et al.*, 2010), a lot remains to be done (Borja *et al.*, 2009c), conceptually as well as empirically, before IBIs reliably quantify ("measure") ecosystem health over a broad range of ecosystems and regions (Roset *et al.*, 2007).

From a statistical point of view, once candidate metrics are proposed on subject matter grounds (Bailey *et al.*, 2004), one of the major challenges of index building is the selection of an optimal set of metrics. In its simplest form, an IBI is an average (or, equivalently, a sum) of scored metrics, ecological attributes scoring the functioning of the ecosystem as a "distance to target". However, as it is not known theoretically, at least not without doubt, which ecological attributes are essential and/or indicative for a good ecological condition, also this average score model (AVG) has to be calibrated by screening potential candidate metrics for their response to an independent precursor assessment of ecosystem degradation (Degerman *et al.*, 2007; Aubry and Elliott, 2006), we refer to as the preclassification of the sites.

According to a methodological overview of Roset *et al.* (2007), only very few studies use objective criteria and rigorous statistical procedures for the metric selection. Important unresolved questions include how many metrics to select optimally (Seegert, 2000) and how to choose properly between metrics (Johnson *et al.*, 2006; 2007). Most index builders are well aware that they should avoid redundant and/or highly correlated metrics, however, the statistical method is seldom appropriate. For instance, it is common practice, as advocated in one of the cookbooks to construct IBIs (Hering *et al.*, 2006a), to choose between redundant metrics based on pairwise correlations (Angermeier and Schlosser, 1987). However, bivariate correlations can miss complex multivariate relationships. For instance, three mutually weakly correlated metrics, can be highly collinear (Kutner *et al.*, 2005; Seber, 1984). More importantly, investigation of correlations is not the most appropriate technique to retrieve metrics responsive to human alterations of the environment, because it does not take into account the relation with the endpoint, i.e. the ecosystem condition (Jongman *et al.*, 1995).

Moreover, the attitude towards metric redundancy is somewhat ambivalent. Some authors (Alden *et al.*, 2002; Roth *et al.*, 1998) argue in favour of (some) redundancy to improve the robustness of the index. To our knowledge, robustness is nowhere defined explicitly, but the connotation is that an IBI with more metrics is better buffered against exceptional events. The argument is that, if one metric in the index contains an outlier, this is less a problem if there are many metrics in the index.

Also it is assumed that more metrics in the index contributes to detect pressures not thought about at the development of the index. For all these reasons, some authors prefer to include an excess number of metrics even if the classification efficiency did not improve (Alden *et al.*, 2002) or even decreased (Roth *et al.*, 1998). Another factor sometimes inflating the number of metrics is the ambition that an index should help to uncover the cause why the ecosystem is degraded (Chessman and McEvoy, 1997). With more metrics in the basket, we have additional information for a causal analysis. To cover the main functions of an ecosystem, it is sometimes argued the index should contain metrics of all classes as prescribed by Karr & Chu (1997) or by another rationale (Schmutz *et al.*, 2000).

However, more is not always better. It is well known in statistical literature about model building that a more complex (and more correct) model can have worse statistical properties than a simpler (less accurate) variant if too many parameters have to be estimated (Linhart and Zucchini, 1986). If the model is too complex in relation to the data available, it models noise and particularities in the data set instead of the underlying process. The fit is too close to the sample. This overfitting undermines the generality of the model, but also gives an overoptimistic picture of the diagnostic accuracy (Zucchini, 2000).

In this paper, we concentrate on these issues of metric selection and the risk of overfitting by developing a coherent strategy to determine the optimal number of metrics composing an IBI. An important observation is that the traditional format of an IBI, the average score model (AVG), is very close to a logistic regression model, but with fixed and equal regression coefficients. As a consequence, concepts, strategies and techniques from the field of (statistical) model building can be used to build an IBI. Specifically, the selection of the optimal composition of metrics is equivalent to subset selection in regression (Miller, 2002). For this chapter, we assume the candidate metrics (i.e. the explanatory variables of the model) are given and well thought of, although a full optimisation should critically evaluate the starting set (Noble *et al.*, 2007). Also we suppose that the preclassification ranking the calibration sites with respect to anthropogenic impact (i.e. the response variable of the model) is sufficiently accurate and the sample covers the full spectrum of pressures, a point we return to in the discussion. In this respect, it is important to recognise that no single index can be a 'silver bullet' assessing surface water health unequivocally (Boulton, 1999), but we should define sharply at which combination of pressures the index is aimed (Van Stickle and Paulsen, 2008).

To define the optimisation criterion, our approach links IBIs to diagnostic or prognostic models in a medical context, assessing the health status of a patient from one or more diagnostic or prognostic test variables (Pepe, 2003; Zhou *et al.*, 2002). Similar to IBIs, these models are calibrated with respect to a "preclassification" of the patients, preferably a "gold standard" assessment of the health condition, based on intensive clinical investigations, expert judgement of clinicians and/or additional information only available at a later disease stage. With respect to this gold standard, the model is optimised by searching for an optimal suite of predictors from a candidate set of plausible diagnostic measures ("metrics") minimising the misclassification error of the health status of the patients. For this optimisation, Receiver Operator Characteristic curves (ROC curves), a concept (and tool) borrowed from statistical decision theory (Swets, 1988; Murtaugh, 1996), are

routinely used. Also, in medicine, gold standards do not always exist and should be standardised (Hertzman *et al.*, 2001).

We illustrate our approach by revising the Estuarine Biotic Index (EBI) for the mesohaline part of the Zeeschelde Estuary in Flanders, Belgium (Breine *et al.*, 2007). Specifically, we demonstrate how statistical modelling techniques, such as bootstrapping and best subset regression, combined with optimisation criteria derived from the ROC curve, can be forged together to a powerful and transparent tool to select the optimal basket from a candidate lists of metrics. We also compare the traditional average score model (AVG) with the proportional odds model which is a specific case of ordinal logistic regression (OLR). This extension to generalised linear models (GLM) opens a window to formulate more flexible index models better adapted to the sampling design and able to incorporate background variables to adjust for differences between the sites. As a side result, we investigate the impact of the scoring. Quite often, it is chosen for discrete scoring, while continuous scoring is more powerful if possible (Blocksom, 2003). Finally, in the discussion, we make some suggestions to improve data collection for the design of IBIs. A representative sampling scheme aimed at covering the full spectrum of anthropogenic pressures, is an important prerequisite to retrieve the optimal set of responsive metrics.

5.2 Material & Methods

5.2.1 The case study: the estuarine biotic index (EBI)

5.2.1.1. The study area: the mesohaline part of the Zeeschelde estuary

We borrow our data from a study in the Zeeschelde estuary in Belgium, for which we developed a fish-based IBI (Breine *et al.*, 2007): the estuarine biotic index (EBI). The Zeeschelde is a single-channel, macrotidal estuary with intertidal area (Baeyens *et al.*, 1998). Commonly, three salinity zones are distinguished: a mesohaline zone between Zandvliet (Dutch/Belgian border) and Antwerpen; an oligohaline zone from Antwerpen to Temse, including the Rupel tributary; and a freshwater zone further upstream until Gent, including the Durme tributary.

The study area was restricted to the mesohaline part of the Zeeschelde estuary. For many studies, this zone can be considered as an ecological unity with distinct ecological features (Baeyens *et al.*, 1998; Meire *et al.*, 2005). However, as we acknowledge in Breine *et al.* (2007), a key problem for the development of an IBI is that the gradient of human impact in the mesohaline zone coincides with the salinity gradient. Near the Dutch/Belgian border, the salinity is maximal and harbour activities are minimal. Near Antwerp, the opposite is true. In fact, any other variable correlated with this gradient and/or the distance to the sea, can be the real cause for the change of the ecological community. However, as such, this confounding does not have an impact on the methodology presented here, but it hampers an unequivocal interpretation of the EBI in terms of a response to human impact.

With the data available, it was impossible to avoid the confounding. It is important to realise that the development of an IBI for the Zeeschelde estuary is much harder than for freshwater rivers (Belpaire *et al.*, 2000; Breine *et al.*, 2004). For the latter, it was possible to rely on a survey of

many different rivers to diminish confounding and to realise replication (Lindsay and Ehrenberg, 1993; Thompson, 1994; Schafer, 2001), while for former, we only have one single system resulting in pseudo-replication (Hurlbert, 1984; Heffner *et al.*, 1996). A more intelligent design disentangling different gradients (Declerck *et al.*, 2005; 2007) to uncover relations that can be generalized (Millar and Anderson, 2004) would be hard to realise. In fact, similar to the freshwater rivers, it is necessary to collect a sample of similar estuaries, for instance mouthing in the North Sea, ranked for a gradient of anthropogenic impact, to construct a truly reference based IBI.

5.2.1.2. The response variable: the habitat quality class (HQC)

As worked out in Chapter 2, the response variable of the statistical model underlying an IBI is the condition of the ecosystem. To calibrate the model, an a priori assessment of the true state of the ecosystem is necessary, preferably independent of the ecological community to avoid circularity. This preclassification is the Achilles' heel of index development as rather seldom a gold standard can be achieved. Also in medicine, the development of a gold standard is the hardest part of the exercise (Zhou *et al.*, 2002). For IBIs, in most instances, the preclassification is derived from a sum of scores expressing the presence and/or impact of human activities and pressures (Van Stickle and Paulsen, 2008). The underlying rationale is that absence of human activity is indicative for a pristine situation and that exposure can be used to construct a gradient of impactness. Clearly, this approach involves rather strong assumptions (Yuan and Norton, 2004) including knowledge of dose-response curves (how much is the ecosystem affected by the human pressures/activities) and the additivity of these pressures ignoring differences in impact of the pressures and interaction (synergism).

As specified in Table 5.1 a series of human impact variables H_k is scored and their weighted sum (weights w_k) gives an indication of the overall human impact. Formally:

$$HIS = \sum_{k=1}^{K} w_k H_k$$

Most often, the weights are simply set to 1 or guessed by expert judgement. This total human impact score (HIS) is subsequently categorised by setting thresholds (lower part of Table 5.1), resulting in an ordinal variable, the human quality class (HQC), ranking the sites along a gradient of human impact. This approach mirrors the calculation of an IBI, but now with data independent of the ecological community. Table 5.2 gives the frequency distribution of HQC for the case study. HQC ranges from 3 to 5. The lower limit corresponds to a moderate habitat quality, which gives a true picture of reality, compliant with the Water Framework Directive (WFD). The implication is the reference is not the pristine situation. This is not a problem for the method proposed, but it limits the scope of the index to this range. For the index development, it does not matter whether we classify the sites from 3 to 5 or from 1 to 3, as it is based on a gradient. However, because data is lacking about class 1 and 2, we cannot extrapolate the index beyond class 3.

By no means, this system can be perfect. However, the preclassification is not a purpose in itself, but a device to rank the sites in a reasonable way with respect to an anthropogenic gradient of pressures enabling construction of an IBI. To put in perspective what is achievable, a well-

controlled exercise of Falcone *et al.* (2010) testing the capacity of an a priori ranking of watersheds with an extensive set of GIS-variables demonstrated that the diagnostic accuracy of the classification in least- and most-degraded sites was about two-thirds. Further standardisation is surely needed. In this respect, it is preferable to use existing schemes (Borja *et al.*, 2009a). Therefore, we derived our preclassification (Breine *et al.*, 2007) from an existing and validated framework (Aubry and Elliott, 2006) resulting from an extensive collaboration of field experts.

Table 5.1 **The preclassification: habitat indicators and threshold values.** Scores for the habitat indicators range from high (1) to bad (5) quality (adapted from Aubry and Elliott (2006)). For each site, the scores are summed resulting in an overall human impact score (HIS). This sum is categorised in the human quality class (HQC)

Parameter	**Score**				
+ GROUP 1: EXPOSURE INDICATORS (anthropogenic state & habitat alterations)					
	1	**2**	**3**	**4**	**5**
Minimum dissolved oxygen (DO) (%)	>80	≤80 & >70	≤70 & >50	≤50 & >30	≤30
Benthos (Not in Freshwater zone)		Classification of Brys *et al.*, 2005			
Intertidal area loss (%)	0	<20	≥20 & <30	≥30 & <50	≥50
Land reclamation (%)	0	<5	≥5 & <40	≥40 & <60	≥60
+ GROUP 2: ACTIVITY INDICATORS (anthropogenic activities & pressures)					
	1	**2**	**3**	**4**	**5**
Port & marina activities (absence / presence)	No				Yes
Industrial activities (expert judgement)	Low		Moderate		High
Dredging activities (absence / presence)	No				Yes
+ Human Impact Score (HIS) = sum of all indicators & Human Quality Class (HQC) = ordinal					
Thresholds for Human Impact Score (HIS)	7	8 - 14	15 - 21	22 - 28	29 - 35
Human Impact Level	Very low	Low	Moderate	High	Very high
Human Quality Class (HQC)	1	2	3	4	5

Table 5.2 **The preclassification:** frequency distribution of the habitat quality score (HQC): absolute and relative frequency for the full gradient and the two binary contrasts obtained by grouping the classes (number of fishing occasions = 130).

Frequency	**Ordinal classification (ranking)**			**Contrast 3 / 4-5**		**Contrast 3-4 / 5**	
	3	4	5	3	4-5	3-4	5
Absolute (N)	83	22	25	83	47	105	25
Relative (%)	63.8%	16.9%	19.2%	63.8%	36.2%	80.8%	19.2%

To have a balanced picture, we selected our indicators from two complementary groups of human pressures (Table 5.1): "exposure" and "activity" indicators (Bedoya *et al.*, 2009). The "exposure" indicators address factual anthropogenic alterations of the environment directly affecting the ecological community. The first two variables in Table 5.1 (dissolved oxygen and benthos) are factors supporting fish life (Turnpenny *et al.*, 2006); the next two variables express the loss of habitat availability for fish (Madon, 2008). For oxygen, the year median of monthly measurements was used for scoring. Benthos was scored on a yearly basis (Brys *et al.*, 2005). Intertidal area loss (%) and land reclamation (%) were determined with respect to the intertidal surface in 1960 and old maps from 1890 respectively. There is some overlap between both variables, however as land reclamation is more linked to industrial development, we gave it extra weight.

The "activity" indicators make an inventory of anthropogenic activities possibly deteriorating the ecosystem. Boat traffic and construction activities have a negative effect on fish life (Tull, 2006). The presence of marinas was assessed with aerial photographs. Industrial activities (e.g. bank reinforcement) decrease habitat diversity and occasional pollution has a negative impact on fish assemblages (Sindilariu *et al.*, 2006; Wheeler, 1969). The degree of industrial activity (low, moderate or high) was provided by experts. Dredging negatively influences benthic communities, a principal food resource for some estuarine fish species (Elliott *et al.*, 1998; Gard, 2002; Kennish, 2002). The Maritime Access Division of the Flemish Ministry provided data about the channel dredging activities.

5.2.1.3. The predictors: choice and motivation of the candidate metrics

The metrics are the explanatory variables or predictors of the IBI model. To compose a good starting set of metrics, the candidates should be inspired on ecological hypotheses about how the ecosystem will respond to ecosystem impairment (Bailey *et al.*, 2004; Olden *et al.*, 2006). To cover a broad spectrum of pressures (Noble *et al.*, 2007), we composed a set of candidate metrics from two complementary rationales: (i) the more classic "generic" metrics quantifying global functioning of (aquatic) ecosystems (Karr and Chu, 1999) and (ii) more specific "estuarine" metrics expressing functions of the estuary (Elliott *et al.*, 2007; Franco *et al.*, 2008).

Table 5.3 gives an overview of the metrics and specifies the underlying (alternative) hypothesis H_a: "generic" (G) or "estuarine" (E), and positive (+) or negative (-). Positive metrics are positively correlated with ecological quality and are expected to decrease under anthropogenic pressure. The opposite holds for negative metrics. The underlying hypothesis of generic metrics (G) is that, with increasing anthropogenic disturbance, species having narrow and/or specific habitat and/or biotic requirements (Pis: piscivores as top predators) will become less abundant, and, conversely, generalists (Omn: omnivores, often opportunists with a wide tolerance of feeding and other conditions) will become more dominant. However, as an estuary is a naturally stressed ecosystem with characteristics very similar to ecosystems subjected to human disturbance (McLusky and Elliott, 2004; Martinho *et al.*, 2008), it could be that the generic metrics are not very responsive.

To cope with this difficulty, known as the Estuarine Quality Paradox (Dauvin and Ruellet, 2009), we complemented the classical metrics with "estuarine" metrics specifically aimed at the composition and functioning of estuarine species (Elliott and Quintino, 2007). We included three metrics

incorporating the diverse ecological functions of estuaries such as providing spawning and nursery area (Mjm: marine juvenile migrating species, Ers: estuarine resident species) or the connection between the sea and upstream zones of the river (Dia: diadromous species). We also added three metrics reflecting the environmental circumstances of an estuary (Bra: brackish species) or species considered to be very characteristic of an estuary (Flo: Flounders, Sme: Smelt). In addition, we selected benthic species (Ben) because they are very sensitive to physical disturbance as dredging associated with harbour activities (MacDonald *et al.*, 1996).

Table 5.3 **Definition of the metrics.** The metrics are grouped by numerical format (ns, pi, va, vd). H_A = type of alternative hypothesis: G = generic metric, E = estuarine metric; + = positive metric, – = negative metric. Quality (of the metric): left = univariable diagnostic accuracy (3 = high, 2 = moderate & 1 = low) and right = final decision (X = excluded after screening (step 1), C = core metric (retained after step 1 but not included in final index), F = included in final index, is also a core metric). aucF & aucP = full & partial AUC for individual metric. BCa-limits = confidence limits of aucP based on the bias-corrected, accelerated percentile method (bootstrapping).

Code	Definition	H_A	Quality		aucF	aucP (+ BCa-limits)	
ns = metrics based on the number (#) of species in an ecological group or guild							
nsBen	# of benthic species [0 – 10]	E+	3	C	0.768	0.594	(0.453-0.717)
nsBra	# of brackish species [0 – 17]	E+	3	F	0.864	0.783	(0.652-0.889)
nsDia	# of diadromous species [0 – 5]	E+	1	X	–	–	–
nsErs	# of estuarine resident species [0 – 5]	E+	1	X	–	–	–
pi = metrics based on the proportion (%) of individuals in an ecological group or guild							
piDia	% of diadromous individuals	E+	1	X	–	–	–
piErs	% of estuarine resident individuals	E+	1	X	–	–	–
piMjm	% of marine juvenile migrating individuals	E+	3	F	0.880	0.770	(0.602-0.879)
piFlo	% of Flounder individuals	E+	2	F	0.664	0.507	(0.372-0.660)
piSme	% of Smelt individuals	E+	2	F	0.793	0.657	(0.478-0.829)
piPis	% of piscivorous individuals	G+	2	F	0.682	0.495	(0.341-0.639)
piOmn	% of omnivore individuals	G–	3	C	0.847	0.735	(0.578-0.848)
piExo	% of invasive individuals	G–	3	C	0.776	0.693	(0.537-0.809)
va = metrics based on the average value of indicative values of species							
vaTol	Average intolerance value	G+	3	C	0.846	0.770	(0.628-0.885)
vd = metrics based on diversity indices							
vdDiv	Simpson diversity index	G+	2	C	0.735	0.638	(0.477-0.782)
vdSha	Shannon diversity index	G+	1	X	–	–	–
vdSim	Simpson dominance index	G+	1	X	–	–	–

For the generic positive metrics (G+), we complemented the metrics based on ecological guilds with diversity metrics (vdSha, vdSim, vdDiv) and an intolerance index (vaTol). As explained in Chapter 2, species richness and evenness of the species distribution tends to decrease under anthropogenic stress which can be captured by diversity indices. The intolerance value is an average of sensitivity scores of the species, compiled in Breine et al. (2001), weighted by their abundance. Finally, we added invasive species (Exo) as a general negative metric of disturbance.

In principle, the metrics are based on proportions derived from abundance data which is more informative than presence/absence data. In addition, with metrics based on the number of species, the range of values is restricted diminishing the discriminatory power. Yet, for the estuarine metrics, some metrics on the number of species as they are informative indicators of the number of species associated with the estuary: brackish species (nsBra), diadromous species (nsDia) and estuarine resident species (nsErs).

5.2.2 The IBI model

In Chapter 2, we demonstrated an IBI is in fact a statistical regression model. This holds even for the commonly used average score model (AVG) which is simply an average of scored metrics.

5.2.2.1. The generic four-step format

Although many variants exist, in essence, most multi-metric indices (MMIs) have the same four-step format as described in the first line of Box 5.1. (i) The model equations start with extracting from the community data the relevant ecological information by calculating a set of metrics, i.e. indicator variables representing ecosystem attributes (composition, structure or function) that are sensitive to the anthropogenic alterations of the environment. (ii) Subsequently, the metrics are scored, to express how (dis)similar the metric observations are in comparison to type-specific or site-specific reference conditions by taking into account the site typology and/or correcting for differences in environmental conditions. (iii) In a third step, the individual scores are combined (traditionally by simply summing or averaging) into one single measure assessing the global impact, the ecological quality measure (EQM). As such, EQM is hard to interpret because it does not tell very much how impaired the ecosystem is and whether restoration is necessary. (iv) Hence, the fourth and final step compares the EQM with decision thresholds resulting in the ecological quality class (EQC), an ordinal class variable appreciating the degree of degradation of the ecosystem (or, expressed positively, the level of biotic integrity). This simple one-line format is very flexible and hides a lot of complexity. Each of the four steps can be described by one or more mathematical functions. (v) At the calibration stage, the appropriate functions should be derived and the (unknown) parameters estimated by matching EQC as closely as possible with the human quality class (HQC), an independent gold standard assessment of the true state of the ecosystem.

5.2.2.2. The average score model

A special case of the one-line format is the average score model (AVG). This simple model is representative for many IBIs. Box 5.1 presents AVG along the four transformation steps of the generic index format. (i) The metric functions are weighted averages or diversity measures of the

community data (see Chapter 2). (ii) The scoring functions are z-scores scaling the metrics with respect to the reference distribution. (iii) The predictor function calculates EQM and is simply a (weighted) average of scored metrics. (iv) Finally, the decision function specifies how to derive EQC by comparing EQM to two decision thresholds. The first makes a distinction between the baseline class 3 (moderate) and the more degraded classes, the second is a further refinement between class 4 (poor) and 5 (bad).

Box 5.1 **The average score models (AVG).** C = community data, M_j = metrics (j = 1,2, …, J), S_j = metric scores, EQM = ecological quality measure, EQC = ecological quality class, HQC = human quality class, qROC = diagnostic quality derived from the ROC curve (e.g., aucF & aucP = full & partial area under the ROC curve). E_R & $Stdev_R$ = expected value and standard deviation of the metric under reference conditions. $T_{3/4\text{-}5}$ = threshold between class 3 and higher (4-5), $T_{4/5}$ = threshold between class 4 and 5. β_j = regression coefficients (β_0 = intercept, does not add to discrimination as independent of the metrics).

$$\boxed{C \overset{(i)}{\rightarrow} M_j \overset{(ii)}{\rightarrow} S_j \overset{(iii)}{\rightarrow} EQM \overset{(iv)}{\rightarrow} EQC} \quad \overset{(v)\ qROC}{\leftrightarrow} \quad HQC$$

$$(i)\ metric\ \ functions: M_j = C \otimes species\ characteristics \quad (j = 1,2,...,J)$$

$$(ii)\ scoring\ \ functions: S_j = \delta_j \frac{M_j - E_R[M_j]}{Stdev_R[M_j]} \qquad \delta_j = \begin{cases} +1 & if\ \ positive\ metric \\ -1 & if\ \ negative\ metric \end{cases}$$

$$(iii)\ predictor\ \ function: EQM = \sum_j w_j S_j \overset{w_j = \frac{1}{J}}{=} \frac{1}{J}\sum_j S_j$$

$$(iv)\ decision\ \ function: EQC = \begin{cases} 3 & if \quad T_{3/4-5} < EQM \\ 4 & if \quad T_{4/5} < EQM \le T_{3/4-5} \\ 5 & if \quad EQM \le T_{4/5} \end{cases}$$

Relation with logistic regression models

$$\boxed{logit(EQM) = \log\left(\frac{EQM}{1-EQM}\right) = [\beta_0] + \sum_j \beta_j S_j} \quad \overset{\beta_0 = 0;\ \beta_j = \frac{1}{J}}{\rightarrow} \quad logit(EQM) = \frac{1}{J}\sum_j S_j$$

Box 5.2 extends the diagnostic accuracy measures to a tri-state ordinal model. Because of the ordinal nature, we can speak about false positives and false negatives and pool classes in binary contrasts. The latter allows to apply the ROC curve to construct indices, for instance with respect to aucP(3,4-5). As discussed in Chapter 2, the original discrete scoring of Karr (1981; 1986) remains

to be used although it implies a loss of power (Blocksom, 2003). Therefore we shortly will investigate the impact of the scoring.

Box 5.2 **Confusion matrix of FP and FN extended to three ordinal classes.** The matrix on the left is for the full impact gradient, the matrix on the right is for the binary contrast 3 / 4-5. Note: TCF = true classification fraction (correct classification).

	$EQC = 3$	$EQC = 4$	$EQC = 5$
$HQC = 3$	$TCF(3,3)$	$FPF(3,4)$	$FPF(3,5)$
$HQC = 4$	$FNF(4,3)$	$TCF(4,4)$	$FPF(4,5)$
$HQC = 5$	$FNF(5,3)$	$FNF(5,4)$	$TCF(5,5)$

$\leftrightarrow$

	$EQC = 3$	$EQC = 4\text{-}5$
$HQC = 3$	$TNF(3,3)$	$FPF(3,4\text{-}5)$
$HQC = 4\text{-}5$	$FNF(4\text{-}5,3)$	$TPF(4\text{-}5,4\text{-}5)$

5.2.2.3. The proportional odds model

As the response variable is an ordinal variable, we can use ordinal logistic regression models to estimate the regression parameters. At the end we compare AVG with the proportional odds model. The latter model assumes that for each cumulative binary contrast of the response variable (3/4-5 and 3-4/5), the regression coefficients are equal. Only the intercepts differ. As explained in Chapter 2, the underlying assumption of the proportional odds model is that the classes of biotic integrity are a discretisation of a "latent" continuous variate of which the distribution is shifted by an increasing disturbance (McCullagh and Nelder, 1989).

5.2.3 Optimisation criteria

5.2.3.1. The optimisation contrast

We calibrate the index with respect to the habitat quality class (HQC), which is an ordinal variable with three levels. We primarily optimise the model for the distinction between the baseline class 3 (least degraded) and the other two more degraded classes 4-5. Except for the initial screening of the metrics, we do not use the information distinguishing class 4 and 5. Yet, as the separation between class 4 and 5 is important, afterwards the model is validated by controlling whether the diagnostic increases with increasing degradation. Also we compare the results with an ordinal logistic regression model.

5.2.3.2. The full and partial area under the ROC curve (aucF & aucP)

As the optimisation criterion, we use the area under the ROC curve for the contrast 3/4-5. We complement the full AUC (aucF) with a focused variant: the partial AUC (aucP) in the range of 10 % to 30 %. We expect that for FPF < 10 %, the sensitivity will be too low and for FPF > 30 %, the burden of FPs becomes very high. Figure 3.3 compares both aucF and aucP with the sensitivity at different fixed levels of FPF. The advantage of aucP over the sensitivity at fixed levels is its better statistical properties as it is an integrated measure (Dodd and Pepe, 2003). This general result will be tested in our specific case. Chapter 2 gives more details.

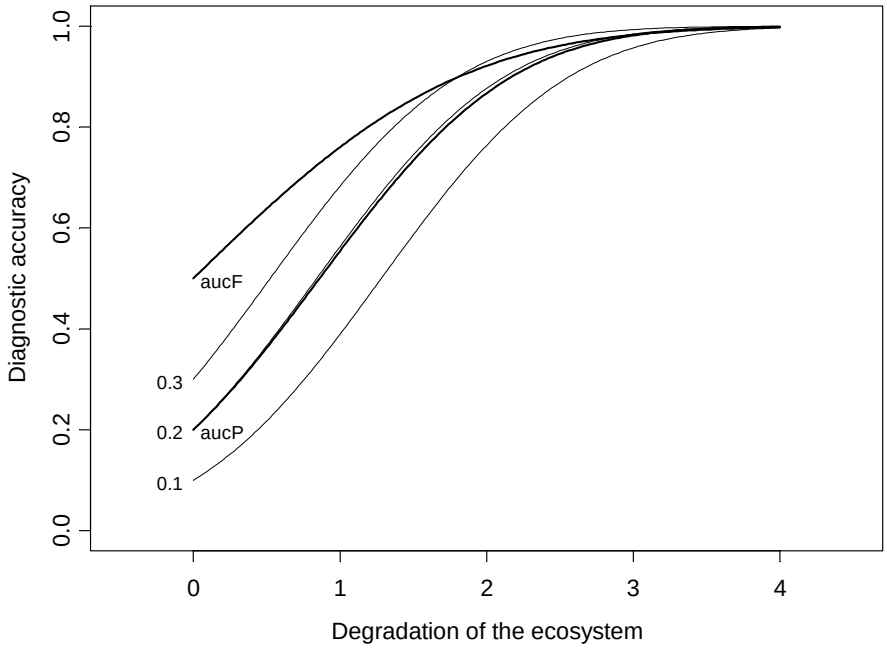


Figure 5.1 **Comparison of the measures of diagnostic accuracy.** Evolution of aucF and aucP($0.1 \leq FPF \leq 0.3$) (lines in bold) compared with TPF (the sensitivity) at fixed values of FPF (0.1, 0.2 and 0.3) for an increasing degradation of the ecosystem (modelled as an increasing distance between two normal distributions with equal variance).

5.2.3.3. Statistical properties and bootstrapping

Both aucF and aucP are rank based statistics quite robust against outliers. However, their statistical sampling distributions are complex. Also the theoretical asymptotic results depend crucially on underlying hypotheses, because, in contrast to for example the sample mean which is quite robust against distributional assumptions, these AUC statistics depend on the whole distribution. In this situation, resampling methods as bootstrapping are ideal for simulating the sampling distribution and its corresponding statistics (Efron and Tibshirani, 1993; Shao and Tu, 1995; Lunneborg, 1999; Davison and Hinkley, 2006).

In addition, measures of diagnostic accuracy are bias prone. We expect that, because of the fitting, the model will be closer to the calibration data than can be expected from future samples and, as a consequence, statistics of diagnostic accuracy will be biased upwards. Bootstrapping allows to estimate bias of a statistic by comparing the parameter calculated from the EDF with the mean of its resampled distribution. As explained in Figure 5.2, this estimate can be used to correct for bias (by subtracting the estimated bias from the observed statistic) and to calculate the corresponding confidence limits. Based on this principle, Efron (1987) developed the accelerated bias-corrected percentile method (BCa) improving the consistency of the estimator for small samples.

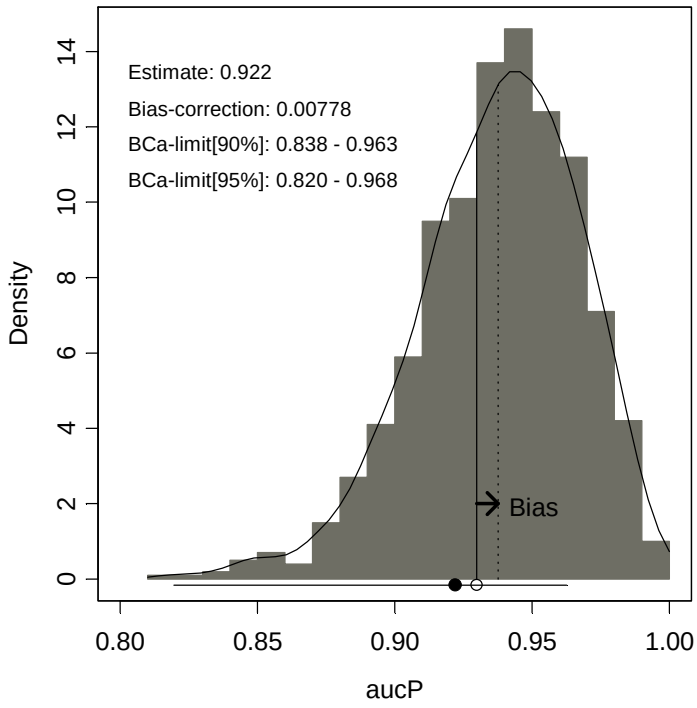


Figure 5.2 **Principle of the accelerated bias-corrected percentile method** (BCa method (Efron, 1987); n_B = 1000) to correct for bias. The white point equals aucP estimated from the original sample. The density distribution represents the resampling distribution of aucP. The difference between the mean of the resampling distribution (dotted line) and the white point is an estimate of the bias. By subtracting the bias estimate, a bias-corrected measure of aucP is obtained (black point). The percentiles of the bias-corrected resampling distribution are used to calculate the confidence limits as described by Efron (1987).

5.2.4 Description of the modelling steps

5.2.4.1. Step 1. Screening the diagnostic accuracy of the individual metrics

The number of possibilities increases exponentially with the size *s* of the candidate set; there are 2^s combinations possible (or $2^s - 1$ excluding the null model with no metrics). Therefore, it is important to limit the candidate set of metrics (by careful ecological motivation) and to drop the less relevant metrics with exploratory data analysis (EDA) techniques. The first remedy is not the topic of this paper. We take for granted the proposed metrics are well motivated. The preliminary EDA screening is a usual part of any statistical analysis, but requires here some special attention. The univariable preselection of the metrics may not be too severe. Predictors which are individually poor can offer complementary information in combination with other variables. Therefore, we only

eliminate metrics not showing any response at all to anthropogenic pressure. The candidate metrics retained we call the core set.

5.2.4.2. Step 2. All possible subsets investigation

After the preliminary stage, we determine the level of complexity by investigating the trend of the diagnostic accuracy of all possible subsets of the core set as a function of the number of metrics in the index, visually with boxplots, and, more formally, with bootstrapping. We resample all pairwise differences of the diagnostic accuracy of the best models to assess the confidence limits. If a zero is included in the confidence intervals, the models do not differ statistically from each other (test of the null hypothesis, but without correction for multiple testing). However, we also look at the limits of the confidence limits to judge how different the models can be as confidence limits also give information which alternatives are compatible with the data. In this way, we hope to balance the type I error (selecting too many metrics or overfitting) and the type II error of eliminating good metrics (selecting too few metrics or underfitting).

