EFI+ Project. Improvement and spatial extension of the European Fish Index Deliverable 4.1 : Report on the modelling of reference conditions and on the sensitivity of candidate metrics to anthropogenic pressures. Deliverable 4.2: Report on the final development and validation of the new European Fish Index and method, including a complete technical description of the new method. 6th Framework Programme Priority FP6-2005-SSP-5-A. N° 0044096. Rapport final

Didier Pont, Pierre Bady, Maxime Logez, Jacques Veslot

SIXTH FRAMEWORK PROGRAMME
PRIORITY FP6-2005-SSP-5-A
"Integrating and Strengthening the European Research Area" – "Scientific Support to Policies"

Project no.: 0044096

Project acronym: EFI+

Improvement and spatial extension of the European Fish Index

Instrument: STREP
Thematic Priority: Scientific Support to Policies (SSP) - POLICIES-1.5

WORKPACKAGE 4

Deliverable 4.1: Report on the modelling of reference conditions and on the sensitivity of candidate metrics to anthropogenic pressures

Deliverable 4.2: Report on the final development and validation of the new European Fish Index and method, including a complete technical description of the new method.

Authors:

Pierre Bady*, Didier Pont*, Maxime Logez**, Jacques Veslot**

* CEMAGREF (HBAN), Parc de Tourvoie, BP 44, 92163 Antony Cedex, FRANCE
** CEMAGREF (HYAX), 3275 Route de Cézanne, CS 40061, 13182 AIX EN PROVENCE Cedex 5, FRANCE

Keywords: EFI+, Fish index, multimetric index, modelling procedure, model, GLM, scoring, prediction, diagnostic, goodness of fit, predictive error, zonation, spatial organisation, metric aggregation.

Contents

Summary and Recommendations

Objectives

The European Fish Index (EFI) is a multimetric index based on a predictive model that derives reference conditions from abiotic environmental characteristics of individual sites and quantifies the deviation between the predicted fish community (in the "quasi absence" of any human disturbance) and the observed fish community (described during a fish sampling occasion). The metrics used are based on species guilds describing the main ecological and biological characteristics of the fish community.

The objective of the index is to evaluate the ecological status of sites at the European scale. One of our main objectives during the development phase was to define a calibration dataset (to calibrate models) and to model and select metrics in a way that the index could be correctly calibrated for all or most of ecoregions and environmental situation, i.e. in the absence of any significant pressures, the index values must be high (close to 0.80) and comparable between ecoregion, river zone and local environment. The sensitivity of the final indices to morphological pressures has to be considered first at such large scale.

In the same way, and at the opposite of the previous European index (FAME project), the ecological classes boundaries are based on the distributions of indices values for undisturbed sites in two types of rivers (see below). In other words, the main objective was to optimize first the specificity (capacity of the indices to correctly classify an undisturbed site as undisturbed, i.e. with a high index value) and in a second step the sensitivity (i.e. the capacity of the indices to detect the effect of a pressure).

Two indices, each composed of 2 different metrics, can be computed depending of the river zone classification of a given site:

- Salmonid Dominated Fish Assemblage Index (Salm.Fish.Index) for sites classified as Salmonid Dominated Fish Assemblage River Type (Salmonid river zone)
- Tolerant Fish Assemblage Index (Cypr.Fish.Index) for sites classified as Cyprinid Dominated Fish Assemblage River Type (Cyprinid river zone)

Salm.Fish.Index = (Ni.Hab.150 + Ni.O2.Intol) / 2
Cypr.Fish.Index = (Ric.RH.Par + Ni.LITHO) / 2

Metric names	Detailled names
Ric.RH.Par	Rheophilic reproduction habitat species richness
Ni.O2.Intol	Oxygen depletion intolerant species abundance (Nb. individuals)
Ni.LITHO	Lithophilic reproduction habitat species abundance (Nb. Individuals)
Ni.Hab.Intol.150	Abundance of individuals < 15 cm of Habitat intolerant species

One metric is expressed in term of richness, two in abundance of individuals and one in abundance per size class. Two metrics are based on tolerance responses, and two on reproduction. The four metrics decrease when exposition to human pressures increases. The correlations between metrics are relatively low (Pearson's coefficients less than 0.65)

Species classified as oxygen depletion intolerant, habitat alteration intolerant, lithophilic and rheophilic reproduction habitat are listed in Annex.

The final scoring is presented below in the Method description section.

The distinction between the 2 river types is based on the proportion of typical species belonging to Salmonid dominated fish communities (or Salmonid type species) - denominated ST-species - which are oxygen depletion intolerant, habitat alteration intolerant, stenothermic, lithophilic or speleophilic reproduction type species and with a rheophilic reproductive habitat. These 19 species are the following:

Alburnoides.bipunctatus	*Cobitis.calderoni*	*Coregonus.lavaretus*
Cottus.gobio	*Cottus.poecilopus*	*Eudontomyzon.mariae*
Hucho.hucho	*Lampetra.planeri*	*Phoxinus.phoxinus*
Salmo.salar	*Salmo.trutta.fario*	*Salmo.trutta.lacustris*
Salmo.trutta.macrostigma	*Salmo.trutta.trutta*	*Salmo.trutta.marmoratus*
Salvelinus.fontinalis	*Salvelinus.namaycush*	*Salvelinus.umbla*
Thymallus.thymallus		

List of intolerant species typically belonging to Salmonid dominated fish communities

Typically, an undisturbed salmonid river type site is dominated by ST-species which represent more than 80% of the number of individuals caught (more than 90 % most of time). At the opposite, the relative abundance of these species is less than 20% (most of time 10%) for a typical cyprinid type site.

Due to the fact that human pressures impact significantly the fish community structure, it is not possible to directly use this fish community based criteria to discriminate between salmonid type sites and cyprinid type sites. A solution to classify the sites was to use a typology based on abiotic variables. Melcher et al. (2007) produced such a typology at the European scale during the FAME project (EFT classification). Using 7 environmental variables, the authors differentiate between 15 fish-based river types. These types can be gathered in our two main river types, considering our criteria related to the relative abundance of ST-species.

This typology has been used during the process of metric standardization and selection. Nevertheless, in several situations, sites are misclassified:

- Some undisturbed sites classified as cyprinid river zone sites have a high relative abundance of ST-species.
- At the opposite, but more seldom, undisturbed sites classified as salmonid river zone site have a too low relative abundance of typical upstream intolerant species.

Then the proportion of upstream intolerant species has to be evaluated by the user a posteriori to check the correctness of the river type proposed for each site and to attribute the correct index to the considered site (Salm.Fish.Index or Cypr.Fish.Index). Depending of the situation (see after), recommendations are given to the user.

The general description of the method is first summarized, followed by the limitations of the 2 indices, the procedure used for metric modelling, metric selection and standardisation. Finally, the main results related to indices performance (specificity versus sensitivity) and evaluations of uncertainties are presented.

Method description:

The procedure is summarized in Figure a. Each number (from 1 to 7) refers to one of the different steps of the procedure presented below.

1. Data needed

EFI+ uses 2 types of data.

- Data from single-pass electric fishing catches to calculate the assessment metrics. Individuals have to be measured separately (to the next mm) to compute the observed values of the metrics. The results are given in number of individual caught per species.
- variables describing environment at the site scale or river segment scale, and the sampling method (Tables b and c).
- Additional information on location (longitude and latitude), site name and sampling date is required.

Ecoregion classification is the one of Illies, but with the addition of a Mediterranean region. Spatial coordinates are used to define the corresponding ecoregion.

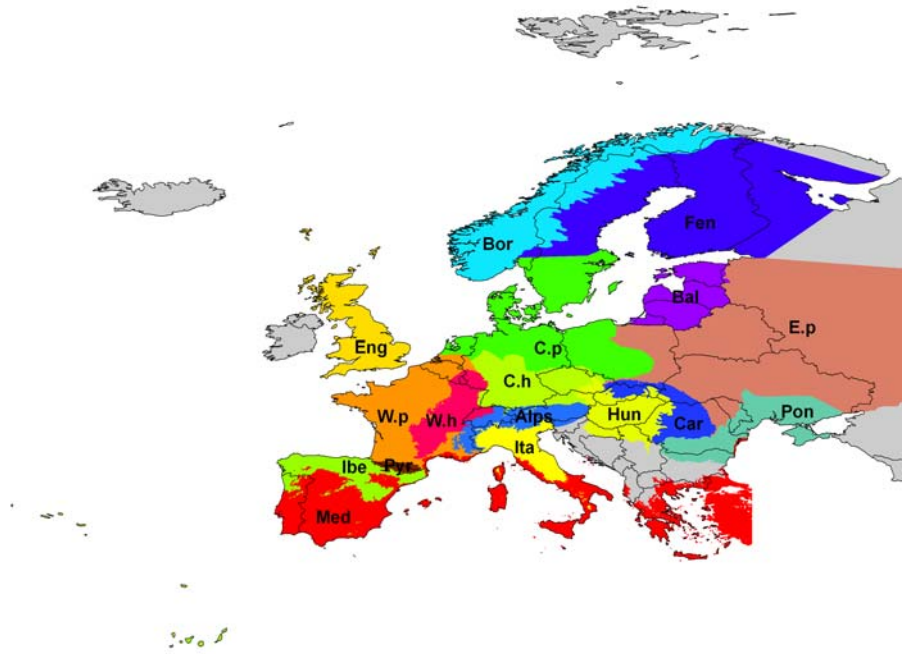


Figure a. Ecoregion and Additonal Mediterranean region.

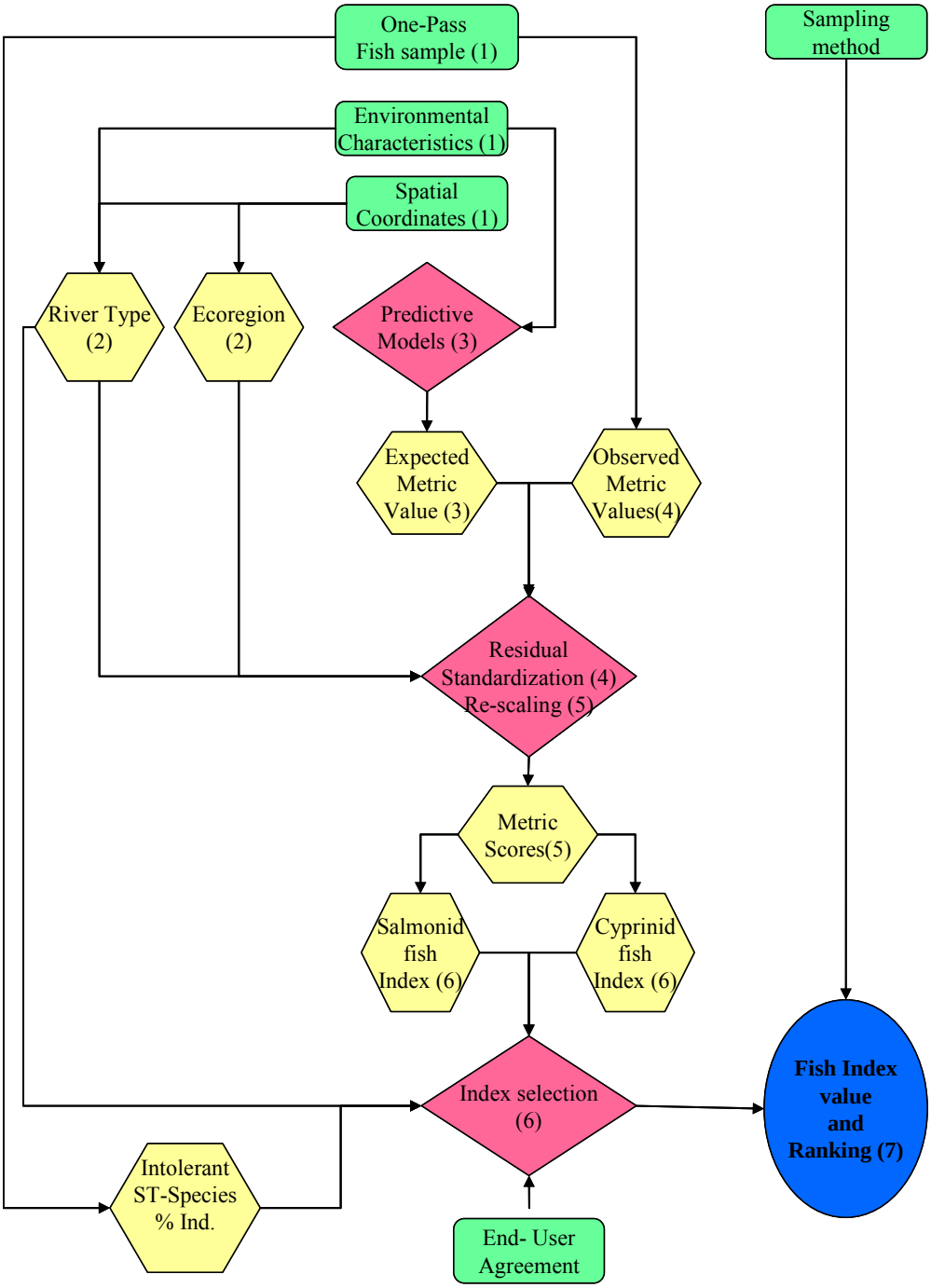


Figure b. Flow chart describing the procedure. Green rectangles: input data and end-user intervention. Pink rhomb: computation and process. Yellow hexagon: intermediate results available in the software output. Blue oval: fish index value and ranking in five classes.

Table a. Covered ecoregions. Abbreviation, full name and corresponding number.

Alp	Alps (4)	Car	The Carpathians (10)
Pyr	Pyrenees (2)	Eng	Great Britain (18)
Hun	Hungarian Lowlands (11)	Ibe	Iberian Penisula (1)
E.p	Eastern Plains (16)	Ita	Italy, Corsica and Malte (3)
Pon	Pontic Province (12)	W.p	Western Plains (13)
Fen	Fenno-Scandian Shield (22)	W.h	Western Highlands (8)
Bor	Borealic Uplands (20)	C.h	Central Highlands (9)
Bal	Baltic Province (15)	C.p	Central Plains (14)
Med	Mediterranean region		

Table b. Description of the numerical variables used in the procedure. Median value, minimum and maximum values for sites slightly impacted. These values indicate the range of environmental conditions for which the models can be considered as calibrated (N=2526 sites)

Variable	Median	Minimum	Maximum
Latitude	46.26	36.77	68.80
Longitude	5.24	-9.25	29.65
Drainage area (km^2)	56.02	0.72	208,106.00
Distance.from.source (km)	13	0.50	1454.00
Actual.river.slope ($m.km^{-1}$)	9.13	0.001	323.63
Wetted.width (m)	6.00	0.70	658,00
Fished.area (m^2)	372	100	32,500
temp.jan (°C)	1.60	-16.00	11.40
temp.jul (°C)	17.80	8.60	25.10

Table c. Description of the categorical variables used in the procedure. For each variable, the number of sites per modality is indicated. These values indicate the range of environmental conditions for which the models can be considered as calibrated (N=2526 sites).

Variable	Modality	Number of sites
Water.source.type	Glacial	12
	Groundwater	78
	Nival	539
	Pluvial	1897
Floodplain.site	No	2120
	Yes	406
Natural.sediment	Boulder/Rock	432
	Gravel/Pebble/Cobble	1853
	Organic	12
	Sand	197
	Silt	32
Geomorph.river.type	Braided	86
	Meand regular	236
	Meand tortous	121
	Naturally constraint no mob	1053

	Sinuous	1030
Sampling Method	Wading or wading/boating	2362
	Boating	164

2. River zone and ecoregion

The definition of the river zone for each site (salmonid river zone or cyprinid river zone) is based on the European Fish typology (EFT) classification (Melcher et al. 2007). Using this typology, each site is first classified in one among 15 types. These 15 types are after gathered in two river zone, based on the dominance of intolerant species belonging to Salmonid dominated fish communities (see before).

Ecoregions are defined based on the spatial coordinates of the site

3. Predictive models

Models are used to predict for each metric and for a given site a predicted value in the absence of quasi-absence of human disturbance (i.e. a value corresponding to "a reference condition". These predicted metric values (also called expected values), are computed from environmental parameters (see table before) using generalized linear models. These models have been calibrated with "undisturbed" sites.

The parameters retained for each of the 4 models are given in Table. Details about the modelling procedure are given in Annex.

Table d. Parameters associated with the four selected metrics. The environmental variables describing the hydro-morphological characteristics of the river site are synthesized in two new descriptors (Syngeomorph1 and Syngeomorph2) using a multivariate analysis. The annual temperature range is the difference between Mean July temperature and mean January temperature.

	Ni.O2.Intol	Ni.Hab.Intol.150	Ric.RH.Par	Ni.LITHO
Temp July	+			
Annual Temp Range		+		+
Actual river slope	+	+	+	+
Natural Sediment	+	+	+	+
Syngeomorph1	+	+		
Syngeomorph2		+	+	+

4. Metric score

The metric score is basically a standardized distance (M_{iq}) between the predicted value (T*i*, i.e. the expected one in the absence of any significant human disturbance) and the observed value O*i* (computed from the sampled fish community).

The score (M_{iq}) of each of the 4 metrics in a given river zone $_q$ (salmonid river zone or cyprinid zone) and a given ecoregion *j* is obtained in the following manner for each site:

$$M_{iq} = (R_i - Mjq) / S_q$$

With $R_i = Oi - Ti$

R_i: model residual, i.e. difference between observed and expected metric value for the given site.

M*jq*: Median value of the residuals in the ecoregion *j* and the river zone *q* in the whole undisturbed dataset for a given river zone (salmonid or cyprinid)

S_q: Standard deviation of the residuals in the whole undisturbed dataset for a given river zone (salmonid or cyprinid)

Sites defined here as undisturbed sites correspond to sites (N= 2526) which present no or only slight degree of perturbation (selection based on the pressure variables: channelization, impoundments, water quality, toxic presence, water abstraction, hydropeaking, presence of barrier at the river segment scale).

The value of the median is chosen because it is less sensitive to extreme values than the mean. For the same reason (stability), the variance of residuals of the whole undisturbed dataset is used instead of the variance of the residual distribution in each ecoregion.

5. Standardization and re-scaling of metric scores

Standardized scores vary from -∞ to + ∞. A requirement is that each final metric score varies within a finite interval from 0 to 1. Such "rescaling" is accomplished by using two transformations.

- For a given river zone, all the values over a maximum (Max_j) and below a minimum (Min_j) are replaced by this maximum (Max_j) and this minimum (Min_j). Then the following transformation is applied to each metric score:

$$(Mi - Min_q) / (Max_q – Min_q)$$

Max_q value is the quantile 0.95 of the distribution of standardized residuals (M*iq*) for the undisturbed site dataset in the considered river zone $_q$.

An additional requirement is that, after transformation, each metric must have the same median value in the absence of any disturbance (i.e. in the undisturbed dataset). Such result was obtained by computing with an algorithm for each metric in each river zone the Min_q value corresponding to a median value of 0.80 for the scores of undisturbed sites. Depending of the considered metric and river zone, Min_q values vary from 0.004 to 0.11.

The final result is that, when only considering undisturbed sites, all the 4 metrics have a median value of 0.80 and close values for the 25% quantile (0.61 to 0.73). Then, metrics can be aggregated, each one having a similar distribution in the absence of any significant disturbance.

Impacted sites exhibit a greater deviation from the theoretical value and thus will be characterized by a low metric value and are less likely to belong to the reference residual distribution than unimpacted or only slightly impacted sites (i.e. value << 0.80).