5.2.4.3. Step 3. Exploration in the vicinity of the optimal models

By comparing 1023 models ($2^{10} - 1$), the best model can just be a lucky event (Zucchini, 2000). With an exploration of the models in the vicinity of the optimum and comparing them, we have some control whether this is the case. Absence of a systematic pattern (for instance, a model around the optimum has totally different implications) is an indication the data are not coherent or the family of candidate models is not well chosen. This is not to say that the models may not differ in important points, but these differences should be interpretable and consistent within the (ecological) framework the model building was started. By comparing the alternatives, it should be possible to get a better understanding of the underlying processes and/or some particularities of the data influencing the end result. The final aim of this step is to make a decision about the final model(s), at best one, but not more than a few. In this process, we should be careful for an overinterpretation of the results and we should be aware the procedure is not a full-proof guarantee.

5.2.4.4. Step 4. Tuning of the final model

The final step is the tuning of the proposed index against the full gradient by setting the decision thresholds. We set FPF to 20 % for both the main contrast "3/4-5" and the contrast "4/5" (thus FPF(3,4-5) = 0.2 and FPF(4,5) = 0.2). For this configuration, the confusion matrix can be estimated. Special point of attention is whether FPF(3,5) and FNF(5,3) are sufficiently small (as they represent misclassification of two categories) and whether the overlap of classes 3 and 5 with class 4 in between is not too large (FNP(4,3) and FPF(4,5)). We also investigate in more detail the sensitivity for the main contrast (3/4-5) at different cut points with bootstrapping.

5.3 Results

5.3.1 Step 1: screening the univariate response of the candidate metrics

5.3.1.1. Boxplots

Figure 5.3 and Figure 5.4 contain the boxplots to evaluate the potential of the individual candidate metrics for the index model. The four metrics based on the number of species (upper part of Figure 5.3) are discrete. The metrics nsDia (diadromous) and nsErs (estuarine species) have too few distinct values (≤ 5). Preliminary analyses (results not shown) including these two metrics resulted in numerical instabilities because they appeared to fit to particularities of the data in combination with other data. Also, they do not show a consistent response to human pressure. Therefore, we excluded them from further analyses. In fact, only nsBra (brackish species) has sufficient distinct values showing a clear response to HQC. However, we also keep nsBen (benthic species) as it has some discriminatory power with respect to the lowest habitat class. None of the diversity indices (lower part of Figure 5.3) had a clear response with HQC. Only for vdDiv (Simpson's diversity index) some relation was found. The tolerance metric (vaTol) performed well.

For the metrics based on the proportion of individuals (Figure 5.4), two metrics piMjm (marine juvenile species, a positive indicator) and piOmn (omnivores, a negative) showed a highly consistent but opposite response the human impact gradient. Also the response piSme (smelt) is sharply marked, but the range of this metric is rather small (from 0 to 5 %). This metric is based on one species only. Next comes piExo (exotic species) with a considerable overlap between class 3 and 4 of HQC, but class 5 is clearly distinguished. The same holds for piFlo (flounders) and piPis (piscivores) showing a contrast between class 3-4 and 5 in the opposite direction as the exotic species. Finally, the proportions of estuarine resident species (piErs) and of the diadromous species (piDia) do not seem linearly correlated with HQC. Hence these two metrics were dropped after the first step.

5.3.1.2. Synoptic diagram

As explained in section 3.3.2, it is possible to complement the screening based on boxplots with empirical distributions (EDF) and ROC curves. Figure 5.5 synthesizes this information in a "synoptic diagram" of aucF for two contrasts 3/4 and 3/5: aucF(3/4) and aucF(3/5). At one glance, we can compare all metrics with respect to each other and the metrics are grouped according to the type of relation with the pressure gradient. Metrics with coordinates close to (0.5,0.5) are totally insensitive, as aucF = 0.5 corresponds to an indifferent indicator. This is the case for vdSha and vdSim. Two other metrics with low performance are piErs and piDia. Their response to the contrast 3/4 is better, but it disappears for 3/5 suggesting the curvilinear response as found with the boxplots. Therefore, we exclude also these two metrics for further investigation (note that we excluded nsDia and nsErs because of their limited range in the previous step). The other ten metrics in Figure 5.5 are kept in the model.

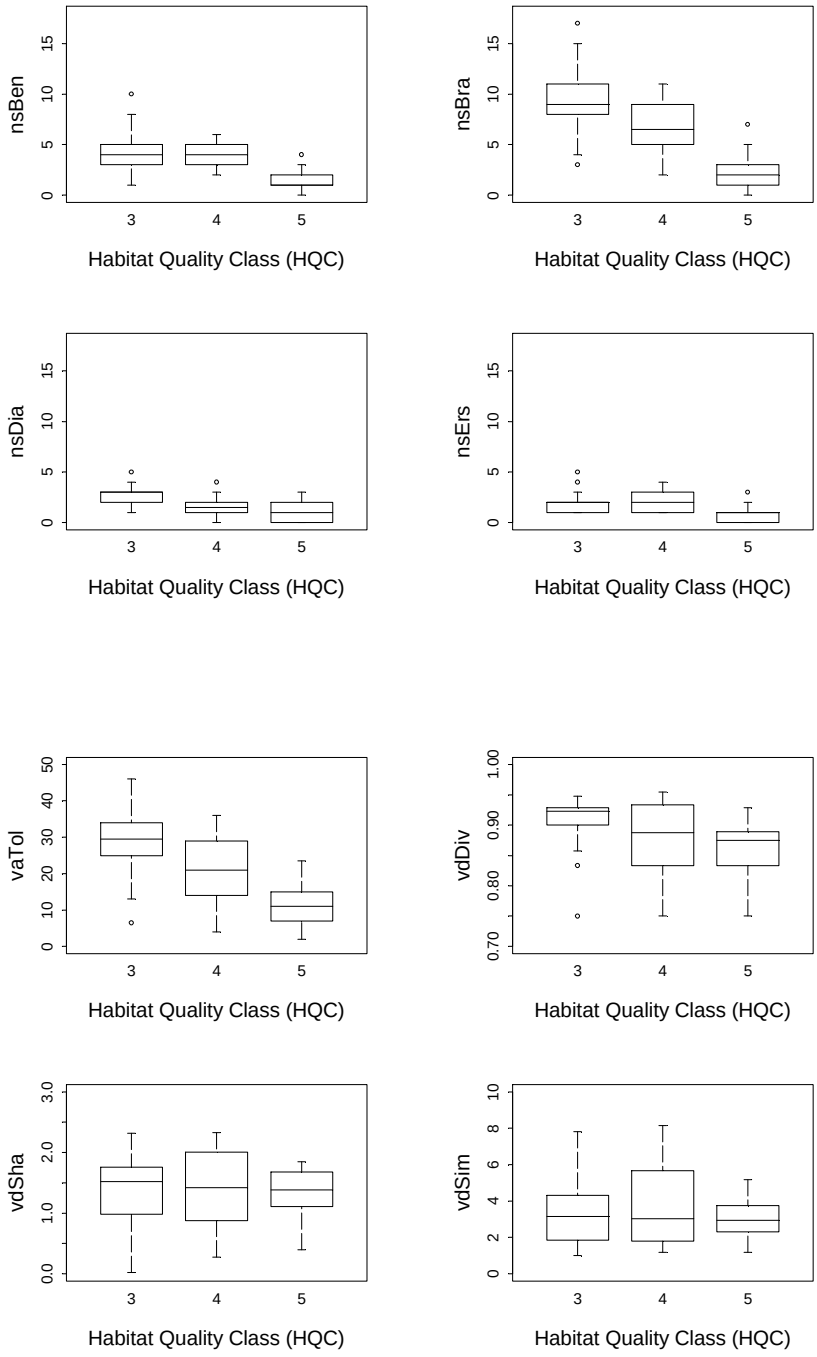


Figure 5.3 **Univariable response of the metrics**. **Boxplots along a gradient of human impact** as preclassified with the habitat quality class (HQC). The metrics (see Table 5.3) are based on the number of species belonging to a certain ecological guild (ns), diversity measures (vd) or express tolerance values (va).

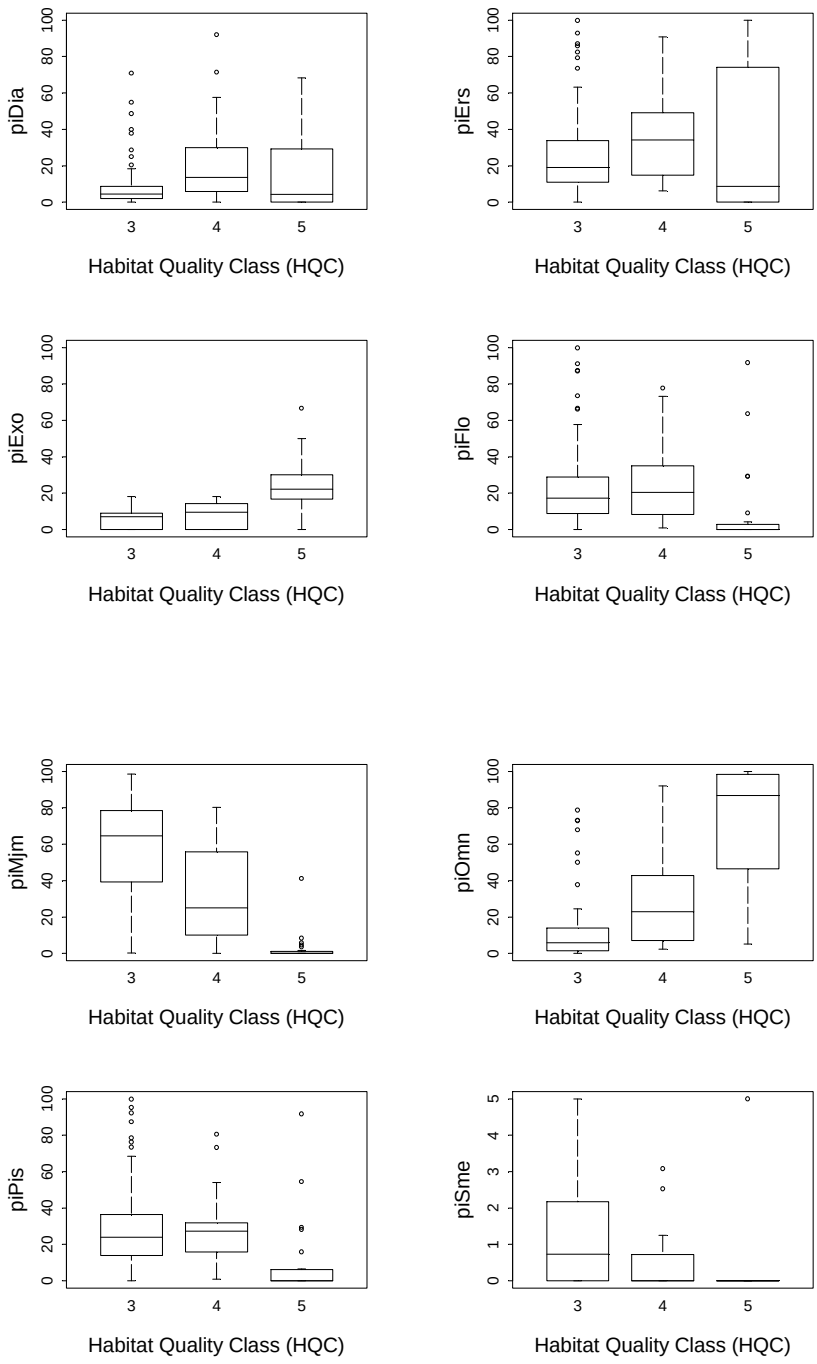


Figure 5.4 **Univariable response of the metrics. Boxplots along a gradient of human impact** as preclassified with the habitat quality class (HQC). All metrics are based on the percentage of individuals (pi) belonging to a certain ecological guild (see Table 5.3).

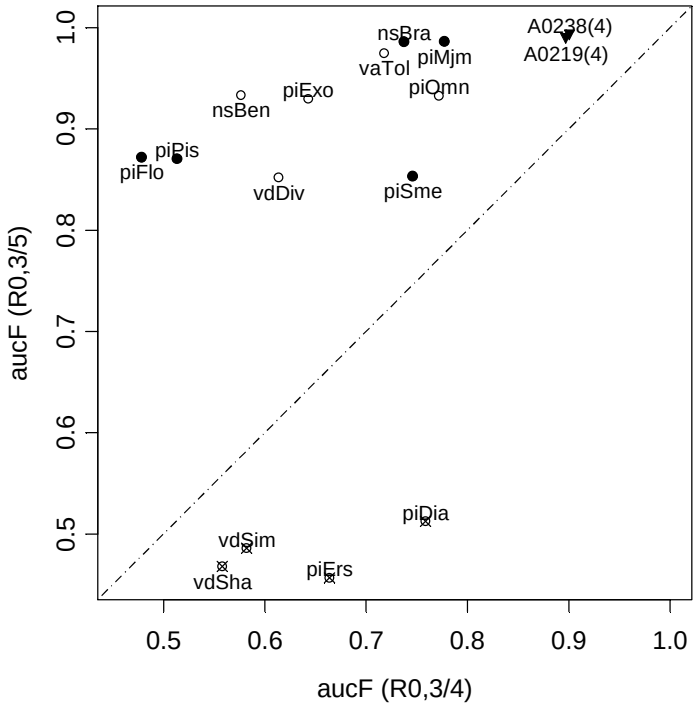


Figure 5.5 **Individual response of candidate metrics compared with the two best models with four metrics: synoptic diagram of the diagnostic accuracy** (aucF for the binary contrasts 3/4 and 3/5). ○ = metric; ● = metrics included in the two best models with four metrics (piFlo & piPis are interchangeable, see text, thus five metrics are marked); crossed (x) bullets: metrics eliminated because of a low discriminating value. ▼ = the two best models with four metrics (R0.A0238 and R0.A0219, see Table 5.5). R0 refers to the type of scoring as proposed by Blocksom (2003): with respect to the reference distribution (R) and continuous (0 = no classes).

All remaining metrics show at least a good response to the contrast 3/5 (the y-axis). They show an increasing response to human impact as they are located above the first diagonal (aucF(3/4) < aucF(3/5)). At the upper right part of the plot, we have the four best metrics nsBra, piMjm, vaTol and piOmn. As is apparent from the boxplots, they have good discriminatory power with respect to the first contrast 3/4 (aucF(3/4) ≈ 0.75) and a monotone relationship to human impact (aucF(4/5) ≈ 0.95 > aucF(3/4)). It is instructive to compare them with the performance of the two best MMIs with four metrics (triangles). For the contrast 3/5, the improvement of the MMIs is minor, but the response of the four-metric MMI with respect to 3/4 is much better. Combining the metrics improves the discriminatory power at lower levels of human impact.

As indicated by the filled bullets in Figure 5.5, only two of the four best univariate metrics are included in the final models (nsBra and piMjm). The other two metrics are piSme and piFlo or piPis

(the latter two are interchangeable). This finding illustrates (as is well known) that the best model does not necessarily contain the best metrics from a univariate perspective, but that one should search for the best complementary set. The two interchangeable metrics (piFlo or piPis) are lying in the upper right corner of the diagram. These two metrics are insensitive to smaller deviations from the baseline (aucF(3/4) ≈ 0.5) but show a high response to larger deviations (aucF(3/5) ≈ 0.85). This different response results in complementary information, which is possibly the reason why they are selected for the model. Three other metrics with a similar behaviour are vdDiv, nsBen and piExo. A special case is piSme lying close to the first diagonal: the contrast of 3/4 and 3/5 is about equal. This is not directly apparent from the boxplot (Figure 5.4), but the small range of values is probably due to this fact.

In summary, four metrics are not responsive at all: two diversity indices (vdSha & vdSim) and two estuarine related metrics (piDia and piErs). We can (safely) drop these four nonresponsive metrics together with nsDia and nsErs (because of their limited range) for the subsequent analyses resulting in 10 core metrics reducing the workload considerably for the next step screening all possible combinations ($2^{10} - 1 = 1023$ instead of $2^{16} - 1 = 65536$ models to investigate).

5.3.2 Step 2: determination of the dimension of the model

5.3.2.1. Best subset regression

To determine the required model complexity, we considered all possible subsets of metrics and plotted the boxplots (upper panels of Figure 5.6) of the diagnostic quality as a function of the size of the model. As expected, with increasing complexity, the boxplots shift upwards and they become more compact, i.e. the difference between the best and the worst model decreases. There are many competing models with interchangeable metrics. However, after a while, the maxima bend down (bottom panels of Figure 5.6). For aucF, the maximum is reached at four metrics, and afterwards the diagnostic quality of the optimal model gradually levels off. The aucP criterion gives the same message, but the pattern is more pronounced. The increase is most important from one to three metrics, levels off reaching its maximum at five metrics and then goes sharply down. The message is essentially the same: the best index is situated in the range from three to five metrics.

5.3.2.2. Bootstrapping the difference in diagnostic accuracy

To evaluate whether the optimal models are significantly different from each other, we used bootstrap to assess the confidence limits of the difference in diagnostic accuracy of the best models for an increasing number of metrics in the model (Figure 5.7). At three metrics, none of the more complex models significantly differs anymore, but the point and interval estimates of the difference with four or five metrics is shifted far to the right. Hence, we cannot totally exclude that a model with four or five metrics is better. At four metrics only the model with five metrics is slightly better, but the point estimate is small compared to the confidence interval which indicates they are essentially equivalent. From five metrics on, the message is clear. All models with more metrics are significantly worse. From this analysis we infer that the most parsimonious model that predicts well is a model with three metrics. However, given the small number of observations (and hence relatively small power), also the model with four (and to a lesser extent five) metrics is defensible.

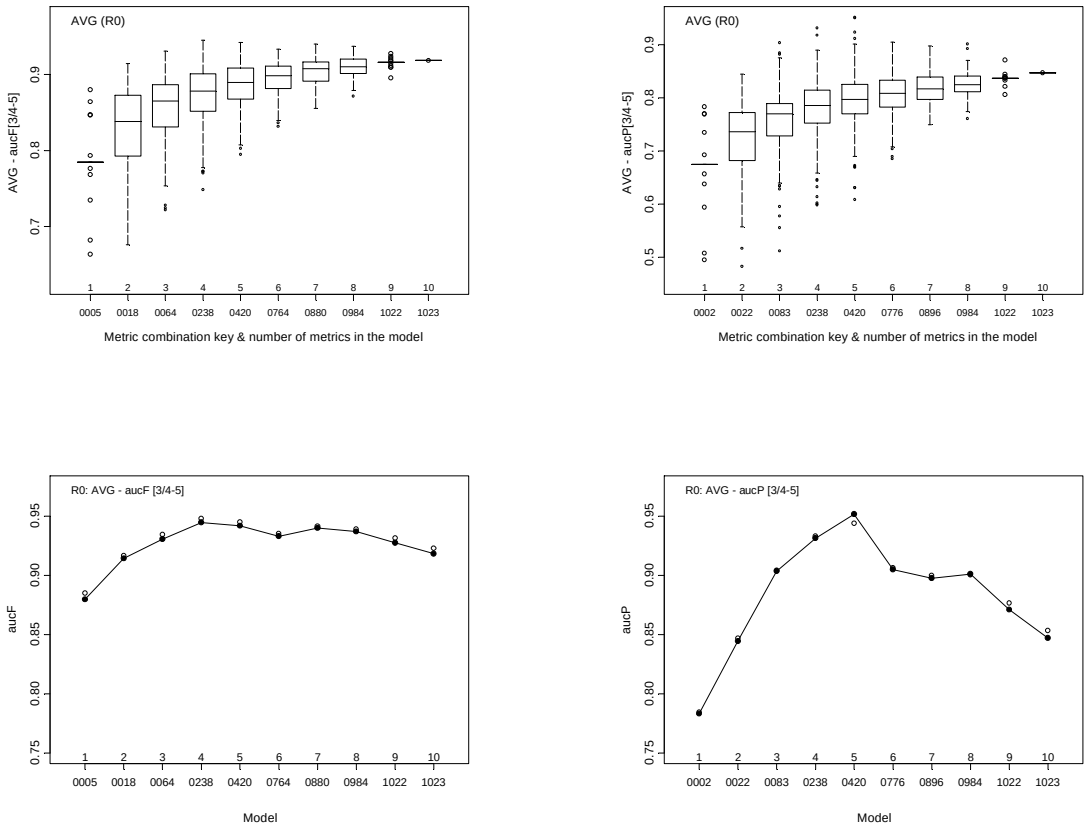


Figure 5.6 **Determination of the model dimension.** Top: evolution of the diagnostic accuracy (left: aucF and right: aucP for the contrast 3/4-5) of all possible models as a function of the number of metrics in the index (x-axis). Bottom: optimal model (maximum value of the boxplots) for each number of metrics in the index. White points = uncorrected estimates; black points = (bootstrapped) bias-corrected estimates (n_B = 1000). The labels of the x-axis refer to the model keys in Table 5.5 and the number of metrics.

5.3.2.3. Note 1: the bias-correction

The bias-correction in Figure 5.6 appears to be small. A similar exercise based on logistic regression estimating the weights (Box 5.1) resulted in a more pronounced bias-correction. Figure 5.8 gives the trend of the median bias on both aucF and aucP for all possible models as a function of the number of metrics. As expected, the bias increases with the number of metrics in the model for logistic regression. In contrast, for AVG it remains constant (except for the transition form a model with one to two metrics). A possible explanation is that the logistic regression is more flexible and can be better fit to the idiosyncrasies of the data resulting in an overfitting and overestimating the diagnostic accuracy. Note that the logistic regression models are fit with the maximum likelihood criterion, while the AVG models are optimised directly to aucP or aucF. Pepe (1997; 1998) and Pepe and Thompson (Pepe and Thompson, 2000) developed algorithms to optimise logistic regression models with respect to AUC but this procedure is not readily available for testing.

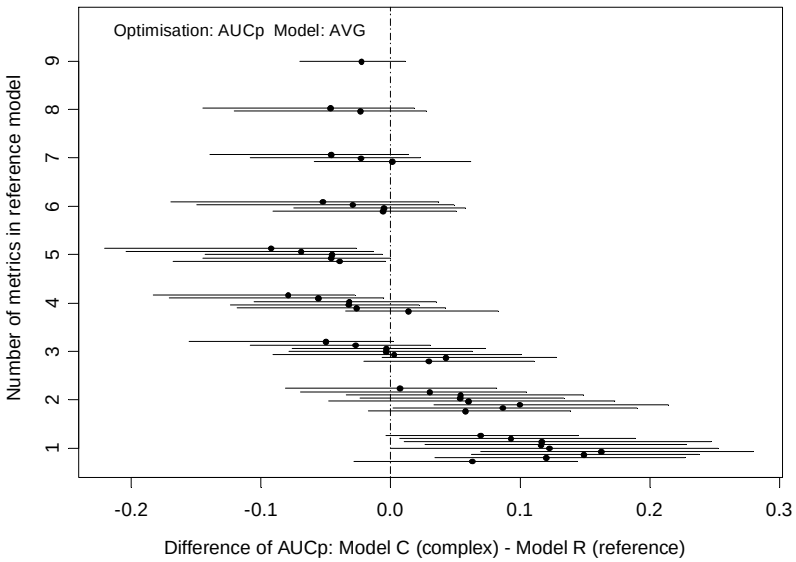


Figure 5.7 **Determination of the model dimension.** BCa confidence limits (n_B = 1000) of the pairwise differences of diagnostic accuracy of the best models with respect to the main contrast 3/4-5 (aucP). Each group of segments compares the best model for a given number of metrics with all other best models of higher complexity. For instance, at position 6, the segments compare the optimal model for six metrics, with the optimal models for 7, 8, 9 and 10 metrics.

5.3.2.4. Note 2: the evolution of the diagnostic accuracy

For a better understanding of the mechanism explaining the difference in trend of aucF and aucP, we visualised the evolution of the ROC curves (Figure 5.9) and estimated the sensitivity at different cut points of FPF in the range from 0.1 to 0.3 (Table 5.4). From one to four metrics, all indicators steadily increase (note that TPF(0.1) first goes down, before increasing). At first, the highest progress is in the lower range of the curve. This explains the sharp increase of aucP in comparison to aucF (Figure 5.6). Going from the fourth to the fifth metric, aucP still increases, but aucF decreases because the ROC becomes worse below 0.1. In Table 5.4 some of the sensitivity values still increase. The last ROC curve with six metric confirms this downwards tendency, but now the quality also deteriorates in the range from 0.1 to 0.3 which is reflected in a decrease of aucP. Hence, there is an indication that with the entry of the fifth metric, there is overfitting.

Table 5.4 **90 % BCa limits (n_B = 1000) of the diagnostic accuracy measures** (aucF, aucP, TPF at FPF from 0.1 to 0.3) for the optimal model with respect to the binary contrast (3/4-5). The shaded cells express where a (local) maximum is reached.

	R0.A0002(1)	R0.A0022(2)	R0.A0083(3)	R0.A0238(4)	R0.A0420(5)
aucF	0.862 [0.797,0.919]	0.906 [0.859,0.942]	0.93 [0.877,0.96]	0.945 [0.903,0.971]	0.942 [0.902,0.967]
aucP(0.1,0.3)	0.779 [0.659,0.876]	0.843 [0.705,0.924]	0.905 [0.781,0.969]	0.932 [0.829,0.979]	0.953 [0.859,0.991]
TPF(0.1)	0.758 [0.587,0.866]	0.677 [0.473,0.848]	0.641 [0.497,0.905]	0.844 [0.590,0.946]	0.752 [0.536,0.952]
TPF(0.15)	0.798 [0.683,0.902]	0.843 [0.624,0.926]	0.926 [0.663,0.973]	0.947 [0.803,1.000]	0.949 [0.768,0.992]
TPF(0.2)	0.767 [0.635,0.872]	0.858 [0.699,0.937]	0.921 [0.793,0.979]	0.948 [0.839,0.992]	0.924 [0.792,0.982]
TPF(0.25)	0.824 [0.681,0.915]	0.871 [0.741,0.947]	0.93 [0.823,0.982]	0.95 [0.860,0.998]	0.97 [0.884,1.000]
TPF(0.3)	0.885 [0.774,0.952]	0.928 [0.814,0.982]	0.964 [0.876,1.000]	0.986 [0.911,1.000]	0.990 [0.911,1.000]

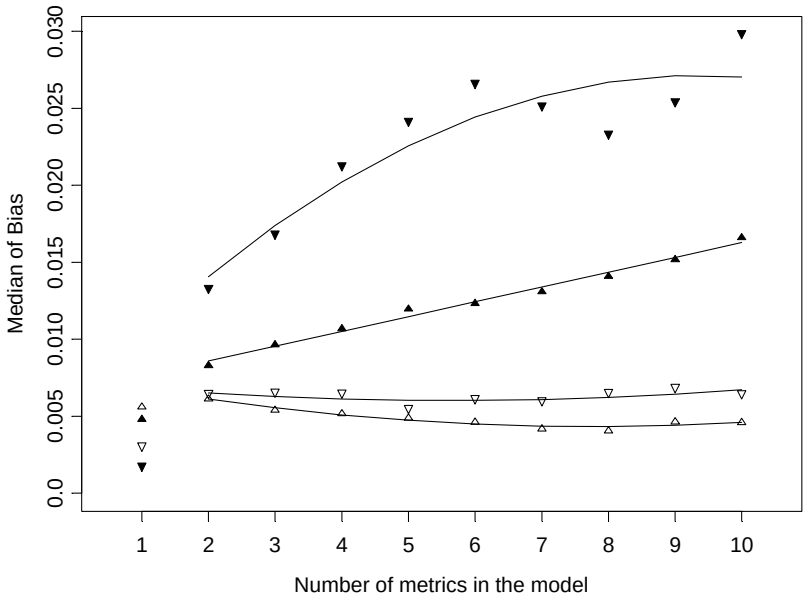


Figure 5.8 **Median of bias on AUC** for all possible models as a function of the number metrics in the model. Legend: ▼ = aucP and ▲ = aucF for logistic regression and ▽ = aucP and △ = aucF for AVG. The smooth curves are based on locally weighted regression models (Cleveland and Devlin, 1988).

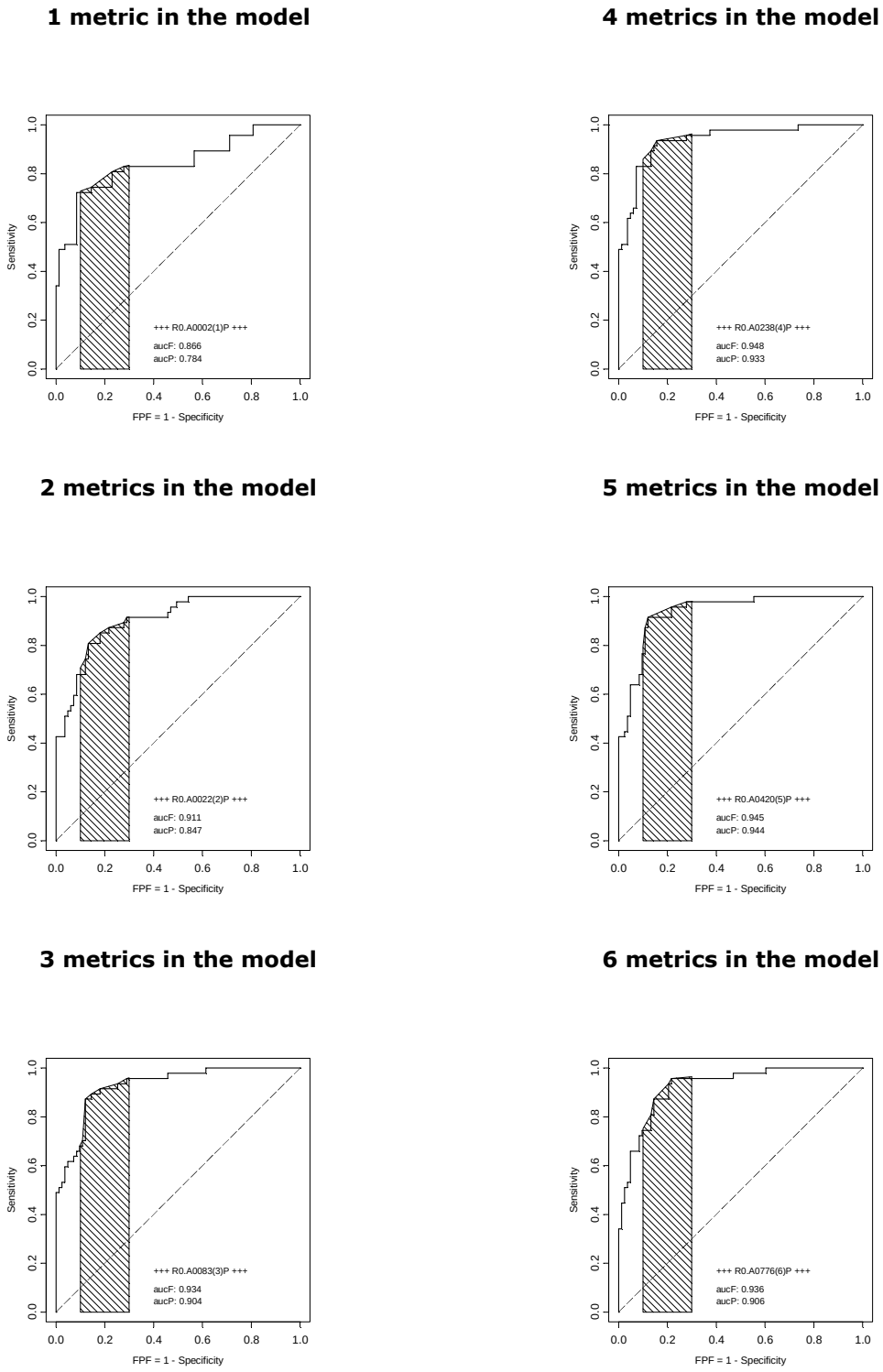


Figure 5.9 **Stepwise improvement of the ROC-curves optimised with respect to aucP**: aucP improves up to 5 metrics. Yet, by adding the fifth metric, aucF decreases because the ROC curve becomes worse for FPF ≤ 0.1.

5.3.3 Step 3: exploration in the vicinity of the optimum

5.3.3.1. Comparison of the seven best models

For the exploration in the vicinity of the optimum, we selected models with a diagnostic accuracy at least as good as the best model with three metrics. This resulted in eight models: one pair with three metrics, one pair with four metrics, and two pairs with five metrics. Figure 5.10 gives a synoptic overview of the diagnostic accuracy of these model. Table 5.5 makes clear why we talk about pairs. At first, no consistent pattern in the metric composition was apparent. However, piPis and piFlo appeared to be interchangeable: for each model with piSme, there is a "twin model" with piFlo with about the same quality. By interchanging piFlo and piPis, the successive models become nested: nsBra + piMjm + (piPis or piFlo) + piSme + piOmn. The synoptic diagram Figure 5.5 reveals that the interchangeable metrics do not discriminate well along the first baseline contrast (3/4), but have a good resolution for the second baseline contrast (3/5). Compared with many other candidate metrics, they are weak, but they are selected for the model because they offer complementary information.

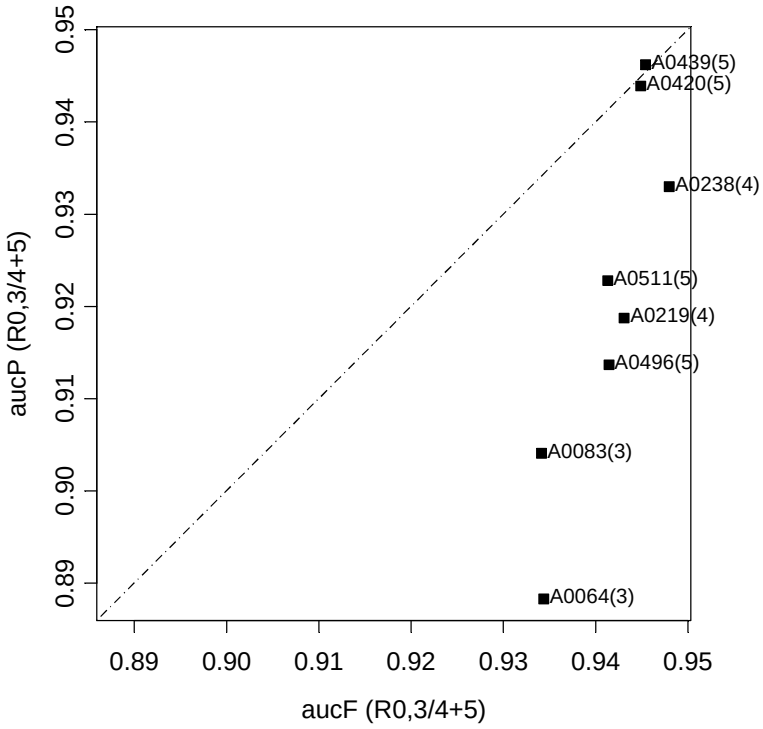


Figure 5.10 **Synoptic diagram of the seven best models** (both aucF and aucP)**.** For the model keys, see Table 5.5 (within brackets, the number of metrics in the index is specified). The models do not differ significantly as based on a bootstrapped confidence limits of aucP (Figure 5.5).