Table e. Summary of the 4 selected metrics distribution for undisturbed sites

	River zone	Min.	25% quantile	Median	Mean	95% quantile.	Max.
Ni.Hab.Intol.150	Salmonid	0.000	0.69	0.80	0.74	0.87	1.000
Ni.02.Intol	Salmonid	0.000	0.73	0.80	0.77	0.86	1.000
Ric.RH.Par	Cypr.	0.000	0.70	0.80	0.77	0.86	1.000
Ni.LITHO	Cypr.	0.000	0.71	0.80	0.73	0.83	1.000

6. Fish Index

Two indices, each composed of 2 different metrics, are computed for each site, depending on the river type classification.

Salm.Fish.Index = (Ni.Hab.150 + Ni.O2.Intol) / 2
Cypr.Fish.Index = (Ric.RH.Par + Ni.LITHO) / 2

Indices values vary between 0 and 1. As for each metric, an undisturbed site would have an index value close to 0.80, and a highly disturbed site a value lower than the 25% quantile of the index distribution for undisturbed sites.

A critical point to use the method is the classification of sites in one of the two river zone (salmonid river zone versus cyprinid river zone). From our definition, in the absence of any human disturbance, a salmonid river zone site is characterized by a very high proportion of the intolerant ST-species (most of them with more than 80% of individuals belonging to these species). At the opposite, a typical cyprinid site is characterized by a relative low abundance of these species (lower than 20%, in most of cases 10%).

The classification is more efficient to identify the salmonid river type than the cyprinid one. Concerning the salmonid river type, only a small number of sites can be considered as misclassified (i.e. with a very low relative abundance of ST-species). At the opposite, a larger amount of sites classified as “cyprinid river type” are dominated by ST- species.

It is clear that the consequences of a misclassification are quite different, depending of the river type.

- For sites misclassified as salmonid river sites (i.e. with a low relative abundance of ST-species), and in the absence of any disturbance, the salmonid fish index cannot be used, and has to be replaced by the cyprinid fish index.

- For undisturbed sites misclassified as cyprinid sites with a high relative abundance of ST- species, the values given by the cyprinid index are quite close to the one given by the salmonid index when the site is not disturbed. However, in case of disturbance, the impact would not be correctly evaluated if the cyprinid index is used instead of the salmonid index.

Considering the risk of misclassification and the associated consequences on the evaluation of sites the best solution is to give systematically to the user the initial classification of the site (cyprinid or salmonid river zone), the relative abundance of ST-species and the value of both indices (salmonid fish index and cyprinid fish index) when they can be computed.

Very often, the proposed river zone type is correct and the user has to consider the corresponding index. In other cases, the users, as expert, will have to evaluate the situation and to confirm the proposed classification or will have to make their own choice between the two fish indices.

There are several possibilities and associated recommendations:

Sites classified by the EFT classification as Salmonid river zone site

The site is classified as a "Salmonid" site and the % of ST- species is high (i.e. > 80%). The classification is correct and the Salmonid fish index has to be used.

The site is classified as a "Salmonid" sites and the % of ST-species is relatively high (from 50 to 80%). The reduction of the relative abundance of ST-species could be due to a human disturbance of the river ecosystem. The risk of misclassification is relatively low but the user has to check the proposed typology.

The site is classified as a "Salmonid" sites and the % of ST-species is relatively low (from 20 to 50%) to very low (less than 20%). The reduction of the relative abundance of these intolerant species can only be due to a very severe human disturbance (i.e. heavy impoundment, high level of water quality degradation,…).. The risk of misclassification is important and the user has to evaluate the proposed typology and to confirm or reject the choice of the adapted fish index. A warning is included in the output of the software.

Sites classified by the EFT classification as a Cyprinid river zone site

The site is classified as a "Cyprinid" site and the % of ST-species is very low (less than 20 %). The classification is correct and the Cyprinid fish index has to be used.

The site is classified as a "Cyprinid" sites and the % of ST-species is relatively high (from 20 to 50%). The increase of the relative abundance of these intolerant species can be due to some particular human disturbance of the river ecosystem (extreme channelisation and huge increase of the water velocity, water cooling downstream from a dam,…). Nevertheless, in most of cases, a misclassification of the site is possible. The software proposes to classify the site as a salmonid river zone type and to use the

Salmonid.Fish.index. The user has to evaluate the proposed typology and to confirm or reject the choice of the adapted fish index. A warning is included in the output of the software.

The site is classified as a "Cyprinid" sites and the % of ST-species is quite high (from 50 to 80%) or very high (more than 80%). The increase of the relative abundance of these intolerant species can also be due to particular severe human disturbances (see upper § for examples) but the risk of misclassification is very important. A correction for the river zone is included in the output of the software (site reclassified as a Salmonid river type site) and the value of the Samonid fish index is proposed. The software proposes to classify the site as a salmonid river zone type and to use the Salmonid.Fish.index. The user has to evaluate the proposed typology and to confirm or reject the choice of the adapted fish index. A warning is included in the output of the software.

The different options are summarized in Table.

Table f . Summary of the different options to select the appropriate fish index.

	% of ST-species (intolerant salmonid type species)			
Initial site classification	[0% – 20%]	]20% - 50%]	]50% - 80%]	]80% - 100%]
Salmonid zone	Risk of misclassification **Salmonid index proposed** User has to confirm the river zone and the index choice	Risk of misclassification **Salmonid index proposed** User has to confirm the river zone and the index choice	**Salmonid Index recommended** User has to check the classification	Correct classification **Salmonid Index should be used**
Cyprinid zone	Correct classification **Cyprinid Index should be used**	Increase of % of intolerant species can be linked to a human disturbance **Salmonid Index proposed** User has to confirm the river zone and the index choice	Increase of % of intolerant species can be linked to particular extreme disturbance **Salmonid Index proposed** User has to confirm the river zone and the index choice	High risk of misclassification **Salmonid Index proposed** User has to confirm the river zone and the index choice

In particular ecoregions, the possibilities for a site to be a salmonid river zone site are very low (see section 1.1.1). This is the case for Hungarian lowlands, Eastern plains, Pontic province, Baltic province and Mediterranean region.

In particular ecoregions, the possibility for a site to be a cyprinid site is very low (see section 1.1.1). This is the case for Alps, Pyrenees, Fenno-Scandian shield and Borealic uplands.

7. Ecological class boundaries

Ecological class boundaries are only based on the distributions of indices values for undisturbed sites in the two river types (table g).

As the sampling method greatly influences the score value especially in the cyprinid zone, class boundaries have been computed separately for sites sampled by boating and wading in the cyprinid zone (see Indices limitations section below).

The limits between class 1 and 2 correspond to the value of the 95% quantile of the index distribution for undisturbed sites.

The limits between class 2 and 3 correspond to the value of the 25% quantile of the index distribution for undisturbed sites.

The limits between classes 3-4 and 4-5 are defined in a way that the ranges between classes 3, 4 and 5 are similar.

The specific scoring for cyprinid zone sites sampled by boating has to be considered as a preliminary one. A more specific work is needed in the future, by using enough undisturbed or slightly disturbed boating sites and being able to correctly handle these parameters in the different models.

Table g. Ecological class boundaries for the 2 indices.

	Salmonid Zone index	**Cyprinid Zone Index**	
		Wading	**Boating**
Class 1	[0.911 -1]	[0.939 -1]	[0.917 - 1]
Class 2	[0.755- 0.911[	[0.655- 0.939[	[0.562 - 0.917[
Class 3	[0.503 -0.755[	[0.437 -0.655[	[0.375 - 0.562[
Class 4	[0.252 -0.503[	[0.218 -0.437[	[0.187 - 0.375[
Class 5	[0 - 0.252[	[0 - 0.218[	[0 - 0.187[

Limitation of the Index in relation with the environment

The statistical models that are used for the EFI reflect the average response of fish communities to environmental conditions. The application of the EFI for particular environmental situations might cause problems.

This index has been developed for sites located in the ecoregions presented in Annex. Therefore, the index should not be applied in areas with a fish fauna deviating from those of the tested ecoregions.

The model was developed using data from sites with environmental characteristics ranging between specific limits. These values are given in Table b and c. Your site should have characteristics within these ranges in order to obtain a confident EFI.

Some environmental situations are not correctly handled by the two indices. These situations are:

- presence of a natural lake upstream from the site
- presence of a winter dry period
- case of "organic" rivers

Even if no clear effect have been observed, the indices must be used with caution for intermittent/ summer dry rivers due to the low number of undisturbed sites used to test the index.

River size: The metrics have been mainly calibrated for rivers with an upstream drainage area less than 10,000 km2. Independently from the sampling method, the river size seems not to significantly influence the index values for undisturbed sites when the upstream drainage area is less than 10,000 km2.

The index should be used with caution in the lowland reaches of very large rivers as no reference sites from these reaches have been used for the calibration of the index. In those cases the index uses only extrapolated predictions based on the trends observed in the models.

Limitations in the use of the Index due to the number of fish caught

When few specimens were caught the software still allows you to calculate the index, but the results must be considered with care. The same applies when the sampled area is smaller than 100 m². Consequently, when no fish occur at a site, this method is not applicable.

The index seems relatively independent from the number of fish caught. This could be directly related to our modelling methods. All the 4 selected metrics are modelled after taking into account the sampling effort (i.e. the total richness or the number of fish caught depending of the metric). Nevertheless, a too low number of fish caught would alter the capacity of the index to respond correctly to a pressure. The user has to be careful when the number of fish caught is less than 30 individuals and a warning has to be included in the output of the software in such a situation.

Two cases could be problematic and the EFI should be used with care:

(1) undisturbed rivers with naturally low fish density and (2) heavily disturbed sites where fish are nearly extinct. In the first case, fish are close to the natural limits of occurrence and therefore might not be good indicators for human impacts. The occurrence of fish in those rivers is highly coincidental and therefore not predictable. If the very low density is caused by severe human impacts more simple methods or even expert judgement are sufficient to assess the ecological status of the river.

Limitations in the use of the Index due to the sampling method

Only fish data obtained with single-pass electric fishing may be used to calculate the EFI. If data from multiple passes are used (i.e. same site fished several times and catches cumulated) the EFI produces erroneous results.

The sampling method (boating or wading) has a strong impact on the index values. Most of our calibration sites were sampled by wading and it was not possible to include the variable describing the sampling method as a potential explanatory variable.

The number of sites sampled by boating in the salmonid river zone is limited. But their range is not too different from the range sites sampled by wading. At the opposite, there is a clear effect of the sampling method on the index values for the cyprinid zone. Most of low index values are related to boating sites. These low value boating sites are not belonging to any particular region or country.

As a first conclusion, it seems that the fish index, at the present state, could be used only with caution when sites have been sampled by boating, especially in the cyprinid zone, i.e. for larger and deeper rivers. The boating effect is not only to reduce the mean value of the cyprinid index but to increase its variability.

Nevertheless, as additional information, we propose to the user a classification of sites sampled by boating in 5 specific classes, defined in a different way than for wadeable sites (see next section). This specific scoring has just to be considered as a preliminary one and a more specific work is needed in the future if enough undisturbed or slightly disturbed sites sampled by boating are available.

Limitations in the use of the Index due to the sampling location

We also examined the case of fishing occasion where the lateral water bodies from the floodplain were sampled with or without the main channel. In such case, the indices values are significantly and clearly lower in comparison with sites where only the main channel is sampled.

The fish index, at the present state, cannot be used for fishing occasion realized in lateral water bodies of the flood plain and is only calibrated correctly for sites sampled in the main river channel.

Index development

Dataset description

The initial database contains 14221 sites corresponding to 29509 sampling occasions distributed in 15 countries: Austria, Switzerland, Germany, Spain, Finland, France, Hungary, Italy, Lithuania, Netherlands, Poland, Portugal, Romania and Sweden. In our study, we only conserve one sampling occasion by sites and we exclude the ill-informed sites. The final working table contains 9948 sites. From this table, we define two specific datasets:

- The first corresponds to the slightly disturbed sites (SID, N= 2526) which present no or slight degree of perturbation (selection based only on the pressure variables). This one is used to explore and to test the response of metrics among ecoregions in the « quasi » absence of pressure.

-The second called calibration dataset (CD) are included in SID and it's used to model the metric. The selection process of the calibration sites is relatively strict and it's extended to the effects of pressure (i.e. modification of the hydrological regime). In addition, we impose that the caught fish number must be superior to 50 individual for reducing the potential effect of the sampling effort. Finally, the site selection is completed by the exclusion of neighbouring sites for limiting the spatial autocorrelation and by a subsampling procedure to limit the over-representativeness of calibration sites located to North of Poland, Romania and to North of Spain (e.g. Galicia and Asturias). After these operations, we obtained 533 calibration sites to model the metrics (5.3% of the initial dataset). However, a strict selection is essential to obtain an unbiased calibration dataset and unbiased models.

Modelling process

The metric are modelling by generalised linear model (Nelder Wedderburn 1972, McCullagh & Nelder 1989) and stepwise procedure based on AIC (Venables & Ripley 1999). This approach appears to be a good compromise between over-learning and predictive error. The metrics based on the species number are modelled by Poisson model with logarithmic link and we prefer to use negative binomial distribution for the ones based on fish number because these lasts are largely over-dispersed. An offset parameter is systematically used to impose a baseline corresponding to total richness or total number of fish (e.g. McCullagh & Nelder 1989, Cameron & Trivedi 1998).

The environmental variables integrate several aspects of the river characteristics such as morphologic, climatic (more details on the description of this variables are available in the precedent report). We select 6 environmental variables: actual river slope (log-transformed, m/km), July temperature (°C), Thermal amplitude (Tdif=Tjul-Tjan, °C), natural sediment (coded in 3 categories) and two latent variables based on linear combination of geomorphological variables. In addition, a specific weighting stratified by Strahler order and ichtyoregions (Reyjol et al. 2007) reduced the unbalanced organisation of the calibration dataset. To consider the non-linear responses of metric to environmental condition, we compute orthogonal polynomial of degree 2 for slope and July temperature.

Model selection

The selection model process is based on two main steps:

- The first selection based on the simple criteria such as the residuals structure, good adjustment of the fitted value enables the reducing of the number selected models. This first screening is essential, because for each modality of a given trait, we can compute above 5 different metrics (e.g. binary, count proportion data based on species number and count and proportion date based on fish number).

- The second selection step involves more the consideration of more complex criteria. The selected models are characterized by a satisfactory stability, satisfactory adequacy between expected and observed values, low residuals structure and quasi-normal residuals distribution. The consideration of these criteria is strongly required to increase the extrapolation capacity of models and to limit bias of predictions based on environmental conditions in outside the calibration environment.

After a few conservative and strict modelling process, the number of candidate metrics is relatively low. We only conserve 13 metrics: five metrics based on species number (Ric.O2.Intol, Ric.Hab.Intol, Ric.Hab.RH, Ric.INSV and Ric.RH.Par), four metrics based on fish number (Ni.O2.Intol, Ni.hab.Intol, Ni.INSV and Ni.LITHO) and four metrics based on individual number inferior to 150 mm (Ni.O2.Intol.150, Ni.Hab.Intol.150, Ni.RH.150 and Ni.INSV.150).

Metric selection

Three criteria were used to select metrics:

- Correlation between metrics (Pearson coefficient < |0.70|,

- Representativeness of the metric in the different ecoregions. In some particular ecoregions and/or countries, species belonging to some of the candidate guilds are never abundant, even in undisturbed sites. This is in particular the case for the cyprinid zone and eastern or Mediterrenean regions. Several tests and previous analysis demonstrated that in such situation, the score is always underestimated for sites belonging to the lowest pressure group: the median value of sites is not close to 0.80, as for other metrics, but below 0.50 and the score of all sites, whatever the level of human disturbance, are always underestimated..

- Sensitivity to the index of pressure

In all case, the metrics based on the guild of insectivorous species are insensitive to pressure.

In the salmonid river zone, the most sensitive metrics are based on oxygen depletion and habitat intolerant guild species, and expressed in “relative” abundance of individuals The 2 corresponding metrics considering all the size class are highly correlated between them (Ni.O2.Intol and Ni.Hab.Intol.150). Among the metrics expressed in term of abundance of small-sized individuals, the 2 based on these species guilds are also highly correlated (Ni.O2.Intol.150 - Ni.Hab.Intol.150). In order to not use the same guilds with 2 different metrics, and following complementary evaluation of metrics responses, the 2 following metrics are selected:

In the cyprinid zone, the metrics based on oxygen depletion and habitat intolerance cannot be used due to their lack of representativeness in several ecoregions. Among the others and considering the high correlation between Ric.Hab.RH and Ric.RH.Par, we selected two metrics. Ric.RH.PAR has been preferred to Ric.Hab.RH in relation with its higher relative abundance in undisturbed Mediterranean sites.

The metrics are finally selected:

Salmonid zone: Ni.O2.Intol and Ni.Hab.Intol.150

Cyprinid zone: Ric.RH.PAR and Ni.LITHO

Table h. Parameters associated with the four selected metrics. The term 'poly' indicates that we used orthogonal polynomials of degree 2 (e.g. Venables & Ripley 1999).

metric	Ni.O2.Intol	Ni.Hab.Intol.150	Ric.RH.Par	Ni.LITHO
(Intercept)	-0.27832	-0.61978	-0.41193	0.07676
poly(Tjul, 2)1	-3.21395			
poly(Tjul, 2)2	-0.01514			
poly(Islope, 2)1	1.36557	-1.20607	4.12612	2.12899
poly(Islope, 2)2	-2.13935	-1.10176	-0.67042	-0.73348
natsedmedium	-0.05848	-0.25953	0.08545	0.0485
natsedsmall	-0.5376	-0.09464	-0.12434	-0.37812
syngeomorph1	0.13998	0.15807		
syngeomorph2		-0.07988	0.05822	0.03907
Tdif		0.01061		-0.01644

Performance

To quantify the performance of altered site detection, we used the slightly disturbed sites as unexposed sites, and the sites classified in the classes 4 and 5 by the pressure index as exposed sites. A specific dataset which integrates previously described limitations of index is required to reduce the potential bias induced mainly by river zone misclassification of sites.

Under the hypothesis that our pressure index is an acceptable measure of the site alteration, we observe that the indices, particularly in cyprinids zone, are typical “rule-in” tests. In spite of low sensitivities, we note that the measures of specificity (spec) and positive predictive (ppv) are relatively high in the cyprinid zone (spec=0.89 and ppv=0.78). The less significant results in the salmonid river zone (spec=0.93 and ppv=0.54) can be explained by the low prevalence of pressure (prev=0.16). The both indices are optimised to recognize undegraded sites in most cases. Consequently, the detection of an altered situation efficiently confirms the high degraded level of this site. From an economical/management point of view, this objective corresponds to the idea that the risk for managers to invest in restoration measures for undegraded sites is low.

Error estimation

To estimate the predictive error associated with individual and global fish bio-indicator scores, we propose a hybrid approach based on three elements: i) theoretical knowledge on generalized linear model (GLM), ii) simulation procedure and iii) principle of the error propagation.

For one single metric, the model provides expected values and standard errors. By extending the classical regression propriety, a random sampling procedure based on normal, expected values and standard errors produces an empirical distribution of the expected values in the link space. After the inverse link transformation, the computation of the standardized distance between the quantile values (e.g. 0.1 and 0.9) and the observed ones provides a good approximation of the predictive intervals. For the metric based on species with preference to spawn in running waters (Ric.RH.Par), we observe that the size of 80 % tolerance intervals are close to 0.39 units (+/- 0.11 units). For the one based on Lithophilic Fish (Ni.LITHO), the tolerance interval appears to be larger and it’s close to 0.42 (+/- 0.14 units).