The criteria aucF and aucP at first select different metrics: aucF starts with piMjm while aucP takes nsBra, revealing that nsBra has a (somewhat) better local performance than piMjm. However, from three metrics on, both nsBra and piMjm are included in the model. These two metrics are the best individually (Figure 5.5). The third metric entering is piPis or piFlo (whichever), the fourth is piSme and the fifth is piOmn. Although piOmn is better individually (Figure 5.5), piSme is entered first.

Table 5.5 **Optimal IBIs (and close competitors) up to six metrics.** Both optimisation criteria are bias-corrected with bootstrap (n_B = 1000). The arrows indicate the trend in comparison to the previous model with one metric less. Model keys: the number in the centre refers to a (unique) combination of metrics as indicated by the '+' signs; the number within brackets specifies the number of metrics in the index; the prefix gives information about the model format: R0.A = continuous scoring (0 classes) with respect to the reference distribution (R), average score model (A). (All the models are the same here, but see Figure 5.11 for another scoring and Figure 5.8 for logistic regression models). The suffixes specify for which criterion the model is optimal (F: aucF / P: aucP). Models with no suffix are competing models in the vicinity of the optimum (see text and Figure 5.10).

Model key	Metrics in model										aucF		aucP	
	nsBen	nsBra	piExo	piFlo	piMjm	piOmn	piPis	piSme	vaTol	vdDiv				
R0.A0005(1)F					+						0.881		0.770	
R0.A0002(1)P		+									0.864		0.782	
R0.A0018(2)F		+			+						0.914	↑↑	0.803	↑↑↑
R0.A0022(2)P		+				+					0.906	↑↑	0.844	↑↑↑
R0.A0064(3)F		+		+	+						0.930	↑	0.882	↑↑↑
R0.A0083(3)P		+			+		+				0.929	↑	0.904	↑↑↑
R0.A0219(4)		+		+	+			+			0.941	↑	0.918	↑
R0.A0238(4)FP		+			+		+	+			0.945	↑	0.931	↑
R0.A0420(5)FP		+		+	+	+		+			0.942	↓	0.952	↑
R0.A0439(5)		+			+	+	+	+			0.942	↓	0.950	↑
R0.A0496(5)				+	+	+		+	+		0.938	↓	0.914	↓
R0.A0511(5)					+	+	+	+	+		0.938	↓	0.923	↑
R0.A0764(6)F		+	+		+		+	+		+	0.933	↓	0.888	↓↓↓
R0.A0776(6)P			+		+	+	+	+		+	0.932	↓	0.905	↓↓↓

5.3.3.2. Metrics not included in the model

The metrics not included in the model are: nsBen, piExo and vdDiv. Although these metrics are not indicative for degradation, they can be of interest for other objectives, for instance, piExo for

monitoring exotic species. With respect to vdDiv, it is instructive to notice that none of the three diversity indices "survived". This is in agreement with some authors questioning the inconsistency of these concepts (Hurlbert, 1971). Yet, they remain popular because they are easy to calculate. However, this simplicity can be the reason for their weakness. The power of metrics is precisely due to augmenting the community data with knowledge about the ecological properties of the species. Another interesting feature results from the comparison of the two pairs of five metric models. The pairs have four metrics in common. Substituting nsBra with vaTol results in a model with metrics only depending on the abundance of the species with a small loss of diagnostic accuracy. Or, it allows to exclude nsBra because it is suspected to be linked with the salinity gradient. Thus, if we decide for some good ecological reason to exclude some metrics, it is possible to evaluate whether important information is lost.

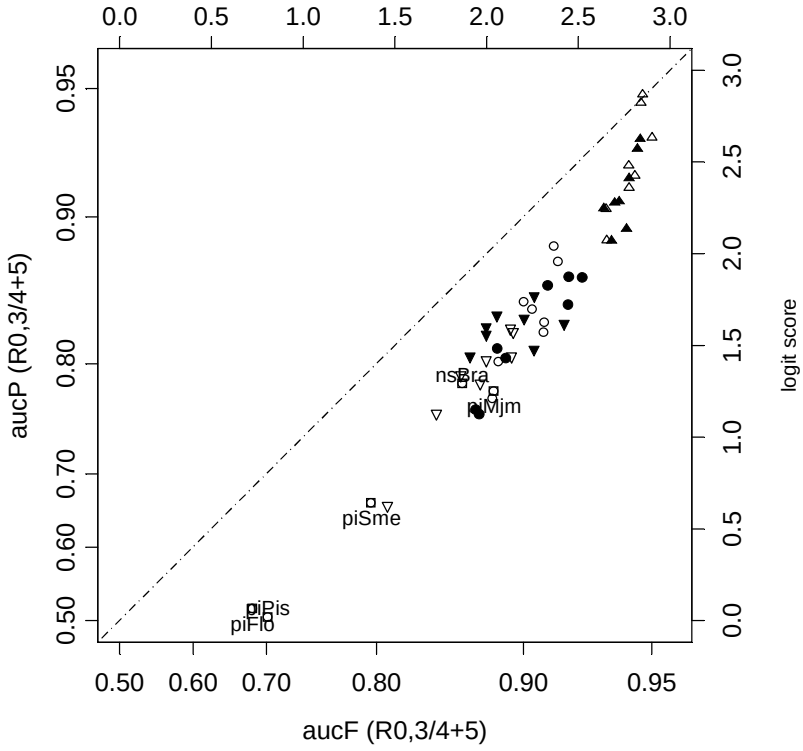


Figure 5.11 **The influence of scoring on the diagnostic accuracy.** Synoptic diagram for the seven best models (Table 5.5), but with metrics scored differently. Range of scoring: R = reference (white or open symbols), F = full distribution (black symbols). Scale of scoring: 0 = continuous (upper triangles), discrete with 5 (bullets) or 3 classes (lower triangle). With decreasing resolution going from continuous to discrete with 5 and 3 classes, the diagnostic accuracy of the IBI deteriorates. The metrics of the optimal models with four metrics are added as a reference. The logit-scale of the diagram log(AUC)/log(1-AUC) is chosen to have a better discrimination between the models. (Symbols: ▲ = F0, ▵ = R0, ● = F5, ○ = R5, ▾ = F3, ▿ = R3, □ = metric).

5.3.3.3. Sensitivity analysis with respect to the scoring

We refitted the seven best models with a different scoring to investigate its impact on the diagnostic accuracy. As Figure 5.11 makes clear, discretisation of the metrics results in a considerable loss of the diagnostic accuracy, especially of the classical trisection method is used.

5.3.4 Step 4: Tuning and validation with respect to the full gradient

5.3.4.1. Response to the gradient

Although we did not optimise directly for the full gradient of the habitat quality, the optimal model with four metrics shows a good response. There is nearly linear separation of the three degradation classes (Figure 5.12). The first decision threshold distinguishes class 3 and 4+5 at an FPF of about 20% (= FPF(3/4-5). The second threshold makes a further distinction between class 4 and 5, (again) at an FPF of 20% for class 4 now (=FPF(4/5)). With respect to this configuration, the confusion matrix was calculated (the percentages along the boxplots).

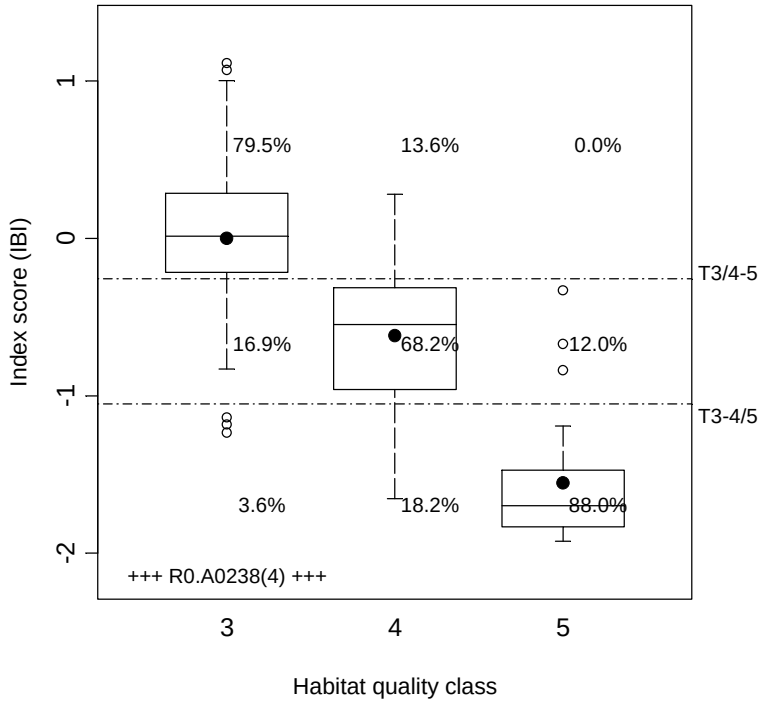


Figure 5.12 **Boxplots testing the response of the final model to the pressure gradient.** The model is R0.A0238(4) with nsBra, piMjm, piPis and piSme as metrics. Thresholds $T_{3/4-5}$ and $T_{3-4/5}$ are fixed to FPF = 20%. The percentages on the (negative) diagonal represent the true classification (e.g. 79.5% = TCF(3,3), the percentages above the diagonal are the false negatives (i.e. classification too mild, e.g. 12.0 % = FNF(5,4)), and the values below the diagonal the false positive fractions (classification too severe, e.g. 16.9 % = FPF(3,4)).

With respect to the main contrast (3/4-5), 16.9% of class 3 is misclassified in class 4 (= FPF(3,4)) and 3.6% in class 5 (= FPF(3,5)), a monotone decrease. The sensitivity for this first threshold is close to 93 % or only about 7% of the more degraded classes (4 or 5) has a value larger than this threshold. This is an average of 13.6% for class 4 (= FNF(4,3)) and 0% for class 5 (= FNF(5,3)), again a monotone decrease. With respect to the secondary threshold (4/5), the estimated sensitivity is 88% or 12% of the sites of class 5 are misclassified (= FNF(5,4)). However, as we signalled already in the previous paragraph, a few class 3 sites (3.6%) also have value below this threshold (= FPF(3,5)).

5.3.4.2. Confidence limits of the diagnostic accuracy measures

Figure 5.13 gives the bootstrapped bias-corrected confidence limits of the sensitivity for the main contrast (3/4-5) and, at the same time, we investigated the sensitivity at other settings of the FPF in the range from 10% to 30% and compared them with aucP and aucF. The sample is rather small (N = 130) and not optimally balanced (2/3 of the observations is in class 3 and about 1/6 in each of the other two classes). As a consequence, the confidence limits are wide. Yet some patterns appear.

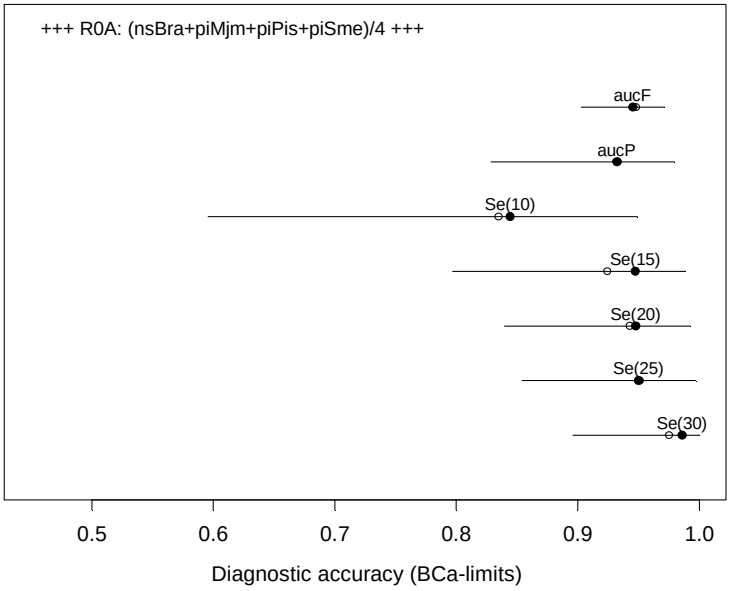


Figure 5.13 **Bootstrapped confidence limits for the diagnostic accuracy measures.** 90 % BCa limits (n_B = 1000) for model R0.A0238(4), one of the optimal models with four metrics. aucF, aucP and Se (TPF) at different settings of FPF in the range from 10% to 30%.

The lower limit for the sensitivity at an FPF of 20% is about 85%. This is what we can guarantee at a confidence level of 90%. Because of the noise on the data the evolution from FPF = 10% to 30% is not smooth and should be interpreted carefully. Yet, the broad confidence limit at an FPF of 10%

is an indication that beyond this value the sensitivity goes sharply down, a feature we also observed in the ROC curves of Figure 5.6. Conversely, at an FPF of 30% the sensitivity approaches unity. From this pattern, setting the threshold at a FPF of 20% seems a reasonable choice.

The confidence interval for aucP (over an FPF-range from 0.1 to 0.3) is very comparable to that of TPF(20) (the sensitivity at an of FPF of 20%) as could be expected from the theoretical example (Figure 3.3). It also seems to give a good summary for the range although the endpoints at 10% and 30% start to diverge. The confidence interval for aucF (the full area under the ROC curve) is about one half smaller, but aucF is less focused criterion, not having a clear interpretation except that it can be used as a global measure of the diagnostic accuracy. However, in combination with aucP, it is a powerful tool, aucF controlling the total shape of the ROC and aucP the more focused criterion.

5.3.4.3. Ordinal logistic regression

The regression coefficients of the proportional odds model are negative (Table 5.6) because, by convention, the underlying binary logistic regression models compare the most degraded sites with the less degraded ones (4-5 is contrasted with 3, and 5 is contrasted with 3-4). As the signs of the regression coefficients all have the same (negative) sign, the (scored) metrics are positively correlated with a high ecological condition. The coefficients are (in absolute value) very close to one. Nor the individual t-tests nor the likelihood ratio test setting all the coefficients equal to one, indicate the more simplified model fits the data worse. In this situation the average score model has a similar performance as the more flexible proportional odds model.

Table 5.6 **Comparison of the proportional odds model with the average score model.** The model considered is model R0.A0238(4) (one of the two best models with four metrics). The t-values in the right column test whether the coefficients significantly differ from one in a univariate way, while the deviance makes a global comparison between the two models.

	Proportional odds model			Testing for coefficients equal to one		
	Value	Std. Error	t-value	Value	Std. Error	t-value
piMjm	-1.339	0.318	-4.21	-0.339	0.318	-1.07
nsBra	-0.928	0.270	-3.43	0.072	0.270	0.27
piPis	-1.038	0.280	-3.71	-0.038	0.280	-0.14
piSme	-0.645	0.331	-1.95	0.355	0.331	1.07
Intercept 3/4-5	2.325	0.377	6.16			
Intercept 3-4/5	4.903	0.690	7.10			
Deviance		115			117 (all coefficients = 1)	

5.4 Discussion

5.4.1 The optimisation criteria based on the ROC curve

5.4.1.1. The relevance of ROC curve as an optimisation criterion

From a societal perspective, striving for an index with maximal diagnostic accuracy is a relevant objective. For the manager to take operational decisions about restoration as well as for the policy maker to choose between strategic options, it is very important to differentiate between false positive and false negative errors because these two errors imply a different ecological and societal cost. Although it is seldom possible to precisely quantify the costs and profits associated with FPs and TPs, rough paper-and-pencil calculations can already give an idea about the usefulness of the optimised index. As we discussed in Chapter 3, the true restoration fraction (TRF), the proportion of the budget used properly, depends strongly on the ratio of TPF and FPF. If we have to accept a high FPF to achieve a sufficiently high TPF, then a large part of the resources will be detracted for restoring sites in a (relatively) good condition. This is only acceptable if the expected ecological gain is sufficiently large and the society is willing to pay (WTP).

5.4.1.2. The full and partial area under the ROC curve

The ROC curve of an index shows the tradeoff between the false and true positive fraction. The curve represents the intrinsic discriminative capacity of the index (Zhou *et al.*, 2002). Hence, it is logical to optimize the ROC curve of the index. As it is not evident to optimize curves as a whole, it is common to take a summary characteristic. An often used criterion is the "full" area under the curve (aucF) ranging from 0.5 (an indifferent index) to 1 (the optimal index). This statistic can be interpreted as the average sensitivity (TPF) for all possible values of the FPF and gives an indication of the overall usefulness of the index. Although a valid criterion, aucF has the drawback it evaluates the diagnostic accuracy for the full FPF range. For practical applications, only a part of the curve is useful as for very small values of FPF the sensitivity is too small, and, very high values of FPF are equivalent to a low specificity.

A solution is to focus on a particular range of the curve, for instance for an FPF from 10 % to 30 %, with the so-called partial area under the ROC curve (aucP) (Dodd and Pepe, 2003). This measure can be understood as the average sensitivity in the range specified (Figure 5.1 and Figure 5.13). A technical advantage of the aucP statistic is that is has better statistical properties. It has a lower variability than the sensitivity at a fixed point and it is less influenced by particularities and/or discreteness of the data. In Figure 5.13, the confidence limits for Se(10)=TPF(10) are exceptionally large, possibly because of the specific configuration of the data collected.

To build the model, we found it advantageous to consider aucF and aucP simultaneously. The former statistic controls the quality of the curve as a whole, the latter gives information about the index in the lower region of the ROC curve giving a better focus. The evolution of the aucP statistic (Figure 5.6) gave a better picture how the introduction of additional metrics improved the strength of the index. However, by focusing the search for an optimal combination of metrics in a smaller region, the risk for both overfitting and "gerrymandering" (i.e. choosing the aucP window to

optimise the fit) increases. We can select a combination of metrics that performs well in a narrow range, but not outside. This was revealed informally with Figure 5.6 showing a divergence between the aucF and aucP statistic; while the maximum value of the former started to decrease from 4 metrics on, the latter still increased. The more formal confidence limits comparing the optimal models of Figure 5.7 confirmed that the model with five metrics was not significantly better than model with four metrics at a 10% significance level.

5.4.1.3. Cost considerations

We choose the aucP window from 10% to 30%. From our FAME experience (Quataert *et al.*, 2007), we know that controlling FPF below 10 % is seldom achieved without sacrificing the sensitivity. A realistic setting is to aim for a FPF of about 20 % and we choose a symmetric interval of ±10 %. This resulted in an upper value of 30 % which can be defensible if the gain of restoration is high in comparison to the risk and costs of unnecessary restoration. A possible extension of the strategy could be to take three optimisation windows for aucP and to choose smaller windows (e.g. 5 – 15%, 15 – 25% and 25 – 35%) to perform a sensitivity analysis. Our main message is that working with a complementary set of windows is to be preferred in contrast to optimisation at a single FPF value (Figure 5.13).

A further extension of the approach is to consider the utility functions introduced in Chapter 3 and further explored in Chapter 4. If information is available about the costs, we can directly optimise the utility function. In this respect, an interesting idea could be to include the cost of the metrics. For this Zeeschelde case study, there is no additional cost for a metric as they are all derived from the same community data. However, to decide about the taxonomic resolution (e.i. determination on a species, genus or still higher level) or how many taxa to combine (e.g. fish, macroinvertebrates, macrophytes and/or diatoms), a full cost approach can be helpful.

5.4.1.4. Synoptic plots

The synoptic plots of aucP or aucF allow to graphically investigate the diagnostic accuracy for multiple binary contrasts (e.g. 3/4 and 3/5) extending the application of ROC curves from binary to ordinal variables. We used synoptic plots to exclude non-informative metrics and/or to select the core metrics (Figure 5.5), to explore competing multi-metric indices in the vicinity of the optimum (Figure 5.10), and, to evaluate the effect of the metric scoring (Figure 5.11). Figure 5.5 was also instructive to understand why the metrics are chosen (the black points). The first two metrics are also the best ones individually. However the metrics entered next (piPis or piFlo) are rather weak, but in combination with the metrics in the model, they are more complementary than the others, possibly because their discrimination is high for the contrast between the third and the fifth class.

5.4.1.5. Trend of aucF and aucP as a function of the number of metrics

To find the optimal number of metrics (Figure 5.6), the combined use of aucF and aucP gave good results. The aucP criterion was the more sensitive and increased till five metrics, while, with aucF, the increase was less pronounced and stopped one step earlier. Investigating this discrepancy (Figure 5.9) revealed that indeed the ROC curve slightly improved in the FPF range from 10% to

30% when going from four to five metrics, but that there were signs of overfitting in the range from 0% to 10%. By observing both the evolution of aucP and the more focused aucF, it is possible to obtain a picture with a higher resolution.

5.4.1.6. The optimisation with respect to the baseline

We primarily optimised with respect to this first binary contrast (3/4-5). In the final stage, the response to the full gradient was checked and turned out to be close to linear (Figure 5.12). This simple relationship lead us to try out an proportional odds model with all regression coefficients set equal to one. This simplified model was not significantly different from the full model (Table 5.6), indicating the average score model is acceptable.

5.4.2 The statistical model building

5.4.2.1. The general model building strategy

We inspired our strategy to construct the index mainly on Hosmer & Lemeshow (2000). These authors stress their method is not to be used as a substitute for, but rather as an addition to, clear and careful thought, taking into account the constraints of the available data. In their opinion, successful modelling of a complex data set is part science, part statistical methods, and part experience and common sense, an idea expressed by many other model builders, e.g. McCullagh & Nelder (1989). The latter emphasize data rather seldom points unequivocally to one single model and one should acknowledge this uncertainty and unravel its consequences. Further exploring the environment of the optimum is an occasion to learn more from the data (Kutner *et al.*, 2005). All these considerations forged our four step approach: (1) univariable screening, (2) sorting out the optimal number of metrics in the index, (3) exploration of the vicinity of the optimum and (4) final tuning and validation of the model.

5.4.2.2. Step 1: the univariable screening and exploratory data analysis

Exploratory data analysis (EDA) is always the first step of any statistical analysis (Hoaglin *et al.*, 1983). EDA is necessary to control and improve the data quality (outliers, ill defined variables, …), to get a feeling with the data (the main patterns, its particularities, its anomalies) as a safety net to not overinterpret the data, and to synthesize and to document the data for later reference. The result of the EDA step should be an improved dataset ready for statistical analysis and a first elementary analysis of the data. For instance, in the case study, to avoid numerical instability, we could directly eliminate two ill-defined metrics of the original study (Breine *et al.*, 2007) because they had too few distinct values.

More specifically, for model building, the univariable screening of the individual predictors for their relation with the response variable is part of EDA. Commonly, there are (too) many candidate predictors and a first selection of predictors can reduce the workload considerably. The problem is that some predictors turn out to have an impact only once other variables are included in the model. Therefore, Hosmer & Lemeshow (2000) propose to keep any variable in the starting model having a correlation with the response variable with a p-value > 0.25. Therefore, we only dropped

the four metrics not showing any response at all and we retained other metrics for the next steps of the modelling. This was indeed necessary as one of two rather weak metrics (piPis or piFlo, they are interchangeable) was accepted in the model as the third metric (Figure 5.5).

As a general remark, it is important to realise that the parsimony of the model should start when composing the basket with the candidate metrics. Each candidate should be well founded. It does not make much sense to try out blind attempts or many variations, without a clear motivation or hypothesis in mind. Rather seldom ecological data is sufficiently strong to compare subtle variations, and, if a weakly motivated metric turns out to be highly significant, there is no rationale to interpret the result and to discriminate it from an artefact (Rothman, 1990). This is not to promote overly simple starting models, but we should actively seek for a balance (Myung, 2000).

5.4.2.3. Step 2: the assessment of the optimal number of metrics

Overfitting is an important problem in model building (Hosmer and Lemeshow, 2000; Harrell *et al.*, 1996; Claeskens and Hjort, 2008). Overfitting occurs when too many predictors are entered in the model in comparison to the available data. The fit seems good, however the predictive quality of the model is lower than a more parsimonious counterpart, because the model fits noise and/or the particularities of the sample instead of the underlying process generating the data. Therefore we first investigated the trend of the diagnostic accuracy of all possible subsets as a function of the number of metrics to assess the optimal balance between complexity and prediction quality. Bootstrapping was used to get unbiased estimates of the diagnostic accuracy.

The number of metrics is an important issue in metric selection. Most index builders are aware they should avoid redundancy in the metrics, but sometimes they argue for a limited redundancy to increase the robustness of the index (without defining this concept sharply) and to improve the ecological interpretation. However, Figure 5.6 illustrates that adding too much metrics can be harmful defining five metrics as the upper limit. Multiple comparisons based on bootstrapping (Figure 5.7) indicate that the most parsimonious model has three metrics, defining the under limit. Based on a more detailed investigation we tend to favour a model with four metrics. From five metrics on the ROC curve starts to deteriorate in its lower part (Figure 5.9).

5.4.2.4. Step 3: exploration in the vicinity of the optimum

It is possible that many models with different ecological implications are compatible with the data (McCullagh and Nelder, 1989). Moreover, the optimal model can be just (bad?) luck (Zucchini, 2000) not representing the underlying processes. For both reasons, we look at competing models in the neighbourhood of the best model to evaluate their differences and ecological implications. This diagnosis can contribute to a better understanding of the underlying process and reveal particularities of the dataset. In our case, looking at the competing models revealed that piPis and piFlo were to a large extent interchangeable and with this information it became clear that the consecutive models are nested. First the two most important metrics (nsBra and piMjm) are entered in the model, then piPis or piFlo, followed by piSme and finally piOmn. Also we detected an alternative close to optimal variant with metrics only based on the percentage of individuals (pi). This can be interesting if we prefer to exclude the metrics only based on species counts.

5.4.2.5. The validation step and tuning for the full gradient

The optimisation does not take into account the full gradient of degradation and concentrates on the most important contrast with the baseline. An advantage of this approach is that the optimisation does not depend on the division between class 4 and 5. For the preclassification, it is sometimes more easy to define the baseline (absence of any human impact), than to rank the sites with respect to disturbance because this implies combination of multiple pressures and a good knowledge of the cause effect relations and possible interactions. Also, as the full gradient is not used to optimise the model, we can use the extra information as a kind of validation of the model.

Although we did not optimise directly for the full gradient, the optimal four metric index has a nearly linear response to the three degradation classes (Figure 5.12). Based on bootstrapping, we can guarantee a sensitivity with respect to main contrast of at least 85%. Misclassification with a difference of two classes is small (FPF(3,5) = 3.6% and FNF(5,3) = 0.0%). Also for class 4 the false negative fraction remains below 15% (FNF(4,3) = 13.6%). Thus the optimisation procedure seemed to be successful. Within the data constraints, this is probably the best we can obtain.

5.4.3 The model format

5.4.3.1. Sensitivity analysis with respect to the scoring

We scored the metrics with respect to the least impact sites as advocated by Seegert (2000). Baseline scores are directly interpretable as standardised "distances to target" and the standardisation makes the (implicit) assumption of equal weights of the AVG model plausible. Yet, there exist many other scoring systems and many index builders prefer a scoring in classes to reduce the data uncertainty and variability in the data inspired on the original trisection method of Karr (1981). In his approach, the entire distribution of metrics is subdivided in three equal sections with the scores 5 (high quality), 3 (moderate quality) and 1 (low quality). Although this practice has its value, discretisation often implies a loss of information.

To our knowledge, there is little guidance and recommendations are not thoroughly investigated nor quantified except for a small study of Blocksom (2003) who proposes a scoring typology on two axes: the range of metric values used to score (derived from the reference or entire distribution) and the scale of scoring (continuous or discrete). She recommends a continuous scoring with respect to the entire or full distribution. Our results confirm the first part, but not the second part. Scoring with respect to the full distribution (F0) did not improve model performance (Figure 5.11). More pronounced, the impact of discrete scoring (5 and 3 classes) is negative for the diagnostic accuracy (Figure 5.11).

However, we should be careful before extrapolating the result blindly. Scoring in classes can have advantages (for instance to model nonlinear relationships) and sometimes it is the only information available when the assessment on the field is directly in classes. From this perspective, we can interpret the results also positive. For instance, with a scoring in three classes, in spite of the simplification, we still can obtain a sensitivity of about 90% with the best model.

5.4.3.2. The proportional odds model

We did not take into account the full gradient of degradation. This can be felt as suboptimal, and, indeed models exist directly taking into account the ordinal classification. As an example, we fitted the proportional odds model to the data, resulting in a regression model that was very close to the average score model. The regression coefficients are not significantly different from one (nor based on the individual t-tests nor by comparing the deviance setting all the coefficients equal to one at once). The proportional odds model and other ordinal regression models are more flexible and have far greater potential. The average score model is simple but has few degrees of freedom. Therefore, it is at least interesting to test, as we have done here, not a better model is possible by estimating the weights of the metrics. With a regression model we can investigate more complex nonlinear effects of the metrics, add other extra environmental variables and, not at least, incorporate information about the sampling design. Also the opposite is true. The fact that we can fit the proportional odds model to the data, leads to the interesting interpretation that there is an underlying (latent) variable representing the increasing degradation shifting the (cumulative) distribution downwards.

5.4.4 Bootstrapping and the validation of the index

5.4.4.1. Bootstrapping and measures of diagnostic accuracy

Bootstrapping is an ideal instrument to estimate standard errors and confidence limits if the theoretical distribution of the statistic is not well known or the assumptions are not fulfilled. Less known is that bootstrapping also can be used to estimate bias and to correct for it. Measures of diagnostic accuracy are bias prone because we use the same data for modelling and evaluation. As a consequence, the sample and the model are closer to each other than will be true with future samples. For this reason, we used the BCa (bias-corrected, accelerated percentile) method to estimate the confidence limits as developed by Efron (1987). In principle this method is very simple. From the sample we calculate a measure of diagnostic accuracy, say aucP. Next, with resampling from the observed distribution (i.e. the sample), we reconstruct the sampling distribution of aucP. The difference between observed aucP and the average of the resampled distribution is an estimate of the bias. After adjusting for bias, we can use the percentiles of the distribution to estimate the confidence limits (Figure 5.2). For our application, bias was minimal, but still the method was useful to calculate the confidence limits. Possibly, the reason for the smaller bias is that AVG model is rather inflexible in contrast to a logistic regression which can be fit more closely to the data.

5.4.4.2. Bootstrapping and model selection

It is less known that bootstrapping can also be used for model selection and validation. A common practice is split-sample validation, in which one part of the dataset (typically one third) is put aside to validate the model. However, this is rather "inefficient" unless the database is very large. Only two thirds of the data are used to construct the model, and only one third is available for validation. Steyerberg *et al.* (2001) found that, based on a simulation study sampling from a large realistic dataset, split-sample validation is indeed inefficient, while bootstrapping provided stable

estimates with low bias. Hence, he recommends bootstrapping for estimation of internal validity of predictive logistic regression models.

5.4.4.3. Internal and external validation. The role of follow-up studies.

Bootstrapping is an internal validation procedure: a wrong context, a flaw in the study design or bad data collection cannot be corrected for. In fact, for bootstrapping, the sample is assumed to be the best possible representation of the entire population. Hence, nearly by definition, the resampling technique has the same flaws as the original study. In practice, the diagnostic accuracy of the index will be lower than predicted. For instance, Paul *et al.* (2001) reported that in a follow-up study the sensitivity of their index decreased from 90% to 80%. External validation is necessary in follow-up studies (Hosmer and Lemeshow, 2000) to find out whether there are no fundamental design mistakes and to refine the model. Perhaps most crucially, the construction of an IBI strongly depends on the preclassification of the sites. Ideally, it should be a gold standard, i.e. measured with a high precision, say one order of magnitude higher than the index. In practice, this is seldom realised and there will be "imperfect gold standard" bias (Zhou *et al.*, 2002). There exist methods to acknowledge this extra source of uncertainty (Hui and Walter, 1980), but the real solution is to set up well-designed follow-up studies to refine the index by controlling its performance in different situations and alternative definitions of degradation.

5.4.5 Redundancy, also a design problem

As described in the introduction, it is common practice is to include more metrics than strictly necessary for the diagnostic accuracy. Some authors signal that extra metrics did not improve the diagnostic accuracy (Alden *et al.*, 2002) or even resulted in a decrease (Roth *et al.*, 1998). Some redundancy in the metric combination is encouraged to increase the robustness and scope of the index. Although these arguments are legitimate from an ecological and practical perspective, the very question is whether this is good scientific practice. Why should we believe a theoretical construct, if the empirical analysis suggests that the index does not improve or even gets worse? In the subsections to follow, we explore two alternatives.

5.4.5.1. Coping with spectrum bias with probability-based sampling

A first series of comments has to do with the design for the index construction. An IBI has the potential and the ambition to cover a broad range of impacts on the ecosystem. If, however, the sample available for the index construction is not representative for the impacts in the region and does not cover the full gradient of pressures, there is little hope the biotic index will fulfil its aims. It is unrealistic to hope that the relevant metrics sensitive to the pressures not present in the dataset will be selected. Therefore it is necessary that the data collection covers the full spectrum of impacts at least in the region where the index will be applied. Then, it is more plausible that the model building will select the relevant metrics increasing the scope of the index.

In this respect, an ongoing discussion is whether the sample should be probability-based. An evaluation of Fore (2003) showed that a probability sample was superior to cover a broader spectrum of pressures in contrast to the intuition of the researchers involved in the project. A

probability sample is not only a way to get an unbiased picture of the ecological condition, but also of the relations between variables (Kish, 1987). We acknowledge in practice a random sample is hard to achieve and is not cheap. However, in the long run the investment pays back as argued by many authors (Dauer and Llansó, 2003; Hughes *et al.*, 2000; Llansó *et al.*, 2003; Overton, 1993; Paul *et al.*, 2008; Southerland *et al.*, 2009).

Evidently, also all other aspects of the design are important to improve the metric selection including the preclassification (its problems to achieve a gold standard already discussed above), the determination of reference conditions, the theoretical foundation of the metrics, the sample size, and so on. Considering these points, Southerland *et al.* (2007) were capable to improve substantially the Maryland fish and benthic macroinvertebrate index.