The estimation of predictive error for the both indices is more complex because it involves the addition of non-independent variables. For example, for the cyprinid index, the correlation between Ric.RH.Par and Ni.LITHO is equal to 0.51. Consequently, the computation of theoretical variances is extremely complicate. To reduce these difficulties, we

propose to generalize the previous results and we adapt a simulation strategy to estimate empirical distribution of aggregated scores. At each step, we compute an empirical value from normal distribution based on expected value and expected variance for each metric and we calculate the new scores. Afterwards, we aggregate these ones to obtain the final indices. The consideration of quantile values (e.g. 0.1 and 0.9) easily completes the construction of the tolerance interval. For cyprinid index, the size of the 80% tolerance interval is close to 0.30 units (+/- 0.06 units, Figure c). This corresponds more or less to one class. The error estimation is presently in an experimental phase and it requires some additional tests before the implementation the software.

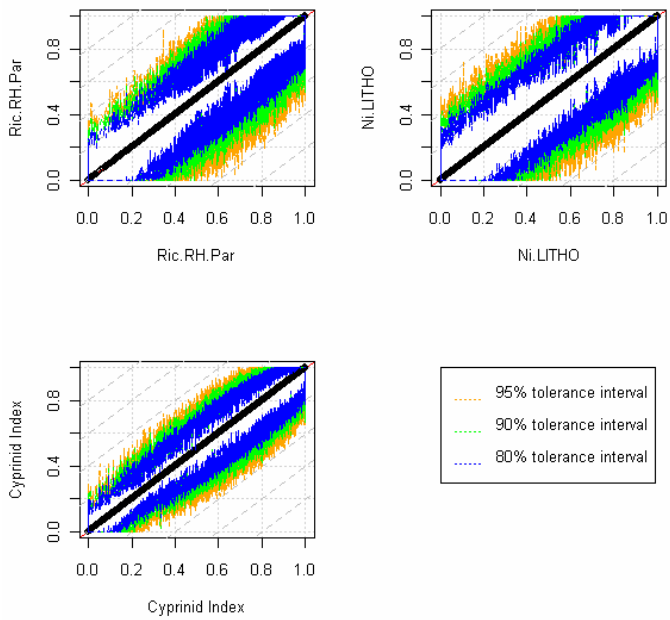


Figure c. Simulated tolerance error associated with the cyprinid index and metrics based on species with preference to spawn in running waters (Ric.RH.Par) and based on Lithophilic Fish (Ni.LITHO). Red, orange, green and blue lines correspond to the tolerance intervals based on percentiles (80%, blue; 90% , green; 95%, orange).

References

Cameron A. C. & Trivedi P. K. (1998), Regression Analysis of Count Data, Econometric Society Monograph No.30, Cambridge University Press, pp. 432.

McCullagh P. & Nelder J. A. (1989) Generalized Linear Models, second edition edn. Chapman & Hall/CRC, London, pp. 532.

Nelder J. A. & Wedderburn R. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), **135**, 370-384.

Venables W. N. & Ripley B. D. (1999) Modern Apllied Statistics with S-plus, Third edition, Statistics and computing, Springer-Verlag, New York, pp. 501.

Melcher, A., S. Schmutz, G. Haidvogl & K. Moder (2007): Spatially based methods to assess the ecological status of European fish assemblage types. Fisheries Management and Ecology, 14, 453-463

1 Introduction

Since last years, the assessment of the ecological integrity of ecosystems has become a key issue for environmental management decisions. Many diverse biomonitoring tools using the concept of biological indicators of environmental conditions have been designed for all types of ecosystems (e.g. Bonada et al. 2006, Statzner & Mog 1998). For fish assemblage and Freshwater system, the multi-metric indices initially proposed by Karr (1981), Karr et al. (1986), etc. have spread widely throughout the scientific community (e.g. Hugues et al 1998, Oberdorff et al. 2001, 2002, Pont et al. 2006, 2007). Multi-metric index calculation involves the use of several metrics reflecting different aspects of fish assemblage integrity (e.g. tolerance guilds, habitat guilds, trophic guilds), taxonomic richness and individual abundance (Oberdorff et al. 2002, Pont et al. 2006, 2007).

The first European Fish Index is a multimetric index based on a predictive model that derives reference conditions from abiotic environmental characteristics of individual sites and quantifies the deviation between the predicted fish community (in the "quasi absence" of any human disturbance) and the observed fish community (described during a fish sampling occasion). The metrics used are based on species guilds describing the main ecological and biological characteristics of the fish community. To develop a tool applicable to large scale, the authors intensively used statistical model to predict the biological and functional characteristics of fish assemblages (e.g. Oberdorf et al. 2002, Pont et al. 2006, 2007).

In general, the construction of multi-metric based on modeling process can be decomposed in three parts: the definition of calibration sites which include references or low-impacted sites, the modeling step and scoring step. In previous project (FAME project), Quataert et al. (2004, 2007) also stressed on the difficulties to establish correct measures of pressures and appropriate calibration dataset. The calibration dataset was relatively different to the site population and it contained a high proportion of sites with a low Strahler order (e.g. small rivers). Moreover, the distribution of calibration sites was relatively unbalanced in Europe and large rivers were under-represented in the sample. In present project, we devoted many efforts and times to understand the dataset and to limit sources of bias such as unrepresentative sampling, instability of variable over space and time, interference (e.g. historical effect), and contamination from any number of sources (e.g. Magurran 1988, Hellmann & Fowler 1999).

In this project, our modelling strategy is based on previous works on the Fish indices: Oberdorf et al. (2002), Pont et al. (2006, 2007) and the FAME project (http://fame.boku.ac.at/, Schmutz et al. 2007). Generalised linear models (GLM, Nelder and Wedderburn 1972, McCullagh & Nelder 1989) were used to model the biological and functional metrics. This tool is routinely used in various fields of ecology (e.g. Austin et al. 1984, Austin 1987, Austin, 2002, Candy, 2003; Eyre et al., 1993; Freeman et al., 2003; Guisan et al., 2002; Oksanen & Minchin, 2002). Initially, GLM was introduced in ecology by Austin (1980) to model and predict the response of plant species to varying environmental conditions. A procedure based on GLM appears us to be a good compromise between predictive power and interpretability. In addition, GLM offer interesting theoretical property for the estimation of predictive interval error (e.g. McCullagh & Nelder 1989, Cameron & Trivedi 1998, Collett 2003, McCulloch & Searle 2001). Concerning our decision to use GLM, we send the readers to the very substantial review proposed by Austin (2007) on the good modelling practice in ecology. Indeed, we agree with Austin (2007) who writes "it is clear that there is no standard for current best practice when modelling species environmental niche or geographical distribution, whether plant or animal. Numerous incompatibilities between the ecological, data and statistical models can be identified". These Ideas and conclusions converge to the remarks of several famous statisticians on the role of models and on the philosophy of the modelling approach (e.g. McCullagh & Nelder 1989, van Tongeren 1995, Buja 2000, Mease &

Wyner 2008). According to McCullagh & Nelder (1989) and Austin et al. (2006), there is no absolute model and the results presented in the intermediate report and in annex of this document on the comparison of the modeling approaches tend to confirm our point of view. The good modeling practice involves the consideration of these following unexhaustive points: Explanatory variable must be linked to the outcome; Data must be representative; Parsimony principle (without unnecessary variables); Good model must be checked with independent data; Several methods to validate models are necessary because none is perfect; Only numerical models based on theoretical equations or/and experimentation provide explications and predictions; In all cases, we need a good quantification of incertitude.

This document on the construction European Fish Index is structured on five main parts: the data descriptions, the metric modelling, the metric selection, the metric aggregation and scoring procedure and the predictive error estimation.

The first section will be devoted to describing the environmental and biological Data. It contains the definition of our working dataset, calibration dataset used to model the functional and biological metrics and the slightly disturbed dataset used to explore and test the responses of metrics in the « quasi » absence of pressure. We also describe the main environmental variables included in the model (e.g. thermal amplitude, geomorphological latent variables) and used in the construction of the final scores (e.g. river zone, ecoregions). In addition, we present the construction of a new pressure index based only on exposure to main seven pressures: impoundment, hydropeaking, water abstraction, presence of toxic substance, water quality, modification of river section associated with channelization level and the present of downstream barriers on the segment. This section is completed by the description of the metrics based on species guilds describing the main ecological and biological characteristics of the fish community.

The second section contains the description of the predictive model that derives reference conditions from abiotic environmental characteristics of individual sites. An offset parameter is used to impose a baseline corresponding to the total richness or the total number of fish (i.e. the expected value is less dependant from the sampling effort). These models allow to standardize the metric responses to natural environment variability (air temperature, river slope, sediment size, drainage area, river regime and river morphology …). The criteria used to judge the quality and model capacity are also described in details. A large number of metrics have been tested, each available species guild being expressed in term of species richness, individual abundance and individual biomass. A specific subsection will focus on the development of new metrics, specific to low species rivers, based on the age classes and fish.

The third part includes the final selection of metrics based on three criteria (low correlations between metrics, response to pressure and representativeness for the different ecoregions and river zones) and the technical elements on the construction of individual score (standardization and rescaling between 0 and 1).

The fourth part is devoted to the metric aggregation and to the assessment of the index performance. At the opposite of the fish index developed during the FAME project, the ecological classes boundaries are only based on the distributions of indices values for undisturbed sites in the two river zones. Examination of the fish index responses for specific environmental and methodological situations will be shown the limitations and the potential bias related to specific variables as such sampling method or sampling location. These results will allow to establish recommendations on the use of the fish indices on the base of these results

Finally, to conclude, we will present a procedure to estimate the predictive error associated with individual and global fish bio-indicator scores. This hybrid approach is based theoretical knowledge on generalized linear model, simulation procedure and principle of the error propagation.

2 Environmental and biological Data

2.1 Dataset Description

2.1.1 Global description of the dataset

The initial database contains 14221 sites corresponding to 29509 sampling occasions distributed in 15 countries: Austria, Switzerland, Germany, Spain, Finland, France, Hungary, Italy, Lithuania, Netherlands, Poland, Portugal, Romania and Sweden (see Table 1 and Figure 1). More complete information on national dataset[1] and descriptions of sampling methods, type of fish data, environmental and pressures variables[2] are described in greater details in the previous reports. Fish assemblage described in our working table (N=9948, Table 1 and Figure 1) contains 161 species and 1.938.339 individuals that corresponds to about 43.8 (+/- 11.5) species by country.

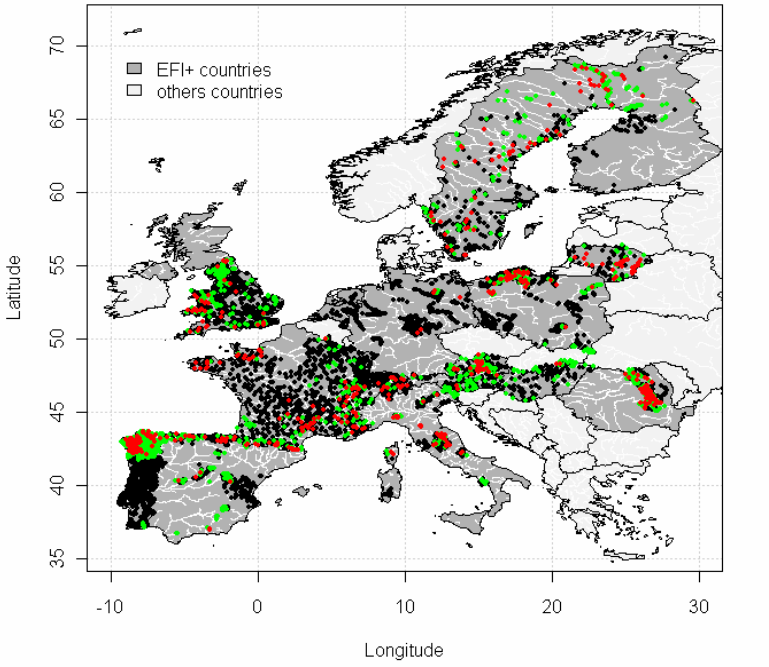


Figure 1. Localisation of the sites contained in the three datasets (N=9948): The green (N=533), red (N=2526) and black (N=7244) points correspond to the calibration, slightly disturbed and disturbed datasets, respectively. The light and dark grey polygons identify the countries included in the EFI+ project (see for more details, http://efi-plus.boku.ac.at/).

[1] EFI+ 0044096 Deliverable 3_4 New metrics development:
http://efi-plus.boku.ac.at/downloads/EFI+%200044096%20Deliverable%20D3_4.pdf
[2] EFI+ 0044096 Deliverable D1_1-1_3 Lists and descriptions of sampling methods, type of fish data, environmental variables and pressure variables:
http://efi-plus.boku.ac.at/downloads/EFI+%200044096%20Deliverable%20D1_1-1_3.pdf

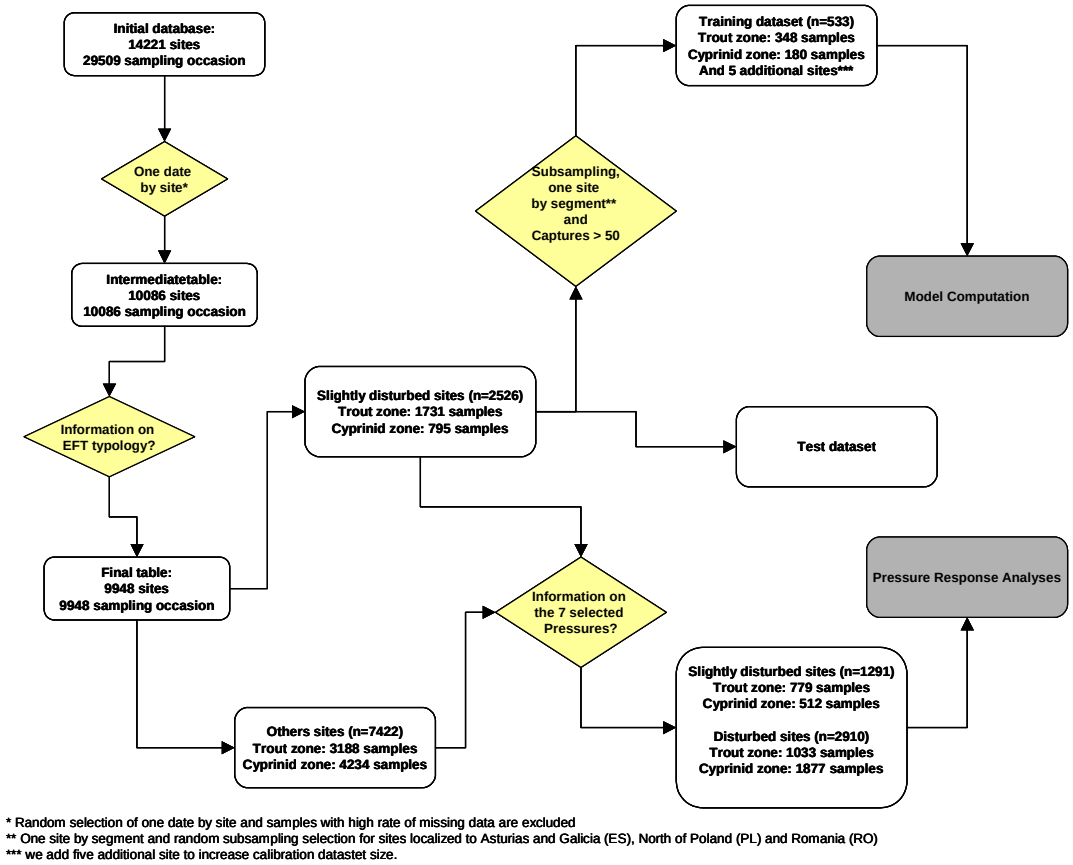


Figure 2. Description of the dataset selection process.

The selection procedure to obtain working tables is described in the Figure 2. In a first step, we selected only one sampling occasion by site with a random procedure to limit the potential effect of the temporal autocorrelation (e.g. Pinheiro & Bates 2000, Legendre et al. 2002, Schabenberger & Gotway 2005, Dormann et al. 2007). The exclusion of sites with high rate of missing values provides a new table composed by 10086 sites and 10086 sampling occasions (Figure 1). This simplifies the manipulation and the organisation of the data. In a second step, we keep only sites with complete information on the river typology (Melcher et al. 2007). Finally, our working table includes 9948 sites: 2526 slightly disturbed sites (SID) and 7244 disturbed sites. The calibration dataset (CD) are directly issued of slightly disturbed sites.

Table 1. Distribution of the sites by country in initial database, final table, slightly disturbed sites (definition given in the next section) and calibration datasets (definition given in the next section).

Country	Initial database (N=14221)	final dataset (N=9948)	SID (N=2526)	CD (N=533)
AT	938	840	172	33
CH	717	601	48	26
DE	803	760	33	5
ES	4239	1659	901	97
FI	530	220	137	13
FR	1145	971	185	84
HU	193	146	36	2
IT	652	498	152	33
LT	115	109	54	30
NL	182	105	0	0
PL	919	866	208	53
PT	923	866	8	0
RO	263	239	178	60
SE	615	504	179	56
UK	1987	1564	235	41

2.1.2 Description of the calibration (CD) dataset and description of the slightly disturbed dataset (SID)

The calibration and slightly disturbed datasets are characterized by low level of pressure (Figure 3). The first dataset is used for metric modelling and the second is used to explore and test the responses of metrics in the « quasi » absence of pressure. In contrast to calibration dataset, SID includes between ecoregions sites from all countries and ecoregions (Table 1 and Table 2). In addition, SID, after omitting CD sites can be used as a « quasi » independent validation dataset.

Table 2. Description of calibration dataset (Missing Data were accepted for the variables Acidification and Toxic). The slightly disturbed sites (SID) are selected only by the pressure variables ().

Pressure Variable	Selection criteria
Barriers.river.segment.down	No or Partial
Impoundment	No
Hydropeaking	No
Water.abstraction	No or Weak
Channelisation	No or Intermediate
Cross.sec	No
Colinear.connected.reservoir	No
Toxic.substances	No or Intermediate
Acidification	No
Water.quality.index	inferior to 3

Table 3. Supplementary criteria for the selection of calibration sites.

Variables type	Variables	Selection criteria
Effects	Hydro.mod	No
	Velocity.increase	No
	Sedimentation	No or Weak
	Instream.habitat	No or Intermediate
	Embankment	No or Local
	Riparian.vegetation	No or Slight
	Temperature.impact	No
Water quality	Eutrophication	No or Low
	Organic.pollution	No or Weak
	Organic.siltation	No
Other criteria based on indices	Water.alteration	No or Medium
	Habitat.index	No or Slight
Exclusion of Specific sites	Water.source.type	exclusion of sites characte by Groundwater water s type
	Flow.regime	only permanent river
	Spatial dependence limitation	only one fishing occasio segment
Sampling constraints	captures	superior or equal to 50 fish
	farea	superior or equal to 100 m
	month	between July and Nove (included)

The slightly disturbed sites present no or slight degree of perturbation for impoundment, hydropeaking, water abstraction, channelization, cross section, water quality, toxic substances, acidification and collinear connected reservoir. The high rate of missing values is an important limitation in the site selection. As results, to reduce the impact of these ones on the sample size of calibration and slightly disturbed datasets, we accept their presence in two cases: acidification and presence of toxic substance. Then we postulates that the experimenter knows if these two pressures are present and that missing values correspond to no impact.

Calibration sites are included in the SID, but we extend the selection process to the effects of pressure (i.e. modification of the hydrological regime). We only select sites characterised by no (or very few) effects such as embankment, sedimentation, organic pollution or eutrophication (Table 3). We integrate two supplementary indices in the selection process which describe habitat degradation and water alteration. The first, called “Habitat.index”, is based on the aggregation of the value of the three following pressure descriptors: Instream.habitat, Riparian.vegetation and Embankment. The second, called “Water.alteration”, is based on the aggregation of the value of the three following pressure descriptors: Eutrophication, Organic pollution and Organic pollution. More details on these two indices are available in the annex.