Our argumentation is very close to the concept of spectrum bias in diagnostic medicine (Ransohoff and Feinstein, 1978). If the sites included in the calibration dataset are not representative for the regional distribution, the index will not be as effective as found in the calibration set. In addition, the risk is that not the ideal combination of metrics will be selected.

5.4.5.2. Better separation of diagnosis and causal analysis

Another proposal is to separate the assessment of the level of human impact (diagnostic modelling) and the investigation of the causes of impairment (causal analysis). Now, many authors give (explicitly or implicitly) two functions to an IBI: signalling whether something is wrong, and, judging the cause from the metrics. Our results confirmed by other research (Roth *et al.*, 1998; Paul *et al.*, 2001) suggest that a limited number of metrics can be sufficient to detect impairment. For a causal analysis additional information is necessary including extra metrics, species level information (e.g. which characteristic species are missing, having a special relationship with environmental pollution), concentration of toxic substances, physical and chemical characteristics of the environment together with their ecologically relevant threshold values, and indications of habitat characteristics at local or landscape level. Our suggestion is that it is a more cost-effective strategy to distinguish both functions and optimize each basket of indicators separately. Three consecutive books spanning one decennium authored by Simon (Davis and Simon, 1995; Simon, 1999; Simon, 2003) seem to confirm this trend. Whereas the first two books primarily deal with signalling the problem, the last book is devoted mainly to causal analysis.

5.5 Conclusion

An important bottleneck of the construction of IBIs remains the selection of an optimal subset of metrics from the candidate set. Two factors are responsible for this situation: clear and transparent criteria are missing and it is not recognised an IBI is a regression model. In fact, we demonstrated the traditional average score model is a special case of a generalised linear regression model. Because of this link, we can use statistical model building strategies and techniques as bootstrapping to optimise an IBI. By combining the ROC curve concept and statistical model building techniques, we were able to develop a coherent strategy to retrieve the optimal number of metrics and the best possible metric sets (plural!). By doing so, index construction has become a less isolated technique and can be framed in a broader theoretical perspective.

6 Evaluation of the European Fish Index (EFI). The false positive fraction and false negative fraction to detect disturbance and consistency with alternative fish indices

This chapter is an adaptation of a paper published in a special issue of Fisheries Management and Ecology about the Fame project (Schmutz et al.*, 2007b) under the title "Evaluation of the European Fish Index: false-positive and false-negative error rate to detect disturbance and consistency with alternative fish indices" (Quataert* et al.*, 2007).*

Abstract

An important requirement for meeting obligations under the European Water Framework Directive (WFD) is the development of a fish-based index (FI) that is able to predict the ecological status of surface waters, and particularly to distinguish between (nearly) pristine and disturbed conditions. For this purpose, the EU funded FAME project (Development, evaluation and implementation of a standardised fish-based assessment method for the ecological status of European rivers) developed the European Fish Index (EFI) based on the concept of the Index of Biological Integrity (IBI), alongside alternative models such as the Spatially Based Method on a European level (SBM-EU). A critical issue about these models is that they are simple to use but nevertheless are able to predict whether a site is disturbed with a high degree of precision. From a decision making perspective, two prediction errors need to be small: falsely declaring a site disturbed when it is not (false positive error; FP) and wrongly classifying a disturbed site as undisturbed (false negative error, FN). For the EFI, the overall FPF was 22 % and the FNF 19 %. The performance was better for the SBM-EU method with a smaller FPF of 7 % and an FNF of 20 %. However EFI is preferred because, with only marginal loss of precision, it is far less complex. EFI consists of a single model based on ten fish metrics, while SBM-EU comprises 12 models covering 49 metrics. Comparison of EFI and SBM-EU with existing national or regional fish-based assessment methods (Ex-M) revealed major discrepancies making intercalibration between these methods infeasible.

Keywords

error curve; false negative error; false positive error; fish index; precision.

6.1 Introduction

An important requirement for meeting obligations under the European Water Framework Directive (WFD; EU 2000) is the development of a fish-based index (FI) that is responsive to human pressures and is able to predict the ecological status of surface waters, but particularly to distinguish between (nearly) pristine and disturbed conditions. Such an index, the European Fish Index (EFI), based on the concept of the Index of Biological Integrity (Pont *et al.*, 2007) was developed under the EU, FAME (Fish-based Assessment Method for the Ecological Status of European Rivers) project (see www.fame.boku.ac.at) to classify rivers. The EFI was developed alongside the Spatially Based Method on a European level (SBM-EU) (Melcher *et al.*, 2007). Also, existing nationally and regionally methods (ExM) were available for comparison.

A critical issue with these models is that they are simple to use but are able to predict whether a site is disturbed with a high degree of precision. From this perspective, two prediction errors need to be small: falsely declaring a site disturbed when it is not (false positive error; FP) and wrongly classifying a disturbed site as undisturbed (false negative error, FN). The aim of this paper was to evaluate whether the EFI and SBM-EU models respond effectively to a 5-tiered classification system, the Pressure Status (PS), derived from hydromorphological and physico-chemical pressures of anthropogenic origin ranging from very low (1) to very high (5) impact (Degerman *et al.*, 2007). The paper also compares the outputs of the EFI and SBM-EU indices with those derived from national and regional methods, and examines how successful EFI and SBM-EU are to predict PS in terms of false positive and false negative misclassification errors.

6.2 Material and methods

6.2.1 The use of the pressure status as a common reference stick

To examine the precision of the various indices to detect disturbance, PS was chosen as a proxy for ecological quality. The underlying assumption is that high ecological quality at a site is achieved when there is no or little evidence of hydromorphological and physico-chemical pressures of anthropogenic origin. Such a site is termed pristine (or nearly pristine) or reference and is given a PS score of 1-2. According to the WFD, a fish index should be able to discriminate between (nearly) pristine (PS 1-2) and impacted sites (PS 3-5), or more specifically in terms of ecological status, to separate "high/good" from "moderate/poor/bad". In addition, PS ranks the sites from good to bad ecological quality and the fish index should be responsive to this gradient (second level of detail).

6.2.2 Contrast between 1-2 (undisturbed) and 3-5 (disturbed)

6.2.2.1. False positives and false negatives

Two misclassification errors are possible with the outputs of a fish index: a false positive (FP) error that classifies a reference site as disturbed, i.e. PS = 1-2 and FI = 3-5; and a false negative (FN) error that classifies a disturbed site as undisturbed, i.e. PS = 3-5 and FI = 1-2. The proportion or fraction of both errors should be small. A large false positive fraction (FPF), will cause too many false alarms, and resources may be used to recover a site when they are not necessary or even

harmful. With a high false negative fraction (FNF), too many disturbed sites will be left unnoticed, and the necessary measures will not be taken which implies a large ecological cost. When evaluating a fish index, one should look at both errors and assess the associated costs (Field *et al.*, 2004) to get an overall picture. If no information is available about these (relative) costs or there is no management decision about the relative importance of both errors, as is the case here, an option is to declare both errors of equal importance and require them to be more or less equal or balanced.

Misclassification fractions are calculated following the procedure outlined in by Motulsky (1995) (Table 6.1); each cell contains the number of the sites from cross-classification of the pressure status (PS) and the fish index being tested (FI), after regrouping the five classes into undisturbed (1-2) and disturbed (3-5) classes only. The proportion of FP for PS = 1-2 estimates FPF and the proportion of FN for PS = 3-5 equals FNF. The difference between the fractions for the different fish indices was tested statistically based on the chi-square test for independence (Fleiss, 2003). Confidence limits for these fractions were calculated based on the binomial distribution (Fleiss, 2003) to evaluate the magnitude of the difference between the fractions (Chambers *et al.*, 1983).

Table 6.1 **Formulas to estimate the false positive or false negative (error) fraction.** The cross-classification of the pressure status (PS) and a fish index (FI) gives the number N of the matches (true positives and negatives) and mismatches (false positives and negatives)

(Reference Index)	Fish Index (FI) = 1-2	Fish Index (FI) = 3-5	Estimated Error
Pressure Status (PS) = 1-2	N_{TN} = # TN	N_{FP} = # FP	FPF = N_{FP} / (N_{TN}+N_{FP})
Pressure Status (PS) = 3-5	N_{FN} = # FN	N_{TP} = # TP	FNF = N_{FN} / (N_{TP}+N_{FN})

6.2.2.2. The analogy with a medical laboratory test

To explain the procedure an analogy is made with a medical laboratory test based on a continuous measurement of a variable indicative of the presence of a disease (e.g., a substance in the blood). If the value of that indicator (or index score) is below (or above depending on the variable measured) a certain threshold, then the patient is declared to have the disease. As a result of biological variation between people, a person can have a low score without being ill (a FP) or a person with the disease can have a score higher than the threshold (a FN). The upper part of Figure 6.1 shows the situation for two laboratory tests, A and B, designed to detect the same disease, but based on a different variable or method. Because the impact of the disease is larger for the variable used by B, the overlap between the density curves (upper part of Figure 6.1) is smaller and, as a consequence, laboratory test B is better. The shaded area under the density curves equals the misclassification error (left for the FP and right for the FN). Since it is difficult to assess the area under the curve by eye, cumulative curves are generated that specify directly the proportion of the population below or above the threshold (lower part of Figure 6.1) (Sokal and Rohlf, 1995).

6.2.2.3. The negative relation between the FPF and the FNF

It is important to recognise that FP and FN are negatively linked. Increasing one error, will decrease the other. In Figure 6.1, the thresholds are chosen such that FPF is 10 % , which fixes the FNF for A and B at 41.4% and 4.3%. This setting corresponds with the points a' and b' in Figure 6.2. Alternatively, one can balance the errors (a & b) or fix FNF to 10 % (a" and b") and the curves show all possible combinations for each test. Typically these "error" curves are hyperbolic (Field *et al.*, 2004) and the practical implication is that decreasing one error too much, can be at the expense of the other. Hence, balance between both errors is important and this will be an important criterion to evaluate the existing indices.

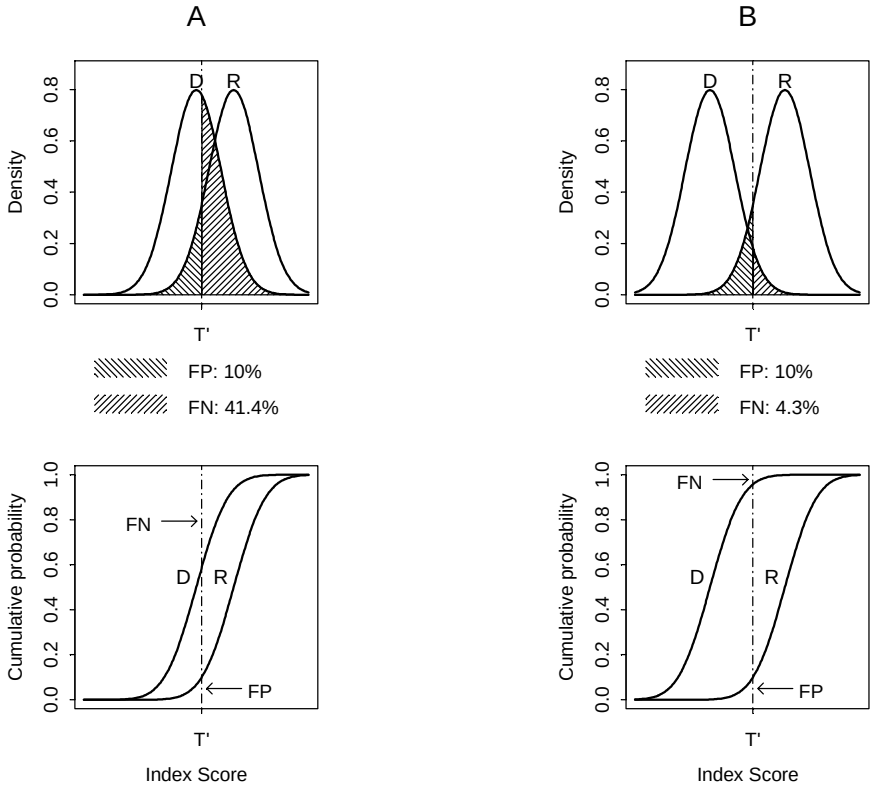


Figure 6.1 **Response of an index to human impact.** Shift in the distribution (R = Reference, D = degraded) of the index score for two hypothetical laboratory tests or fish indices (A and B). The shaded areas in the density function (top) represent the error fractions with respect to the threshold T' (chosen with a FPF of 10 %). The cumulative distributions (integrals of the density) allow the error fractions to be read directly from the graph.

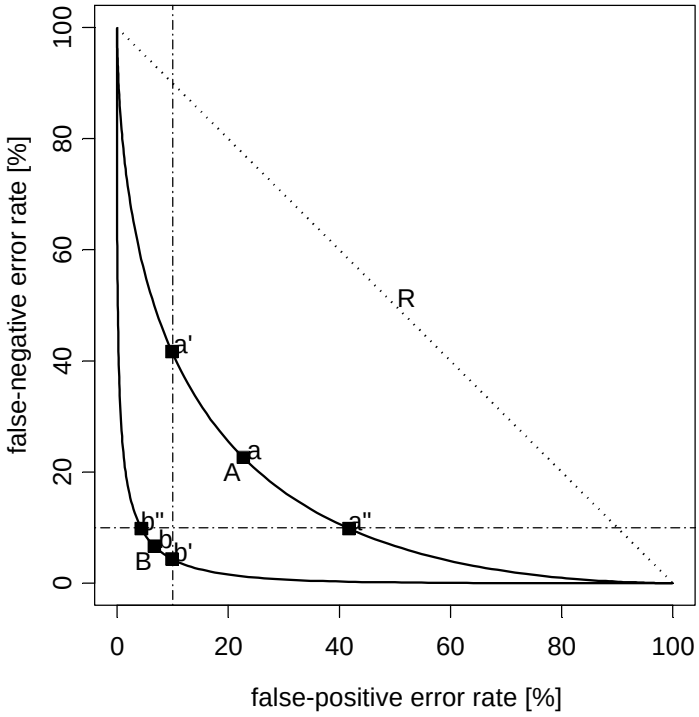


Figure 6.2 **Error curves** corresponding to Figure 6.1 showing the tradeoff between FPF and FNF. For an explanation of the points on the curves, see text (e.g. points a' and b' correspond to a fixed FPF of 10 %; the threshold T' in Figure 6.1). R is the error curve if the index is not influenced by disturbance (no discrimination possible).

6.2.3 Investigation of the gradient from class 1 to 5

6.2.3.1. Extension to a gradient of disturbance

Figure 6.1 shows how the distribution of an index score is shifted to the left because of disturbance. We can extend easily this idea to a gradient of disturbance. Figure 6.3 gives the situation for three classes and illustrates how the density and cumulative distribution of the index scores "respond" to an increasing impact from reference R (no or small impact, undisturbed situation; categories 1-2), over moderate M (category 3) to disturbed D (high or very high impact; categories 4-5). The ranking of the density curves conditional on PS is consistent with the ranking of PS and the higher the pressure, the more the curves are apart. In practice, these cumulative curves can be estimated from the data by the empirical distribution function (EDF); i.e. the proportion of the index scores below a certain value as a function of that value (D'Agostino and Stephens, 1986). This offers a graphical tool to control to what extent a continuous scoring system is consistent with another index. To test whether curves differ significantly, the (two sample) Kolmogorov-Smirnov test is used (D'Agostino and Stephens, 1986).

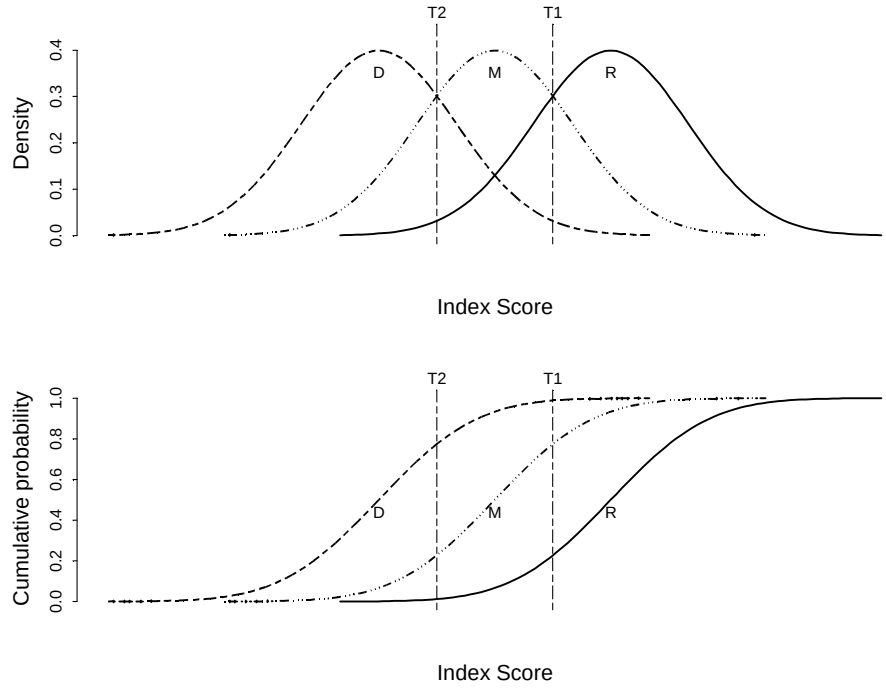


Figure 6.3 **Consistency of the index to an increasing gradient of human impact.** Graphical exploration of the ranking of the distributions to test if the (fish) index is responsive to increasing pressure (R = reference, M = moderately disturbed, D = highly disturbed) or is consistent with another index (see text). The vertical lines T1 and T2 represent the two decision or detection thresholds.

6.2.3.2. Extension of FPF and FNF with respect to multiple thresholds

The two vertical lines in Figure 6.3 represent the detection thresholds: T1 distinguishes between reference or not (i.e. the main contrast 1-2/3-5); T2 separates high to moderate from poor to bad (1-3/4-5). As before, the cumulative curves enable estimation of the misclassification errors directly, but a distinction between small and severe FP or FN errors can be made. For instance, about 20 % of the reference sites (R) lie below T1 (a small FP error), but only a small proportion of these reference sites is found below T2, implying that few of the high quality sites are severely misclassified. As in the more simple case with two disturbance classes, changing the threshold cannot minimise FP and FN at the same time and a compromise is necessary.

6.2.3.3. The continuous index score behind EFI

The methodology outlined can be applied to the evaluation of the EFI because it is based on a continuous score. To get an index on a five-point scale, as required by the WFD, four thresholds (at 0.67, 0.45, 0.28 and 0.19) were chosen for optimal discrimination with the pressure status: between 1 and 0.67 the EFI class is 1, between 0.67 and 0.42 it is 2, and so on (Pont *et al.*, 2007).

Table 6.2 **Overview of the existing national or local fishing indices** (ExM)**.** N = number of sites where all the indices are available, split in reference and disturbed sites as determined by PS (1-2 / 3-5). The percentages in the columns (between brackets) specify the relative contribution of each region to the total of the column (bottom line). The percentage in the bottom line are calculated with respect to the total number.

Country	Label	N	N reference (PS = 1-2)	N disturbed (PS = 3-5)	Short characterization (Reference)
Austria	ATm	177 (3.4%)	130 (3.7%)	47 (2.8%)	Mulfa method: expert judgement: (Schmutz *et al.*, 2000)
	ATn	80 (1.5%)	75 (2.1%)	5 (0.3%)	National method: expert judgement: (Haunschmid *et al.*, 2006)
Belgium Flanders	BF	796 (15.2%)	0 (0%)	796 (46.7%)	Adapted from Karr et al. (1986): (Breine *et al.*, 2004; Belpaire *et al.*, 2000)
Belgium Wallonia	BW	94 (1.8%)	83 (2.4%)	11 (0.6%)	Adapted from Karr et al. (1986): (Goffaux *et al.*, 2001; Kestemont *et al.*, 2000)
France	FR	1547 (29.6%)	869 (24.7%)	678 (39.8%)	Adapted from Karr et al. (1986), similar to EFI: (Oberdorff and Hughes, 1992; Oberdorff *et al.*, 2002)
Lithuania	LT	193 (3.7%)	165 (4.7%)	28 (1.6%)	Adapted from (Karr *et al.*, 1986): (Kesminas and Virbickas, 2000)
Sweden	SE	2144 (41.0%)	2031 (57.7%)	113 (6.6%)	Swedish Electric fishing RegiSter; SERS): (Appelberg *et al.*, 2000)
U.K.	UKs	194 (3.7%)	167 (4.7%)	27 (1.6%)	Two Welsh salmonid indices based on the presence of Salmon or Trout: (Strange *et al.*, 1989)
TOTAL		5225	3520 (67.4%)	1705 (32.6%)	

6.2.4 The dataset and indices

6.2.4.1. The FIDES database

For the FAME project, the consortium compiled the common database FIDES (Fish Database of European Streams) from existing electric fishing and environmental data complying to a minimal set of quality criteria mainly related to the fish sampling (Beier *et al.*, 2007). From FIDES, a total of 5225 sites were available to compare the pressure status (PS) with the predicted ecological status from the three fish indices: EFI, SBM-EU and ExM (Table 6.2).

6.2.4.2. The Pressure Status (PS)

Based on the environmental variables in the dataset, it was possible to assess the pressure status (PS) by a scoring system integrating hydromorphological and physico-chemical anthropogenic pressures into one single preclassification representing the overall pressure (Degerman *et al.*,

2007). This was a hard exercise, because it was not evident to find a common denominator of the pressure information for all participating countries. About two thirds (67%) of these sites were classified reference sites (PS = 1-2).

It should be recognised that large imbalances exist in the volume of data between regions and between disturbed and reference sites (Table 6.2) possibly decreasing the efficiency of the tests or the precision of the estimates. For instance, about 60 % of reference sites came from Sweden, and 25 % from France, but none were from Flanders (Belgium): in comparison to European standards, reference sites seem to be rare in Flanders, which is not unexpected because it is a densely populated region. Conversely, > 85 % of the disturbed sites came from Flanders and France, whilst other regions have few such sites, e.g. only 5 sites for the national method of Austria (ATn) and 11 sites for Wallonia in Belgium (BW).

6.2.4.3. The two FAME models: EFI & SMB-EU

Both the EFI and SBM-EU model were developed within FAME. As explained in Chapter 2, for the scoring, it is crucial to control for the natural variability (Oberdorff *et al.*, 2002; Irz *et al.*, 2008). With EFI, this is realised by developing global regression models on a pan-European scale predicting the expected value of the metrics given the environmental conditions (Pont *et al.*, 2007). The scores are the p-values of the observed metric values with respect to this model. In contrast, the SMB-EU model first classifies a site in a fish type based an a discriminant analysis model and then scores the metric values with respect to the type-specific average (Melcher *et al.*, 2007). As a consequence, the SBM-EU model is much more complex than EFI. It comprises 12 models covering 40 metrics, while EFI consists of one single model based on ten metrics.

6.2.4.4. The existing national or regional fish indices

The existing national and regional fish indices (ExM) are all published, but are very different in nature (Table 6.2). Three countries (Belgium, France and Lithuania) are a variant of the method of Karr (Karr *et al.*, 1986). In addition, the method in France was further extended for EFI (Oberdorff *et al.*, 2002; Pont *et al.*, 2006). The MuLFA method (multi-level concept for fish-based assessment) in Austria also uses the concept of the Reference Condition, but the reference situation is derived from an expert judgement compilation of historical data, data of reference sites and reference models (Schmutz *et al.*, 2000). The same holds for the alternative Austrian fish index (Haunschmid *et al.*, 2006). Interestingly, the MuLFA index has an explicit ecological rationale to compose the metric basket. Metrics are selected out of five different biological organisation levels (fauna, community, guild, population and individual) to cover disturbances at different spatial and temporal scales (Schmutz *et al.*, 2000). The U.K. approach is of a totally different nature and is based on two salmonids (trout and salmon). The concern of the index developers was to calibrate a semi-quantitative technique in comparison to a fully quantitative sampling to estimate the stock of salmonids in a site at a reduced cost (Strange *et al.*, 1989). Because the semi-quantitative technique is less intensive than quantitative sampling, it is possible to survey whole catchments based on representative sampling with respect to stock classes for decline. The Swedish fish index is primarily tuned with respect to acidification (Appelberg *et al.*, 2000) and was not RCA-based. An extensive sample of Swedish rivers and lakes was analysed as a whole to derive thresholds.

6.3 Results

6.3.1 Contrast between 1-2 (undisturbed) and 3-5 (disturbed)

The overall FPF for the EFI was 22 % and the FNF 19 %. The performance was better for the SBM-EU method with a smaller FPF of 7 % and an FNF of 20 %. The performance of the ExM was similar to the EFI with overall FPF of 17 % and an FNF of 20 % (Table 6.4). The FNF (about 20%) do not differ statistically, while the FPF do. SBM-EU is superior to the other methods: for a similar FNF it has a smaller FPF. These global percentages, however, mask large regional differences. This is partly because of a small sample size (Table 6.3) so confidence limits are broad, although it should be recognised that in most cases the fractions differ statistically, so this variation is real.

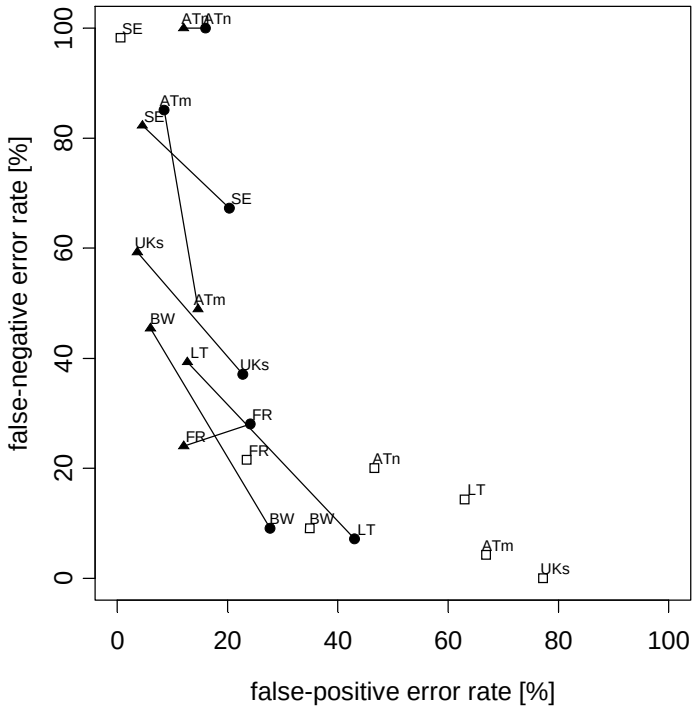


Figure 6.4 **Empirical relationship between the FPF and FNF.** The labels indicate the region; the symbols the type of fish index (FI): ExM (□), SBM-EU (▲), EFI (●). The points of the SBM and EFI indices are connected to show the difference in balance, but not necessarily the "distance" (see text).

The relationship between these errors (Figure 6.4) is hyperbolic and similar to the pattern shown in Figure 6.3. A possible explanation is that the points lie on the same error curve, but with a difference in tradeoff between FP and FN. For instance, existing fish indices are grouped in the lower right part of Figure 6.4 while the EFI and SBM-EU are located on the left side indicating a low FPF and a high FNF. There is a tendency for SBM-EU to have a higher FNF combined with a lower

FPF, although there are two exceptions. The existing Swedish Fish Index has a very small FPF (nearly 0%) but FNF is nearly 100%. This local index is not responsive to disturbance as measured by PS. In France, a national index was based on a similar approach to that used for the EFI and the three fish indices gave similar outputs, with SBM-EU scoring best: for a similar FNF (confidence limits overlap in Table 6.3), the FPF is smaller.

Table 6.3 **Measures of diagnostic accuracy by region and index.** Interval estimates of false positive fraction (FPF) and false negative fraction (FNF) for each region or country where a local index was given. See figure 4 for a graphical representation. The columns with $P(\chi^2)$ give the *P*-value (%) of the chi-square test of independence testing for the difference (see methods) between all fish indices (1) and between the FAME indices separately (2)

	FPF			$P(\chi^2)$		FNF			$P(\chi^2)$	
	EFI	SBM-EU	ExM	(1)	(2)	EFI	SBM-EU	ExM	(1)	(2)
ATm	8 % (4 – 15)	15 % (9 – 22)	67 % (58 – 75)	< 0.1	17	85 % (72 – 94)	49 % (34 – 64)	4 % (1 – 15)	< 0.1	< 0.1
ATn	16 % (9 – 26)	12 % (6 – 22)	47 % (35 – 59)	< 0.1	64	100 % (0 – 100)	100 % (0 – 100)	20 % (1 – 72)	0.4	-
BF	-	-	-	-	-	1 % (0 – 2)	3 % (2 – 5)	9 % (7 – 12)	< 0.1	0.1
BW	24 % (13 – 37)	4 % (0 – 13)	33 % (21 – 27)	< 0.1	< 0.1	10 % (0 – 45)	50 % (19 – 81)	10 % (0 – 45)	5.5	15
FR	24 % (21 – 27)	12 % (10 – 14)	23 % (21 – 26)	< 0.1	< 0.1	28 % (25 – 32)	24 % (21 – 27)	22 % (18 – 25)	2.0	11
LT	43 % (35 – 51)	13 % (8 – 19)	63 % (55 – 70)	< 0.1	< 0.1	7 % (1 – 24)	39 % (22 – 59)	14 % (4 – 33)	0.7	1.1
SE	20 % (19 – 22)	5 % (4 – 6)	1 % (0 – 1)	< 0.1	< 0.1	67 % (58 – 76)	82 % (74 – 89)	98 % (94 – 100)	< 0.1	1.4
UKs	23 % (17 – 30)	4 % (1 – 8)	77 % (70 – 83)	< 0.1	< 0.1	37 % (19 – 58)	59 % (39 – 78)	0 % (0 – 13)	< 0.1	17
Global	22.1 % (20.7 – 23.5)	7.3 % (6.5 – 8.2)	17.0 % (15.8 – 18.3)	< 0.1	< 0.1	19.4 % (17.5 – 21.3)	20.1 % (18.2 – 22.0)	19.9 % (18.0 – 21.9)	87	64

6.3.2 Investigation of the gradient from class 1 to 5

To evaluate the response of the EFI to an increasing gradient of disturbance, the cumulative distributions of the EFI index score were plotted for the classes as defined by the other indices

(Figure 6.5 – Figure 6.7). All cumulative curves were statistically different (Kolmogorov Smirnov test Table 6.4), but not all cumulative curves separated well. The curves for PS 1 and 2 intertwined and those for PS 4 and 5 were close. Regrouping PS into three categories (1-2, 3 and 4-5), places the curves in a logical order and improves the separation. At the threshold 1-2/3-5, FNF for sites of poor or bad status (PS = 4-5) was close to 0% but for PS = 3, FNF was 30%. Few of the highly disturbed sites were misclassified severely as undisturbed, but about 30 % of the moderately disturbed sites were classified as undisturbed by the EFI. Conversely, about 20% of the good sites (PS = 1-2) were misclassified as 3 (small FPF), but few as 4-5 (severe FPF). At the next threshold 1-3/4-5, the (small) FNF for class 4-5 was 20 % and the (small) FPF for class 3 was 30%.

Table 6.4 **Test whether the difference between two successive distributions differs significantly.** Probability values (%) for the Kolmogorov-Smirnov test.

	Comparisons between pressure states			
	1/2	2/3	3/4	4/5
PS	< 0.1	< 0.1	< 0.1	< 0.1
SBM-EU	17.8	< 0.1	< 0.1	< 0.1
ATmulfa	18.7	14.7	< 0.1	
ATnat	89.2	3	88.5	93.1
BF	-	1.7	< 0.1	-
BW	1	34.5	0.9	0.4
FR	< 0.1	< 0.1	< 0.1	0.5
LT	-	2.3	1.3	-
SE	< 0.1	98.9	-	-
UK	83.0	23.7	83.1	< 0.1

Correspondence between the EFI and SBM-EU was high (Figure 6.6), with a similar picture to that between EFI and PS (Figure 6.5), again with class 3 (now of SBM-EU) at an intermediate position between 1-2 and 4-5. By contrast with the previous situation, discrimination between classes 1 and 2 disappears. The relation between the existing fish indices with EFI was poor with no consistent pattern (Figure 6.7). The EFI gave systematically lower scored for Austria, with few sites lying above the threshold for 1-2/3-5. For Flanders (Belgium), the local fish index scored the sites from 2 to 4, while the EFI scores them between 4 or 5. The curves derived from French data were well ordered, but the separation was small. Furthermore, the local index scored more than half of the sites 4 or 5, while the EFI positioned most in categories 3 or 2. The differentiation made by the Lithuanian fish index disappeared with the use of the EFI: 40% of class 2 sites received a score of 3 and 80 % of class 4 a score of 3 or 2. In Sweden, class 2 and 3 cannot be separated. In the UK only class 5 was different, but the EFI scored about 50 % of the highly disturbed sites (from the perspective of the local index) as class 2.

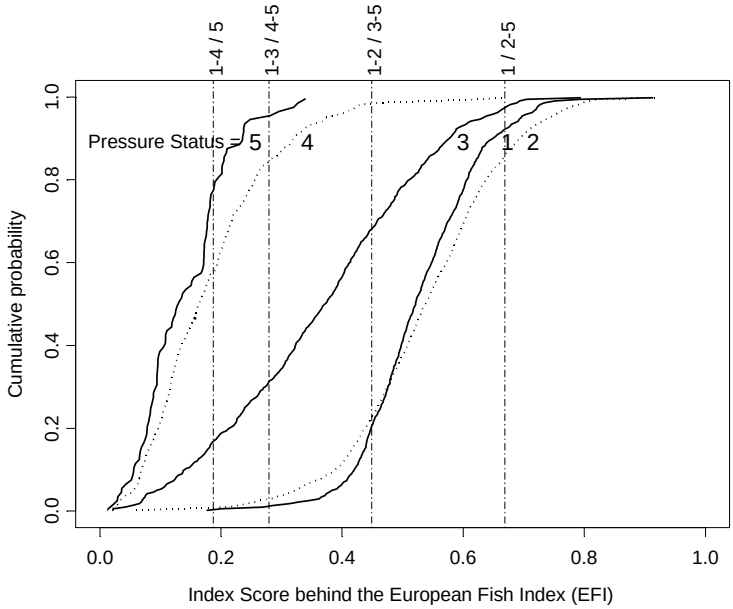


Figure 6.5 **Cumulative distribution of EFI conditional on the Pressure Status** (solid lines: class 1, 3 & 5 and dotted lines: class 2 & 4): The vertical lines (labeled 1/2-5, 1-2/3-5, …) are thresholds categorizing the index score of EFI into the WFD five-class system.