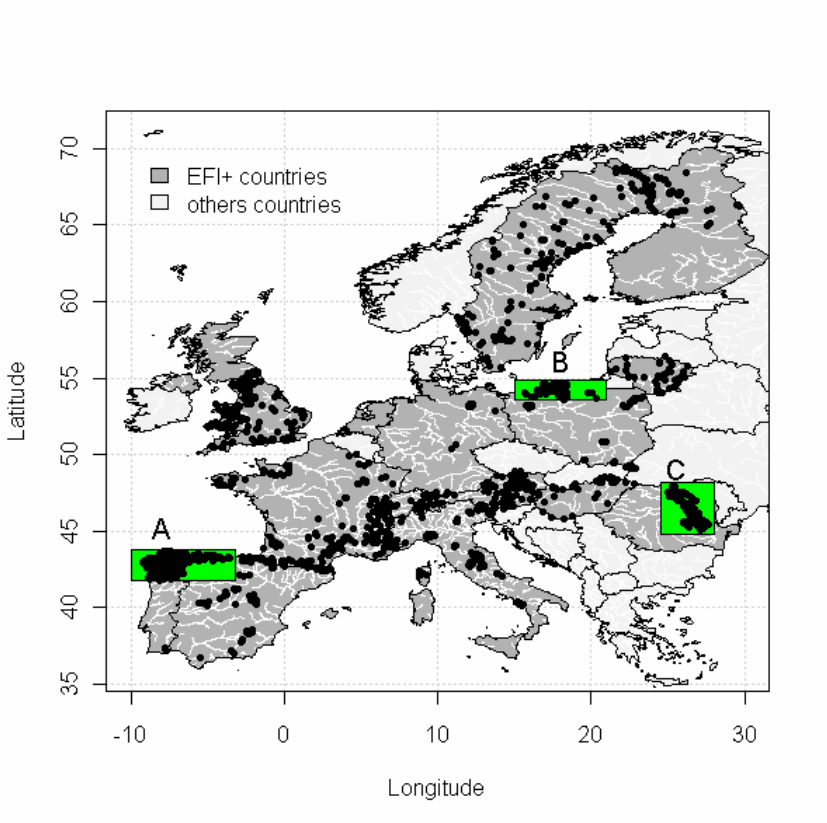


Figure 3. Representation of slightly disturbed sites and the three subsampling area (green rectangle): A: Asturias and Galicia (ES); B: North of Poland (PL); C: Romania (RO). The light and dark grey polygons identify the countries included in the EFI+ project (see for more details, http://efi-plus.boku.ac.at/).

To complete the selection of CD, supplementary criterion based on the quality of the sampling is considered to reduce the potential effect of the sampling effort on the estimation of the metric: we only conserve site with caught fish number superior to 50 individual (Table 3). The integration of a constraint on sampling effort is an essential point, because it's clear that if the number of individuals is too small, estimations of the richness and a metrics are biased (Magurran 1988, Hughes et al. 2002, Reynolds et al. 2003, Hughes et al. 2007).

The last selection constraints concern the spatial organisation of the site at local and large scale. To limit the potential effect of spatial autocorrelation, we excluded the neighboring sites in conserving only one site by segment[3] and we use a subsampling procedure for reducing the over-representativeness of calibration sites located to North of Poland, Romania and to North of Spain (e.g. Galicia and Asturias, Figure 3). The high sites concentration in these three regions could excessively bias the estimation of model parameters and reduced the extrapolation capacity of our models.

After this all operations, we obtained 533 calibration sites to model the metrics. Sample size is appreciably reduced and we only conserve 5.3% of the initial dataset. However, as written previously, a strict selection of the calibration site is crucial to obtain an unbiased calibration dataset and unbiased models.

Table 4. Description of the numerical environmental variables. For each variable, we indicate the mean values and the standard deviation (in parenthesis). Additional non-parametric tests (Wilcoxon's test) provide a comparison between disturbed and slightly disturbed sites.

variable	CD (N=533)	Disturbed (N=7422)	SID (N=2526)	W	p-value
Latitude	49.392 (6.801)	48.469 (5.542)	48.523 (7.346)	10074079	< 0.001
Longitude	9.692 (11.634)	5.424 (9.552)	6.192 (12.62)	9291873	0.51
Altitude	357.462 (338.308)	250.839 (301.311)	369.397 (325.65)	6860636.5	< 0.001
AREA.ctch	817.278 (3551.31)	6257.813 (24565.186)	1200.665 (8308.391)	12290546	< 0.001
Distance.from.source	33.167 (59.817)	83.548 (178.882)	36.713 (84.531)	11927344	< 0.001
Actual.river.slope	17.67 (26.565)	11.520 (26.715)	16.865 (28.807)	6338475.5	< 0.001
Wetted.width	10.893 (18.972)	24.016 (68.63)	14.427 (38.93)	10899858	< 0.001
Fished.area	815.874 (1628.629)	1339.3 (3554.936)	811.178 (2482.506)	12011474.5	< 0.001
temp.jan	-0.695 (5.828)	1.576 (4.674)	0.507 (6.444)	9984302	< 0.001
temp.jul	17.38 (2.207)	18.150 (2.517)	17.588 (2.327)	10234937	< 0.001
Tdif	18.075 (4.988)	16.573 (3.78)	17.080 (5.34)	9074447	0.0163
syngeomorph1	0.996 (1.215)	0.087 (1.534)	0.955 (1.193)	6090048.5	< 0.001
syngeomorph2	0.153 (1.27)	-0.079 (1.157)	0.112 (1.249)	8682156.5	< 0.001

The comparison between slightly disturbed sites and disturbed sites clearly indicate that the both datasets are relatively different (Table 4 and Table 5). In general, we observed that the disturbed sites appear to be localized in bigger rivers than the slightly disturbed sites. For example, in means, drainage area and distance from the source are higher in disturbed dataset that in the slightly disturbed dataset. In spite of the unbalanced effective between the both datasets and the presence of several outliers, we observe that the disturbed dataset clearly presents more heterogeneous environmental conditions. These results justify our weighting strategy presented in the next section 3.1 (Statistical models) to limit the prediction bias and they confirm that we need models with a good extrapolation capacity.

Table 5. Description of the categorical environmental variables. For each variable, we indicate the counts of sites in each modality for a given variable. Additional Chi-Squared tests provide a comparison between disturbed and slightly disturbed sites.

variable	Modality	CD (N=533)	Disturbed (N=7422)	SID (N=2526)	X-squared	df	p-value
Water.source.type	Glacial	6	41	12	42.655	3	< 0.001
	Groundwater	0	212	78			
	Nival	123	1169	539			
	Pluvial	399	6000	1897			
Flow.regime	Intermittent	0	27	5	59.377	3	< 0.001
	Permanent	528	7063	2476			
	Summer dry	0	318	32			
	Winter dry	0	14	13			
Floodplain.site	No	442	4762	2120	344.428	1	< 0.001
	Yes	86	2660	406			
Lakes.upstream	No	495	7101	2412	0.118	1	0.732
	Yes	33	321	114			
Natural.sediment	Boulder/Rock	90	811	432	300.88	4	< 0.001
	Gravel/Pebble/Cobble	376	4754	1853			
	Organic	1	76	12			
	Sand	56	1384	197			
	Silt	5	397	32			
Geomorph.river.type	Braided	29	337	86	383.394	4	< 0.001
	Meand regular	60	929	236			
	Meand tortous	20	748	121			
	Naturally constraint no m	234	1649	1053			
	Sinuous	185	3759	1030			
geotype	Calcareous	180	3007	690	230.098	2	< 0.001
	Organic	12	100	128			
	Siliceous	336	4315	1708			

2.1.3 Spatial organization (regionalization and zonation)

The influence of environmental structure at large has been considered using two typologies: one based on Illies ecoregions and the second considering the type of fish community (.i.e. the dominance of intolerant fish species).

Illies ecoregions:

Illies ecoregion delimitations are mainly based on geographical criteria like altitude, climate, dominance of mountains or plains, catchments boundaries (in some case) and river valleys. They are contiguous geographical units and are for a part characterised by some of the environmental variables used in our modelling approach. For example, local air temperature is for a large part governed by regional-based processes. River slope and upstream drainage area will tend to be respectively higher and smaller in mountainous ecoregions but, nevertheless, the variability of such parameters remains important in a given ecoregion.

Then ecoregion can be considered as relatively homogeneous geographical unit which gives a description of the general type of environment around a site at the regional level. In addition, Illies ecoregions are officially recognized as a basic European typology in the Water Framework Directive.

River zone:

Preliminary analysis demonstrated that the sensitivity of metric to pressure differs markedly between river types characterized by different fish fauna. Two rivers are mainly distinguished:

- Upstream and/or Nordic and/or alpine rivers dominated by salmonid type species (named by convention "salmonid" rivers in this report).
- Downstream type species and/or lowland plain rivers and/or Mediterranean rivers where the fish fauna is dominated by cyprinid (named by convention "Cyprinid" rivers in this report).

In association with the ecoregions, this river typology will be used:

- to standardize the scores between ecoregions and for each of the 2 river types
- to test the metric sensitivity and select metrics for each and the two river-type
- and finally to define one index per river type.

2.1.3.1 Ecoregion typology

The ecoregion classification retained is the Illies classification. In addition, we defined a Mediterranean ecoregion using the criteria retained by P. Segurado (see report form the Mediterranean group).
The sites belonging to the Mediterranean ecoregion in our dataset are those corresponding to the Mediterranean type 1 from Segurado. In order to avoid any serious misclassification of sites, the latitude of a site classified as Mediterranean must be < 45°. In practice, areas defined as Mediterranean in the project cover south part of Illies' ecoregion Ibe, Ita and the Provencal french mediterranean coast (ecoregion W.p.).

The full name list and abbreviation list of the 17 considered ecoregions and their corresponding number) are given below:

Alp	Alps (4)	Car	The Carpathians (10)
Pyr	Pyrenees (2)	Eng	Great Britain (18)
Hun	Hungarian Lowlands (11)	Ibe	Iberian Penisula (1)
E.p	Eastern Plains (16)	Ita	Italy, Corsica and Malte (3)
Pon	Pontic Province (12)	W.p	Western Plains (13)
Fen	Fenno-Scandian Shield (22)	W.h	Western Highlands (8)
Bor	Borealic Uplands (20)	C.h	Central Highlands (9)
Bal	Baltic Province (15)	C.p	Central Plains (14)
Med	Mediterranean region		

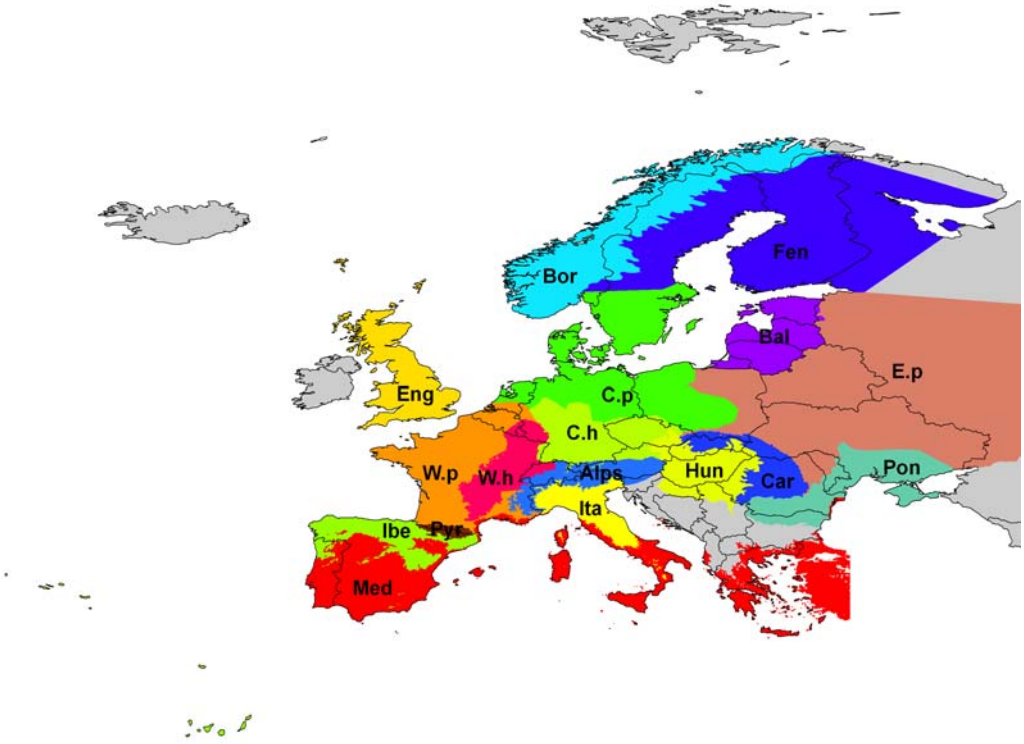


Figure 4. Upper: Representation of Illies' ecoregions (Illies 1978). Lower: Modified ecoregions used in the EFI+ project (creation of a Mediterranean ecoregion).

2.1.3.2 River zonation

In comparison with several fish based river classifications (Huet classification, rhytral-potamal system …), our 2 types-typology is quite simple but it reflect one of the most important feature of fish communities at the European scale. The distinction between the both river types is based on the proportion of typical species belonging to Salmonid dominated fish communities (or Salmonid type species) - denominated ST-species - which are oxygen depletion intolerant, habitat alteration intolerant, stenothermic, lithophilic or speleophilic reproduction type species and with a rheophilic reproductive habitat. These 19 species are the following:

Alburnoides.bipunctatus	*Cobitis.calderoni*	*Coregonus.lavaretus*
Cottus.gobio	*Cottus.poecilopus*	*Eudontomyzon.mariae*
Hucho.hucho	*Lampetra.planeri*	*Phoxinus.phoxinus*
Salmo.salar	*Salmo.trutta.fario*	*Salmo.trutta.lacustris*
Salmo.trutta.macrostigma	*Salmo.trutta.trutta*	*Salmo.trutta.marmoratus*
Salvelinus.fontinalis	*Salvelinus.namaycush*	*Salvelinus.umbla*
Thymallus.thymallus		

These 19 species are the only species (represented in the dataset, i.e. among 160 species) which are oxygen depletion intolerant, habitat alteration intolerant, lithophilic or speleophilic type for reproduction (sensus Balon classification), and stenothermic.

These species (ST-species) are considered as typical for the salmonid river zone. They will be named salmonid-type species in this report.

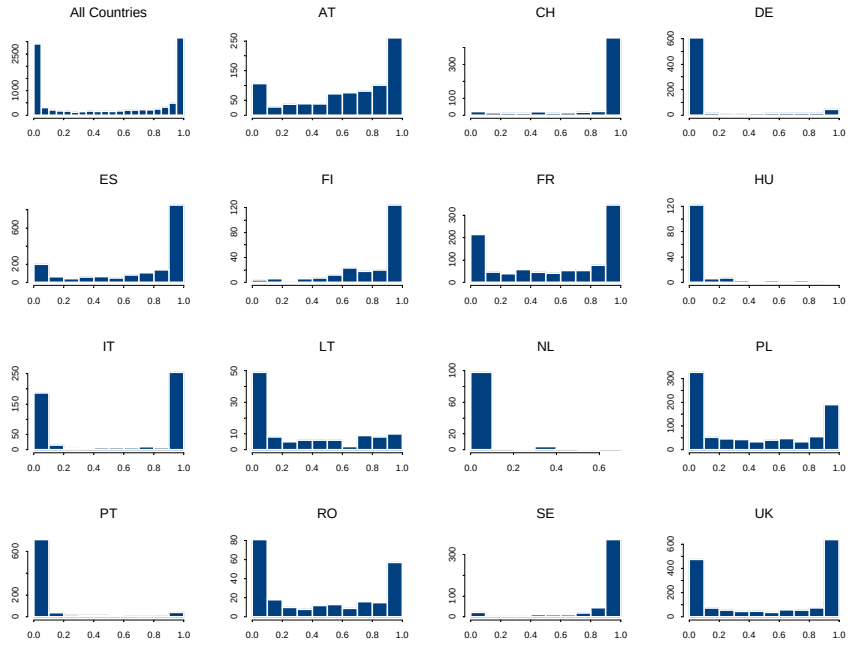


Figure 5. Distribution of the relative abundance of salmonid-type species

The distribution of the relative abundance of salmonid-type species highlights the two types of fish community: less than 10-20% of salmonid-type species and more than 80-90%.

Considering the complete dataset, 79% of sites are characterized by a fish community with more than 90% or less than 10% of salmonid-type species, i.e. most of sites are clearly salmonid river type sites or cyprinid type sites. Nevertheless, there are some differences among countries and ecoregions.

Most of countries/ecoregions where cyprinid type sites are highly dominant-(e.g. NL, PT, DE, HU, ecoregions Pon, Hun, Med) have a very low proportion of sites with a mixed fish fauna, but with the exception of LT (ecoregion Bal).

For countries/ecoregions dominated by salmonid-type Rivers, the situation is contrasted. Some ecoregions (Pyr, Bor, Alp) have a low proportion of sites with a mixed fish fauna. At the opposite, countries like AT and FI have a high proportion of mixed-fish fauna sites. This is in particular the case for AT.

Table 6. Distribution of the relative abundance of salmonid-type species per country (in blue: high proportion of salmonid dominated sites, in yellow: high proportion of cyprinid dominated sites.

Countries	[0-10%]	]10-20%]	]20-50]%	]50-80%]	]80-90%]	90-100%]	10-90%
All Dataset	32.4	3.9	9.3	11.8	5.9	36.7	21.1
NL	93.3	1	4.8	1	0	0	5.8
IT	37.6	3	2.6	4.4	1.2	51.2	7
PT	81.9	4.6	5.1	2.9	0.8	4.7	8
DE	79.9	2	4.1	6.1	2	6.1	10.2
HU	83.6	4.1	7.5	4.1	0	0.7	11.6
SE	4.4	1	4.2	7.7	8.7	74	11.9
CH	3.5	2.3	7.2	7.3	3.7	76	14.5
UK	30.4	4.7	9.7	9.8	4.7	40.8	19.5
ES	12.1	3.7	10.1	14.3	8.4	51.4	24.4
PL	37.8	6	14	13.9	6.4	22.1	27.9
RO	33.9	7.5	12.6	15.9	6.3	23.8	28.5
FR	22	4.7	14.5	15.1	7.9	35.6	29.6
FI	1.8	2.7	5.9	24.1	9.1	56.4	30
LT	45	7.3	15.6	15.6	7.3	9.2	31.2
AT	12.7	3.3	13.6	27.3	12	31.1	40.9

Table 7. Distribution of the relative abundance of salmonid-type species per ecoregion (in blue: high proportion of salmonid dominated sites, in yellow: high proportion of cyprinid dominated sites.

Ecoregions	[0-10%]	]10-20%]	]20-50]%	]50-80%]	]80-90%]	90-100%]	10-90%
All ecoregions	32.4	3.9	9.3	11.8	5.9	36.7	21.1
Pyr	0	0	0	0	0	100	0
Bor	4.6	0	1.5	0	4.6	89.2	1.5
Alp	3.9	1.8	6.7	10.6	7.2	69.8	17.3
Fen	1.7	1.5	4.2	14.6	8.6	69.4	18.8
W.h	12	3.5	12.3	11.3	7.1	53.8	23.6
Ibe	11.9	4.5	10.4	14.6	8.2	50.4	25
Ita	39.5	3.5	3.2	5.1	2.2	46.5	8.3
Car	7.4	6	12.6	20.5	8.4	45.1	33.1
Eng	30.4	4.7	9.7	9.8	4.7	40.8	19.5
C.p	49.4	4	7.8	10.1	5.7	22.9	17.9
C.h	35.2	3	11.9	22.7	7.1	20.1	34.6
W.p	32.5	5.6	16.3	18.1	7.6	19.9	34.4
Bal	45	7.3	15.6	15.6	7.3	9.2	31.2
Med	82.9	3.1	4.3	2.5	1.2	6	6.8
E.p	63	5.6	14.1	10.2	2.8	4.2	24.3
Hun	71.9	4.5	10.6	8.5	1.5	3	19.1
Pon	89.5	5.3	5.3	0	0	0	5.3

Due to the fact that human pressures impact significantly the fish community structure, it is not possible to directly use this fish community based criteria to discriminate between salmonid type sites and cyprinid type sites.

A better solution to classify the sites is to use a typology based on abiotic variables. Melcher et al. (2007) produce such a typology at the European scale during the FAME project. It is a discriminant analysis approach based on 7 environmental variables defined at the site scale: altitude, wetted width, mean of air temperature, river slope, distance from

source and spatial coordinates (longitude and latitude). Using this tool, it is possible to predict the type of fish community in a given river sites.

The authors differentiate between 15 fish-based river types. This type can be gathered in two main river types, considering our criteria related to the relative abundance of salmonid-type intolerant species (see table below).

- Salmonid river type: types 1, 2, 3, 4, 7, 8, 9, 11 and 12
- Cyprinid river type: types 5, 6, 10, 13, 14 and 15

Based on the dataset available in the European project FAME, this method is able to correctly classify more than 70% of sites.

	European Fish Type														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Sites	148	365	553	229	1130	832	69	84	7	81	9	446	148	67	432
Fish species															
Salmo trutta fario	**94**	**81**	**43**	**37**	11	5	**45**	7	**25**	14		9	4	1	3
Cottus gobio	0	**14**	**38**	5	**19**	**12**	4	**13**				**17**	0	1	0
Phoxinus phoxinus	0	1	7	**17**	**21**	**31**		9				7	**15**	3	2
Barbatula barbatula		0	3	**13**	**14**	**13**	1	1				1		1	3
Anguilla anguilla	3	0	0	0	**16**	1		0		3		0	9		1
Leuciscus souffia			0	12		0									
Thymallus thymallus		1	1	0	0	1	**45**	11	**18**			1		0	0
Salmo salar	2		1	0	7	0		**45**	9			3		0	
Cottus poecilopus		2	0	1				5	**47**			0			4
Leuciscus carolitertii	0									**36**					
Chondrostoma polylepis	1									**23**					
Rutilus arcasii	0									**14**					
Barbus bocagei										10					
Salmo trutta lacustris		0				0	0				**100**	6			2
Salmo trutta trutta			0		0	0		0	1			**40**		0	
Barbus meridionalis		0		1		0							**53**		
Leuciscus cephalus		0	1	4	2	5	1	2				0	11	10	8
Barbus haasi													8		
Gasterosteus aculeatus			0		2	1		1				1	0	**39**	1
Alburnoides bipunctatus			0	0	0	3	0							**15**	1
Rutilus rutilus		0	0	0	3	6		1				2		10	**37**
Alburnus alburnus			0		0	4	0					0		6	7
Gobio gobio	0	0	1	6	1	5	0	0				1		4	7
Perca fluviatilis		0	0	0	0	3		1				4		1	6
Lota lota		0	0	0	0	0		1				4		0	2
Leuciscus leuciscus			0	0	1	4		1						2	4
Esox lucius		0	0		0	1		1				1		1	3
Barbus barbus		0	0	1	0	1	1	0						0	2

Using this method, it was possible to classify 9948 sites as salmonid type or cyprinid type sites. The variable “River wetted width” was lacking for some sites and they were excluded from the dataset.

This typology will be used during the process of metric selection. Nevertheless, the proportion of salmonid type species will be used a posteriori to check the correctness of the river type attributed to each sites. This checking could be used in particular in two situations.

In case of undisturbed sites, this proportion can be directly used as a criterion to check the river classification. In such case, one could expect that a site classified as cyprinid must not

have a proportion of salmonid-type species higher than at least 80%. At the opposite, one could expect an abundance of salmonid type species higher than 80%.

This criterion can also be used in a second situation when sites are exposed to human pressure. Relative abundance of salmonid-type species would be much more altered by human disturbances in salmonid-type river than in cyprinid-type river, due to the fact that these salmonid-type species (ST-species) are intolerant (impoundment, water quality alteration …). An opposite effect (increase of intolerant species) could be linked for example to channelization or cold-water release downstream from a dam. But the effect is in general less important.

2.1.3.3 Combined typology

To ensure that the standardization of residuals will be correctly done (after the modelling process) using the "undisturbed sites (n=2526), the number of sites per combination (river type * ecoregion) must be at least close to 30.

Table 8. Combined typology between ecoregion and river zonation.

Ecoregion	Ecoregion name	Grouped ecoregion	Trout river	Cyprinid river
Alp	Alps (4)	Alp	YES	No sites
Pyr	Pyrenees (2)			
Hun	Hungarian Lowlands (11)	Est	No sites	YES
E.p	Eastern Plains (16)			
Pon	Pontic Province (12)			
Fen	Fenno-Scandian Shield (22)	Nor	YES	No sites
Bor	Borealic Uplands (20)			
Bal	Baltic Province (15)	Bal	No sites	YES
Med	Mediterranean region	Med	No sites	YES
Car	The Carpathians (10)	Car	YES	YES
Eng	Great Britain (18)	Eng	YES	YES
Ibe	Iberian Penisula (1)	Ibe	YES	YES
Ita	Italy, Corsica and Malte (3)	Ita	YES	YES (included in W. p.)
W.p	Western Plains (13)	W.p	YES	YES
W.h	Western Highlands (8)	W.h	YES	YES (included in W. p.)
C.h	Central Highlands (9)	C.h	YES	YES (included in C. p.)
C.p	Central Plains (14)	C.p	YES	YES

Then, several ecoregions are gathered and in some river types not considered. This process is realized on the "undisturbed" sites dataset (table below)

In some ecoregions, the salmonid zone is not considered when no sites are classified as salmonid river type (Ecoregions Bal, Med, Hun, E.p. and Pon). All sites from these ecoregions will be considered as cyprinid sites.

At the opposite, when undisturbed sites classified as cyprinid are few in a given ecoregion and shows a high proportion of salmonid-type species, they are re-classified as salmonid-type river sites (ecoregions Alp, Pyr, Fen and Bor).

For some ecoregions, the number of "undisturbed" sites classified as cyprinid type sites is too low. They are then aggregated to the closest ecoregion (cyprinid sites of ecoregions Ita and W.h. included in W.p, cyprinid sites of ecoregion C.h included in C.p.).

The classification below is finally obtained for "undisturbed" sites (N=2526) per ecoregion

```
      Alp Bal C.h C.p Car Eng Est Ibe Ita Med Nor W.h W.p
Cypr   0  54   0 123  68 189 112 134   0  83   0   0  32
Salm 160   0  87 117 103  46   0 701 121   0 260  64  72
```

For all sites per ecoregion

```
      Alp Bal C.h  C.p Car  Eng Est  Ibe Ita Med Nor W.h W.p
Cypr   0 109   0 1213  86 1290 502  364   0 913   0   0 552
Salm 817   0 403  500 129  274   0 1310 337   0 470 448 231
```

For all sites per country:

```
       AT  CH  DE   ES  FI  FR  HU  IT  LT  NL  PL  PT  RO  SE   UK
Cypr 173  28 623  530   0 453 144  85 109 105 626 678 173  12 1290
Salm 667 573 137 1129 220 518   2 413   0   0 240 188  66 492  274
```

2.1.3.4 Match between river typology and relative abundance of salmonid-type species (undisturbed sites)

The rate of misclassification for undisturbed sites is significantly higher in the cyprinid zone than in the salmonid zone. 41% of sites from the cyprinid zone have more than 80% of relative abundance of salmonid type species. The rate of misclassification in the cyprinid zone is very high in most of ecoregions and especially in Eng, Ibe and W.p.

Table 9. Undisturbed sites distribution per classes of relative abundance of salmonid type species in the two river zones.

% salmonid type species	salmonid zone	cyprinid zone
[0–20%]	93	237
]20-80%]	281	226
]80-100%]	1357	332

Table 10. Undisturbed sites distribution per ecoregion in the cyprinid river zone per ecoregion.

% salmonid type species	Alp	Bal	C.h	C.p	Car	Eng	Est	Ibe	Ita	Med	Nor	W.h	W.p
[0–20%]	0	15	0	44	15	27	88	8	0	34	0	0	6
]20-80%]	0	25	0	48	30	33	23	34	0	19	0	0	14
]80-100%]	0	14	0	31	23	129	1	92	0	30	0	0	12

In the salmonid zone, the equivalent rate of misclassification is only 7% (less than 20% of salmonid type species). Problems mainly occur in Car, C.h and W.p.

Table 11. Undisturbed sites distribution per ecoregion in the salmonid river zone per ecoregion.

% salmonid type species	Alp	Bal	C.h	C.p	Car	Eng	Est	Ibe	Ita	Med	Nor	W.h	W.p
[0-20%]	4	0	7	2	6	1	0	22	45	0	1	0	5
]20-80%]	22	0	31	27	24	0	0	120	10	0	24	5	18
]80-100%]	134	0	49	88	73	45	0	559	66	0	235	59	49

The consequences of this asymmetrical rate of misclassification on the efficiency of the selected metrics will be discussed later.

2.1.4 Pressure indices

To quantify the level of exposure to pressures, we proposed to use two pressures indices. The first was developed by the Austria Team (see previous report[4]) and the second is a simplified version based only on seven pressures: impoundment, hydropeaking, water abstraction, presence of toxic substance, water quality, modification of river section associated with channelization level and the present of downstream barriers on the segment. To summarize the information contained in this pressure table, we used Multiple Correspondence Analysis (MCA[5], Tenenhaus & Young 1985, Venables & Ripley 2002), a specific method for analysis of multi-dimensional categorical table.

The eigenvalues graphic shows that the table inertia is mainly summarized by the first axis and to a lesser extent by the second one. The exploration of row coordinates shows that the first axis of MCA is a good candidate to summarize the pressure levels. Actually, this axis clearly corresponds to the "size effect" which integrates the pressure intensity (or accumulation). Moreover, we observe that the relationship between pressures and the first axis are relatively similar, same trend and same direction (Figure 7). In contrast, the second axis doesn't integrate pressure intensity, but it mainly provides a typology of sites based on the association between the pressures. For example, we observed a high relationship between the presence of toxic substance and the level of channelization (Figure 8).

[4] EFI+ 0044096 Deliverable 3_1_3_2 Pressure analysis and global pressure index (http://efi-plus.boku.ac.at/downloads/EFI+%200044096%20Deliverables%20D3_1_D3_2.pdf).

[5] We used the function 'mca' implemented in the Package MASS of R software, see in Venables & Ripley (2002) for more details.

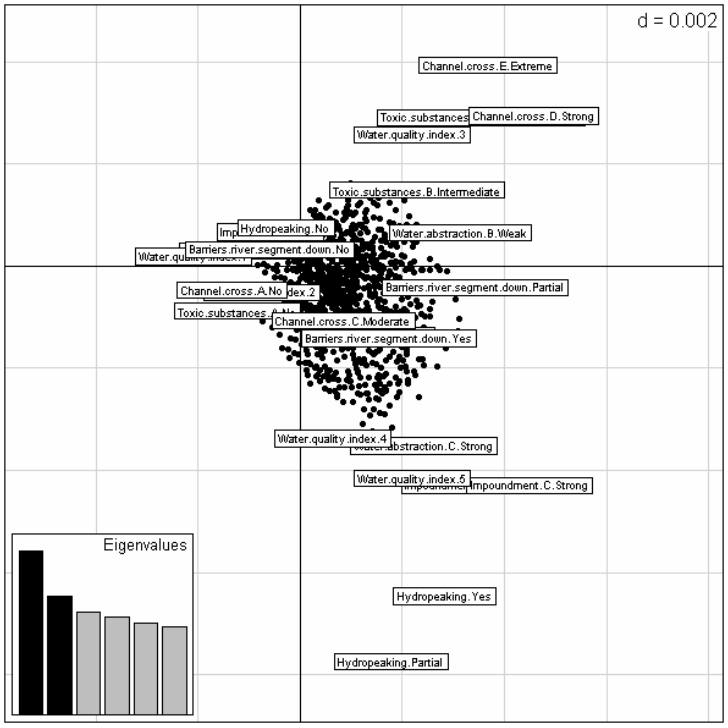


Figure 6. Representation of the results obtained by MCA based on restricted pressure index. The coordinates of rows and columns are projected on the first plan of the analyses which corresponds to 42.7% of the total inertia.

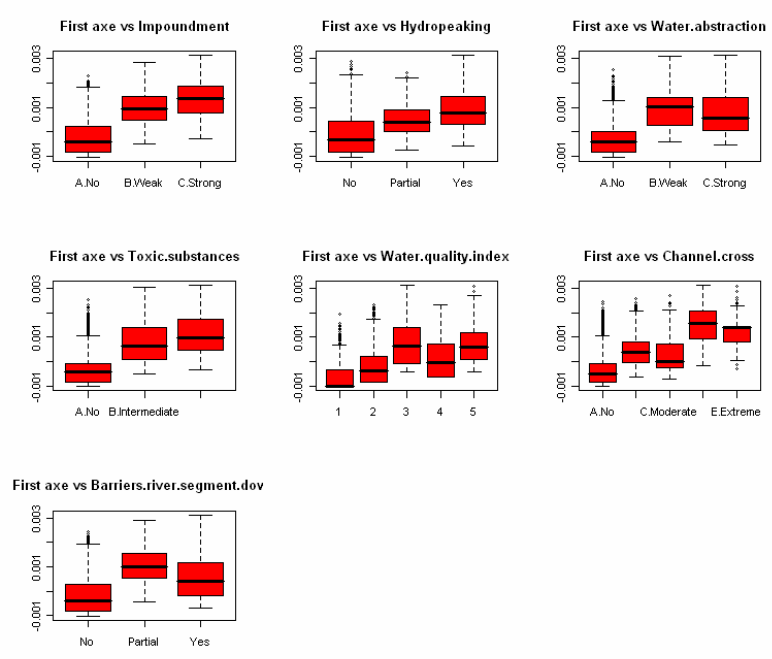


Figure 7. Boxplot representation of the first axis of MCA in function of the seven pressure variables. Graphics are separated by pressure variables.

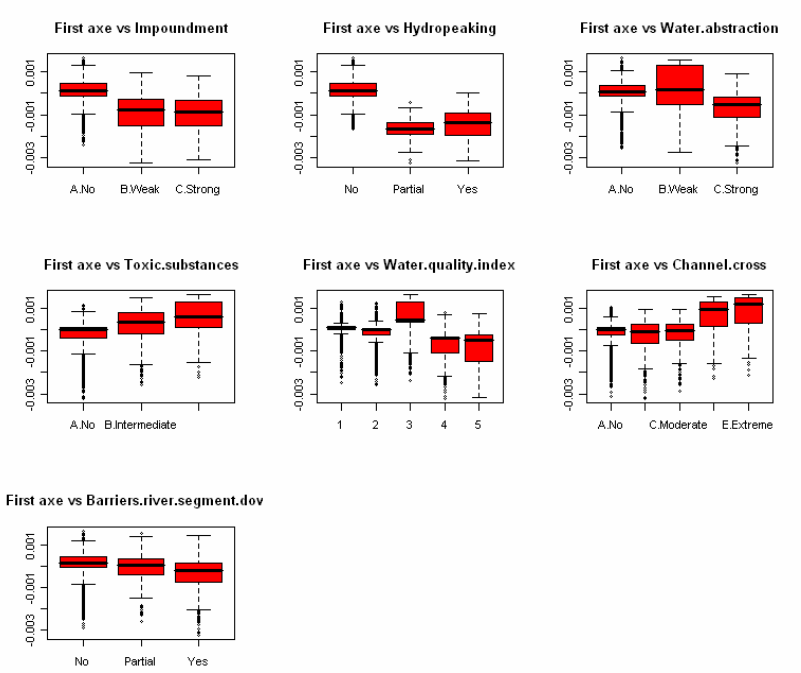


Figure 8. Boxplot representation of the second axis of MCA in function of the seven pressure variables. Graphics are separated by pressure variables.

To complete the construction of the new pressure index, we used a classical Min-Max transformation (based on uniform distribution, Legendre & Legendre 1998, Saporta 2006) to rescale the index between 0 and 1. The equation is defined as follows:

$$\text{Press.Index.B}=\frac{RS_1-\min(RS_1)}{\max(RS_1)-\min(RS_1)}$$

Finally, we categorize the pressure index by a k-means clustering based on the algorithm proposed by Hartigan & Wong (1979). To stabilize our classification, we performed an additional iterative procedure to find the k-means solution which minimizes the total within-cluster sum of squared distances. The distribution of the pressures index by categories is illustrated in the Figure 9. In a same way, we use similar procedure to categorize the initial pressure index (Press.Index.A, see the previous report[6] on the pressure analysis).

[6] EFI+ 0044096 Deliverable 3_1_3_2 Pressure analysis and global pressure index (http://efi-plus.boku.ac.at/downloads/EFI+%200044096%20Deliverables%20D3_1_D3_2.pdf).

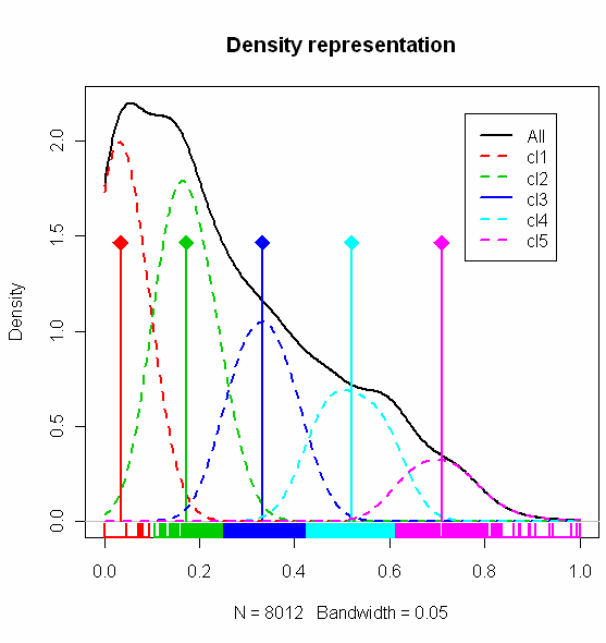


Figure 9. Kernel Estimation of the score distributions by classes obtained by K-means methods.

The comparison between the both indices shows that a common trend in the group organisation (Kappa=0.35), but, in details, we note that there are important overlaps among the contiguous classes in the confusion matrix (Table 12). The divergence among the both pressure indices are mainly induced by the integration of the several additional variables and the combination of effects and exposures to pressures in the previous index (Press.Index.A).

Table 12. Confusion Matrix and coherence among the both pressure indices.

	Press.Index.A				
Press.Index.B	**class 1**	**class 2**	**class 3**	**class 4**	**class 5**
class 1	1286	606	148	17	0
class 2	336	828	626	204	30
class 3	43	222	458	400	137
class 4	0	20	208	462	203
class 5	0	2	27	115	244

2.2 Environmental Variables for modelling

The environmental variables included in models take in account several aspects of the river characteristics such as geomorphology or climatic condition (more details on the description of these variables are available in the previous report[7] on the description of environmental variables). We select 6 environmental variables: actual river slope (log-transformed, m/km), July temperature (°C), Thermal amplitude (Tdif=Tjul-Tjan, °C), natural sediment (coded in 3 categories) and two latent variables based on linear combination of geomorphological variables.