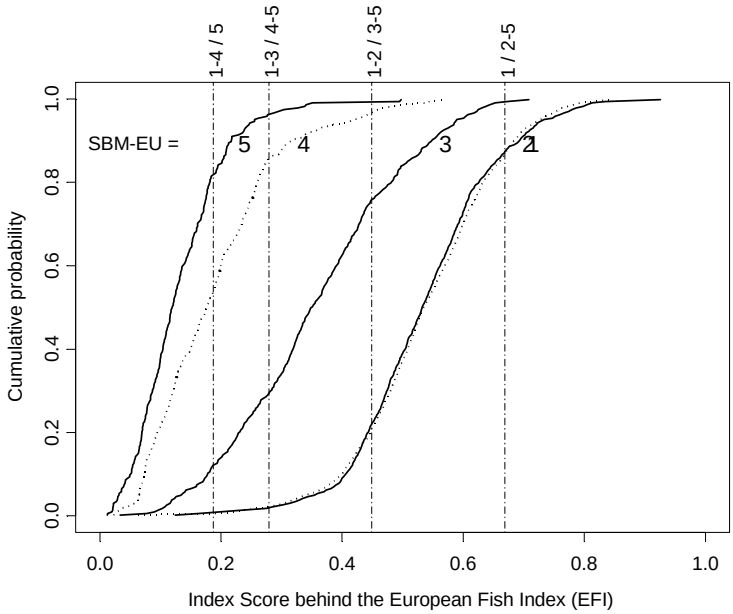


Figure 6.6 **Cumulative distribution of EFI conditional on SBM-EU** (solid lines: class 1, 3 & 5 and dotted lines: class 2 & 4). The vertical lines (labeled 1/2-5, 1-2/3-5, …) are thresholds categorizing the index score of EFI into the WFD five-class system.

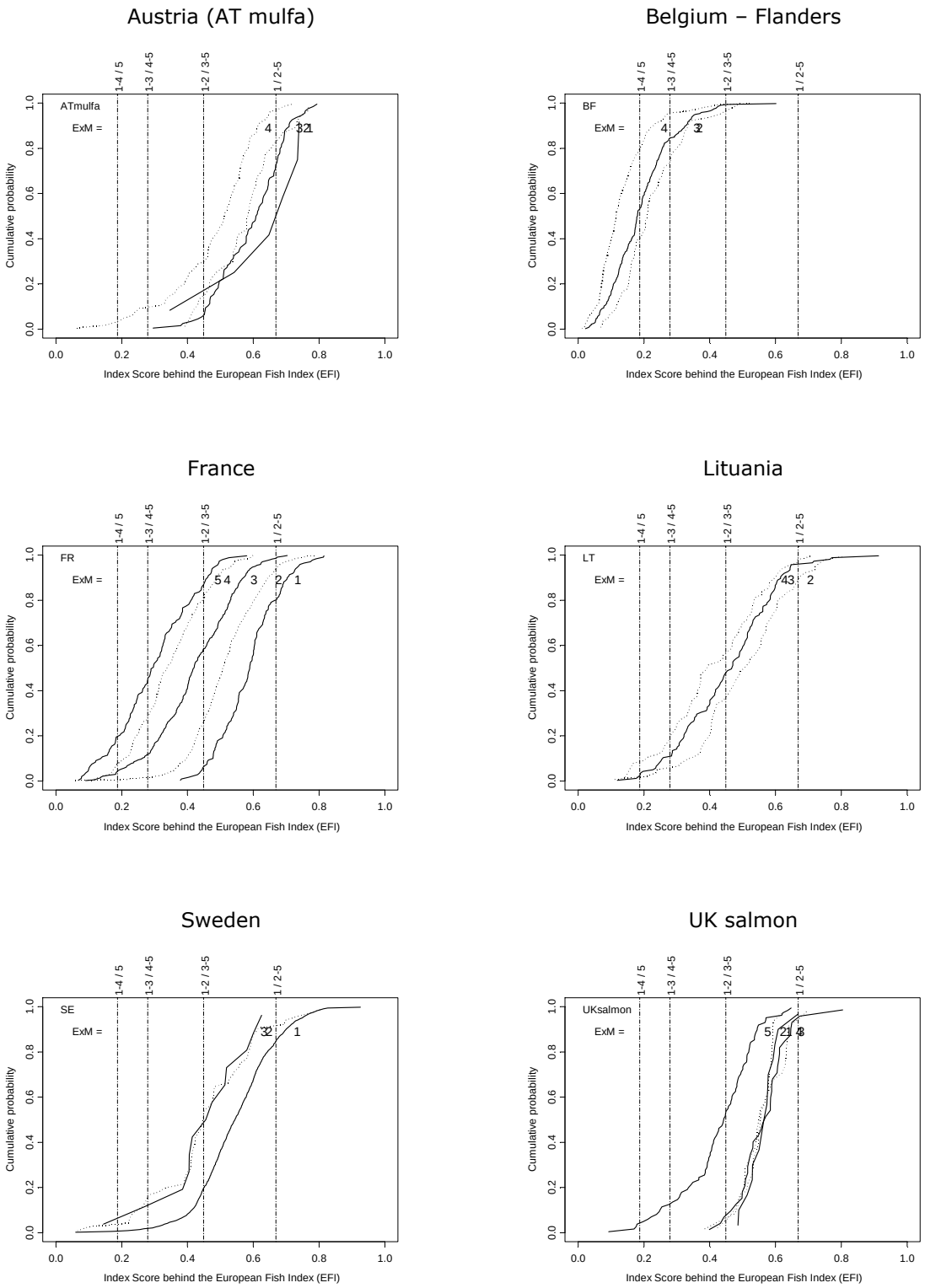


Figure 6.7 **Cumulative distribution of EFI conditional on six existing fish indices** (solid lines: class 1, 3 & 5 and dotted lines: class 2 & 4). The vertical lines (labeled 1/2-5, 1-2/3-5, …) are thresholds categorizing the index score of EFI into the WFD five-class system.

6.4 Discussion

This evaluation of EFI must be considered internal because it was based on the same dataset used to construct the fish indices. With internal data, the error fraction cannot account for study bias and risks underestimating the true misclassification fraction (Hosmer and Lemeshow, 2000). In some regions there were insufficient data resulting in broad confidence intervals, especially for FNF. All fish indices were compared to the same common measurement stick, but PS was not a common standard and was not without error. Despite these limitations, the exercise provides a preliminary assessment of the performance of the EFI.

A first aim was to examine the two types of misclassification error with respect to the most important contrast 1-2 / 3-5 for the WFD. As it is always possible to make one error very small at the expense of the other, fish indices with very different FNF and FPF can have a similar performance if tuned differently. The negative correlation between FNF and FPF of the national or regional indices found in Figure 6.4 suggests that at least part of the differences observed between the local fish indices is because of this effect. It implies that the performance of the existing indices is less different than is apparent from the misclassification errors. In contrast, the EFI is well balanced. For both misclassification errors, the misclassification fraction was about 20 %. However, there were important regional differences and this should be investigated further.

The second aim was to test the response to the full disturbance gradient. As a whole, EFI is responsive to a gradient of disturbance. The cumulative curves of EFI were well separated with respect to PS between class 1-2, class 3 and class 4-5. Thus no clear distinction between class 1 and 2 and between class 4 and 5 is possible. This is possibly due to the fact that it is very hard to make a distinction between the corresponding PS classes.

The two main fish indices developed, performed in a similar manner with respect to PS. Despite the higher performance of SBM-EU, EFI was preferred because it is more simple and transparent with a relatively small loss in terms of misclassification with respect to PS. It is important to realise that the EFI is only a single model with ten metrics in common, while the SBM-EU comprises twelve different discriminant models (one for each stratum) with 49 different metrics in total. Adding explanatory variables improves the fit of the regression model but increases the complexity of the model and at a certain point the gain in fit is too small compared with the added complexity.

No intercalibration of EFI with existing fish indices seems possible because the local indices classify sites differently. The performance of the local fish index was similar only for France, which adopted a similar approach to that for the development of EFI. This was not unexpected as the local fish indices were developed by very different methods ranging from expert judgement to model building as advocated by Karr (Karr, 1981), Fausch *et al.* (1990) and Hughes *et al.* (1998).

7 General discussion and conclusions

We first assemble the main findings of the thesis and present a global narrative piecing together all elements in three key formula (Figure 7.1). In section 7.2, we comment on the sampling design to improve IBI construction. The last section 7.3 is devoted to the cost perspective.

7.1 The main findings

7.1.1 The test variable or yardstick of an IBI

7.1.1.1. Intactness of the biological community as a proxy

An index of biotic integrity (IBI) is an (ecological) indicator. The test variable or yardstick of an indicator is a proxy or surrogate for some underlying complex feature or process of interest which is more expensive, harder or even impossible to measure directly (Murtaugh, 1996):

$$property\ of\ interest \quad \leftrightarrow \quad test\ variable$$

For an IBI, the property of interest is the ecosystem condition and the yardstick is the ecological quality measure (EQM) which assesses the intactness of a part of the biological community in comparison to a well-defined reference situation (Boulton, 1999; Meyer, 1997):

$$ecosystem\ condition\ (integrity\ \&\ health) \quad \leftrightarrow \quad intactness\ biological\ community \quad \leftrightarrow \quad EQM$$

EQM is a "distance to target" measure. The ecological rationale is that anthropogenic alterations of the environment and/or ecosystem ultimately provoke shifts in the species balance in comparison to an unimpaired (pristine) reference situation (Attrill and Depledge, 1997).

7.1.1.2. The Reference Condition Approach (RCA)

Assessing the intactness of a biological community is hard, not at least because the species community is highly variable even under reference conditions (Bailey *et al.*, 2004). Theoretical knowledge is generally lacking to evaluate directly the state of the biological community at a site. The RCA philosophy provides an empirical strategy to overcome this problem. A first crucial step is to sample reference sites of high ecological quality. Therefore, we should carefully operationalize what constitutes a reference site (Wright *et al.*, 2000) and set up a sufficiently broad survey, spatially as well as temporally, to prevent from spectrum bias, a too narrow window on the natural variability. From this reference data, we can derive the multivariate reference distribution $\Phi_C(X|R)$, representing the intrinsic and sampling variability of the species composition under reference conditions as a function of the type and environmental characteristics of the site. Next, in comparison to $\Phi_C(X|R)$, a "distance to target" measure Δ is constructed, to assess the intactness of the observed ecological community at a test site:

$$\textit{intactness biological community} \quad \leftrightarrow \quad EQM = -\Delta(C, \Phi_C(X \mid R))$$

We discriminate between natural and human factors by matching the test site with reference sites of the same type and/or with similar environmental characteristics X and by taking into account the intrinsic and sampling variability as modelled by the multivariate distribution Φ_C.

7.1.1.3. Multimetric indices (MMIs)

Community data is highly multi-dimensional, variable and sparse containing many zeros because many species are rare, making statistical analysis hard. One possible solution is to study attributes of the community data reflecting essential ecological features and therefore hypothesized to be sensitive for human alterations of the ecosystem. These attributes are the so-called metrics M. Metrics incorporate information about the ecological "strategy" of species in the community data (Jørgensen *et al.*, 2005). For instance, ecological guild information (Wilson, 1999) allows to group species according to their ecological niche and hence we can study changes in the niche. We say that the community data is augmented with species information to enhance the diagnostic power:

$$M_j = C \otimes \textit{species info}$$

Metrics are often linear combinations of the properties of the species weighted by their abundance. Another important class is the diversity indices related to the species richness and evenness of the species distribution. From a numerical point of view, we replace the original community data C with a few well-chosen attributes, the metrics M, with more stable statistical properties facilitating analysis. The equation for EQM can be rewritten as follows by replacing C by M:

$$\textit{intactness biological community} \quad \leftrightarrow \quad EQM = -\Delta(M, \Phi_M(X \mid R))$$

7.1.1.4. The average score model (AVG)

A further simplification is possible by evaluating the metrics separately. To appreciate the outcome of a metric, we score each metric by making a comparison with its expected value under reference conditions in relation to the natural variability, for instance, by deriving the standardised residuals of a regression model predicting the expected value of M_j under reference conditions R from the environmental conditions X of a site:

$$S_j = \delta_j \frac{M_j - E_R[M_j \mid X]}{Stdev_R[M_j \mid X]} \qquad \delta_j = \pm 1 \sim \textit{positive / negative metric}$$

These "z-scores" are unitless "distance to target" measures which can be simply averaged because they are expressed in the same scale. This average score model (AVG) can be generalised to a regression equation in which the coefficients are the weights of the metrics. This last step defines the yardstick of an IBI as a regression model:

$$\textit{intactness biological community} \quad \leftrightarrow \quad EQM = \frac{1}{J}\sum_{j=1}^{J} S_j \quad \rightarrow \quad [\beta_0] + \sum_{j=1}^{J} \beta_j S_j$$

7.1.2 The calibration of IBIs

7.1.2.1. The ecological quality class (EQC)

The EQM expresses an overall "distance to target" in comparison to a reference condition. If the value at a site is lower than expected, it is concluded that the site is not reference. This results in a binary two-state classification: degraded or not (reference). However, this information insufficiently reflects the biological significance of a certain difference. Clearly, large differences are more severe than small differences, but the question is from which point on, EQM really signals a problem. To derive thresholds T discriminating between different degrees of human impact in ecological quality classes (EQC), it is necessary to calibrate the yardstick with respect to a sample of both pristine and degraded sites ranked according to a gradient of human impact (Davies and Jackson, 2006).

7.1.2.2. The degree of human impact

The EQC classes are linked to stages of the degradation process. At first, ecosystem changes are reversible. If the stress factors disappear, the ecosystem recovers easily. We are in the reference situation R corresponding to WFD class 1 (high) and 2 (good). However, once critical thresholds surpassed, the modifications of the ecosystem become more and more irreversible and/or require a very long time to recover (Bailey *et al.*, 2004; Davies and Jackson, 2006). Restoration programs are necessary to overcome hysteresis and/or definitive destruction of the ecosystem. In this model, WFD class 3 represent a situation where there is a moderate (M) but clear impact. WDF classes 4 (poor) and 5 (bad) represent a definitely degraded ecosystem (D). By 2015, the EU aims to restore or rehabilitate all waterbodies to the first two classes and to keep the sites of high quality (class 1) intact (stand-still principle). It should be clear that this classification should be scientifically justified to prevent from false reassurance.

7.1.2.3. The a priori classification

To prevent from a circular system with preconceived ideas about what constitutes an intact biological community, for the calibration, an alternative a priori classification of the ecosystem is necessary independent of the community data. As we cannot measure the ecosystem condition directly, a common approach is to score the anthropogenic activities and pressures at a site. Sites with no or no discernable impact are classified as reference sites (the anchor), while the other sites are ranked according to a degree of exposure (the pressure gradient). This scoring results in the human quality class (HQC). The index is calibrated to match the ecological quality class (EQC) with HQC:

$$HQC \leftrightarrow EQC$$

7.1.3 Evaluation of the diagnostic accuracy and validity of IBIs

7.1.3.1. The analogy to diagnostic tests in medicine

A key idea was that IBIs are very similar to clinical tests judging the health condition of a patient from one or more biomarkers indicative for a certain disease. Currently, the Receiver Operator Characteristic (ROC) curve is the best-developed statistical tool for analysing and describing the

performance of test variables in medicine (Pepe, 2003; Zhou *et al.*, 2002). Although advocated already about fifteen years ago by Murtaugh (1996), the ROC concept has not been systematically applied to IBIs and ecological indicators in general.

7.1.3.2. The cost implications of false positives and false negatives

We should distinguish false positives (FPs) and false negatives (FNs) as they imply different costs. A high false positive fraction (FPF) detracts resources from where they are necessary. Besides, treatment of pristine sites can be harmful. Conversely, with a high false negative fraction (FNF), many degraded sites are not restored or rehabilitated continuing a bad ecosystem situation and a loss of ecosystem services and goods. To realise the maximal benefit, we strive to maximise TPF keeping the FPF as small as possible. This tradeoff is precisely what it shown with an ROC curve. A steep curve tells that we can realise a high TPF at a small FPF (the strength of the index).

7.1.3.3. Receiver Operating Characteristic (ROC) curves

ROC curves plot the sensitivity or the true positive fraction (TPF) of a test variable as a function of the false positive fraction (FPF) which is the complement of the specificity (FPF = 1 – specificity). The ROC curve gives the combinations of (FPF,TPF) for all possible decision thresholds T. For a given index, we cannot "escape" the ROC line: when tuning the index, i.e. choosing the decision threshold to discriminate between degraded and pristine sites, there is only one degree of freedom. Fixing FPF to keep the FP burden below a certain limit, fixes TPF. Conversely, fixing TPF to realise a certain objective, fixes FPF. It is not possible to optimise both FPF and TPF at the same time. If no optimal balance can be found, we have to search for a better index.

7.1.4 The usefulness of IBIs

7.1.4.1. Decision analysis with utility curves

Utility curves allow to study the tradeoff between FPF and FNF taking into account the costs. They quantify the cost consequences of management decisions based on the index. In essence, utility curves are ROC curves modified by parameters describing the decision context:

$$usefulness \quad = \quad u(ROC(FPF,TPF),\pi^{+},b,...)$$

The formula contains two critical parameters: the prevalence of degradation (π^{+}) and the benefit ratio b. A low prevalence implies that many sites should be screened before a degraded site is found which increases the FP burden. Then, a strong index is necessary to realise a high TPF at a low FPF. The benefit ratio b expresses the tension between gain (because a TP) and loss (because of a FP). If the gain is large compared to the loss, the optimal decision point on the ROC curve shifts upwards implying we set TPF high at the expense of a high FPF. From a purely monetary point of view, this can be a problem as increasing FPF implies higher management costs. This is not a problem if society is willing to pay (WTP). In this situation, it is required the benefit calculation is transparent for society in order to obtain an agreement for making higher costs.

7.1.4.2. Choosing the optimal index

Comparison of utility curves in combination with the assessment costs allows to determine which index is optimal by minimising the entire cost C_E, defined as the sum of the decision cost C_D (cost consequences of the index) and the assessment cost C_A (to collect data for the index):

$$C_E = C_A + C_D$$

With increasing complexity and/or ambition level of the decision, we have to choose for a stronger index, a point we discuss in the last section 7.3 of this general discussion.

7.1.5 Three key formulas

Figure 7.1 further condenses the main ideas and steps in one scheme.

7.1.5.1. Formula 1: from the RCA principle to a regression model

The first formula links the regression model with the original RCA principle. To avoid working with the highly multivariate, variable and sparse community data (negative reason) and to augment the data with knowledge about the ecological strategy of the species (positive reason), the community data is replaced with metrics. In the next step, the model is further simplified to an average of 'distance-to-target' scores. From this average score model (AVG), the step to regression is small.

Interestingly, the scheme also links the multimetric approach to alternative approaches. In fact, as discussed in Chapter 2, there are two groups of index developers: the multimetric and multivariate school. Some IBI developers prefer to work directly on the species data C with multivariate techniques avoiding to collect information or to make assumptions about the ecological properties of the species which can misguide the analysis (Reynoldson *et al.*, 1997). On the other hand, ecological knowledge possibly strengthening the analysis is not incorporated (Fore, 2003). In fact, both approaches are highly complementary and should be used more simultaneously.

7.1.5.2. Formula 2: from ROC curves to utility curves

Currently, the ROC curve is the best documented tool to evaluate the diagnostic accuracy of clinical tests (Pepe, 2003). An important question of this thesis was why they are so pivotal. Searching for more insight, we arrived at utility functions quantifying the usefulness of (ecological) indicators to decision making. We tackled the problem from two angles: a cost effectiveness analysis (CEA), considering the monetary costs only, and a cost benefit analysis (CBA), also valuing ecological and societal benefits in monetary terms. The advantage of a CBA is that we can determine the optimal decision point, i.e. the point on the ROC curve optimising the societal benefit. The optimum is where the marginal gain equals the marginal loss. The hard point is to value the benefits correctly.

From both a CEA and CBA perspective, the crucial property determining the strength and usefulness of an IBI appears to be its capacity to realise a high TPF keeping FPF low, which is closely linked to the steepness of the ROC curve. In fact, by deriving utility curves, we perform a decision analysis analyzing the cost consequences of decisions. Our approach is based on a similar reasoning as Vickers *et al.* (2006; 2008; 2008) in a clinical context.

7.1.5.3. Formula 3: how to budget monitoring?

With increasing strength of an index, the assessment cost C_A increases but the decision cost C_D decreases. The optimal index is where the entire cost C_E, defined as C_A + C_D, is minimal. With increasing complexity of the decision context and/or ambition level, the optimal strength of the index increases. As we will discuss further on, this simple equation allows to make a tradeoff between the costs and benefits of monitoring (Caughlan and Oakley, 2001).

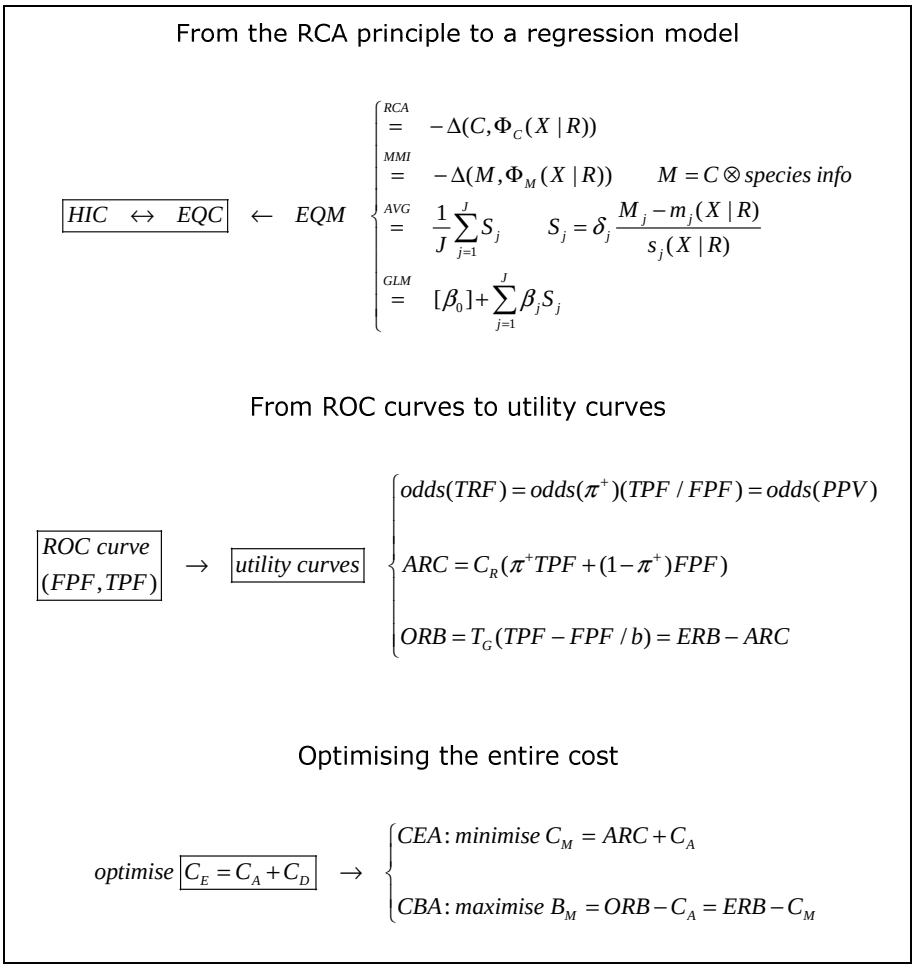


Figure 7.1 **The three key formulas.** Main symbols: (i) HIC = human impact class, EQC = ecological quality class, EQM = ecological quality measure. (ii) TRF = true restoration fraction, ARC = average restoration cost, ORB = overall restoration benefit, ERB = ecological restoration benefit. (iii) C_E = entire cost, C_E = monetary cost because of restoration, B_M = restoration benefit corrected for monetary cost, C_A = assessment cost.

7.2 How to improve the design of indices of biotic integrity?

Figure 7.2 offers a scheme based on the index model to guide the questions and choices when designing an IBI. The scheme is linked to the model as described in previous paragraph. Ideally,

these points should be considered in depth before starting the construction of an index. In practice however, quite often, because of budgetary constraints, index development is based on available data collected for other purposes. Yet, we believe that investment in a specific data collection pays of in the long run, evidently on the condition that the study design is appropriate. Therefore, in this section, we suggest some improvements from a statistical point of view.

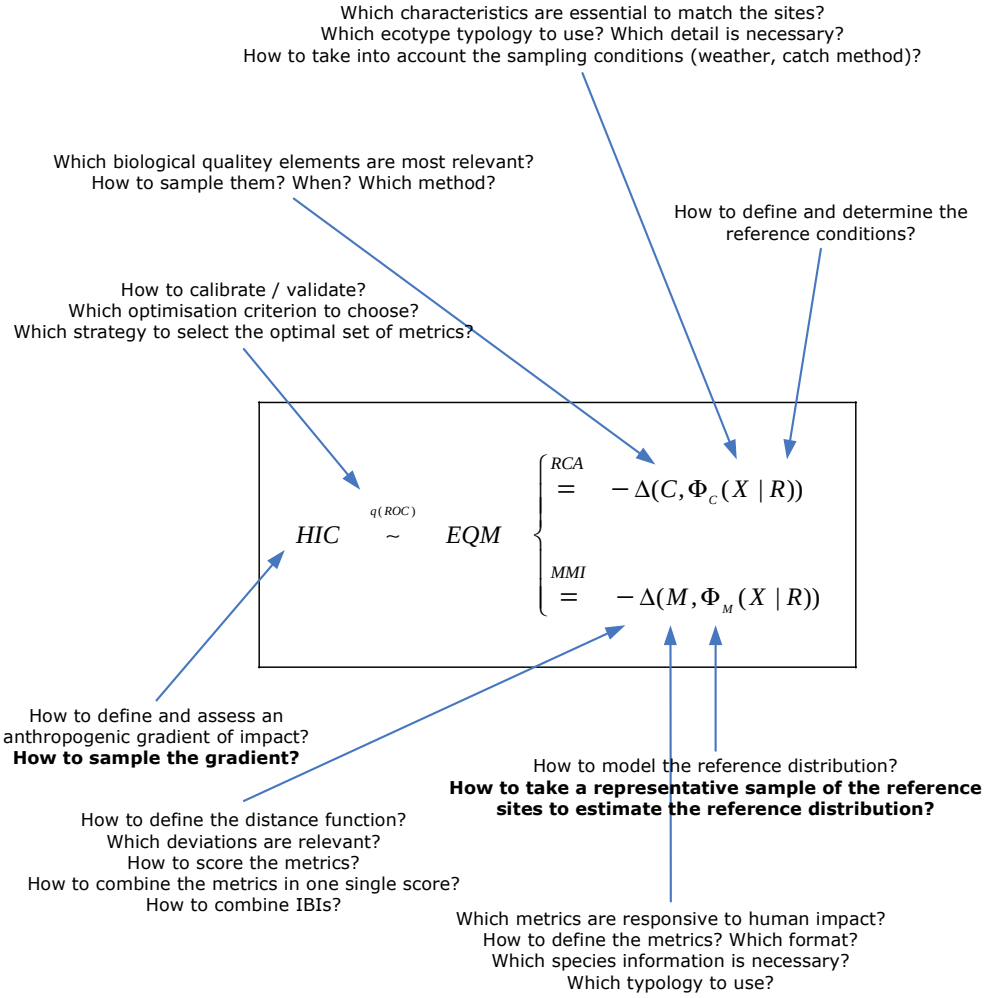


Figure 7.2 **Design questions when constructing an IBI as based on the RCA principle.** The questions are organised according to the model equations representing the RCA principle (see Figure 7.1).

A crucial aspect is to work on a sufficient large spatial level covering a broad range of gradients. The WFD allows EU member states to develop their own index if they intercalibrate, i.e. compare their indices afterwards to guarantee an equal classification. From a policy point of view, given the very different traditions in the European countries – proven assessment methods at a national scale are hard to change – , this freedom of choice was perhaps the only possibility, but it is a missed opportunity. In an evaluation paper at the tenth anniversary of the WFD, Hering *et al.* (2010) observe a proliferation of incompatible datasets, methods and indices.

Already at the onset of the WFD, Hughes *et al.* (2000) hinted, in vain, based on their US experience with a long IBI tradition, that "a common probability-based approach has distinct advantages for monitoring that may be applicable to European Communities seeking to assess the ecological integrity of waters". In contrast, a plethora of methodologies with hundreds of indices, metrics and evaluation tools are presently available (Borja and Dauer, 2008). One of the main challenges for the next decade is to stop this proliferation (Borja *et al.*, 2009b). Instead of creating new indices for local needs, we should spend more resources on developing indices on a larger scale and/or to build further on existing indices. In this section, we will make some remarks with respect to this issue. Especially, we are convinced that it is necessary to pay more attention to the sampling of the sites. Random selection of the sites remains a contested issue (Fore, 2003), because it not easy to perform and it is expensive in comparison to convenience sampling. Yet in the long run, good sampling pays off.

7.2.1 Data collection and sampling

7.2.1.1. More attention is necessary for the site selection

Traditionally, considerable attention is spent on the sampling of biological communities to control for their spatial and temporal variability. The EU-funded project STAR (Standardisation of River Classifications) project was mainly devoted to this problem and contains many suggestions to improve this species sampling as reported in a special issue of Hydrobiologia (Furse *et al.*, 2006b). Beyond doubt, the catch method and effort have a large impact and should be standardised. However, too little attention has been paid to survey design, i.e. the sampling strategy with the purpose to obtain a representative sample of sites. Southerland *et al.* (2009) refute in their paper two traditional beliefs: (i) ad hoc (non-random) sampling is adequate to obtain a representative sample on the condition that sample size is sufficiently high and (ii) more intensive sampling at a site level (e.g., collecting more organisms at a higher species resolution) is always better. With respect to the latter, they demonstrate that, if the total budget is fixed (the correct assumption), additional site sampling effort is often at the expense of the overall precision because too few sites can be selected.

7.2.1.2. Representative sampling to select the appropriate metrics

An IBI has the potential and the ambition to detect a broad range of impacts on the ecosystem. However, if the sample of sites does not cover the human impacts in the region and/or the full gradient of pressures, there is little hope the IBI will fulfil its aims as the metrics sensitive to pressures not covered in the dataset will not be selected. In this respect, an ongoing discussion is whether the sample should be probability based. A case study of Fore (2003) demonstrated that a probability sample (Overton and Stehman, 1995) was superior to cover a broad spectrum of pressures against the intuition of many researchers involved in the project. In general, most people prefer to select sites which are "representative" according to preset criteria; see Kruskal and Mosteller (1979a; 1979b; 1979c; 1980) for a four paper series about the different meanings of representativeness. A random sample does not only assure an unbiased picture of the ecological condition, but also of the relations between variables (Kish, 1987). In practice a random sample is

hard to achieve. For instance, it is quite costly to construct a sampling frame matching the target population. However, in the long run, the investment pays back as argued by many authors (Dauer and Llansó, 2003; Hughes *et al.*, 2000; Llansó *et al.*, 2003; Overton, 1993; Southerland *et al.*, 2009). For instance, Paul *et al.* (2008) show that in contrast to ad hoc samples, it is possible to combine probability-based small samples to make more global inferences.

7.2.1.3. Spectrum bias

Spectrum bias (Begg and Greenes, 1983; Ransohoff and Feinstein, 1978; Zhou *et al.*, 2002) occurs when the calibration sample does not cover the total variability of the target region (the statistical population of the index). Because of the underestimation of the total variability, the diagnostic accuracy assessed at the calibration stage will be higher than in reality. To control for spectrum bias, again representative sampling of the target region is necessary. Also, data should be collected over a sufficiently long time to cover the year-to-year variability.

7.2.1.4. The preclassification

The preclassification is the Achilles' heel of index development. Also in medicine (Zhou *et al.*, 2002), the construction of a gold standard is reported to be the hardest part of the exercise. For IBIs, the preclassification is derived from a combination of scores expressing the effect of human activities and pressures (Van Stickle and Paulsen, 2008). This approach involves rather strong assumptions (Yuan and Norton, 2004) including knowledge of dose-response curves (to which extent is the ecosystem affected by the human pressures/activities) and the additivity of these pressures (ignoring impact differences and interaction). By no means, this is a perfect system and cannot be.

Yet, the preclassification is not a purpose in itself, but a device to rank the sites in a reasonable way with respect to an anthropogenic gradient of pressure enabling IBI construction. To put in perspective what is achievable, a well-controlled exercise of Falcone *et al.* (2010) testing the capacity of an a priori ranking of watersheds with an extensive set of GIS-variables demonstrated that the diagnostic accuracy of the classification in least- and most-degraded sites was about two-thirds. Further standardisation is surely needed in this area. In this respect, it is preferable to use existing schemes (Borja *et al.*, 2009c). Hence, we derived our preclassification (Breine *et al.*, 2007) from an existing and validated framework resulting from an extensive collaboration of field experts (Aubry and Elliott, 2006).

On the other hand, further improvement is possible by considering totally different approaches. An alternative to a pressure inventory is to follow a thermodynamic approach (Kay, 1991; Kay and Schneider, 1992; Schneider and Kay, 1994) which is a more fundamental way to gauge the ecosystem condition independent of the ecological community (Jørgensen and Svirezhev, 2004; Jørgensen *et al.*, 2005). Currently, thermodynamic indicators are hard to measure and cannot be used routinely just as many other new ecological insights (Proulx, 2007). However, in a calibration context, new approaches can serve as the gold standard and / or as an external validation tool.

7.2.2 Model building

7.2.2.1. The modelling approach

To our knowledge, we are the first to demonstrate that the average score model is a regression model. This opens perspectives to improve index construction. As the response variable is ordinal, ordinal logistic regression models are the logical choice as they maximally exploit the ordinal character of the data (McCullagh and Nelder, 1989; Agresti, 2002). Yet they are seldom applied in ecology. We found one paper in the context of IBIs (Maloney *et al.*, 2009) constructing an index on benthic macroinvertebrates based on ordinal regression. In Chapter 5 we chose for the proportional odds model, but probably in some cases this model will be too restrictive. Liu and Agresti (2005) give an extensive up to date overview of ordinal regression techniques. For instance, one of the discussants of the overview paper suggests to model sequential logits such that each change of class is modelled separately which could be appropriate to model the non-linear degradation thresholds of the ecosystem. Mixed ordinal models allow to incorporate the sampling structure to model repeated sampling, longitudinal data and spatial autocorrelation. This would allow to analyse more complex survey data correctly and to avoid pseudo-replication (Hurlbert, 1984).

7.2.2.2. Model building strategy

In Chapter 5, we presented a model building strategy to select the optimal combination of metrics. An important point was to prevent from overfitting. It is well known in statistical literature that a more complex model can have worse statistical properties than a simpler variant because more parameters have to be estimated (Linhart and Zucchini, 1986). If the model is too complex in relation to the data available, it fits noise and data particularities instead of the underlying process (Zucchini, 2000). We advocate to explore the models in the vicinity of the optimum as emphasized by McCullagh & Nelder (1989). Rather seldom data points unequivocally to one single model and one should acknowledge this uncertainty. Further exploring the environment of the optimum is an occasion to learn more from the data (Kutner *et al.*, 2005).

7.2.3 Validation and follow-up

7.2.3.1. Credibility and acceptability

The credibility and acceptability of IBIs among scientists, policy makers and managers depends on the demonstration of index reliability in a meaningful validation process (Borja and Dauer, 2008). Any new methodology or index without a reliable and replicable validation using independent data should be discarded or used with extreme caution (Borja *et al.*, 2009c). Our impression is that we are far from this ideal. Yet, it is very important. In the context of the European legislation, there can be legal pursuits for stakeholders not complying to the rules. Also uncertainty measures for IBI are generally lacking (Ellis and Adriaenssens, 2006).

7.2.3.2. Internal validation

Internal validation is based on data stemming from the same study as for the model construction. It is an important first step to have an idea of the diagnostic accuracy of the index. In literature,

most index developers validate the index in some way. However, not always a distinction is made between FPs and FNs, sometimes just an overall misclassification rate is reported which depends on the prevalence. A common practice is split-sample validation, in which one part of the dataset is put aside (typically one third) to validate the model. However, unless the database is very large, this approach is inefficient. In contrast, bootstrapping and/or cross-validation provide efficient estimates with low bias for logistic regression (Steyerberg *et al.*, 2001).

7.2.3.3. External validation

Internal model validation cannot correct for fundamental flaws of the design and conduct of the study (Hosmer and Lemeshow, 2000). If the sample is insufficiently representative for the region, we will be faced with spectrum bias and the appropriate metrics will not be selected. Also, the construction of IBIs is an observational correlative study susceptible for confounding. Therefore, to control whether the index effectively works in a practical setting, an external validation with new independently sampled data is necessary. Rather seldom, it is possible to find a budget for external validation. Yet, continuous improvement and updates are necessary to optimise and improve the index. For instance, Southerland *et al.* (2007) substantially improved the Maryland fish and benthic macroinvertebrate index by revisiting their index.

7.2.3.4. Follow-up and QC/QA programs

Index development may not stop after its calibration. It is a continuous process. In combination with the effective use of the index by the manager, a scientific activity is necessary critically examining the performance of the index. As for the previous point, it is hard to find resources for this type of work as the perceived merits are low. In between solutions are possible. For instance, in the context of nature restoration programs, the response of the index can be validated by taking gold standard measurements on a subsample of the sites restored. This strategy does not require a separate manipulative field study, but incorporates and integrates scientific research in daily practice (Underwood, 1995; Underwood, 1998; Underwood, 2009). This approach enables a causal analysis if elements of experimental design are incorporated in management (Underwood, 1997). An alternative strategy is to embed index maintenance in a more general quality assurance and quality control (QA/QC) program. A shadow monitoring program can follow-up a random subsample of the sites in more detail. Analyses of complementary data stemming from the monitoring can help to verify whether the index performs as expected.

7.2.3.5. Documentation

Rules about documenting an index are lacking. Yet documentation is essential for transparency and quality (Flotemersch *et al.*, 2006). Full documentation should include information about the field protocol, measures of diagnostic accuracy, and preferably also cost information. If these elements are available, comparative studies would also be more feasible. For instance, in Chapter 4 we sketched a framework to select an optimal index. For its application, currently, a lot of information is not available. As a consequence, it is not possible to optimally tune an index, or to choose between competitive indices unless by making general assumptions of the diagnostic accuracy of the index. Scientific publication of the index is an important element to prove the scientific quality

of the work, but a scientific paper only discusses the relevant parts for the broader scientific community and does not give sufficient practical details (Holl, 2010).

7.2.3.6. The ROC curve

The most complete characterisation of diagnostic accuracy is the ROC curve. Yet, the ROC concept is not applied until recently (Hale *et al.*, 2007; Quataert *et al.*, 2007; Hale and Heltshe, 2008; Benyi *et al.*, 2009; Dos Santos *et al.*, 2011). ROC curves have many advantages. As they fully characterise the diagnostic accuracy for all possible decision thresholds, it is possible to retune the index in a new context. In addition, they assess the discriminatory capacity independently of the original measurement scale, facilitating comparisons (Zhou *et al.*, 2002).

7.3 The cost perspective

7.3.1 The need for cost calculations

The cost-effectiveness of ecological indicators and environmental monitoring in general is rather seldom investigated. Yet it is crucial. Monitoring is perceived as an overhead cost in competition with the budget required for action (Caughlan and Oakley, 2001). Under budget constraints, quite often monitoring programs are the first to be curtailed. To cope with this omnipresent pressure, we should be more explicit about the costs and benefits of monitoring programs. This is more easily said than done, as it is not evident to value monitoring benefits (Caughlan and Oakley, 2001). Also, in a medical context, Vickers (2008) reported that prognostic models are typically evaluated without addressing clinical consequences as it requires additional information not directly available. To circumvent the problem, in Chapter 4 we developed utility curves which depend only on a few parameters facilitating a mathematical discussion. To our surprise, in circumstances with a high benefit ratio, the optimal assessment cost was rather high leading to the false impression the index inflates the costs, while in fact it maximises the benefits. To get a more profound understanding of this result, we discuss the cost tradeoff from a different and more general angle.

7.3.2 The entire cost

We assume a decision context in which an index is effectively used to guide decisions. We define entire cost (C_E) as the sum of the assessment (C_A) and decision cost (C_D):

$$C_E = C_A + C_D$$

C_A is equal to the monitoring budget required for ascertaining the index. This does not only include fieldwork, but also design and set up of the monitoring program, data storage, analysis, interpretation and reporting, organisation, quality control and assurance (QC/QA). If no index exists, the development cost should be added. As assumed in Chapter 4, we may expect that with increasing complexity and accuracy of the index, C_A will increase fast (Figure 7.3). In reality, there are only a few indices available (i1 < i2 < … < i6) scattered along this line.

On the other hand, the decision cost C_D refers to the cost consequences of the decisions as guided by the index. With increasing diagnostic accuracy, the number of mistakes or suboptimal decisions will go down and hence C_D will decrease (Figure 7.3). Importantly, the position and shape of the decision curve depend on the decision context. If the cost of mistakes and/or the benefit of correct decisions is low, the decision curve will be flat. In contrast, if decisions are critical (there is a lot to gain or loose), the decision curve will be steep.

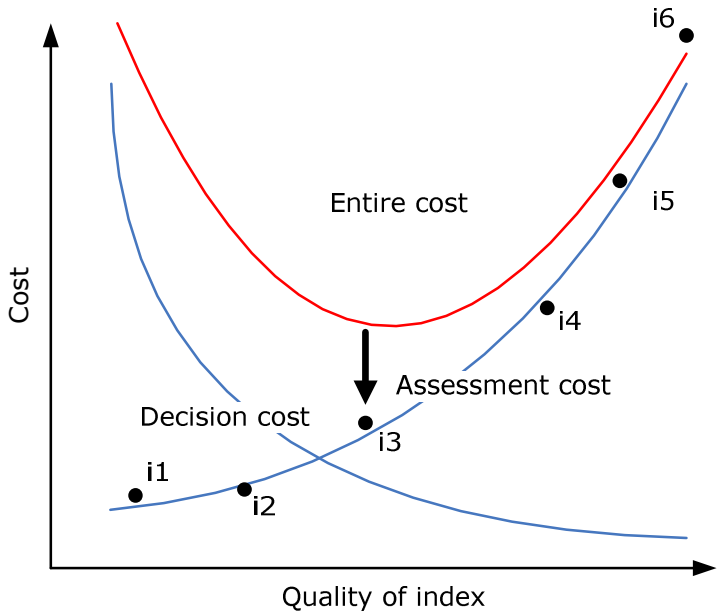


Figure 7.3 **Assessment and decision costs as function of the quality of the index.** The decision cost C_D decreases as function of the index quality, while the assessment cost C_A increases. The entire cost C_E (the red line) results from the sum of C_D and C_A. Theoretically, the optimal index is where the red line reaches its minimum. In practice, we should search among the alternative indices (i1 → i6) available, which are scattered along the C_A line.

7.3.3 Optimisation of the entire cost

As C_D and C_A change in an opposite direction as a function of the index quality, somewhere, the entire cost will reach a minimum. In Figure 7.3, index i3 is closest to this optimum. With the weaker indices i1 and i2 representing personal judgement based on a simple checklist of some essential points, C_A is low (not zero, as superficial investigations require time and budget, which is often underestimated), but C_D is very high, inflating C_E. Conversely, for the stronger indices i4 to i6 (the gold standard index or a full-blown scientific study), C_D is low (not zero), but C_A is too high. In comparison to business as usual (BAU) and a (close to) gold standard, a well-standardised index i3 optimises C_E.

To understand the mechanism, it is instructive to investigate what happens if the negative impact of a wrong decision increases and the decision curve shifts upwards and becomes steeper (Figure 7.4). In this case, the minimum shifts to the right and i4 is now the optimal index. In comparison

to the previous configuration, C_E has increased. This is not because of a higher C_A, but because the shape of C_D describing the decision context has changed. In fact, by choosing for i4 instead of i3, we recuperate part of the extra decision costs. If we would stick to index i3 (arrow 1 in Figure 7.3), C_E would be higher than with index i4 (arrow 2 in Figure 7.3). This result is in agreement with intuition. For more critical decisions, we naturally tend to spend more on investigation. Also, the converse is true. With less critical decisions, a lower quality index suffices. In daily practice, many decisions do not require an index as the additional costs do not warrant this.

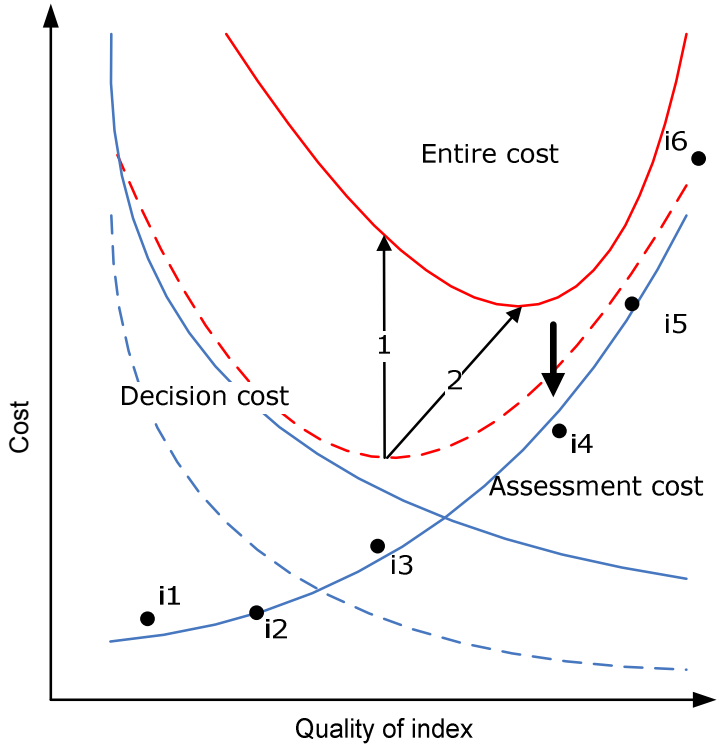


Figure 7.4 **The influence of the decision context on the index choice**. The dashed lines represent the previous decision context (see Figure 7.3). In the new context (full lines), the optimum shifts to the right. Arrow 1 = increase of C_E if we stick to i3; arrow 2 = increase of C_E if we change to i4.

7.3.4 Utility curves are tools to make a decision analysis

Figure 7.5 presents a decision analysis (Vickers, 2008; Steyerberg and Vickers, 2008) for index A and B based on utility curves for the average restoration cost (ARC) for two different values of the prevalence of degradation (20 % and 5 %). We observe that the ARC curves of A and B diverge for an increasing sensitivity (i.e. deciding to restore more sites). Breakeven will be reached in favour of the more expensive index B, when the difference of ARC (in favour of index B) becomes equal to the difference of C_A (in favour of index A): $\Delta ARC = \Delta C_A$. In Figure 7.5, the line segment is set equal to ΔC_A at the point where it is equal to ΔARC. Above this line, it is more favourable to choose for index B; below, index A is to be preferred. This is conform the mechanism explained in Figure 7.4.

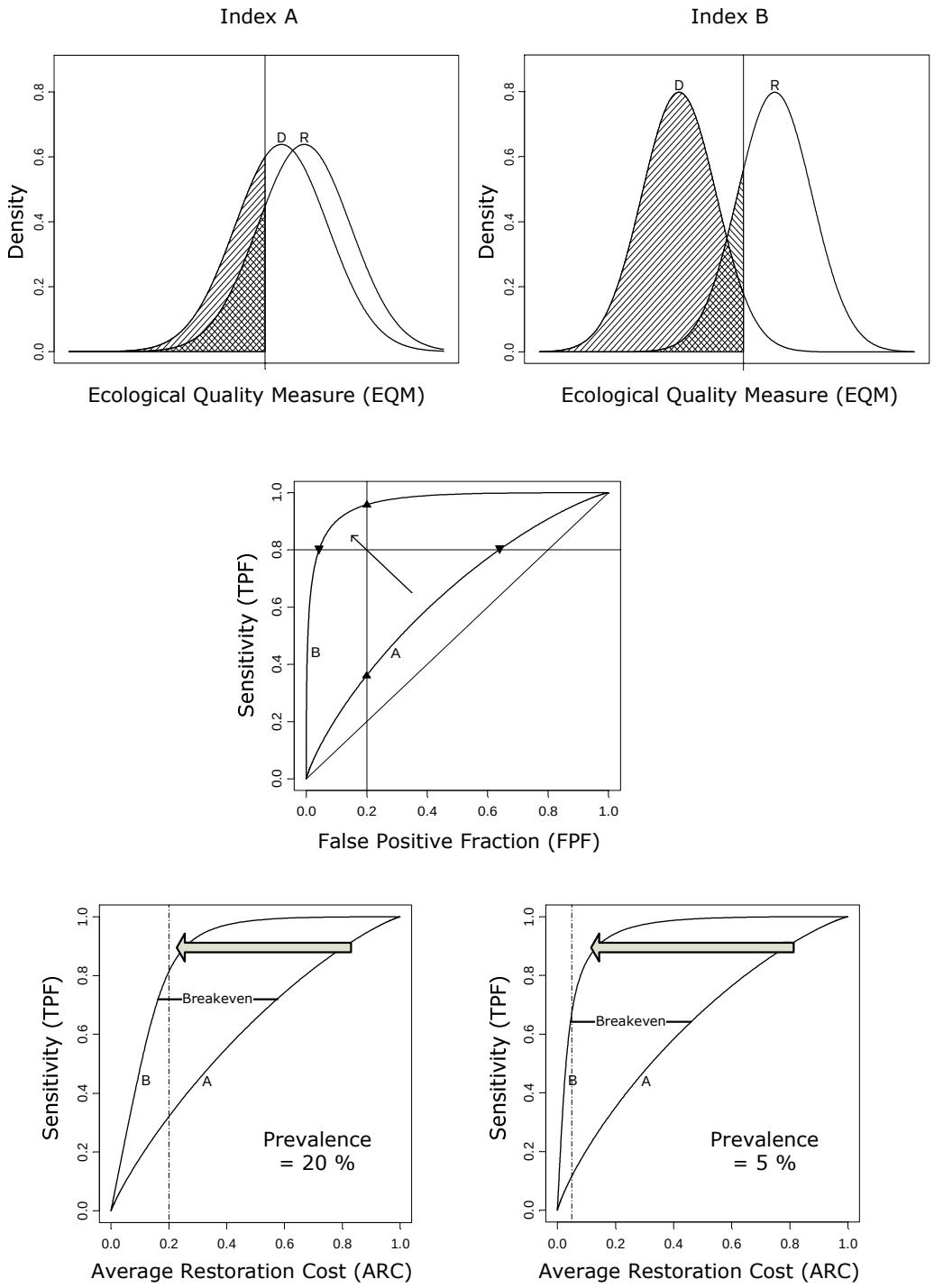


Figure 7.5 **Decision analysis.** Top: Distribution of test variable of index A & B. Middle: ROC curve. Bottom: Average restoration cost (ARC) for a prevalence of 20 % and 5 % (four times as small). The breakeven segment equals the difference in the assessment costs (ΔC_A). Above this segment, it is more advantageous to choose for the more expensive index B as indicated by the grey arrows.

Increasing the ambition level (increasing TPF to restore more sites) is equivalent to moving C_D upwards resulting in a higher quality but requiring a more expensive index. A second observation is that with a smaller prevalence (right panel in Figure 7.5), the breakeven is lower because the ARC curves diverge faster. In a more challenging context with steeper decision curves, strong indices are preferred, again conform the mechanism explained in Figure 7.4.

7.3.5 Marginal costing, an application to IBIs

Marginal costing is a technique to investigate the cost consequences of a change in the production. It makes abstraction of the costs which are fixed in the short run and evaluates whether it is profitable to increase the delivery of goods and services in comparison to the current situation. Our analysis fits in this approach. Increasing the quality of an index implies an increase of the information delivered to the decision maker. We advice to increase the quality as long as the marginal gain. The optimal point is where the decrease of the decision cost is offset by the increase of assessment cost. In a mathematical language, this is where the first derivative of the entire cost equals zero:

$$\frac{dC_E}{dQ} = \frac{dC_D}{dQ} + \frac{dC_A}{dQ} = 0 \quad \Leftrightarrow \quad \frac{dC_A}{dQ} = -\frac{dC_D}{dQ} = \frac{dB_D}{dQ} \quad \Rightarrow \quad \Delta C_A = \Delta B_D$$

In the derivation, $B_D = -C_D$ = the benefit because of decisions. The optimum is where the marginal assessment cost (ΔC_A) equals the marginal decision benefit (ΔB_D). The advantage of working with marginal costs is that no knowledge is necessary about the fixed costs as the first derivative of a constant is zero. We only have to evaluate the changes in comparison to the current situation. For instance, with an existing monitoring program, we should consider only the additional costs and benefits to evaluate whether a new IBI of higher quality is profitable. If there are only minor changes of the fieldwork and the software calculating the IBI, the decision can be based on a tradeoff of these additional costs without considering the full cost picture.

It is important to realise that one should make a distinction between the short and long run. In the long run, many fixed costs and benefits become variable. Not choosing to change something because the costs on the short run seem prohibiting, can be short-sighted. Hence, with marginal costing, it is necessary to carefully define the time horizon to not block off innovation.

In Chapter 5, we searched for the optimal suite of metrics to compose an estuarine biotic index for the Zeeschelde estuary. As all metrics can be derived from the same community data, the cost element does not play a role in the metric selection. Also, changing the index does not involve a major cost on the condition that the software is well developed and can be easily adapted to a new definition, for instance by softcoding the index model instead of hardcoding. In the former case, the model is generically programmed but its parameters can be changed externally. Softcoding has the additional advantage that a wide spectrum of indices can be calculated within the same framework, increasing the efficiency.

In some cases, the cost of additional or new metrics can be substantial. To increase the diagnostic accuracy, we can opt to include metrics that depend on more detailed information and/or a higher catch effort. An example is the decision about the taxonomic resolution. It is relatively easy to

determine fish at the species level, but for other taxa this is not evident. In this case we should investigate whether the genus or even the family is not sufficient. For instance, Gayraud (2003) found that for invertebrate communities in rivers, taxonomic resolution scarcely influenced the quality of the index, hence detailed determinations were not necessary. Many counterexamples can be found. For instance, according to Nijboer *et al.* (2004) and Schmidt-Kloiber *et al.* (2004), the additional information of rare taxa and a higher taxonomic resolution is very valuable in spite of the larger sampling effort. Our point is however that according to the marginal costing principle, we should actively search for the optimal point at which the taxonomic resolution is sufficient for the purpose (Jones, 2008). A similar discussion holds for the combination of IBIs of more than one assemblage. According to the WFD, four biological quality elements should be considered for rivers: fish, macroinvertebrates, phytoplankton and water flora (macrophytes and phytobenthos). For the surveillance monitoring, according to the WFD, all quality elements should be considered, but this is not true for the operational monitoring. In this case, it should be actively searched for an optimal choice.

7.3.6 Activity based cost accounting

Activity based cost accounting (ABC) is an approach to analyse the contribution of the activities of an organisation to its goals. By valuing the activities, ABC aims to obtain a better understanding of the relative importance of the processes of an organisation to prioritise in a more evidence-based way. With our analysis of the tradeoff between the assessment costs of an index and the cost implications of the decisions related to the strength of the index, we are able to quantify – at least to an order of magnitude – how much to invest in the index depending on the management context. Based on this analysis, we do not advocate a strong index in any possible situation. Sometimes, no index is warranted as it does not contribute substantially to better decisions. If the decision costs increase, it pays-off to choose for a stronger index. This general rule is intuitively clear, but not evident to apply in practice. Yet, we offer a calculation scheme to determine the optimal investment in an index.

In our opinion, the crucial point is to assess correctly the contribution of the index to decision making. We only provided a hypothetical model to give insight in the underlying mechanism. To model the fast increasing costs when it is aimed for a gold standard, we assumed a quadratic relationship between the strength of the index and the assessment cost. In the lower region of the parabola, improvements can be realised at a relatively low price. However, gradually the costs to improve the index further become higher and at a certain point no gain is possible. Conversely, reasoning from the high cost end, the parabolic relationship implies that we can gain a lot by simplifying the gold standard. Analysing what is crucial and concentrating on the essential elements, can reduce the costs considerably without loosing much of the diagnostic accuracy. True simplification without loosing the essence, requires a major effort and a thorough understanding, but it can pay off in the long run.

7.3.7 Another view on the precautionary principle (PP)

The Precautionary Principle implies that society should not use lack of full scientific knowledge as a reason to postpone cost-effective preventive measures (Gollier, 2001). As the debate about the PP

is a never ending story (Buhl-Mortensen, 1996; Gray, 1996; Hansson, 1997; McGarvey, 2007), our intention is not to take a position, but to clarify the issue. Compliant to PP, it is often advocated to set the highest safety standards and to increase the sensitivity at the expense of FPs (Underwood and Chapman, 2003). As shown by the utility curves in Figure 7.5, ARC – here to be interpreted as the monetary cost for preventive measures – increases fast above a sensitivity of 90 % for both index A and B. Sometimes, pressure is society is high to attain a zero risk and to control any danger. Figure 7.5 demonstrates that the monetary burden would be very high in this case. However, it is a societal choice to set the risk level. If a sensitivity of 99,9% is chosen to minimise a certain risk, the cost will be disproportionate, unless we succeed in developing a still better (stronger) index C keeping FPF low at a reasonable price. With disproportionate costs, perhaps society avoids one problem, but resources will be exhausted to tackle other risks.

7.3.8 Finally, the potential value of ecological indicators in decision making

Figure 7.6 is inspired on a six-level hierarchical model to assess the efficacy of a clinical test (Fryback and Thornbury, 1991) ranging from the technical efficacy (level 1) to the societal efficacy (level 6). The key feature of the model is that, for a test to be efficacious at a certain level, it must be efficacious at all lower levels (Zhou *et al.*, 2002). First of all, the ecological indicator should be relevant and well constructed (left in Figure 7.6). Without a good theoretical framework, it is naïve to expect powerful indicators. Several authors stress that the success of empirical studies critically depends on a guiding conceptual and/or theoretical scheme (Ford, 2009; Underwood *et al.*, 2000) which should be made explicit (Ford, 2000). The same scientific principle also holds for ecological indicators. It was beyond the scope of this thesis to investigate the ecological value of the concept in depth. However, we recapitulated the ecological rationale and translated the IBI concept in a statistical model (Chapter 2) to improve the construction and hence the diagnostic accuracy of the index (Chapter 5).

At the other end of the chain in Figure 7.6, the question is whether an IBI or any other ecological indicator is effectively used and improves decision making. Turnhout (2003) adopts a science sociology point of view to study this question. She makes clear how an ecological indicator can serve as a boundary object at the science policy interface (Turnhout *et al.*, 2006), facilitating the integration of scientific findings in policy making (Turnhout, 2009). To be successful, the design of ecological indicators should integrate policy objectives and language. We should carefully analyse information needs of policy makers when developing decision tools – boundary objects – in nature conservation (Pullin, 2002; Pullin and Knight, 2003; Pullin *et al.*, 2009). In the same vein, we recommended to pay more attention to the data wishes of policy makers after auditing and revising environmental programmes in a Flemish context (Onkelinx *et al.*, 2006; Onkelinx *et al.*, 2007a; Onkelinx *et al.*, 2007b; Wouters *et al.*, 2008b) resulting in a manual guiding the design of environmental monitoring programmes in a policy context (Wouters *et al.*, 2008a; Onkelinx *et al.*, 2008). In this respect, a strong point of IBIs is that they are asked for and have a place in the management cycle of the Water Framework Directive (Geeraerts and Quataert, 2010). Yet, we should think about how to better integrate IBIs in environmental decision making and how to use them effectively.

In complement to the sociological approach of Turnhout (2003), we focused on the cost-effectiveness of ecological indicators. It is generally stated that indicators are cost-effective (Murtaugh, 1996; Vos *et al.*, 2000), but is this really true? We figured out the mechanism of how ecological indicators indeed are capable to decrease the management costs and/or maximise the realised benefit. By doing so, as depicted in Figure 7.6, we filled in a gap between an ecological approach (left) and a sociological approach (right). Utility curves allow to estimate the potential of the index to improve the cost-effectiveness of the decisions. By taking into account the assessment cost, we can judge the feasibility and choose the best index.

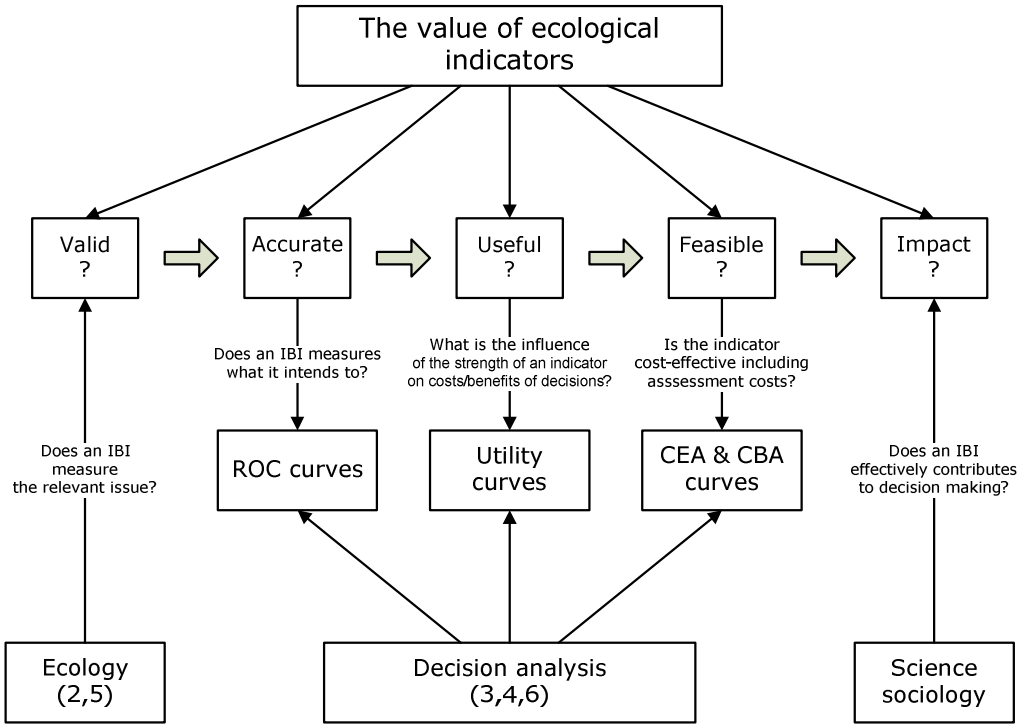


Figure 7.6 **The value of ecological indicators.** Hierarchical scheme to guarantee the efficacy of ecological (and environmental) indicators in environmental policy and conservation biology.

List of publications

Indices of biotic integrity and aquatic ecology

C. Belpaire, G. Goemans, C. Geeraerts, P. Quataert, and K. Parmentier. Pollution fingerprints in eels as models for the chemical status of rivers. *ICES Journal of Marine Science* 65 (8):279-300, 2008.

C. Belpaire, G. Goemans, C. Geeraerts, P. Quataert, K. Parmentier, P. Hagel, and J. De Boer. Decreasing eel stocks: survival of the fattest? *Ecology of Freshwater Fish* 18:197-214, 2009.

J. Breine, I. Simoens, P. Goethals, P. Quataert, D. Ercken, C. Van Liefferinghe, and C. Belpaire. A fish-based index of biotic integrity for upstream brooks in Flanders (Belgium). *Hydrobiologia* 522 (1-3):133-148, 2004.

J. Breine, J. Maes, P. Quataert, E. Van den Bergh, I. Simoens, G. Van Thuyne, and C. Belpaire. A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). *Hydrobiologia* 575:141-159, 2007.

J. Breine, P. Quataert, M. Stevens, F. Ollevier, F. A. M. Volckaert, E. Van den Bergh, and J. Maes. A zone-specific fish-based biotic index as a management tool for the Zeeschelde estuary (Belgium). *Marine Pollution Bulletin* 60 (7):1099-1122, 2010.

E. Degerman, U. Beier, J. Breine, A. Melcher, P. Quataert, C. Rogers, N. Roset, and I. Simoens. Classification and assessment of degradation in European running waters. *Fisheries Management and Ecology* 14 (6):417-426, 2007.

A. S. Harzevili, I. Dooremont, I. Vught, J. Auwerx, P. Quataert, and D. De Charleroy. First feeding of burbot, Lota lota (Gadidae, Teleostei) larvae under different temperature and light conditions. *Aquaculture Research* 35 (1):49-55, 2004.

P. Quataert, J. Breine, and I. Simoens. Evaluation of the European Fish Index: false-positive and false-negative error rate to detect disturbance and consistency with alternative fish indices. *Fisheries Management and Ecology* 14 (6):465-472, 2007.

H. Verreycken, D. Anseeuw, G. Van Thuyne, P. Quataert, and C. Belpaire. The non-indigenous freshwater fishes of Flanders (Belgium): review, status and trends over the last decade. *Journal of Fish Biology* 71:160-172, 2007.

Epidemiology and public health

M. Arbyn, P. Quataert, G. VanHal, and H. Van Oyen. Cervical cancer screening in the Flemish region (Belgium): measurement of the attendance rate by telephone interview. *European Journal of Cancer Prevention* 6 (4):389-398, 1997.

E. Garne, P. Quataert, C. de Vigan, H. Mendizabal, D. Igoe, M. C. Addor, C. M. Bayon, S. Garcia-Minaur, and D. F. Lillis. Congenital diaphragmatic hernia: a European population-based study of epidemiology, prenatal diagnosis and mortality. *Prenatal and Neonatal Medicine* 4 (6):441-447, 1999.

E. Garne, P. Quataert, C. de Vigan, H. Mendizabal, D. Igoe, M. C. Addor, C. Moro-Bayon, S. Garcia-Minaur, and D. Lillis. Congenital diaphragmatic hernia - Prenatal diagnosis and survival. *Pediatric Research* 45 (6):919, 1999.

P. Quataert, B. Armstrong, A. Berghold, F. Bianchi, A. Kelly, M. Marchi, M. Martuzzi, and A. Rosano. Overview on methodological problems associated with cluster detection and investigation. A framework. *European Journal of Epidemiology* 15:821-831, 1999.

S. Van Gucht, K. Van Den Berge, P. Quataert, P. Verschelde, and I. Le Roux. No emergence of echinococcus multilocularis in foxes in Flanders and Brussels anno 2007 - 2008. *Zoonoses and Public Health*, 2010.

H. Van Oyen, J. Tafforeau, H. Hermans, P. Quataert, E. Schiettecatte, L. Lebrun, and L. Bellamammer. The Belgian Health Interview Survey. *Arch Public Health* 55:1-13, 1997.

S. Vyslouzilova, M. Arbyn, H. Van Oyen, S. Drieskens, and P. Quataert. Cervical cancer mortality in Belgium, 1955-1989. A descriptive study. *European Journal of Cancer* 33 (11):1841-1845, 1997.

Forestry and soil research

N. Cools, V. Delanote, X. Scheldeman, P. Quataert, B. De Vos, and P. Roskams. Quality assurance and quality control in forest soil analyses: a comparison between European soil laboratories. *Accreditation and Quality Assurance* 9 (11-12):688-694, 2004.

K. De Cock, B. Lybeer, K. Vander Mijnsbrugge, A. Zwaenepoel, P. Van Peteghem, P. Quataert, P. Breyne, P. Goetghebeur, and J. Van Slycken. Diversity of the willow complex Salix alba - S x rubens - S-fragilis. *Silvae Genetica* 52 (3-4):148-153, 2003.

K. De Cock, K. Vander Mijnsbrugge, P. Quataert, J. Van Huylenbroeck, J. Van Slycken, and E. Van Bockstaele. A morphological study of autochthonous roses (Rosa, Rosaceae) in Flanders. *Acta Horticulturae* 751:305-312, 2007.

P. De Frenne, A. Kolb, K. Verheyen, J. Brune, O. Chabrerie, G. Decocq, M. Diekmann, O. Eriksson, T. Heinken, M. Hermy, Ü. Jõgar, S. Stanton, P. Quataert, R. Zindell, M. Zobel, and B. J. Graae. Unravelling the effects of temperature, latitude and local environment on the reproduction of forest herbs. *Global Ecology and Biogeography* 18:641-651, 2009.