[7] EFI+ 0044096 Deliverable D1_1-1_3 Lists and descriptions of sampling methods, type of fish data, environmental variables and pressure variables (http://efi-plus.boku.ac.at/downloads/EFI+%200044096%20Deliverable%20D1_1-1_3.pdf)

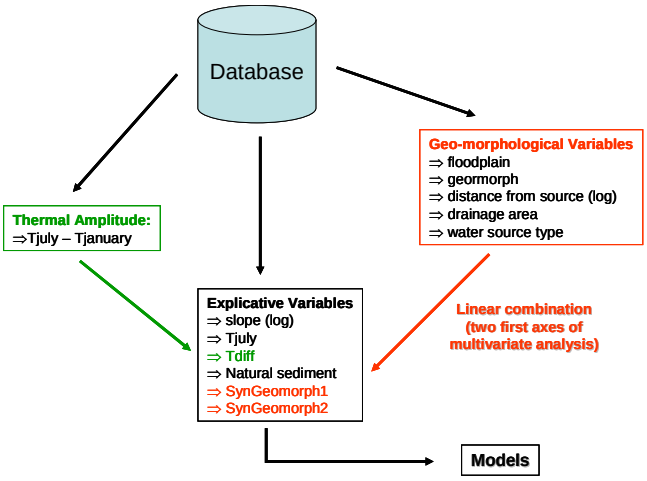


Figure 10. Preparation of the explanatory variables: six variables are considered as explanatory variables in the models.

- **Temperature and thermal amplitude**

The thermal variables allow the opportunity to take in account the climatic condition. For a given site, maximal value is measured by the temperature of July and thermal amplitude is obtained by the difference between the temperature of July and January. In addition, this operation ensures the reducing correlation between these both variables.

- **Natural sediment**

Initially, this variable was coded in five categories: “Slit”, “Organic”, “Sand”, “Boulder/Rock” and “Gravel/Pebble/Cobble”. However, the number of sites in the three first categories is excessively low. To balance the site number among the categories, we define three new categories of sediment size (e.g. small, medium, and large). The modalities “Slit”, “Organic” and “Sand” are grouped together in the modality 'small'. The modalities Boulder/Rock and Gravel/Pebble/Cobble give the large and medium modalities, respectively.

Mis en forme : Surlignage

Table 13. Description of the new variable which codes the natural sediment type. For each dataset, we indicate the count of sites in each modality for a given variable. The calibrations sites (CD) are included in slightly disturbed sites (SID). The Chi-squared test provides a raw comparison among slightly disturbed and other sites.

variable	modality	CD (N=533)	Disturbed (N=7422)	SID (N=2526)	X-squared	Df	p-value
natsed	large	90	811	432	296.201	2	< 0.001
	medium	376	4754	1853			
	small	62	1857	241			

- **Latent Geomorphological variables**

Fluvial *Geomorphology* is an important parameter in the organization of the fish habitat and fish assemblage. However, geomorphological variables were strongly interconnected between them. For example, large rivers (unimpacted) localized in central and Western Europe are quite often associated with the presence of floodplain, meandering structure and high drainage area. For

this reason, a mixed analysis (Hill & Smiths 1976; De Leeuw & van Rijckevorsel 1980, Dray & Dufour 2007) is used to describe and summarize the information contained in geomorphological table made up of the following variables: drainage area (log-transformed), distance from the source (log-transformed), simplified geomorphological type (see Table 14), simplified water source type (with only two modalities, pluvial or non-pluvial) and floodplain information (presence or absence). In this way, we obtained two latent variables based on the two first axes of the mixed analysis. This strategy is widely used in regression modelling (see for example PCR methods, Martens & Naes 1989, Tenenhaus 1998) to reduce the multicollinearity between explanatory variables.

Table 14. Description of qualitative variables used in the mixed multidimensional analysis. For each dataset, we indicate the count of sites in each modality for a given variable. The calibrations sites (CD) are included in slightly disturbed sites (SID). The Chi-squared test provides a raw comparison among slightly disturbed and other sites.

variable	modality	CD (N=533)	Disturbed (N=7422)	SID (N=2526)	X-squared	Df	p-value
watersource	Non-Pluvial	129	1422	629	37.612	1	< 0.001
	Pluvial	399	6000	1897			
geomorph	braided	29	337	86	372.856	3	< 0.001
	constraint	234	1649	1053			
	meand	80	1677	357			
	sinuous	185	3759	1030			

The first plan of the mixed analysis represents 53.4% of the total inertia. The barplot of egienvalues clearly shows that the table structure is mainly summarized by the two first axes (Figure 1).

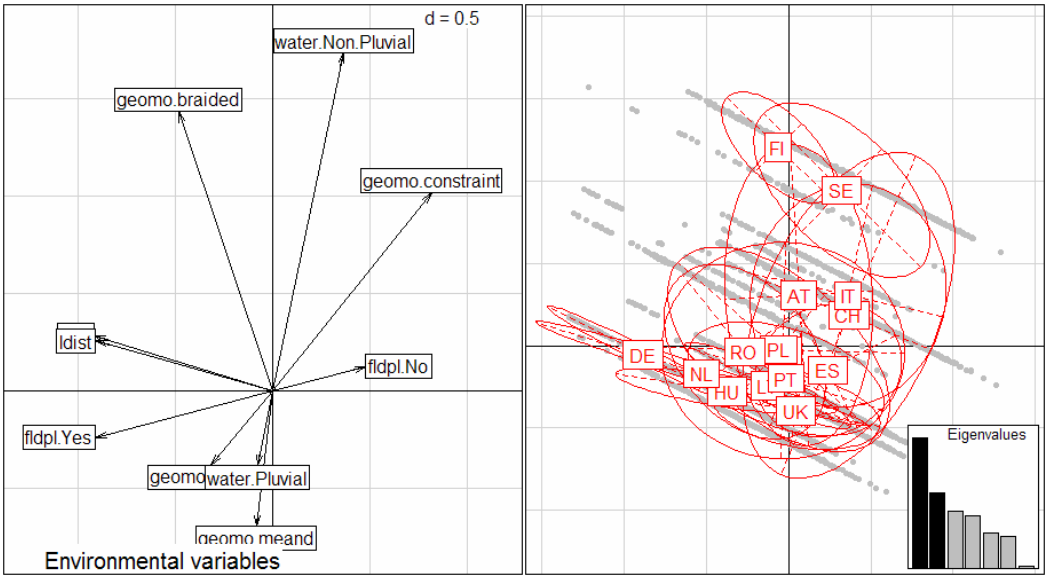


Figure 11. Representation of variables coordinates on the first factorial plan of Hill & Smith analysis and representation of site coordinates on the first factorial plan of Hill & Smith analyses. The inertia ellipses (in red) are added for each country. Bar plot of eigenvalues associated with mixed analysis.

The first axis is explained by the variables drainage area (table 2, Figure 2), the presence of flood plain and distance from the source (log-transformed, ldist, table 2, Figure 2). This axis

discriminates small and large rivers characterized by a Floodplain and high distance from the source and high drainage area. The second axis gives a typology of the river based on geomorphological and water source types. The distribution of the site on the second axis is explained by the separation of three main groups: Nordic River, Alps and a mixed group which contains the other rivers (Figure 11).

Table 15. Summarize of the relationship between each variable and the two first axes of the Hill & Smith analysis. The values correspond to the squared correlation coefficients if it is a quantitative variable, the correlation ratios if it is a factor and the squared multiple correlation coefficients if it is ordered.

variable	RS1	RS2
geomorph	0.2307	0.5326
fldpl	0.4352	0.0299
watersource	0.0292	0.6681
lDR	0.8434	0.0810
ldist	0.8327	0.0669

The rebuilding of the latent variables is relatively easy. We can be used linear model to extract the linear combination of variables for each axis (the coefficients given in the Table 16).

Table 16. Regression coefficients associated with the prediction of the two first axes of mixed analysis.

variable	RS1	RS2
(Intercept)	3.3026	1.6154
geomorphconstraint	0.8452	-0.3548
geomorphmeand	0.2599	-1.8144
geomorphsinuous	0.1085	-1.5475
fldplYes	-0.9023	-0.3104
watersourcePluvial	-0.2881	-1.8083
lDR	-0.2647	0.1076
ldist	-0.4524	0.1682

- **Method and sampling strategy**

The variable 'method' is not kept in the models because this variable is highly correlated with river size and the distribution of this variable is very **unbalanced**. For example, large rivers are mainly sampled by boat and represent only 6% of calibration dataset. A high unbalanced design can be important effect on the stability of the model parameters.

Table 17. Description of qualitative variables associated with sampling. For each dataset, we indicate the count of sites in each modality for a given variable. The calibrations sites are included in slightly disturbed sites (SID). The Chi-squared test provides a raw comparison among slightly disturbed and other sites.

variable	modality	CD (N=533)	Disturbed (N=7422)	SID (N=25	X-squared	df	p-value
Sampling.strategy	Partial	83	1795	414	65.844	1	< 0.001
	Whole	445	5627	2112			
Method	Boat	19	1753	164	377.122	2	<0.001
	Mixed	11	168	30			
	Wading	498	5501	2332			

2.3 Functional guilds and metrics

The calculation of the European Fish Index involves the use of metrics reflecting different aspects of fish assemblage integrity (i.e. tolerance guilds, habitat guilds, trophic guilds), taxonomic richness and individual abundance (e.g. documentation of FAME project[8]), Pont et al. 2006, 2007). There are several manners to use biological information contained in the guild table. Usually, biological characteristics can be considered at the species level (metric based on Richness) or at the individual level (metric based on abundance or density).

Table 18. Description of metric tables. For a given guild, we indicate the number and percentage of species included in each modality. The field 'tested' indicates if the metric participates in the modeling process.

Guild/ trait	Modality	N species (%)	Tested	Guild/ trait	Modality	N species (%)	Tested
WQgen	IM	58 (0.384)	x	Atroph	DETR	6 (0.04)	x
	INTOL	35 (0.232)	x		HERB	5 (0.033)	x
	TOL	55 (0.364)	x		INSV	69 (0.457)	x
	NoData	3 (0.02)			OMNI	36 (0.238)	x
WQO2	O2IM	72 (0.477)	x		PARA	3 (0.02)	x
	O2INTOL	44 (0.291)	x		PISC	13 (0.086)	x
	O2TOL	30 (0.199)	x		PLAN	13 (0.086)	x
	NoData	5 (0.033)			NoData	6 (0.04)	
WQTox	TOXIM	65 (0.43)	x	Repro	ARIAD	3 (0.02)	x
	TOXINTOL	34 (0.225)	x		LIPE	2 (0.013)	x
	TOXTOL	33 (0.219)	x		LITH	73 (0.483)	x
	NoData	19 (0.126)			OSTRA	1 (0.007)	x
WQAc	AIM	41 (0.272)	x		PELA	8 (0.053)	x
	AINTOL	59 (0.391)	x		PHLI	20 (0.132)	x
	ATOL	21 (0.139)	x		PHYT	24 (0.159)	x
	NoData	30 (0.199)			POLY	3 (0.02)	x
Temp	EUTHER	120 (0.795)	x		PSAM	5 (0.033)	x
	STTHER	29 (0.192)	x		SPEL	8 (0.053)	x
	NoData	2 (0.013)			VIVI	2 (0.013)	x
HTOL	HIM	55 (0.364)	x		NoData	2 (0.013)	
	HINTOL	53 (0.351)	x	HabSp	EUPAR	36 (0.238)	x
	HTOL	41 (0.272)	x		LIPAR	31 (0.205)	x
	NoData	2 (0.013)			RHPAR	81 (0.536)	x
Hab	EURY	50 (0.331)	x		NoData	3 (0.02)	
	LIMNO	33 (0.219)	x	ReproB	FR	40 (0.265)	x
	RH	67 (0.444)	x		PRO	13 (0.086)	x
	NoData	1 (0.007)			SIN	92 (0.609)	x
FeHab	B	86 (0.57)	x		NoData	6 (0.04)	
	WC	64 (0.424)	x	Mig	LONG-LMA	10 (0.066)	x
	NoData	1 (0.007)			LONG-LMC	3 (0.02)	x
PC	NOP	115 (0.762)	x		POTAD	49 (0.325)	x
	PROT	33 (0.219)	x		RESID	86 (0.57)	x
	NoData	3 (0.02)			NoData	3 (0.02)	

[8] FAME project: http://fame.boku.ac.at/

In our study, we mainly focus on three different types of variables: continuous, count and binary/proportion variables. As results, each variable was associated with particular model and specific distribution (e.g. Binomial, Poisson, etc…) and/or particular transformation (e.g. log-transformation for the count data): for example, we consider the density of insectivorous species, the number of benthic species and relative number of intolerant species. The guild table is described in details in the annex.

The exploration of guild table shows that the rates of missing values are relatively high for several metrics such as the intolerance to acidification (20% of missing data) or to toxic substance (12.6% of missing data). In addition, several modalities are relatively rare (Table 18). For example, the modality 'viviparity' in reproduction mode only concerns two species: Gambusia affinis (detected in 12 sites in France) and Gambusia holbrooki (detected in 10 sites in Spain, 10 sites in Italy and 98 sites in Portugal). It's clear that the particular localization of this metric is not adapted to the objective of our study because it's absolutely necessary to use representative metrics to obtain a efficient index usable at large scale. In addition, models associated with metrics including low species number (e.g. viviparity, herbivorous) are often associated with models characterized by a poor quality, because there is an insufficient number of event or individual to obtain an acceptable fit.

3 Metric Modelling

A first selection based on the simple criteria such as the residuals structure, good adjustment of the fitted value enables the reducing of the model number (Table 18). This first screening is essential, because for each modality of a given trait, we can compute about 5 different metrics (e.g. binary, count proportion data based on species number and count and proportion date based on fish number). Consequently, the consideration of all possible metrics provides about 300 different models. For this reason, we only present usable models which could be used to compute the final multi-metric index.

3.1 Statistical models

3.1.1 Model description

In Generalized Linear Model (GLM), each outcome Y is assumed to be generated from a distribution function in the exponential family and the mean μ of the distribution depends on the independent variables X. The GLM consists of three elements: a distribution function from the exponential family, a linear predictor ($g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon$) and link function g ($E(Y) = \mu = g^{-1}(\eta)$). The parameters estimation is based on the maximum likelihood (Nelder & Wedderburn 1972, McCullagh & Nelder 1989, Faraway 2006). For example, the linear model is a particular case defined by the *Gaussian* family and *identity* link.

Poisson distribution and logarithmic link are frequently used to model the count data. The addition of offset parameters enables the modeling of rate date with Poisson regression. The offset parameter can be seen as a way to impose a baseline value when comparing different population (e.g. McCullagh & Nelder 1989, Cameron & Trivedi 1998, McCulloch & Searle 2001, Hardin & Hilbe 2007). In our study, the offset is defined by total species number for metric based on richness and by total individual number for metric based on abundance.

$$\begin{aligned} \log\left(\frac{y_i}{n_i}\right) &= \alpha + \beta \mathbf{X}_i + \varepsilon_i \\ \log(y_i) - \log(n_i) &= \alpha + \beta \mathbf{X}_i + \varepsilon_i \\ \log(y_i) &= \alpha + \beta \mathbf{X}_i + \log(n_i) + \varepsilon_i \\ \hat{y}_i &= n_i \exp(\alpha + \beta \mathbf{X}_i) \end{aligned}$$

If the metrics are over-dispersed, we prefer to choose negative binomial distribution that is classical alternative to control the over-dispersion in regression analysis of count data. More details on proprieties of this particular model are available in McCullagh & Nelder (1989), Ripley & Venables (1999) and Cameron & Trivedi (1998).

As usual, our calibration dataset contains a high proportion of sites with a low Strahler order (e.g. small rivers) and the distribution of the sites in Europe is relatively unbalanced (see the distribution of calibration sites, Table 1 and Figure 1). Large rivers are clearly under-represented in the sample, but weighted up appropriately in the analysis to compensate. Moreover, the intermediate report on the evaluation of the present European Fish Index showed that the results from models are affected by geographical repartition (see deliverable 3.3). To reduce the potential effect of these potential biases, a specific weighting based on the regionalization proposed in Reyjol et al. (2007) and the Strahler order is systematically integrated in our models. To consider the non-linear responses of metric to environmental condition, we compute orthogonal polynomial of degree 2 for slope and July temperature (e.g. Jongman et al. 1995, Venables & Ripley 2002, Austin 2002).

Initially, we considered a common model for all metric based on the environmental variables. This approach was biologically more interesting because we could explore the relationship between the environmental conditions and the metric. However, in a predictive context, the consideration of all variable into our models could create overfitting and we observed bias in error estimation when using sites outside of environmental range of calibration dataset. Therefore, we give more importance to predictive capacity than to explicative capacity in using a stepwise procedure based on AIC (e.g. Venables & Ripley 2002, Pont el al. 2006). The new results appear to be better in the extrapolation situation.

3.1.2 Diagnostic and goodness of fit

In this section, we present some tools used to evaluate the quality of our models. Linear model and generalized linear model are based on specific assumptions such as homogeneity of residuals variances, non-multicollinearity, etc. We can use several indices tools to evaluate quality of a model. For example, the RSS associated with a good model follows a chi-squared-distribution with *n-p* degree of freedom (McCullagh & Nelder 1989):

$$RSS \sim \chi^2_{df}$$

Where *df, n* and *p* correspond to the degree of freedom, number of observations and number of parameters in the model respectively. Multicollinearity between explanatory variables was estimated by the variance inflation factor (VIF).

$$VIF_j = \frac{1}{1-r_j^2} = (\mathbf{R})^{-1}_{jj}$$

where **R** represents the correlation matrix between explanatory variables. r_j^2 is the coefficient of determination of the regression between the j^{th} explanatory variable and the other explanatory variables. It has been suggested that if *VIF* is higher than 10, multicollinearity may occur and biases the estimates (Belsley et al. 1980; Chatterjee et al. 2000, Fox 2002). We computed the average VIF (noted $\overline{VIF}$) as an index of multicollinearity (Chatterjee et al. 2000) as follows:

$$\overline{VIF} = \frac{\sum_{j=1}^{m} VIF_j}{m}$$

where *m* corresponds to the number of explanatory variables. For all selected models, the values of VIF are inferior to 3 units. It might seem, therefore, that we don't have multicollinearity phenomena in our models.

The interpretability and predictive power are two important characteristics, but other criteria must be considered in the model selection: Natural Handling of data of mixed type, Robustness to outliers, computational scalability (large N), ability to deal with irrelevant inputs and ability to extract linear combinations of variables (e.g. Snee 1977, Collett 2003, Hastie et al 2001, Faraway 2006). Graphical tools give also precious information on stability and model quality:

- QQ-plot representation of standardized residuals against normal theoretical quantiles and histogram of Pearson residuals (e.g McCullagh & Nelder 1989, Ben & Yohai 2004) provide information on **quasi-normality** of residuals.

- The potential influent points can be detected by the representation of the **leverage** values (hat values) against the standardized residuals.

- The **goodness of fit** is evaluated by the plot of observed values against expected values from generalized linear model.

- The graphic based on residuals in function of the fitted values (eta, fitted values in the 'link space') check the potential **heteroskedasticity** of residuals, dependence with fitted values and autocorrelation (see Figure 12).

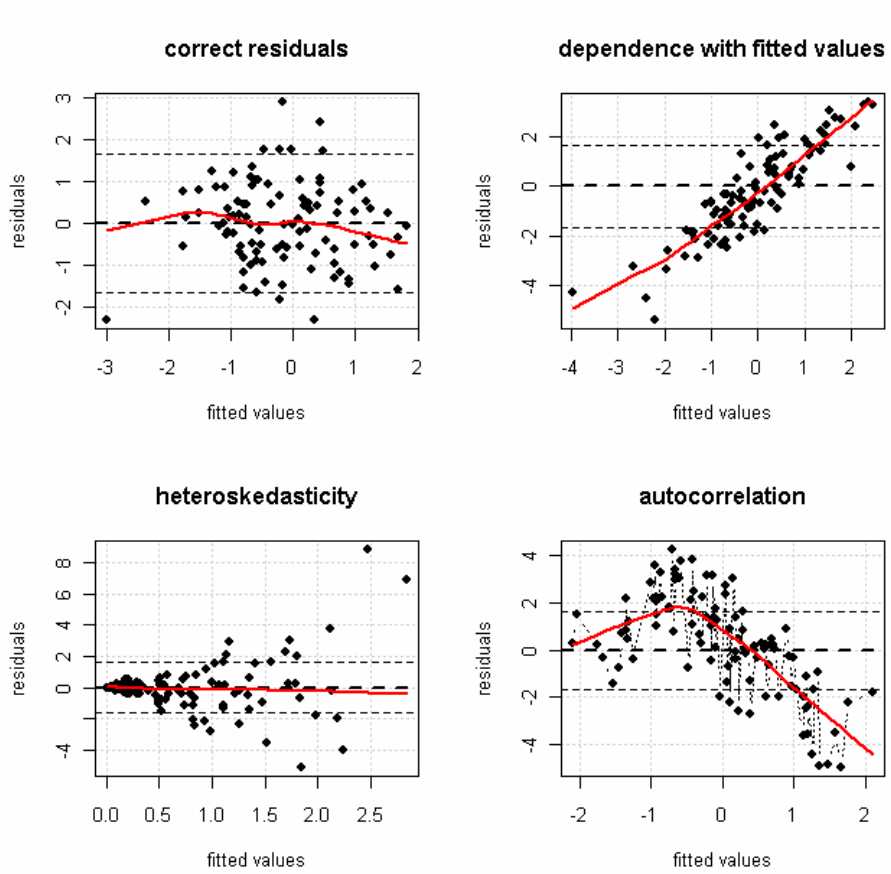


***Figure 12. Representation of four residuals structures: the first corresponds to a classical and right residual structure obtained by GLM. In the second graphic, the dependence between residuals and fitted values is violated. In the third graphic, the assumption of* homoskedasticity is not respected. The increase of the fitted values produces the increase of the residual variance. To conclude, the last graphic presents a typical problem of autocorrelation (lag=6).**

The evaluation of model is completed by internal-validation based on bootstrap technique (Davidson & Hinkley 1997, Efron & Tibshirani 1993). Then, error distribution of each model is estimated by 999 random samples with replace. The cost function (error) corresponds to **RMSE** (Root Mean Square Error) which provides indication on the divergence between the observed and predicted values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j}\left(y_i - \hat{y}_i\right)^2}$$

To illustrate the results of the internal-validation, we can represent the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997). This graphics provides interesting information on model stability (Figure 13).

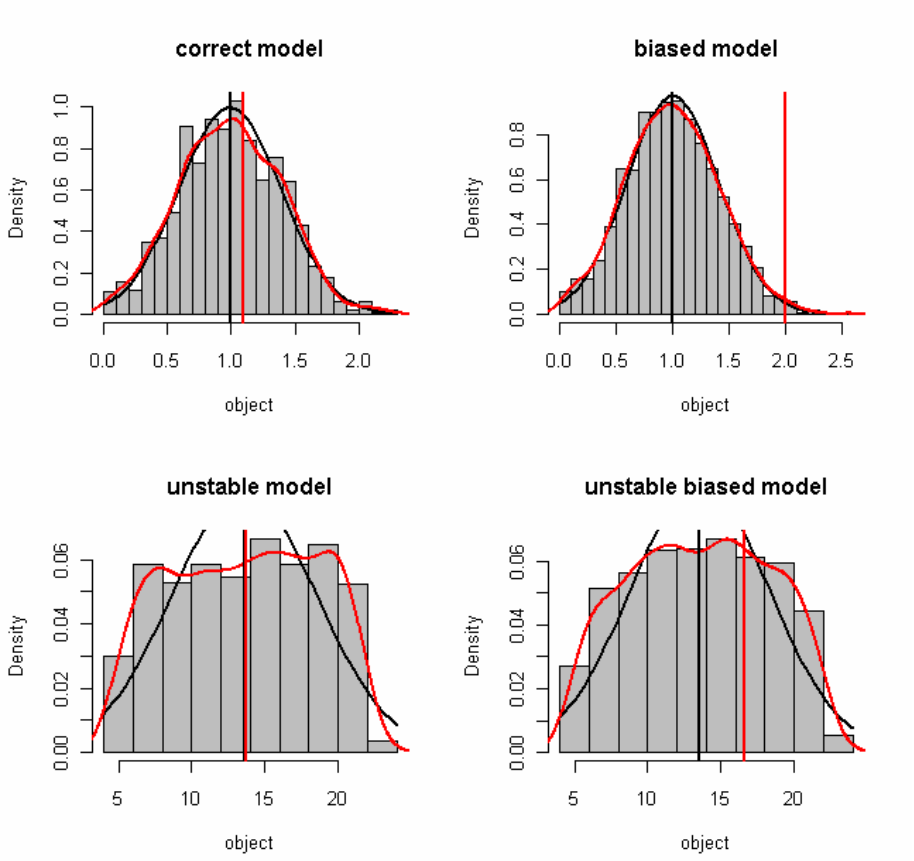


Figure 13. Representation of four bootstrap configurations: the first and second graphics show a classical shape (quasi-normal distribution), but we detect the presence of bias in the second graphic (high distance between the mean of simulation (black vertical line) and the observed value (red vertical lines)). The third and fourth graphic typically indicate that the models are relatively unstable (high variability). In addition, we observe bias between simulation mean (black vertical line) and observed values (red vertical lines) in the last graphic.

3.1.3 Implementation

Generalized linear models are performed with the function 'glm'. To compute bootstrap, we used the library 'boot' and the function 'boot'. For more detail concerning the statistical methods, we encouraged people to consult the appropriate bibliography. All the routines necessary for computing models were implemented in the R software (version 2.7.1, R Development Core Team 2007).

3.2 Metric based on species number

The metrics based on the species number are modelled by Poisson model with logarithmic link. We only observe the slight trend to underdispersion that have few influence on the modelling

process. However, when necessary, the scale parameter for Poisson regression models was held fixed to compensate for the effects of underdispersion.

Table 19. Regression Coefficients associated with metric models. The model selection is based on AIC stepwise procedure (see Venables & Ripley 2002).

metric	Ric.O2.Intol	Ric.Hab.Intol	Ric.Hab.RH	Ric.INSV	Ric.RH.Par
(Intercept)	-0.56668	-0.52566	-0.38701	-0.37452	-0.41193
poly(Tjul, 2)1	-4.28298	-2.62085	-0.94046		
poly(Tjul, 2)2	-0.39941	-0.89592	-1.26907		
poly(lslope, 2)1	2.17753	3.38169	3.1099	3.22686	4.12612
poly(lslope, 2)2	-0.77235	-0.09684	-0.13559	-1.21914	-0.67042
natsed:medium	0.01505	0.07003	0.0703	0.07053	0.08545
natsed:small	-0.45958	-0.33167	-0.28146	-0.31144	-0.12434
syngeomorph1	0.12815	0.04645		0.04483	
syngeomorph2	0.06781	0.05753			0.05822
Tdif			0.00921		

Metrics based on fish number are associated with the six environmental variables. However, we note that the thermal amplitude (Tdif, Table 19) is used only to model the number of rheophilous species. In contrast, river slope and natural sediment type are systematically selected in the models (lslope and natsed, Table 19).

3.2.1.1 Species intolerant to low Oxygen Concentration (Ric.O2.Intol)

The number of Species intolerant to low oxygen concentration is modelled by five environmental variables: July temperature (polynomial function of degree 2), slope (polynomial function of degree 2), natural sediment type and the both geomorphological components. The quasi-normality of the residuals is respected but, we observe that the values of residuals trend to decrease in function of the expected values (defined in the space link). The bootstrap histogram is characterised by the slight asymmetry and we also observe a slight bias between simulation mean and observed values. Despite some leverage points, the goodness of fit is acceptable. This model is an acceptable candidate for the final aggregation.

Table 20. Deviance Analysis of the generalized linear model selected by stepwise procedure. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	382.76	
poly(Tjul, 2)	2	73.77	530	308.99	< 0.001
poly(lslope, 2)	2	69.53	528	239.46	< 0.001
natsed	2	11.14	526	228.32	0.004
syngeomorph1	1	15.15	525	213.17	< 0.001
syngeomorph2	1	5.87	524	207.30	0.015

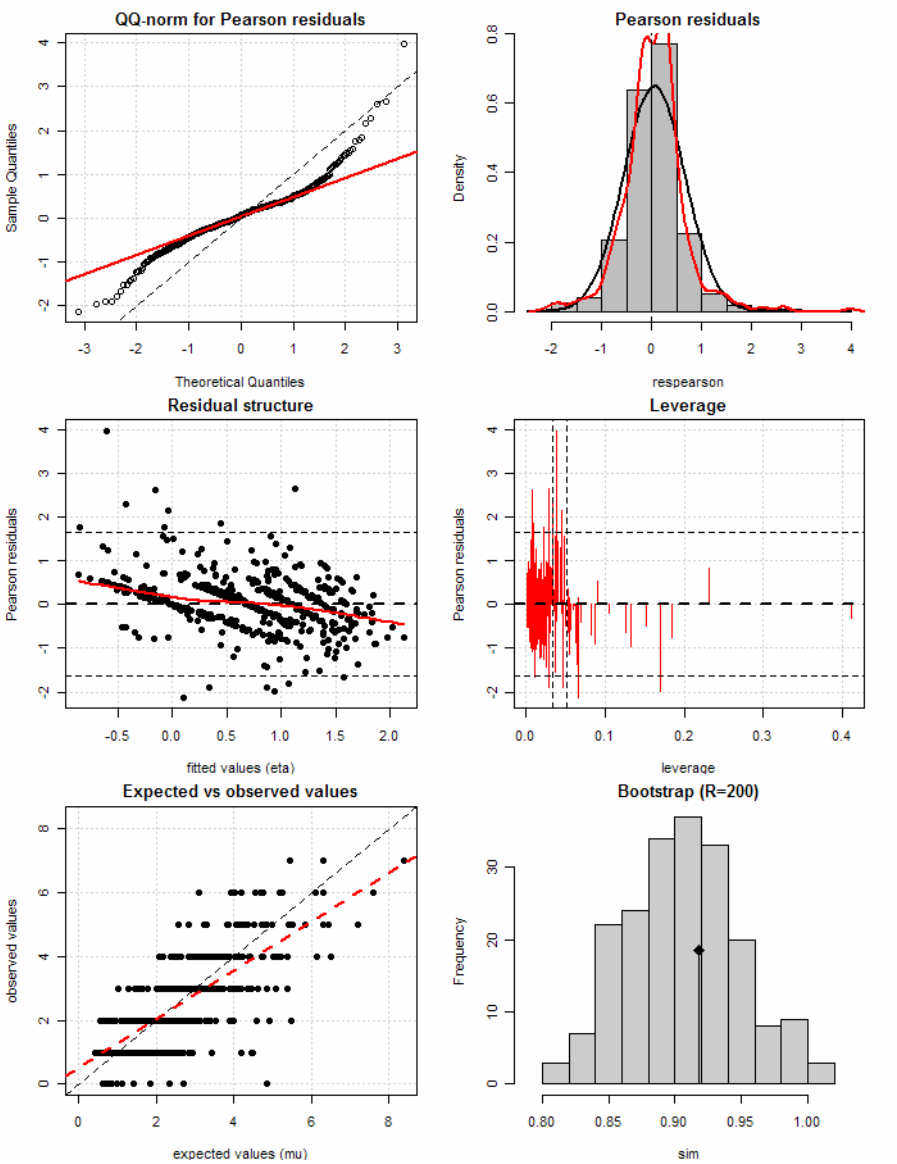


Figure 14. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardised residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

3.2.1.2 Species intolerant to Habitat degradation (Ric.Hab.Intol)

Selected model obtained by the stepwise procedure integrates five variables: slope (polynomial function), July temperature (polynomial function), natural sediment type and the two geomorphological variables (Table 21). The quasi-normality of the residuals is respected but, we

observe that the values of residuals slightly trend to decrease in function of the expected values (defined in the space link). The bootstrap histogram is characterised by the slight asymmetry and we also observe a slight bias between simulation mean and observed values with few effect on model stability. Model adjustment (representation expected and observed values, Figure 15) is satisfactory in spite of slight phenomena of over and underestimates for low and high values, respectively.

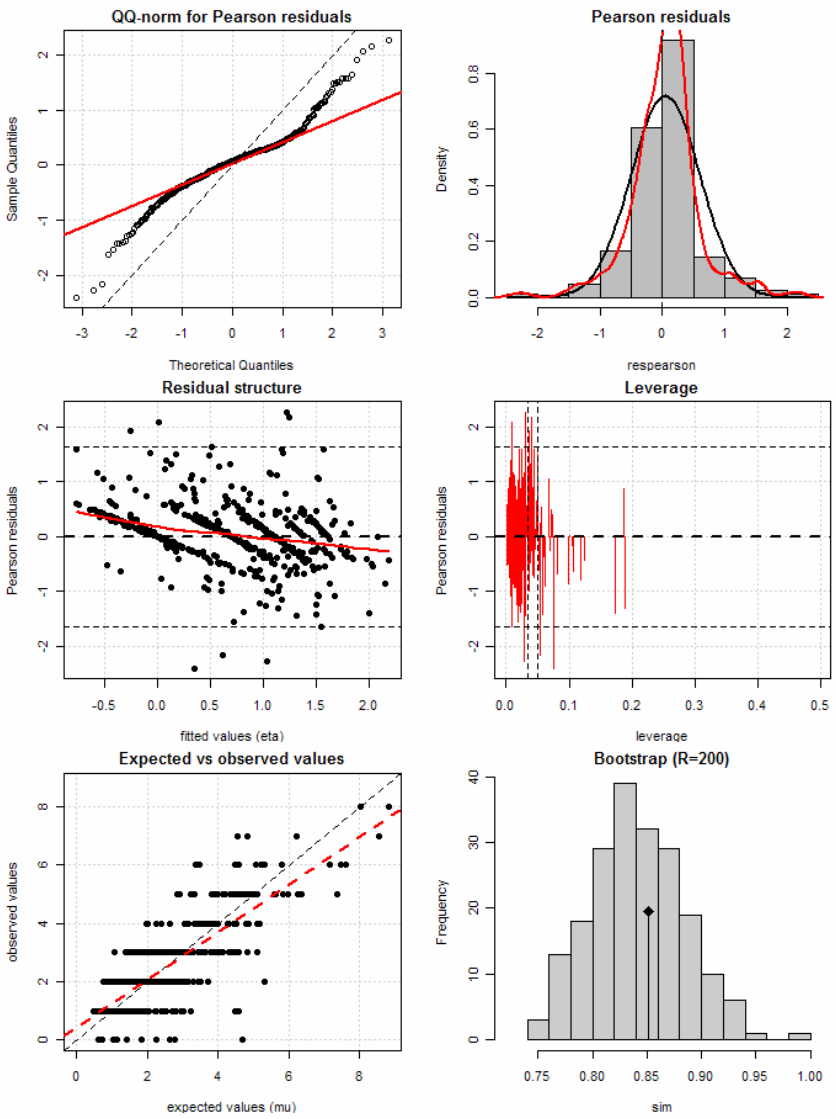


Figure 15. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

Table 21. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	283.54	
poly(lslope, 2)	2	46.76	530	236.78	< 0.001
poly(Tjul, 2)	2	38.61	528	198.18	< 0.001
natsed	2	12.21	526	185.97	0.002
syngeomorph2	1	3.31	525	182.66	0.069
syngeomorph1	1	2.58	524	180.08	0.108

3.2.1.3 Rheophilous Species (Ric Hab RH)

Rheophilous species is modelled by four variables: natural sediment type, slope (polynomial function), July temperature (polynomial function) and the thermal amplitude corresponding to the difference between the July and January temperature (Table 22). For this model, we note that the quasi-normality of the residuals is respected but, we observe that the values of residuals trend to decrease in function of the expected values (defined in the space link). The bootstrap histogram is characterised by the slight asymmetry and we also observe a slight bias between simulation mean and observed values. The coherence between expected and observed values is satisfactory in spite of slight phenomena of over and underestimates for low and high values.

Table 22. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	156.36	
natsed	2	34.93	530	121.42	< 0.001
poly(lslope, 2)	2	11.98	528	109.44	0.002
poly(Tjul, 2)	2	12.46	526	96.98	0.002
Tdif	1	2.17	525	94.81	0.141

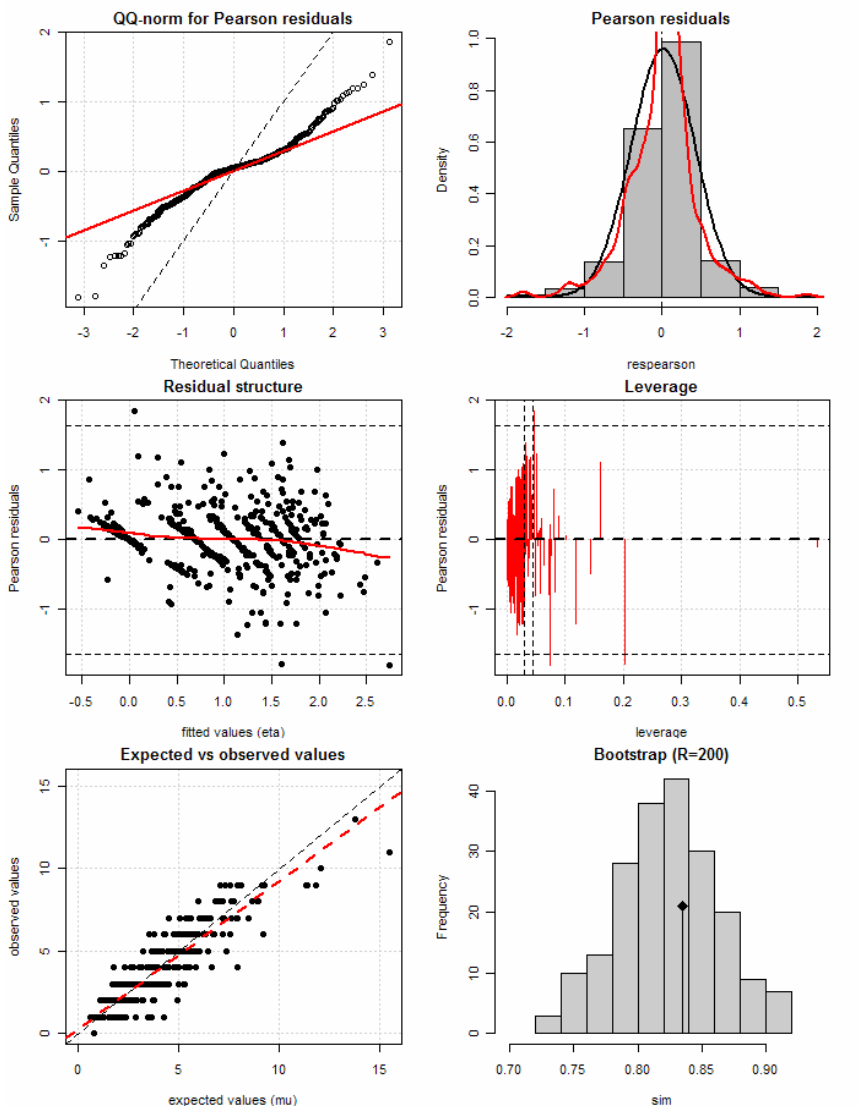


Figure 16. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

3.2.1.4 Insectivorous Species (Ric INSV)

The selected Model associated with the metric based on Insectivorous species only integrates three variables: slope (polynomial function of degree 2), type of natural sediment and the first geomorphological variable (Table 23). The residuals are approximately normal and there is not particular relationship between residuals and fitted values (see figure 1, representation of residuals structure). Adjustment of expected and observed values is acceptable and we just observe the recurrent problem of over and underestimation of the low and high values. A negative point in

favour of low model stability is the particular shape of the RMSE distribution (histogram of simulated RMSE, Figure 17).