B. De Vos, M. Van Meirvenne, P. Quataert, J. Deckers, and B. Muys. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science Society of America Journal* 69 (2):500-510, 2005.

L. Depypere, K. Vander Mijnsbrugge, K. De Cock, P. Verschelde, P. Quataert, J. Van Slycken, and P. Goetghebeur. Indigenous species of Crataegus (Rosaceae-Maloideae) in flanders (Belgium). An explorative morphometric study. *Belgian Journal of Botany* 139 (2):139-152, 2006.

S. Lettens, B. De Vos, P. Quataert, B. van Wesemael, B. Muys, and J. van Orshoven. Variable carbon recovery of Walkley-Black analysis and implications for national soil organic carbon accounting. *European Journal of Soil Science* 58 (6):1244-1253, 2007.

J. Neirynck, I. A. Janssens, P. Roskams, P. Quataert, P. Verschelde, and R. Ceulemans. Nitrogen biogeochemistry of a mature Scots pine forest subjected to high nitrogen loads. *Biogeochemistry* 91 (2-3):201-222, 2008.

B. Vandecasteele, J. Samyn, P. Quataert, B. Muys, and F. M. G. Tack. Earthworm biomass as additional information for risk assessment of heavy metal biomagnification: a case study for dredged sediment-derived soils and polluted floodplain soils. *Environmental Pollution* 129 (3):363-375, 2004.

B. Vandecasteele, P. Quataert, B. De Vos, and F. M. G. Tack. Assessment of the pollution status of alluvial plains: A case study for the dredged sediment-derived soils along the Leie River. *Archives of Environmental Contamination and Toxicology* 47 (1):14-22, 2004.

B. Vandecasteele, P. Quataert, B. De Vos, F. M. G. Tack, and B. Muys. Foliar concentrations of volunteer willows growing on polluted sediment-derived sites versus sites with baseline contamination levels. *Journal of Environmental Monitoring* 6 (4):313-321, 2004.

B. Vandecasteele, E. Meers, P. Vervaeke, B. De Vos, P. Quataert, and F. M. G. Tack. Growth and trace metal accumulation of two Salix clones on sediment-derived soils with increasing contamination levels. *Chemosphere* 58 (8):995-1002, 2005.

B. Vandecasteele, G. Du Laing, P. Quataert, and F. M. G. Tack. Differences in Cd and Zn bioaccumulation for the flood-tolerant Salix cinerea rooting in seasonally flooded contaminated sediments. *Science of the Total Environment* 341 (1-3):251-263, 2005.

B. Vandecasteele, P. Quataert, and F. M. G. Tack. The effect of hydrological regime on the metal bioavailability for the wetland plant species Salix cinerea. *Environmental Pollution* 135 (2):303-312, 2005.

B. Vandecasteele, P. Quataert, and F. M. G. Tack. Uptake of Cd, Zn and Mn by willow increases during terrestrialisation of initially ponded polluted sediments. *Science of the Total Environment* 380 (1-3):133-143, 2007.

B. Vandecasteele, P. Quataert, G. Genouw, S. Lettens, and F. M. G. Tack. Effects of willow stands on heavy metal concentrations and top soil properties of infrastructure spoil landfills and dredged sediment-derived sites. *Science of the Total Environment* 407:5289-5297, 2009.

A. Vanden Broeck, K. Cox, P. Quataert, E. Van Bockstaele, and J. Van Slycken. Flowering phenology of Populus nigra L., P. nigra cv. italica and P. x canadensis Moench. and the potential for natural hybridisation in Belgium. *Silvae Genetica* 52 (5-6):280-283, 2003.

A. Vanden Broeck, V. Storme, J. E. Cottrell, W. Boerjan, E. Van Bockstaele, P. Quataert, and J. Van Slycken. Gene flow between cultivated poplars and native black poplar (Populus nigra L.): a case study along the river Meuse on the Dutch-Belgian border. *Forest Ecology and Management* 197 (1-3):307-310, 2004.

A. Vanden Broeck, J. Cottrell, P. Quataert, P. Breyne, V. Storme, W. Boerjan, and J. Van Slycken. Paternity analysis of Populus nigra L. offspring in a Belgian plantation of native and exotic poplars. *Annals of Forest Science* 63 (7):783-790, 2006.

V. Vandepitte, P. Quataert, H. Derore, and W. Verstraete. Evaluation of the Gompertz Function to Model Survival of Bacteria Introduced Into Soils. *Soil Biology & Biochemistry* 27 (3):365-372, 1995.

Reports on policy oriented monitoring

Project policy oriented monitoring and quality assurance

C. Belpaire, W. De Cooman, G. Goemans, T. Onkelinx, and P. Quataert. Waterbodem- en paling-polluentenmeetnet: een tandem voor de waterbodemsanering. *Water* 26, 2006.

T. Onkelinx, B. Vandecasteele, G. Genouw, and P. Quataert (2006). Kwaliteitszorg meetnetten: gevalstudie 1 - vergelijking van twee meetnetten i.v.m. verzuring (IBW & VMM).

T. Onkelinx, B. Vandecasteele, G. Genouw, and P. Quataert (2007). Kwaliteitszorg meetnetten: gevalstudie 2 - vergelijking van twee meetnetten i.v.m. waterbodems (IBW & VMM).

T. Onkelinx, J. Wouters, D. Bauwens, and P. Quataert (2007). Kwaliteitszorg meetnetten: gevalstudie 3 - verkennende dimensionering Natura 2000.

T. Onkelinx, P. Verschelde, J. Wouters, D. Bauwens, and P. Quataert (2008). Ontwerp en evaluatie van meetnetten voor het milieu- en natuurbeleid. Partim steekproefgrootteberekeningen en analyse van de kosteneffectiviteit. Brussel: Vlaamse Overheid, Departement Leefmilieu, Natuur en Energie & Instituut voor Natuur- en Bosonderzoek. NBO.M.2008.7:1-234.

P. Quataert and D. Maes (1999). Bouwstenen monitoring (1). Ontwikkeling van concepten, methoden en technieken voor natuurkwaliteitszorg in Vlaanderen aan de hand van een multi-soorten benadering. Bestek MINA/121/1999/01.

P. Quataert and H. Heyrman (1999). Bouwstenen monitoring (2). Ontwerp van een monitoringstrategie voor natuurinrichtingsprojecten op basis van vier proefprojecten. Bestek MINA/121/1999/02.

P. Quataert (2000). Bouwstenen monitoring (3). Onderbouwing en uitwerking van een basismonitoring van de natuurreservaten in Vlaanderen conform het uitvoeringsbesluit i.v.m. de erkenning van de natuurreservaten (29 juni 1999; BS 18 september 1999). Bestek MINA/121/2000/01.

J. Wouters, P. Quataert, T. Onkelinx, and D. Bauwens (2008). Ontwerp en handleiding voor de tweede regionale bosinventarisatie van het Vlaamse Gewest. Brussel: Instituut voor Natuur- en Bosonderzoek (INBO). INBO.R.2008.17: 1-174.

J. Wouters, T. Onkelinx, D. Bauwens, and P. Quataert (2008). Ontwerp en evaluatie van meetnetten voor het milieu- en natuurbeleid. Leidraad voor de opdrachtgever. Brussel: Vlaamse Overheid, Departement Leefmilieu, Natuur en Energie (LNE) & Instituut voor Natuur- en Bosonderzoek (INBO). INBO.M.2008.8: 1-90.

J. Wouters, T. Onkelinx, D. Bauwens, and P. Quataert (2008). Ontwerp en evaluatie van meetnetten voor het milieu- en natuurbeleid. Leidraad voor de meetnetontwerper. Brussel: Vlaamse Overheid, Departement Leefmilieu, Natuur en Energie (LNE) & Instituut voor Natuur- en Bosonderzoek (INBO). NBO.M.2008.7: 1-234.

J. Wouters, P. Quataert, and M. Waterinckx. Een steekproef uit het Vlaamse bos. Bosrevue 21:1-4, 2007.

Indices of biotic integrity and aquatic ecology

C. Geeraerts, G. Goemans, P. Quataert, and C. Belpaire (2007). Ecologische en ecotoxicologische betekenis van verontreinigende stoffen gemeten in paling. Studie uitgevoerd in opdracht van de Vlaamse Milieumaatschappij, MIRA. Brussel, Belgium. Instituut voor Natuur- en Bosonderzoek. INBO/R/2007/40:1-207.

C. Geeraerts and P. Quataert (2011). Discussienota Kaderrichtlijn Water. Vraaganalyse in functie van de revisie van het meetnet Zoetwatervis en een betere afstemming met de ecologische monitoring. Brussel. Instituut voor Natuur- en Bosonderzoek. 1-144 (in voorbereiding).

P. Verschelde, L. Denys, I. Jansen, and P. Quataert (2010). Evaluatie, bijsturing en vervollediging van evaluatiemethoden voor biologische kwaliteitskenmerken (cf. Kaderrichtlijn Water), met bijzondere aandacht voor kwaliteitsborging en (Europese) interkalibratie - partim fytobenthos. Evaluatie van het EKR-criterium voor het fytobenthos. INBO.IR.2009.35: 1-120.

S. Vrielynck, C. Belpaire, A. Stabel, J. Breine, and P. Quataert (2002). De visbestanden in Vlaanderen anno 1840-1950 : een historische schets van de referentietoestand. Groenendaal, Belgium:Instituut voor Bosbouw en Wildbeheer. Rapport-IBW - sectie visserij, 2002.089: 1-271.

Epidemiology and public health

W. Aelvoet, K. Bogaert, F. Capet, and P. Quataert (1997). Gezondheidsindicatoren, 1995. Brussel: Ministerie van de Vlaamse Gemeenschap.

P. Quataert and H. Van Oyen (1995). Gegevensinzameling i.v.m. middelengebruik d.m.v. CATI (Computer Assisted Telephone Interviewing). Rapport-IHE (Instituut voor Hygiëne en Epidemiologie).

P. Quataert and S. Munier (1996). Description of the Routine Statistical Methods used to analyse the EUROCAT Registration Data of Anomalies. Rapport-IHE (Instituut voor Hygiëne en Epidemiologie).

P. Quataert (1997). Methodological Report of the EUROCAT Workgroup on Statistics, June 27-28, 1997. Rapport-IHE (Instituut voor Hygiëne en Epidemiologie).

P. Quataert, H. Van Oyen, and J. Tafforeau (1997). Health Interview Survey: Protocol for the selection of the households and the respondents. Rapport-IHE (Instituut voor Hygiëne en Epidemiologie).

P. Quataert and F. Claeys (1997). Epidemiological Surveillance of the General Population. Blood Lead and Cadmium Levels in Belgium 1996: An Exploratory Data Analysis by Linear Regression. Rapport-IHE (Instituut voor Hygiëne en Epidemiologie).

S. Drieskens, P. Quataert, J. Tafforeau, and H. Van Oyen (1997). Age-Period-Cohort models: trends in mortality from lung cancer in women, Belgium 1971-1990. Arch Publ Health 55: 99-117.

Forestry and soil research

A. Verstraeten, K. Vandekerkhove, and P. Quataert. Bosaanplanting of spontane verbossing? Aanbevelingen voor het beleid en het beheer. Bosrevue 20:1-5, 2007.

N. Cools, V. Delanote, B. De Vos, P. Quataert, P. Roskams, and X. Scheldeman (2003). Quality assurance and quality control in forest soil analysis: 3rd FSCC interlaboratory comparison; convention on long-range transboundary air pollution; international co-operative programme on assessment and monitoring of air pollution effects on forests. Geraardsbergen, Belgium: Instituut voor Bosbouw en Wildbeheer. Rapport-IBW - sectie bosbouw, 2003(018). 1-300.

N. Cools, P. Verschelde, P. Quataert, J. Mikkelsen, and B. De Vos (2006). Quality assurance and quality control in forest soil analysis: 4th FSCC interlaboratory comparison; ICP-Forests. Geraardsbergen, Belgium: Forest Soil Coordinating Centre (FSCC), Research Institute for Nature and Forest. INBO.R.2006.6:1-68.

V. Delanote, N. Cools, B. De Vos, P. Roskams, and P. Quataert (2004). Evaluation of the 3rd FSCC interlaboratory comparison and practical recommendations towards the future. Geraardsbergen, Belgium: Instituut voor Bosbouw en Wildbeheer. Rapport-IBW - sectie bosbouw, 2004(005).:1-55.

T. Onkelinx, H. Van Calster, and P. Quataert (2010). Schaduwmeetnet bosinventarisatie.

G. Sioens, P. Quataert, and P. Roskams (2005). Beschrijvende trendanalyse van de kroontoestand in het bosvitaliteitsmeetnet (level I) in de periode 1987 - 2001. Geraardsbergen, Belgium: Instituut voor Bosbouw en Wildbeheer. Rapport-IBW - sectie bosbouw, 2005.002:1-57.

K. Van Den Berge, P. Roskams, A. Verlinden, P. Quataert, B. Muys, D. Maddelein, J. Zwaenepoel (1990). Analyse van een bosreservaat in een 215-jarig bestand in het Zoniënwoud. Werkgroep SEB (Sociale en Economische betekenis van het Bos). SEB-rapport n° 17.

References

Aarts B.G.W. and Nienhuis P.H., 2003. Fish zonations and guilds as the basis for assessment of ecological integrity of large rivers. Hydrobiologia 500, 157-178.

Agresti A., 2002. Categorical Data Analysis. Wiley-Interscience, Hoboken, New Jersey, 710 p.

Alden R.W., Dauer D.M., Ranasinghe J.A., Scott J.M. and Llansó R.J., 2002. Statistical verification of the Chesapeake Bay Benthic Index of Biotic Integrity. Environmetrics 13, 473-498.

Anderson J.A. and Philips P.R., 1981. Regression, Discrimination and Measurement Models for Ordered Categorical Variables. Journal of the Royal Statistical Society. Series C (Applied Statistics) 30, 22-31.

Andreasen J.K., O'Neill R.V., Noss R. and Slosser N.C., 2001. Considerations for the development of a terrestrial index of ecological integrity. Ecological Indicators 1, 21-35.

Angermeier P.L. and Schlosser I.J., 1987. Assessing biotic integrity of the fish community in a small Illinois stream. North American Journal of Fisheries Management 7, 331-338.

Appelberg M., Bergquist B.C. and Degerman E., 2000. Using fish to assess environmental disturbance of Swedish lakes and streams - a preliminary approach. Verhandlungen der Internationalen Vereinigung für Theoretische und Angewandte Limnologie 27, 311-315.

Attrill M.J., 2002. Community-level indicators of Stress in Aquatic Ecosystems. In: Adams S.M. (Ed.)., Biological Indicators of Aquatic Ecosystem Stress. American Fisheries Society, pp. 473-508.

Attrill M.J. and Depledge M.H., 1997. Community and population indicators of ecosystem health: targeting links between levels of biological organisation. Aquatic Toxicology 38, 183-197.

Aubry A. and Elliott M., 2006. The use of environmental integrative indicators to assess seabed disturbance in estuaries and coasts: application to the Humber estuary, UK. Marine Pollution Bulletin 53, 175-185.

Baattrup-Pedersen A., Springe G., Riis T., Larsen S.E., San-Jensen K. and Kjellerup-Larsen L.M., 2008. The search for reference conditions for stream vegetation in northern Europe. Freshwater Biology 53, 1890-1901.

Bady P. and Pont D., 2008. Improvement and Spatial extension of the European Fish Index (EFI+) - Workpackage 3.2 - Evaluation of the existing European Fish Index.

Bady P., Pont D., Logez M. and Veslot J., 2009. Improvement and Spatial extension of the European Fish Index (EFI+) - Workpackage 4., 180 p.

Baeyens W., van Eck B., Lambert C., Wollast R. and Goeyens L., 1998. General description of the Scheldt estuary. Hydrobiologia 366, 1-14.

Bailey R.C., Norris R.H. and Reynoldson T.B., 2004. Bioassessment of Freshwater Ecosystems: Using the Reference Condition Approach. Springer, 170 p.

Bamber D., 1975. The area above the ordinal dominance graph and the area below the receiver operating graph. Journal of Mathematical Psychology 12, 387-415.

Beard G.R., Scott W.A. and Adamson J.K., 1999. The value of consistent methodology in long-term environmental monitoring. Environmental Monitoring and Assessment 54, 239-258.

Bedoya D., Novotny V. and Manolakos E.S., 2009. Instream and offstream environmental conditions and stream biotic integrity: Importance of scale and site similarities for learning and prediction. Ecological Modelling 220, 2393-2406.

Begg C.B. and Greenes R.A., 1983. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39, 207-215.

Beier U., Degerman E., Melcher A., Rogers C. and Wirlöf H., 2007. Processes of collating a European fisheries database to meet the objectives of the European Union Water Framework Directive. Fisheries Management and Ecology 14, 407-416.

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

Benyi S.J., Hollister J.W., Kiddon J.A. and Walker H.A., 2009. A process for comparing and interpreting differences in two benthic indices in New York Harbor. Marine Pollution Bulletin 59, 65-71.

Blocksom K.A., 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands streams. Environmental Management 31, 670-682.

Borg I. and Groenen P.J.F., 2005. Modern multidimensional scaling. Theory and Applications. Springer, New York, 614 p.

Borja Á., Bald J., Franco J., Laretta A., Muxika I., Revilla M., Rodríguez J.G., Solaun O., Uriarte A. and Valencia V., 2009a. Using multiple ecosystem components in assessing ecological status in Spanish (Basque Country) Atlantic marine waters. Marine Pollution Bulletin 59, 65-71.

Borja Á. and Dauer D.M., 2008. Editorial. Assessing the environmental quality status in estuarine and coastal systems: comparing methodologies and indices. Ecological Indicators 8, 331-337.

Borja Á. and Elliott M., 2007. What does 'goed ecological potential' mean within the European Water Framework directive? Marine Pollution Bulletin 54, 1559-1564.

Borja Á., Franco J., Valencia V., Bald J., Muxika I., Belzunce M.J. and Soluan O., 2004. Implementation of the European water framework directive from the Basque country (northern Spain): a methodological approach. Marine Pollution Bulletin 48, 209-218.

Borja Á., Galparsoro I., Solaun O., Muxika I., Tello E.M., Uriarte A. and Valencia V., 2006. The European Water Framework Directive and the DPSIR, a methodological approach to assess the risk of failing to achieve good ecological status. Estuarine, Coastal and Shelf Science 66, 84-96.

Borja Á., Miles A., Occhipinti-Ambrogi A. and Berg T., 2009b. Current status of macroinvertebrate methods used for assessing the quality of European marine waters: implementing the Water Framework Directive. Hydrobiologia 633, 181-196.

Borja Á., Ranasinghe A. and Weisberg S.B., 2009c. Editorial. Assessing ecological integrity in marine waters, using multiple indices and ecosystem components: Challenges for the future. Marine Pollution Bulletin 59, 1-4.

Borja Á., Tueros I., Belzunce M.J., Galparsoro I., Garmendia J.M., Revilla M., Solaun O. and Valencia V., 2008. Investigative monitoring within the European Water Framework Directive: a coastal blast furnace slag disposal, as an example. Journal of Environmental Monitoring 10, 453-462.

Bouckaert G., Van Roosbroek S., Vervaet C. and Demuzere S., 2009. Werken aan kwaliteit: Een praktische gids voor kwaliteitsmanagement in de publieke sector. Vanden Broele, Brugge, 278 p.

Boulton A.J., 1999. An overview of river health assessment: philosophies, practice, problems and prognosis. Freshwater Biology 41, 469-479.

Box G.E.P., 1980. Sampling and Bayes' Inference in Scientific Modelling and Robustness. Journal of the Royal Statistical Society. Series A (Statistics in Society) 143, 383-430.

Breine J., Goethals P., Simoens I., Ercken D., Van Liefferinge C., Verhaegen G., Belpaire C., De Pauw N., Meire P.M. and Ollevier F., 2001. De visindex als instrument voor het meten van de biotische integriteit van de Vlaamse binnenwateren. Instituut voor Bosbouw en Wildbeheer, Groenendaal, 173 p.

Breine J., Maes J., Quataert P., Van den Bergh E., Simoens I., Van Thuyne G. and Belpaire C., 2007. A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). Hydrobiologia 575, 141-159.

Breine J., Quataert P., Stevens M., Ollevier F., Volckaert F.A.M., Van den Bergh E. and Maes J., 2010. A zone-specific fish-based biotic index as a management tool for the Zeeschelde estuary (Belgium). Marine Pollution Bulletin 60, 1099-1122.

Breine J., Simoens I., Goethals P., Quataert P., Ercken D., Van Liefferinghe C. and Belpaire C., 2004. A fish-based index of biotic integrity for upstream brooks in Flanders (Belgium). Hydrobiologia 522, 133-148.

Brosi B.J. and Biber E.G., 2009. Statistical inference, Type II error, and decision making under the US Endangered Species Act. Frontiers in Ecology and the Environment 7, 487-494.

Brown M.T. and Ulgiati S., 2005. Emergy, Transformity, and Ecosystem Health. In: Jørgensen S.E., Costanza R., Xu F.-L. (Eds.), Handbook of Ecological Indicators for Assessment of Ecosystem Health. CRC Press, Boca Raton, London, New York, Singapore, pp. 333-352.

Brys R., Ysebaert T., Escaravage V., Van Damme S., Van Braeckel A., Vandevoorde B. and Van den Bergh E., 2005. Afstemmen van referentiecondities en evaluatiesystemen in functie van de KRW: afleiden en beschrijven van systeemeigen referentieomstandigheden en/of maximaal ecologisch potentieel in elk Vlaams waterlichaamtype, vanuit de - overeenkomstig de Kaderrichtlijn Water - ontwikkelde beoordelingssystemen voor biologische kwaliteitselementen. Instituut voor Natuurbehoud, Brussel, Belgium.

Buffagni A., Erba S., Birk S., Cazzola M., Feld C., Ofenböck T., Murray-Bligh J., Furse M.T., Clarke R.T., Hering D., Soszka H. and Van de Bund W., 2005. Towards European intercalibration for the Water Framework Directive: procedures and examples for different river types from the E.C. Project STAR, 468 p.

Buhl-Mortensen L., 1996. Type-II statistical errors in environmental science and the precautionary principle. Marine Pollution Bulletin 32,

Bunn S.E. and Davies P.M., 2000. Biological processes in running waters and their implications for the assessment of ecological integrity. Hydrobiologia 422-423, 61-70.

Bunn S.E., Davies P.M. and Mosisch T.D., 1999. Ecosystem measures of river health and their response to riparian and catchment degradation. Freshwater Biology 41, 333-345.

Burt T.P., Howden N.J.K., Worrall F. and Whelan M.J., 2008. Importance of long-term monitoring for detecting environmental change: lessons from a lowland river in south east England. Biogeosciences 5, 1529-1235.

Cairns J., McCormick P.V. and Niederlehner B.R., 1993. A proposed framework for developing indicators of ecosystem health. Hydrobiologia 263, 1-44.

Callahan J.T., 1984. Long-Term Ecological Research. BioScience 34, 363-367.

Campbell G., 1994. General methodology I. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statistics in Medicine 13, 499-508.

Carver R.P., 1993. The Case Against Statistical Significance Testing, Revisited. Journal of Experimental Education 61,

Caughlan L. and Oakley K.L., 2001. Cost considerations for long-term ecological monitoring. Ecological Indicators 1, 123-134.

Chambers J.M., Cleveland W.S., Kleiner B. and Tukey A., 1983. Graphical methods for data analysis. Wadsworth, Belmont, California, 336 p.

Chessman B.C. and McEvoy P.K., 1997. Towards diagnostic biotic indices for river macroinvertebrates. Hydrobiologia 364, 169-182.

Claeskens G. and Hjort N.L., 2008. Model selection and model averaging. Cambridge, 320 p.

Clark J.S. and Bjørnstad O.N., 2004. Population time series: process variability, observation errors, missing values, lags, and hidden states. Ecology 85, 3140-3150.

Clarke R.T., Furse M.T., Gunn R.J.M., Winder J.M. and Wright J.F., 2002. Sampling variation in macroinvertebrate data and implications for river quality indices. Freshwater Biology 47, 1735-1751.

Clarke R.T., Wright J.F. and Furse M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling 160, 219-233.

Cleveland W.S. and Devlin S.J., 1988. Locally weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association 83, 596-610.

Coates S., Waugh A., Anwar A. and Robson M., 2007. Efficacy of a multi-metric fish index as an analysis tool for the transitional fish component of the Water Framework Directive. Marine Pollution Bulletin 55 (Spec. Issue 1-6), 225-240.

Cochran W.G., 1983. Planning & Analysis of Observational Studies. John Wiley & Sons, Inc., New York, 145 p.

Common Implementation Strategy (CIS), 2003. Guidance on Monitoring for the Water Framework Directive.

Cummins K.W., 1988. Rapid bioassessment using functional analysis of running water invertebrates. In: Simon T.P., Holst L.L., Shepard L.J. (Eds.), Proceedings of the First National Workshop on Biological Criteria. EPA-905/9-89/003. Chicago, U.S. Environmental Protection Agency, pp. 49-54.

D'Agostino R.B. and Stephens M.A., 1986. Goodness-of-Fit Techniques. Marcel Dekker, New York, 560 p.

Dauer D.M. and Llansó R.J., 2003. Spatial scales and probability based sampling in determining levels of benthic community degradation in the Chesapeake Bay. Environmental Monitoring and Assessment 81, 175-186.

Dauvin J.-C. and Ruellet T., 2009. The estuarine quality paradox: Is it possible to define an ecological quality status for specific modified and naturally stressed estuarine ecosystems? Marine Pollution Bulletin 59, 38-47.

Davies S.P. and Jackson S.K., 2006. The Biological Condition Gradient: a descriptive model for interpreting change in aquatic ecosystem. Ecological Applications 16, 1251-1266.

Davis W.S. and Simon T.P., 1995. Biological Assessment and Criteria. Tools for Water Resource Planning and Decision Making. Lewis Publishers, Boca Raton, London, Tokyo, 415 p.

Davison A.C. and Hinkley D., 2006. Bootstrap Methods and their Application. Cambridge,

Declerck S., Vandekerckhove J., Johansson L., Muylaert K., Conde-Porcuna J.M., Van Der Gucht K., Pérez-Martínez C., Lauridsen T., Schwenk K., Zwart G., Rommens W., López-Ramos J., Jeppesen E., Vyverman W., Brendonck L. and De Meester L., 2005. Multi-group biodiversity in shallow lakes along gradients of phosporous and water plant cover. Ecology 86, 1905-1915.

Declerck S., Vanderstukken M., Pals A., Muylaert K. and De Meester L., 2007. Plankton biodiversity along a gradient of productivity and its mediation by macrophytes. Ecology 88, 2199-2210.

Degerman E., Beier U., Breine J., Melcher A., Quataert P., Rogers C., Roset N. and Simoens I., 2007. Classification and assessment of degradation in European running waters. Fisheries Management and Ecology 14, 417-426.

Deming W.E., 1982. Out of the Crisis. MIT Press, Cambridge, MA, 507 p.

Denys L., 2006a. Assessment of phytobenthos for the Water Framework Directive (WFD) in Flanders, Belgium. Descriptive summary of the proposed methodology. Institute for Nature and Forest Research, Brussels, Belgium

Denys L., 2006b. Validation and revision of the reference concept for river phytobenthos in Belgium – Flanders proposed for the European Water Framework Directive, based on diatom assemblages from reference sites in the Central-Baltic GIG region. Institute for Nature and Forest Research, Brussels, Belgium, 8 p.

Desbiens N.A., 2004. The presence of hypotheses in the medical literature. American Journal of the Medical Sciences 328, 319-322.

Dickson B. and Cooney R., 2005. Biodiversity and the Precautionary Principle: Risk and Uncertainty in Conservation and Sustainable Use. Earthscan, 314 p.

Dodd L.E. and Pepe M.S., 2003. Partial AUC estimation and regression. Biometrics 59, 614-623.

Dos Santos D.A., Molineri C., Reynaga M.C. and Basualdo C., 2011. Which index is the best to assess stream health? Ecological Indicators 11, 582-589.

Duarte C.M., Conley D.J., Carstensen J. and Sánchez-Camacho M., 2009. Return to neverland: shifting baselines affect eutrophication restoration targets. Estuaries and Coasts 32, 29-36.

Dufour S. and Piegay H., 2009. From the myth of a lost paradise to targeted river restoration: forget natural references and focus on human benefits. River Research and Applications 25, 568-581.

Dufrêne M. and Legendre P., 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67, 345-366.

Efron B., 1987. Better Bootstrap Confidence Intervals. Journal of the American Statistical Association 82, 171-185.

Efron B. and Tibshirani R., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York, London, 456 p.

Ellenberg H., Weber H.E., Düll R., Wirth V. and Werner W., 2001. Zeigerwerte der Pflanzen in Mitteleuropa. Verlag Erich Goltze, 216 p.

Elliott M., 2002. The role of the DPSIR approach and conceptual models in marine environmental management: an example of offshore wind power. Marine Pollution Bulletin 44, iii-viii.

Elliott M., Nedwell S., Jones N.V., Read S.J., Cutts N.D. and Hemingway K.L., 1998. Intertidal sand and mudflats & subtidal mobile sandbanks (volume II). An overview of dynamic and sensitivity characteristics for conservation management of marine SACs. Scottish Association for Marine Science (UK Marine SACs project), 151 p.

Elliott M. and Quintino V., 2007. Viewpoint. The Estuarine Quality Paradox, environmental homeostasis and the difficulty of detecting anthropogenic stress in natural stressed areas. Marine Pollution Bulletin 54, 640-645.

Elliott M., Whitfield A.K., Potter I.C., Blaber S.J.M., Cyrus D.P., Nordlie F.G. and Harrison T.D., 2007. The guild approach to categorizing estuarine fish assemblages: a global review. Fish and Fisheries 8, 241-268.

Ellis J. and Adriaenssens V., 2006. Uncertainty estimation for monitoring results by the WFD biological classification tools.

Ellison A.M., 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6, 1036-1046.

Engel B., Backer J. and Buist W., 2009. Evaluation of the Accuracy of Diagnostic Tests From Repeated Measurements Without a Gold Standard. Journal of Agricultural, Biological, and Environmental Statistics

Fairweather P.G., 1999. State of environment indicators of 'river health': exploring the metaphor. Freshwater Biology 41, 211-220.

Falcone J.A., Carlisle D.M. and Weber LC., 2010. Quantifying human disturbance in watersheds: Variable selection and performance of a GIS-based disturbance index for predicting the biological condition of perennial streams. Ecological Indicators 10, 264-273.

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

Fellows C.S., Clapcott J.E., Udy J.W., Bunn S.E., Harch B.D., Smith M.J. and Davies P.M., 2006. Benthic Metabolism as an Indicator of Stream Ecosystem Health. Hydrobiologia 572, 71-87.

Ferreira J.G., Vale C., Soares C.V., Salas F., Stacey P.E., Bricker S.B., Silva M.C. and Marques J.C., 2007. Monitoring of coastal and transitional waters under the EU Water Framework Directive. Environmental Monitoring and Assessment

Fidler F., Burgman M.A., Cumming G., Buttrose R. and Thomason N., 2006. Impact of criticism of null-hypothesis signifance testing on statistical reporting practices in conservation biology. Biological Conservation 20, 1539-1544.

Field S.A., Tyre A.J., Jonzén N., Rhodes J.R. and Possingham H.P., 2004. Minimizing the cost of environmental management decisions by optimizing statistical thresholds. Ecology Letters 7, 669-675.

Fisher I., 1930. The theory of interest. Macmillan, New York

Fleiss J.L., 2003. Statistical methods for rates and proportions. Wiley, New York, 800 p.

Flotemersch J.E., Stribling J.B. and Paul M.J., 2006. Concepts and approaches for the bioassessment of non-wadeable streams and rivers. Office of Research and Development, US Environmental Protection Agency, Cincinnati, Ohio

Ford E.D., 2000. Scientific Method for Ecological Research. Cambridge University Press, Cambridge, 564 p.

Ford E.D., 2009. The importance of a research data statement and how to develop one. Annales zoologici Fennici 46, 82-92.

Fore L.S., 2003. Developing Biological Indicators: Lessons learned from Mid-Atlantic Streams., 42 p.

Fox D.R., 2001. Environmental power analysis - a new perspective. Environmetrics 12, 437-449.

Franco A., Elliott M., Franzoi P. and Torricelli P., 2008. Life strategies of fishes in European estuaries: the functional guild approach. Marine Ecology Progress Series 354, 219-228.

Franklin J.F., Bledsoe C.S. and Callahan J.T., 1990. Contributions of the Long-Term Ecological Research Program. BioScience 40, 509-523.

Fryback D.G. and Thornbury J.R., 1991. The efficacy of diagnostic imaging. Medical Decision Making 11, 88-94.

Furse M.T., Hering D., Brabec K., Buffagni A., Sandin L. and Verdonschot P.F.M., 2006a. The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods., 555 p.

Furse M.T., Hering D., Moog O., Verdonschot P.F.M., Johnson R.K., Brabec K., Gritzalis K., Buffagni A., Pinto P., Friberg N., Murray-Bligh J., Kokes J., Alber R., Usseglio-Polatera P., Haase P., Sweeting R., Bis B., Szoszkiewicz K., Soszka H., Springe G., Sporka F. and Krno I., 2006b. The STAR project: context, objectives and approaches. Hydrobiologia 566, 3-29.

Gard M.F., 2002. Effects of sediment loads on the fish and invertebrates of a Sierra Nevada river, California. Journal of Aquatic Ecosystems Stress and Recovery 9, 227-238.

Gayraud S., Statzner B., Bady P., Haybach A., Schöll F., Usseglio-Polatera P. and Bacchi M., 2003. Invertebrate traits for the biomonitoring of large European rivers: an initial assessment of alternative metrics. Freshwater Biology 48, 2045-2064.

Geeraerts C. and Quataert P., 2010. Discussienota Kaderrichtlijn Water. Vraaganalyse in functie van de revisie van het meetnet Zoetwatervis en een betere afstemming met de ecologische monitoring. Instituut voor Natuur- en Bosonderzoek, Brussel, 144 p.

Gill J., 1999. The insignificance of Null Hypothesis Significance Testing. Political Research Quarterly

Goffaux D., Roset N., Breine J. and De Leeuw J.J., 2001. A Biotic Index of Fish Integrity (IBIP) to Evaluate the Ecological Quality of Lotic Ecosystems - Application to the Meuse River Basin. University of Leuven, Leuven, 170 p.

Gold M.R., Siegel J.E., Russell L. and Weinstein M.C., 1996. Cost-effectiveness in health and medicine. Oxford University Press, New York, 456 p.

Gollier C., 2001. Should we beware of the Precautionary Principle? Economic Policy 16, 301-328.

Gray J.S., 1996. Environmental science and a precautionary approach revisited. Marine Pollution Bulletin 32, 532-534.

Gray J.S., 1990. Statistics and the precautionary principle. Marine Pollution Bulletin 21, 174-176.

Hale S.S., Benyi S.J., Strobel C.J., Kiddon J.A., Hollister J.W., Walker H.A. and Heltshe J.F., 2007. Benthic Indices: Developing, Evaluating, and Using Measures of Benthic Condition for Northeast Coastal Waters (ECO MYP).

Hale S.S. and Heltshe J.F., 2008. Signals from the benthos: Development and evaluation of a benthic index for the nearshore Gulf of Maine. Ecological Indicators 8, 338-350.

Hanley J.A. and McNeil B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29-36.

Hansson S.O., 1997. The Limits of Precaution. Foundations of Science 2, 393-306.

Harrell F.E., Lee K.L. and Mark D.B., 1996. Tutorial in biostatistics: Multivariable prognostic models. Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15, 361-387.

Haunschmid R., Wolfram G., Spindler T., Honsig-Erlenburg W., Wimmer R., Jagsch A., Kainz E., Hehenwarter K., Wagner B., Konecny R., Riedmüller R., Ibel G., Sasano B. and Schotzko N., 2006. Erstellung einer fischbasierten Typologie österreichischer Fließgewässer sowie einer Bewertungsmethode des fischökologischen Zustandes gemäß EU-Wasserrahmenrichtlinie (In German). Bundesamt für Wasserwirtschaft, Wien, 94 p.

Heffner R.A., Butler M.J. and Reilly C.K., 1996. Pseudoreplication revisited. Ecology

Hering D., Borja Á., Carstensen J., Carvalho L., Elliott M., Feld C.K., Heiskanen A.-S., Johnson R.K., Moe J., Pont D., Solheim A.L. and Van de Bund W., 2010. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of the Total Environment 408, 1-4007.

Hering D., Feld C., Moog O. and Ofenböck T., 2006a. Cookbook for the development of a Multimetric Index for biological condition of aquatic ecosystems: Experiences from the European AQEM and STAR projects and related initiatives. Hydrobiologia 566, 311-324.

Hering D., Johnson R.K. and Buffagni A., 2006b. Linking organism groups - major results and conclusions from the STAR project. Hydrobiologia 566, 109-113.

Hering D., Johnson R.K., Kramm S., Schmutz S., Szoszkiewicz K. and Verdonschot P., 2006c. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organisms response due to stress. Freshwater Biology 51, 1757-1785.

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

Hertzman P.A., Clauw D.J., Duffy J., Medsger T.A. and Feinstein A.R., 2001. Rigorous new approach to constructing a "gold standard" for validating new diagnostic criteria, as exemplified by the eosinophilia-myalgia syndrome. Archives of Internal Medicine 161, 2301-2306.

Hill M.O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54, 427-432.

Hill M.O., Mountford J.O., Roy D.B. and Bunce R.G.H., 1999. Ellenberg's Indicator Values for British plants. ITE / DETR, Abbots Ripton, 46 p.

Hill M.O., Mountford J.O., Roy D.B. and Bunce R.G.H., 2009. Extending Ellenberg's indicator values to a new area: an algorithmic approach. Journal of Applied Ecology 37, 3-15.

Hoaglin D.C., Mosteller F. and Tukey J.W., 1983. Understanding robust and exploratory data analysis. John Wiley and Sons, New York, 445 p.

Holl K.D., 2010. Writing for an International Audience. Restoration Ecology 18, 135-137.

Hosmer D.W. and Lemeshow S., 2000. Applied Logistic Regression. John Wiley & Sons, Inc., New York, 375 p.

Hubbard R., Vetter D.E. and Little E.L., 1998. Replication in strategic management: scientific testing for validity, generalizability, and usefulness. Strategic Management Journal 19, 243-254.

Hughes R.M., Kaufmann P.R., Herlihy A.T., Kincaid T.M., Reynolds L. and Larsen D.P., 1998. A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences 55, 1618-1631.

Hughes R.M. and Oberdorff T., 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, Washington, DC, pp. 62-74.

Hughes R.M., Paulsen S.G. and Stoddard J.L., 2000. EMAP-Surface Waters : a multiassemblage, probability survey of ecological integrity in the U.S.A. Hydrobiologia 422/423, 429-443.

Hui S.L. and Walter S.D., 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167-171.

Hurlbert S.H., 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52, 577-586.

Hurlbert S.H., 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54, 187-211.

Irz P., de Bortoli J., Michonneau F., Whittier T.R., Oberdorff T. and Argillier C., 2008. Controlling for natural variability in assessing the response of fish metrics to human pressures for lakes in north-east USA. Aquatic Conservation: Marine and Freshwater Ecosystems 18, 633-646.

Jeppesen E., Søndergaard M., Jensen J.-P., Havens K., Anneville O., Carvalho L., Coveney M.F., Deneke R., Dokulil M.T., Foy B., Gerdeaux D., Hampton S.E., Hilt S., Kangur K., Köhler J., Lammens E., Lauridsen T.l., Manca M., Miracle M.R., Moss B., Nõges P., Persson G., Phillips G., Portielje R., Romo S., Schelske C.L., Straile D., Tatrai I., Wille E. and Winder M., 2005. Lake response to reduced nutrient loading - an analysis of contemporary long-term data from 35 case studies. Freshwater Biology 50, 1747-1771.

Jiang Y., Metz C.E. and Nishikawa R.M., 1996. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 201, 475-750.

Johnson R.K., Furse M.T., Hering D. and Sandlin L., 2007. Ecological relationships between stream conditions and spatial scale: implications for designing catchment-level monitoring programmes. Freshwater Biology 52, 939-958.

Johnson R.K. and Hering D., 2009. Response of taxonomic groups in streams to gradients in resource and habitat characteristics. Journal of Applied Ecology 46, 175-186.

Johnson R.K., Hering D., Furse M.T. and Verdonschot P., 2006. Indicators of ecological change: comparison of the early response of four organism groups to stress gradients. Hydrobiologia 566, 139-152.

Jones F.C., 2008. Taxonomic sufficiency: the influence of taxonomic resolution on freshwater bioassessments using benthic macro invertebrates. Environmental Reviews 55, 45-69.

Jones H.P. and Schmitz O.J., 2009. Rapid recovery of damaged ecosystems. PLoS ONE 4, 1-6.

Jongman R.H.G., ter Braak C.J.F. and van Tongeren O.F.R., 1995. Data analysis in community and landscape ecology. Cambridge University Press, 299 p.

Jordan S.J. and Vaas P.A., 2000. An index of ecosystem integrity for Northern Chesapeake Bay. Environmental Science & Policy 3, 59-88.

Jørgensen S.E., Costanza R. and Xu F., 2005. Handbook of Ecological Indicators for Assessment of Ecosystem Health. CRC Press, Boca Raton, 439 p.

Jørgensen S.E. and Svirezhev Y., 2004. Towards a Thermodynamic Theory for Ecological Systems. Pergamon,

Kaczmarek K. and Ottitsch A., 2004. Programme evaluation in public sector management practice. In: Buttoud G., Tikkanen I., Solberg B., Pajari B. (Eds.), The evaluation of forest policies and programmes, EFI Proceedings 52. European Forest Institute (EFI),

Kagel J.H. and Roth A.E., 1995. The Handbook of Experimental Economics. Princeton University Press, Princeton, New Jersey, 721 p.

Kaika M., 2003. The Water Framework Directive: a New Directive for a Changing Social, Political and Economic European Framework. European Planning Studies 11, 299-316.

Kail J. and Hering D., 2005. Using large wood to restore streams in Central Europe: potential use and likely effects. Landscape Ecology 2005, 755-772.

Kallis G. and Butler D., 2001. The EU water framework directive: measures and implications. Water Policy 3, 125-142.

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

Karr J.R. and Chu E.W., 1999. Restoring life in running waters: better biological monitoring. Island Press, Washington ; California, 206 p.

Karr J.R. and Chu E.W., 1997. Biological monitoring: Essential foundation for ecological risk assessment. Human and Ecological Risk Assessment 3, 993-1004.

Karr J.R. and Dudley D.R., 1981. Ecological perspectives on water quality goals. Environmental Management 5, 55-68.

Karr J.R., Fausch K.D., Angermeier P.L., Yant P.R. and Schlosser I.J., 1986. Assessing biological integrity in running waters. A method and its rationale., 25 p.

Kay J.J., 1991. A Non-equilibrium Thermodynamic Framework for Discussing Ecosystem Integrity. Environmental Management 15, 483-495.

Kay J.J. and Schneider E.D., 1992. Thermodynamics and measurements of ecosystem integrity. In: McKenzie D. (Ed.)., Ecological Indicators. Elsevier, Amsterdam, pp. 159-182.

Keiding N. and Budtz-Jørgensen E., 2005. The Precautionary Principle and Statistical Approaches to Uncertainty. Human and Ecological Risk Assessment 11, 201-207.

Kelly M.G., Haigh A., Collette J. and Zgrundo A., 2009. Effect of environmental improvements on the diatoms of the River Axe, southern England. Fottea 9, 343-349.

Kennish M.J., 2002. Environmental threads and environmental future of Estuaries. Environmental Conservation 29, 78-107.

Kesminas V. and Virbickas T., 2000. Application of an adapted index of biotic integrity to rivers of Lithuania. Hydrobiologia 422 - 423, 257-270.

Kestemont P., Didier J., Depiereux E. and Micha J.C., 2000. Selecting ichthyological metrics to assess river basin ecological quality. Archiv für Hydrobiologie Supplementband Monographic Studies 121, 321-348.

Kish L., 1987. Statistical design for research. John Wiley, New York

Kruskal W. and Mosteller F., 1980. Representative Sampling, IV: the History of the Concept in Statistics, 1895 - 1939. International Statistical Review 48, 169-195.

Kruskal W. and Mosteller F., 1979c. Representative Sampling, III: the Current Statistical Literature. International Statistical Review 47, 245-265.

Kruskal W. and Mosteller F., 1979a. Representative Sampling, I: Non-scientific Literature. International Statistical Review 47, 13-24.

Kruskal W. and Mosteller F., 1979b. Representative Sampling, II: Scientific Literature, Excluding Statistics. International Statistical Review 47, 111-127.

Kutner M.H., Nachtsheim C.J., Neter J. and Li W., 2005. Applied Linear Statistical Models. McGraw-Hill/Irwin, New York, 1396 p.

Laudan L., 1977. Progress and its problems: Toward a theory of scientific growth. University of California Press, Berkeley, CA, 257 p.

Legendre P. and Legendre L., 1998. Numerical ecology. Elsevier, Amsterdam, 853 p.

Lehmann E.L., 1975. Nonparametrics. Holden-Day, Inc., San Francisco

Lemons J., Shrader-Frechette K. and Cranor C., 1997. The precautionary principle: scientific uncertainty and Type I and Type II errors. Foundations of Science 2, 207-236.

Lindsay R.M. and Ehrenberg A.S.C., 1993. The design of replicated studies. The American Statistician 47, 217-228.

Linhart H. and Zucchini W., 1986. Model Selection. Wiley, New York

Liu I. and Agresti A., 2005. The analysis of ordered categorical data: An overview and a survey of recent developments & Discussion. Test 14, 1-73.

Llansó R.J., Dauer D.M., Vølstad J.H. and Scott L.C., 2003. Application of the benthic index of biotic integrity to environmental monitoring in Chesapeake Bay. Environmental Monitoring and Assessment 81, 163-174.

Lücke J.D. and Johnson R.K., 2009. Detection of ecological change in stream macroinvertebrate assemblages using single metric, multimetric or multivariate approaches. Ecological Indicators 9, 659-669.

Lunneborg C., 1999. Data Analysis by Resampling. Duxbury Press,

MacDonald D.S., Little M., Eno N.C. and Hiscock K., 1996. Disturbance of benthic species by fishing activities: a sensitivity index. Aquatic Conservation: Marine and Freshwater Ecosystems 6, 257-268.

Macrory R., 2004. Principles of European Environmental Law. Europa Law Publishing, 256 p.

Maddock I., 1999. The importance of physical habitat assessment for evaluating river health. Freshwater Biology 41, 373-391.

Madon S.P., 2008. Fish community responses to ecosystem stressors in coastal estuarine wetlands: a functional basis for wetlands management and restoration. Wetlands Ecology Management 16, 219-236.

Magurran A.E., 2004. Measuring Biological Diversity. Blackwell Science Ltd,

Magurran A.E., 1998. Ecological diversity and its measurement. Princeton University Press, New Jersey

Maloney K.O., Weller D.E., Russell M.J. and Tothorn T., 2009. Classifying the biological condition of small streams: an example using benthic macroinvertebrates. Journal of the North American Benthological Society 28, 869-884.

Martinho F., Viegas I., Dolbeth M., Leitão R., Cabral H.N. and Pardal M.A., 2008. Assessing estuarine environmental quality using fish-based indices: performance evaluation under climatic unstability. Marine Pollution Bulletin 56, 1834-1843.

Mayer D., 2004. Essential Evidence-Based Medicine.

McClish D.K., 1989. Analysing a portion of the ROC curve. Medical Decision Making 9, 190-195.

McCullagh P. and Nelder J.A., 1989. Generalized linear models. Chapman and Hall, Boca Raton, 511 p.

McCune B., Grace J.B. and Urban D.L., 2002. Analysis of Ecological Communities., 300 p.

McGarvey D.J., 2007. Merging Precaution with Sound Science under the Endangered Species Act. BioScience 57, 65-70.

McLusky D.S. and Elliott M., 2004. The estuarine ecosystem: ecology threats and management. Oxford University Press, Oxford, 216 p.

Meire P.M., Ysebaert T., Van Damme S., Van den Bergh E., Maris T. and Struyf E., 2005. The Scheldt estuary: a description of a changing ecosystem. Hydrobiologia 540, 1-11.

Melcher A., Schmutz S., Haidvogl G. and Moder K., 2007. Spatially based methods to assess the ecological status of European fish assemblage types. Fisheries Management and Ecology 14, 453-463.

Metz C.E., 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8, 283-298.

Metz C.E., 1989. Some practical issues of experimental design and data analysis in radiologic ROC studies. Investigative Radiology 24, 234-245.

Meyer J.L., 1997. Stream Health: Incorporating the Human Dimension to Advance Stream Ecology. Journal of the North American Benthological Society 16, 439-447.

Mickwitz P., 2003. A Framework for Evaluating Environmental Policy Instruments. Context and Key Concepts. Evaluation 9, 415-436.

Millar R.B. and Anderson M.J., 2004. Remedies for pseudoreplication. Fisheries Research 70, 397-407.

Millennium Ecosystems Assessment, 2005. Our Human Planet: Summary for Decision-makers. Island Press, Washington,DC

Miller A.J., 2002. Subset selection in regression. Chapman and Hall, Boca Raton, 238 p.

Moffat A.J., Davies S. and Finér L., 2008. Reporting the results of forest monitoring - an evaluation of the European forest monitoring programme. Forestry

Morris J., 2000. Rethinking Risk and the Precautionary Principle. Butterworth-Heinemann, 294 p.

Motulsky H., 1995. Intuitive Biostatistics. Oxford University Press, Oxford, 386 p.

Mulgan R., 2000. 'Accountability': An Ever-Expanding Concept? Public Administration 78, 555-573.

Muñoz-Erickson T.A., Aguilar-González B. and Sisk T.D., 2007. Linking Ecosystem Health Indicators and Collaborative Management: a Systematic Framework to Evaluate Ecological and Social Outcomes. Ecology and Society 12, 1-19.

Murtaugh P.A., 1996. The statistical evaluation of ecological indicators. Ecological Applications 6, 132-139.

Muxika I., Borja Á. and Bald J., 2007. Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. Marine Pollution Bulletin 55 (Spec. Issue 1-6), 16-29.

Myung J.J., 2000. The importance of complexity in model selection. Journal of Mathematical Psychology 44, 190-204.

Nijboer R.C. and Schmidt-Kloiber A., 2004. The effect of excluding taxa with low abundances or taxa with small distribution ranges on ecological assessment. Hydrobiologia 516, 347-363.

Noble R.A.A., Cowx I.G., Goffaux D. and Kestemont P., 2007. Assessing the health of European rivers using functional ecological guilds of fish communities: Standardising species classification and approaches to metric selection. Fisheries Management and Ecology 14, 381-392.

Nõges P., Van de Bund W., Cardoso A.C., Solimini A.G. and Heiskanen A.-S., 2009a. Assessment of the ecological status of European surface waters: a work in progress (review paper). Hydrobiologia 633, 197-211.

Nõges P., Van de Bund W., Cardoso A.C., Solimini A.G. and Heiskanen A.-S., 2009b. Preface - Assessment of the Ecological Status of European Surface Waters. Hydrobiologia 633, 1-3.

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

Oberdorff T., Pont D., Hugueny B. and Chessel D., 2001. A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. Freshwater Biology 46-399.

Oberdorff T., Pont D., Hugueny B. and Porcher J.P., 2002. Development and validation of a fish-based index for the assessment of "river health" in France. Freshwater Biology 47, 1720-1734.

Olden J.D., Poff N.L. and Bledsoe B.P., 2006. Incorporating ecological knowledge into ecoinformatics: An example of modeling hierarchically structured aquatic communities with neural networks. Ecological Informatics 1, 33-42.

Onkelinx T., Vandecasteele B., Genouw G. and Quataert P., 2007a. Kwaliteitszorg meetnetten: gevalstudie 2 - vergelijking van twee meetnetten i.v.m. waterbodems (IBW & VMM).

Onkelinx T., Vandecasteele B., Genouw G. and Quataert P., 2006. Kwaliteitszorg meetnetten: gevalstudie 1 - vergelijking van twee meetnetten i.v.m. verzuring (IBW & VMM).

Onkelinx T., Verschelde P., Wouters J., Bauwens D. and Quataert P., 2008. Ontwerp en evaluatie van meetnetten voor het milieu- en natuurbeleid. Steekproefgrootteberekeningen en analyse van de kosteneffectiviteit. Vlaamse Overheid, Departement Leefmilieu, Natuur en Energie & Instituut voor Natuur- en Bosonderzoek, Brussel, 234 p.

Onkelinx T., Wouters J., Bauwens D. and Quataert P., 2007b. Kwaliteitszorg meetnetten: gevalstudie 3 - verkennende dimensionering Natura 2000.

Overton W.S., 1993. Probability sampling and population inference in monitoring programs. In:Environmental monitoring with GIS. Oxford University Press, New York, pp. 470-480.

Overton W.S. and Stehman S.V., 1995. The Horvitz-Thompson theorem as a unifying perspective for probability sampling with examples from natural resource sampling. American Statistician 49, 261-268.

Pan X. and Metz C.E., 1997. The "proper" binomial model: Parametric reciever operating characteristic curve estimation with degenerate data. Academic Radiology 4, 380-389.

Paul J.F., Scott K.J., Campbell D.E., Gentile J.H., Strobel C.S., Valente R.M., Weisberg S.B., Holland A.F. and Ranasinghe J.A., 2001. Developing and applying a benthic index of estuarine condition for the Virginian Biogeographic Province. Ecological Indicators 1, 83-99.

Paul J.F., Walker H.A., Galloway W., Pesch G., Cobb D., Strobel C.J., Summers J.K., Charpentier M. and Heltshe J.F., 2008. Combining Existing Monitoring Sites with a Probability Survey Design -- Examples from U.S. EPA's National Coastal Assessment. The Open Environmental & Biological Monitoring Journal 1, 25.

Pepe M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, Oxford, 302 p.

Pepe M.S., 1997. A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. Biometrika 84, 595-608.

Pepe M.S., 1998. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics 54, 124-135.

Pepe M.S., Cai T.X. and Longton G., 2006. Combining predictors for classification using the area under the receiver operating characteristic curve. Biometrics 62, 221-229.

Pepe M.S. and Thompson M.L., 2000. Combining diagnostic test results to increase accuracy. Biostatistics 1, 123-140.

Peterman R.M. and M'Gonigle M., 1992. Statistical power analysis and the Precautionary Principle. Marine Pollution Bulletin 24, 231-234.

Poff N.L., 1997. Landscape filters and species traits. Towards mechanistic understanding and prediction in stream ecology. Journal of the North American Benthological Society 16, 391-409.

Pont D., Hugueny B., Beier U., Goffaux D., Melcher A., Noble R., Rogers C., Roset N. and Schmutz S., 2006. Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. Journal of Applied Ecology 43, 70-80.

Pont D., Hugueny B. and Rogers C., 2007. Development of a fish-based index for the assessment of river health in Europe: the European Fish Index. Fisheries Management and Ecology 14, 427-439.

Primpas I., Tsirtsis G., Karydis M. and Kokkoris G.D., 2010. Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive. Ecological Indicators 10, 178-183.

Proulx R., 2007. Ecological complexity for unifying ecological theory across scales: A field ecologist's perspective. Ecological Complexity 4, 85-92.

Pullin A.S., 2002. Putting the science into practice. In: Pullin A.S. (Ed.)., Conservation biology. Cambridge University Press, Cambridge, pp. 305-328.

Pullin A.S., Báldi A., Can O.E., Dieterich M., Kati V., Livoreil B., Lövei G., Mihók B., Nevin O., Selva N. and Sousa-Pinto I., 2009. Conservation Focus on Europe: Major Conservation Policy Issues That Need to Be Informed by Conservation Science. Conservation Biology 23, 818-824.

Pullin A.S. and Knight T.M., 2003. Support for decision making in conservation practice: an evidence-based approach. Journal for Nature Conservation 11, 83-90.

Quataert P., Breine J. and Simoens I., 2007. Evaluation of the European Fish Index: false-positive and false-negative error rate to detect disturbance and consistency with alternative fish indices. Fisheries Management and Ecology 14, 465-472.

Quataert P., Breine J. and Simoens I., 2004. Comparison of the European Fish Index with the Standardised European Model, the Spatially Based Models (eco-regional and European), and Existing Methods, FINAL REPORT, Development, Evaluation & Implementation of a Standardised Fish-based Assessment Method for the Ecological Status of European Rivers. A Contribution to the Water Framework Directive., 48 p.

Ransohoff D.J. and Feinstein A.R., 1978. Problems of spectrum and workup bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine 299, 926-930.

Rapport D.J. and Whitford W.G., 1999. How Ecosystems respond to Stress: common properties of arid and aquatic systems. BioScience 49, 193-203.

Rényi A., 1961. On measure of entropy and information. In: Neyman J. (Ed.)., Proceedings of the fourth Berkely symposium on mathematical statistics and probability. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Berkeley, pp. 547-561.

Reynolds C.S., 2002. Resilience in aquatic ecosystems - hysteresis, homeostasis, and health. Aquatic Ecosystem Health & Management 5, 3-17.

Reynoldson T.B., Norris R.H., Resh V.H., Day K.E. and Rosenberg D.M., 1997. The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. Journal of the North American Benthological Society 16, 833-852.

Romero J., Martínez-Crego B., Alcoverro T. and Pérez M., 2007. A multivariate index based on the seagrass Posidonia oceanica (POMI) to assess ecological status of coastal waters under the water framework directive (WFD). Marine Pollution Bulletin 55 (Spec. Issue 1-6), 196-204.

Rosenbaum P.R., 2002. Observational studies. Springer, New York, 375 p.

Roset N., Grenouillet G., Goffaux D., Pont D. and Kestemont P., 2007. A review of existing fish assemblage indicators and methodologies. Fisheries Management and Ecology 14, 393-405.

Roth N.E., Southerland M.T., Chaillou J., Klauda R., Kazyak P., Stranko S., Weisberg S., Hall L. and Morgan R., 1998. Maryland biological stream survey: Development of a fish Index of Biotic Integrity. Environmental Monitoring and Assessment 51, 89-106.

Rothman,K.J.. ,1990. A sobering start for the cluster busters' conference. American Journal of Epidemiology 132[1], S6-13.

Schafer W.D., 2001. Replication: a design principle for field research. Practical Assessment, Research & Evaluation 7,

Schmidt-Kloiber A. and Nijboer R.C., 2004. The effect of taxonomic resolution on the assessment of ecological water quality classes. Hydrobiologia 516, 269-283.

Schmutz S., Cowx I.G., Haidvogl G. and Pont D., 2007a. Fish-based methods for assessing European running waters: a synthesis. Fisheries Management and Ecology 14, 369-380.

Schmutz S., Kaufmann M., Vogel B., Jungwirth M. and Muhar S., 2000. A multi-level concept for fish-based, river-type-specific assessment of ecological integrity. Hydrobiologia 422-423, 279-289.

Schmutz S., Pont D., Haidvogl G. and Cowx I.G., 2007b. Preface to the special issue - Fish-based methods for assessing European running waters (FAME). Fisheries Management and Ecology 14, 367.

Schneider E.D. and Kay J.J., 1994. Complexity and thermodynamics : Towards a new ecology. Futures 26, 626-647.

Seber G.A.F., 1984. Multivariate observations. John Wiley & Sons, New York, 686 p.

Seegert G., 2000. The development, use, and misuse of biocriteria with an emphasis on the index of biotic integrity. Environmental Science & Policy 3, 51-58.

Shao J. and Tu D., 1995. The Jackknife and Bootstrap. Springer-Verlag, New York

Shapiro D.E., 1999. The interpretation of diagnostic tests. Statistical Methods in Medical Research 8, 113-134.

Simon T.P., 2003. Biological Response Signatures: Indicator Patterns Using Aquatic Communities., 576 p.

Simon T.P., 1999. Assessing the sustainability and biological integrity of water resources using fish communities. CRC press, Boca Raton, 671 p.

Simpson E.H., 1949. Measurements of diversity. Nature 163, 688.

Sindilariu P., Freyhof J. and Wolter C., 2006. Habitat use of juvenile fish in the lower Danube and the Danube delta: implications for ecotone connectivity. Hydrobiologia 571, 51-61.

Snee R.D., 2007. Adopt DMAIC: Step one to making improvement part of the way we work. Quality Progress 52-53.

Sokal R.R. and Rohlf F.J., 1995. Biometry: the principles and practice of statistics in biological research. W.H. Freeman and Company, New York, 887 p.

Solheim A.L. and Gulati R.D., 2009. Preface: Quantitative ecological responses for the Water Framework Directive related to eutrophication and acidification of European lakes. Aquatic Ecology 42, 179-181.

Southerland M.T., Rogers G.M., Kline M.J., Morgan R.P., Boward D.M., Kazyak P.F., Klauda R.J. and Stranko S.A., 2007. Improving biological indicators to better assess the condition of streams. Ecological Indicators 7, 751-767.

Southerland M.T., Vølstad J.H., Weber E.D., Klauda R.J., Poukish C.A. and Rowe M.C., 2009. Application of the probability-based Maryland Biological Stream Survey to the state's assessment of water quality standards. Environmental Monitoring and Assessment 150, 65-73.

Steyerberg E.W., Harrell F.E., Borsboom G.J.J.M., Eijkemans M.J.C., Vergouwe Y. and Habbema J.D.F., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology 54, 774-781.

Steyerberg E.W. and Vickers A.J., 2008. Decision Curve Analysis: a Discussion. Medical Decision Making 28, 146-149.

Stoddard J.L., Herlihy A.T., Peck D.V., Hughes R.M., Whittier T.R. and Tarquinio E., 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27, 878-891.

Stoddard J.L., Larsen D.P., Hawkins C.P., Johnson R.K. and Norris R.H., 2006. Setting expectations for the ecological condition of running waters: the concept of reference condition. Ecological Applications 16, 1267-1276.

Strange C.D., Aprahamian M.W. and Winstone A.J., 1989. Assessment of a semi-quantitative electric fishing sampling technique for juvenile Atlantic salmon, *Salmo salar* L., and trout, *Salmon trutta* L. in small streams. Aquaculture and Fisheries Management 20, 485-492.

Svarstad H., Petersen L.K., Rothman D., Siepel H. and Wätzold F., 2008. Discursive biases of the environmental research framework DPSIR. Land Use Policy 25, 116-125.

Swets J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285-1293.

Swets J.A., Dawes R.M. and Monahan J., 2000. Better decisions through science. Scientific American 283, 82-87.

Swets J.A. and Pickett R.M., 1982. Evaluation of diagnostic systems: Methods from signal detection theory. Academic Press, New York

Thompson B., 1994. The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality

Trousselier M. and Legendre P., 1981. A functional evenness index for microbial ecology. Microbial Ecology 7, 283-296.

Tull M., 2006. The environmental impact of ports: an Australian case study. Paper presented at the XIV International Economic History Congress (session No. 58), Helsinki, Finland, 21-25 August 2006., 24 p.

Turnhout E., 2003. Ecological indicators in dutch nature conservation: science and policy intertwined in the classification and evaluation of nature. Aksant, Amsterdam: the Netherlands, 172 p.

Turnhout E., 2009. The effectiveness of boundary objects: the case of ecological indicators. Science and Public Policy 36, 403-412.

Turnhout E., Hisschemöller M. and Eijsackers H., 2006. Ecological indicators: Between the two fires of science and policy. Ecological Indicators 7, 215-228.

Turnhout E., Hisschemöller M. and Eijsackers H.J.P., 2008. Science in Wadden Sea policy: from accommodation to advocacy. Environmental Science and Policy 11, 227-239.

Turnpenny A.W.H., Coughlan J. and Liney K.E., 2006. Review of temperature and dissolved oxygen effects on fish in transitional waters. Jacobs Babtie, 81 p.

Underwood A.J., 1997. Experiments in ecology : their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge, 504 p.

Underwood A.J., 1995. Ecological research and (and research into) environmental management. Ecological Applications 5, 232-247.

Underwood A.J., 2009. Components in design in ecological field experiments. Annales zoologici Fennici 46, 93-111.

Underwood A.J., 1998. Relationships between ecological research and environmental management. Landscape and urban planning 40, 123-130.

Underwood A.J., Chapman G. and Connell S.D., 2000. Observations in ecology: you can't make progress on processes without understanding the patterns. Journal of Experimental Marine Biology and Ecology 250, 97-115.

Underwood A.J. and Chapman M.G., 2003. Power, precaution, Type II error and sampling design in assessment of environmental impacts. Journal of Experimental Marine Biology and Ecology 296, 49-70.

Van de Bund W., 2008. Water Framework Directive Intercalibration Technical Report. Joint Research Centre Ispra, Italy

Van Stickle J. and Paulsen S.G., 2008. Assessing the attributable risks, relative risks, and regional extents of aquatic stressors. Journal of the North American Benthological Society 27, 920-931.

Verschelde P., Denys L., Jansen I. and Quataert P., 2010. Evaluatie, bijsturing en vervollediging van evaluatiemethoden voor biologische kwaliteitskenmerken (cf. Kaderrichtlijn Water), met bijzondere aandacht voor kwaliteitsborging en (Europese) interkalibratie - partim fytobenthos. Evaluatie van het EKR-criterium voor het fytobenthos., 120 p.

Vickers A.J., 2008. Decision Analysis for the Evaluation of Diagnostic Tests, Prediction Models, and Molecular Markers. The American Statistician 62, 314-320.

Vickers A.J. and Elkin E.B., 2006. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making 26, 565-574.

Vos P., Meelis E. and Ter Keurs W., 2000. A Framework for the Design of Ecological Monitoring Programs as a Tool for Environmental and Nature Management. Environmental Monitoring and Assessment 61, 317-344.

Westra L., 1997. Post-Normal Science, the Precautionary Principle and the Ethics of Integrity. Foundations of Science 2, 237-262.

Wheeler A.C., 1969. Fish-life and pollution in the lower Thames: A review and preliminary report. Biological Conservation 2, 25-30.

Wholey J.S., Hatry H.P. and Newcomer K.E., 2004. Handbook of Practical Program Evaluation. Jossey-Bass Inc., San Francisco, California, 768 p.

Wilson J.B., 1999. Guilds, functional types and ecological groups. Oikos 86, 507-532.

Wintle B.A., McCarthy M.A., Volinksky C.T. and Kavanaugh R.P., 2003. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conservation Biology 17, 1579-1590.

Wittebolle L., Marzorati M., Clement L., Balloi A., Daffonchio D., Heylen K., De Vos P., Verstraete W. and Boon N., 2009. Initial community evenness favours functionality under selective stress. Nature 458, 623-626.

Wouters J., Onkelinx T., Bauwens D. and Quataert P., 2008a. Ontwerp en evaluatie van meetnetten voor het milieu- en natuurbeleid. Leidraad voor de meetnetontwerper / opdrachtgever. Vlaamse Overheid, Departement Leefmilieu, Natuur en Energie (LNE) & Instituut voor Natuur- en Bosonderzoek (INBO), Brussel

Wouters J., Quataert P., Onkelinx T. and Bauwens D., 2008b. Ontwerp en handleiding voor de tweede regionale bosinventarisatie van het Vlaamse Gewest. Instituut voor Natuur- en Bosonderzoek (INBO), Brussel, 174 p.

Wright J.F., 2000. Chapter 1. An introduction to RIVPACS. In: Wright J.F., Sutcliffe D.W., Furse M.T. (Eds.), Assessing the biological quality of fresh waters. Freshwater Biological Association, Ambleside, Cumbria, UK, pp. 1-24.

Wright J.F., Moss D., Armitage P.D. and Furse M.T., 1984. A preliminary classification of running-water sites in Great Britain base on macro-invertebrates species and the prediction of community type using environmental data. Freshwater Biology 14, 221-256.

Wright J.F., Sutcliffe D.W. and Furse M.T., 2000. Assessing the biological quality of fresh waters. RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK, 373 p.

Yallop M.L., Hirst H., Kelly M.G., Juggins S., Jamieson J. and Guthrie R., 2009. Validation of ecological status concepts in UK rivers using historic diatom samples. Aquatic Botany 90, 289-295.

Yoccoz N.G., Nichols J.D. and Boulinier T., 2001. Monitoring of biological diversity in space and time. Trends in Ecology and Evolution 16, 446-453.

Yuan L.L. and Norton S.B., 2004. Assessing the relative severity of stressors at a watershed scale. Environmental Monitoring and Assessment 98, 323-349.

Zhou X.H., Obuchowski N.A. and McClish D.K., 2002. Statistical methods in diagnostic medicine. Wiley Interscience, New York, 437 p.

Zucchini W., 2000. An Introduction to Model Selection. Journal of Mathematical Psychology 44, 41-61.