Table 23. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	194.59	
poly(lslope, 2)	2	63.85	530	130.74	< 0.001
natsed	2	11.44	528	119.30	0.003
syngeomorph1	1	3.40	527	115.91	0.065

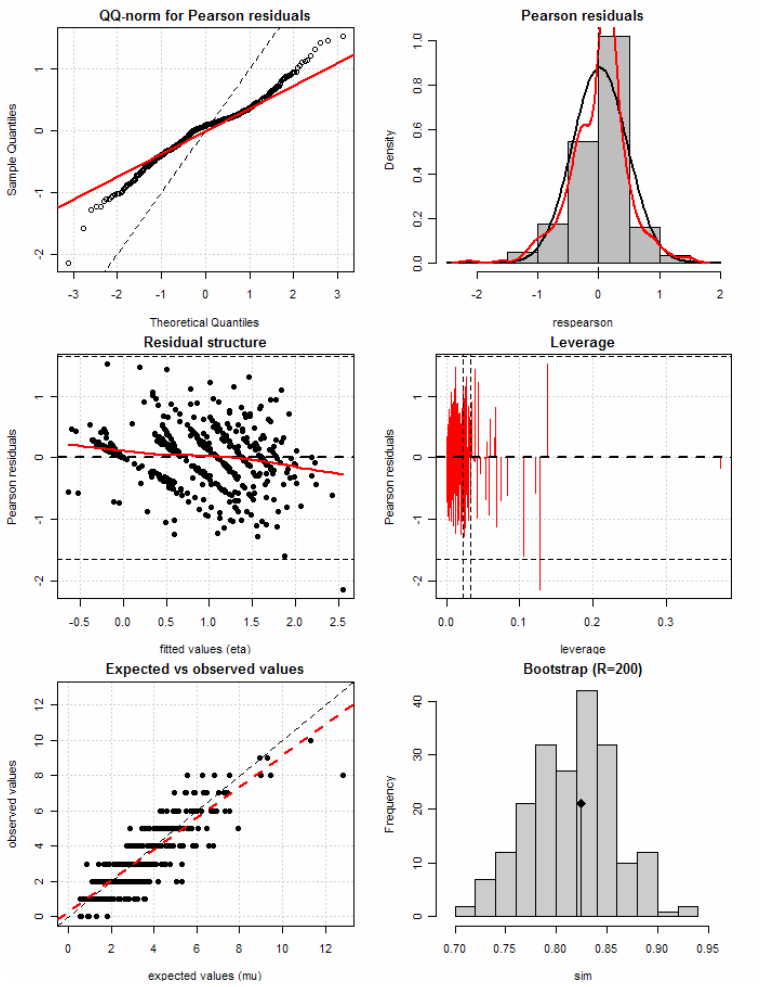


Figure 17. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

3.2.1.5 Species with preference to spawn in running waters (Ric RH Par)

The modelling of number of species intolerant to acidification is relatively good. The selected model obtained by the stepwise procedure integrates three environmental variables: slope (polynomial function of degree 2), natural sediment type and the second geomorphological variables (Table 24). The quasi-normality of residuals is respected and the adjustment between the observed and expected values is correct. The representation residuals in function of the fitted values (eta, fitted values in the 'link space') show a slight relation between residuals and fitted values. As with other models, we only observe some potential leverage points.

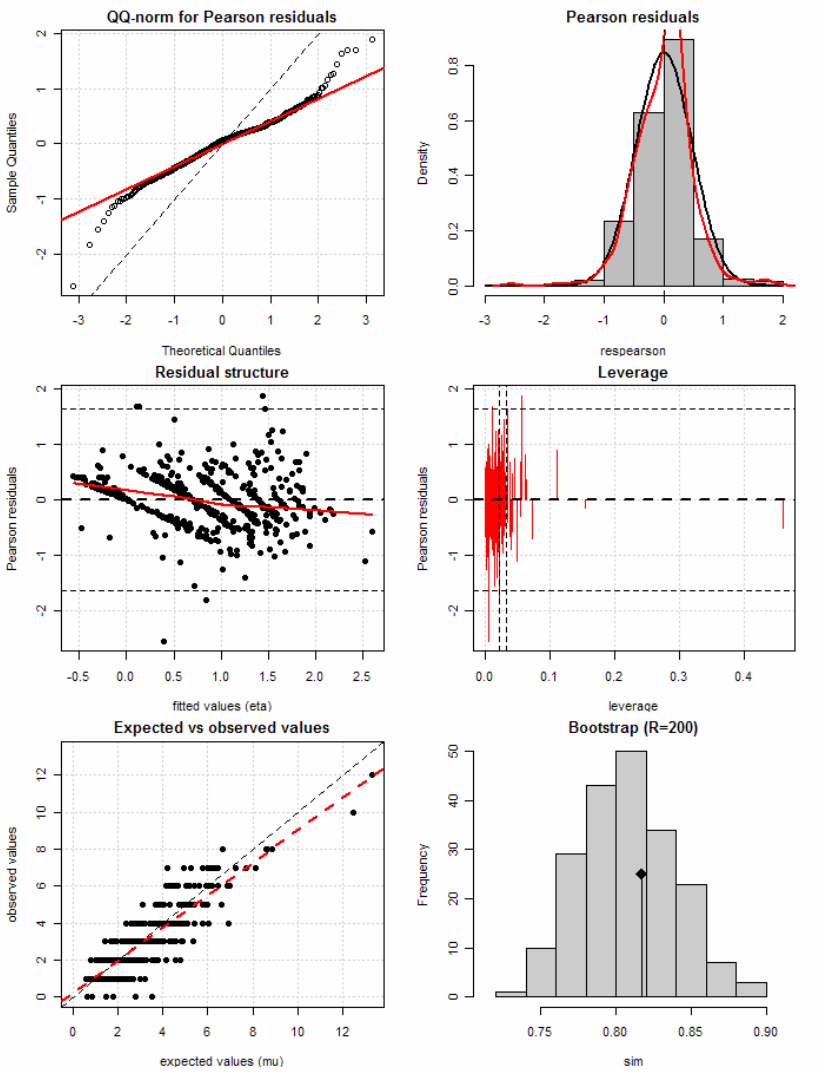


Figure 18. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

Table 24. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	200.51	
poly(lslope, 2)	2	53.91	530	146.60	< 0.001
syngeomorph2	1	7.92	529	138.68	0.005
natsed	2	5.31	527	133.37	0.070

3.3 Metric based on fish number

As mentioned earlier, to model the metrics based on the fish number, we did not used Poisson model regression. We prefer to choose negative binomial distribution because the metric based on fish number are largely over-dispersed. The Negative binomial model is classical alternative to control the overdispersion in regression analysis of count data. More details on propriety of this probability distribution are available in McCullagh & Nelder (1989), Ripley & Venables (1999) and Cameron & Trivedi (1998). To complete the model description, a logarithmic link and an offset parameters based on the total number of fish are added to provide an efficient solution for reducing of the sampling effect.

Table 25. Regression Coefficients associated with metric models. The model selection is based on AIC stepwise procedure (see Venables & Ripley 2002).

metric	Ni.O2.Intol	Ni.Hab.Intol	Ni.INSV	Ni.LITHO
(Intercept)	-0.27832	-0.25906	-0.22434	0.07676
poly(Tjul, 2)1	-3.21395	-2.29768	-0.03183	
poly(Tjul, 2)2	-0.01514	-0.16437	-0.66279	
poly(lslope, 2)1	1.36557	2.08897	2.39186	2.12899
poly(lslope, 2)2	-2.13935	-1.84799	-1.4052	-0.73348
natsedmedium	-0.05848	-0.04238	-0.03202	0.0485
natsedsmall	-0.5376	-0.41005	-0.37429	-0.37812
syngeomorph1	0.13998	0.09535	0.07223	
syngeomorph2		0.02592	-0.02891	0.03907
Tdif				-0.01644

Concerning the results of the stepwise procedure, we once again observe that the thermal amplitude (Tdif) is used only for one model (number of Lithophilic individual, Table 25) and that river slope and natural sediment type are systematically selected (lslope and natsed, Table 25).

3.3.1.1 Fish intolerant to low Oxygen Concentration (Ni.O2.Intol)

The model based on the fish intolerant to low Oxygen concentration is based on Temperature criteria (Tjul), natural sediment, slope and the first geomorphological variables (Table 26).

Table 26. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	738.99	
syngeomorph1	1	44.70	531	694.29	< 0.001
poly(Tjul, 2)	2	28.75	529	665.54	< 0.001
natsed	2	25.16	527	640.38	< 0.001
poly(lslope, 2)	2	8.54	525	631.84	< 0.001

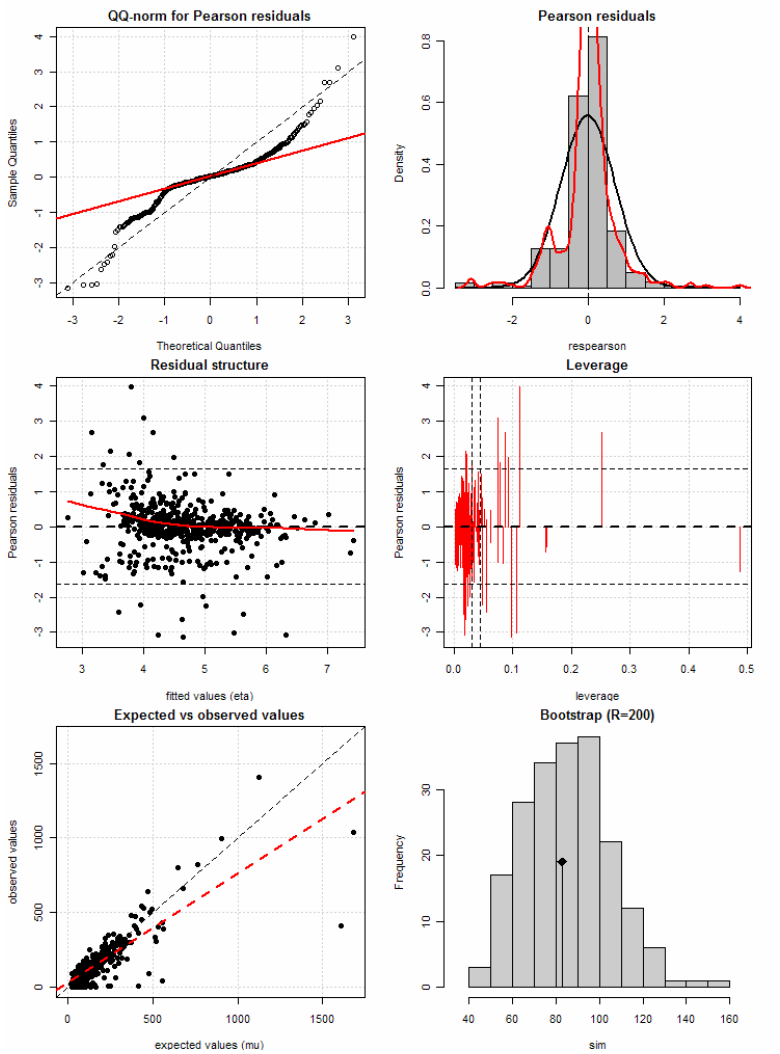


Figure 19. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

After some additional verification at different observation scales (not show in this report), we conclude that the adjustment of expected values to observed values is relatively acceptable. Bootstrap representation highlight an asymmetric distribution of RMSE and a slightly bias between the mean of simulation and observed values induced by extreme values and some leverage points. Pearson residuals approximately follow a normal distribution.

3.3.1.2 Fish intolerant to Habitat degradation (Ni.Hab.Intol)

The modelling of the metric fish number intolerant to habitat degradation involves five environmental variables: slope (polynomial function of degree 2), July temperature (polynomial function of degree 2), natural sediment type and the both geomorphological variables (Table 27). The quasi-normality of the residuals is approximately respected, but we observe a particular structure in these values induced by the reducing of the residuals value associated with the increase of the predicted values (link-transformed). As results, the hypothesis of homokedasticity is poorly respected. However, the model provides acceptable adjustment of expected values to observed values and bootstrap procedure indicates satisfactory model stability (Figure 20).

Table 27. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	725.34	
poly(lslope, 2)	2	39.34	530	686.00	< 0.001
poly(Tjul, 2)	2	31.44	528	654.57	< 0.001
natsed	2	10.27	526	644.30	< 0.001
syngeomorph1	1	11.65	525	632.65	< 0.001
syngeomorph2	1	1.06	524	631.59	0.147

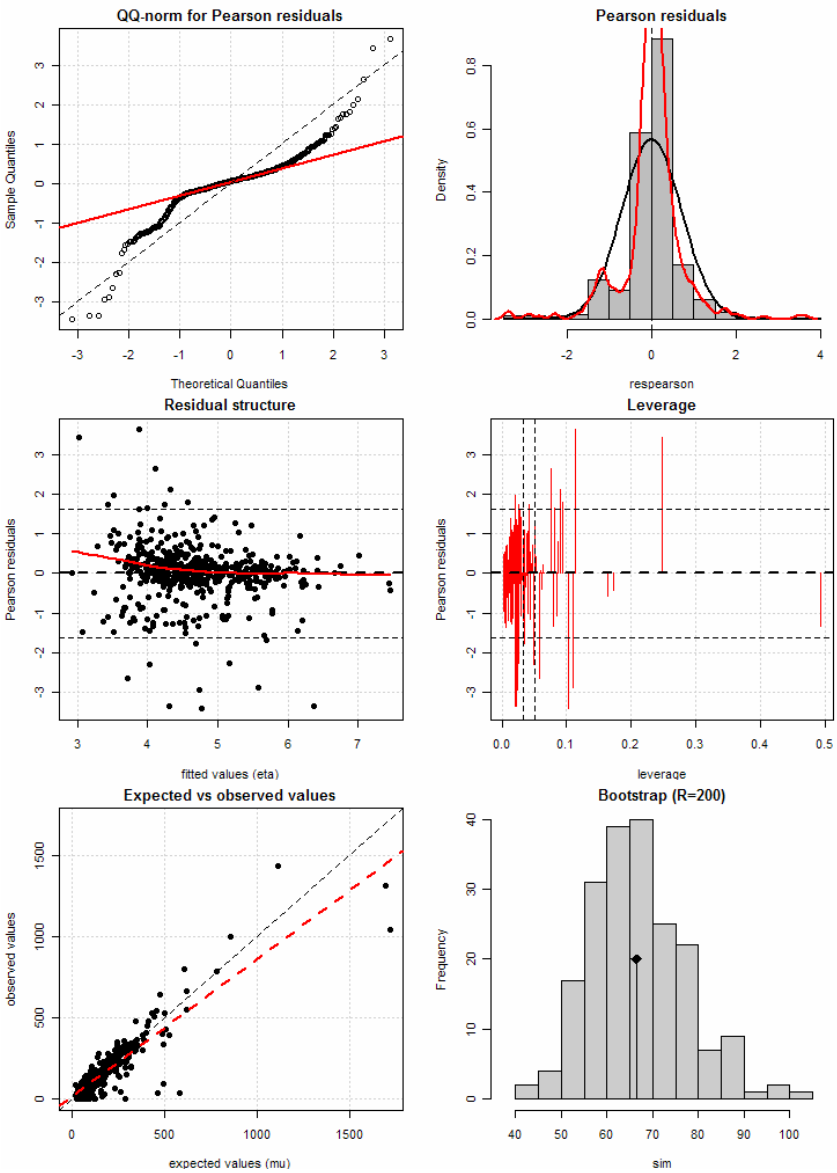


Figure 20. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

3.3.1.3 Insectivorous Fish (Ni.INSV)

Model selected by the stepwise procedure integrates five variables presented in the Table 28 and Table 19: slope (with polynomial function of degree 2), type of natural sediment, temperature (with polynomial function of degree 2) and the both geomorphological variables. For this metric, the graphics proposed in the Figure 21 show that residuals are relatively unstructured

and variance heterogeneity is low. On the other hand, the quasi-normality contraints is poorly respected. With regard to model adjustment, we observe an acceptable appropriateness of expected values to observed values. Resampling procedure based on the RMSE statistic doesn't highlight model instability. As results, we could consider this metrics as a potential candidate for the final aggregation.

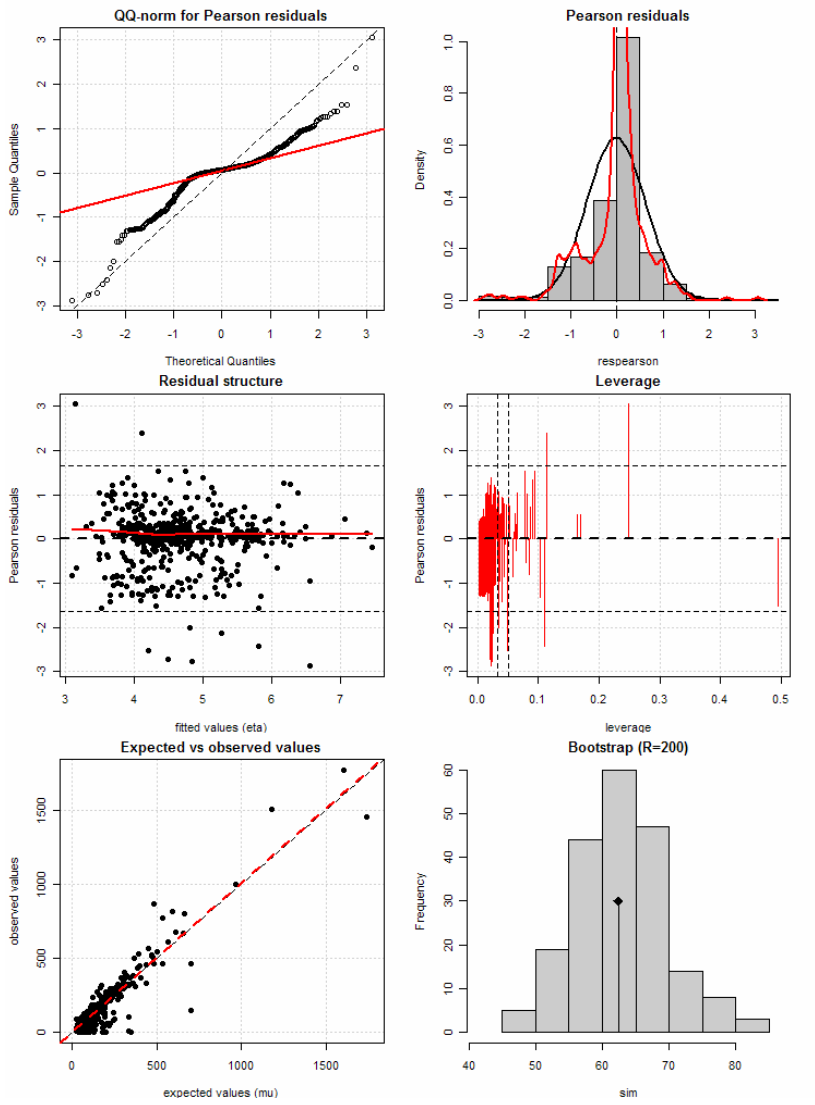


Figure 21. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

Table 28. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	500.04	
poly(lslope, 2)	2	41.52	530	458.52	< 0.001
natsed	2	7.43	528	451.09	< 0.001
syngeomorph1	1	8.07	527	443.02	< 0.001
syngeomorph2	1	2.15	526	440.86	0.022
poly(Tjul, 2)	2	1.84	524	439.02	0.106

3.3.1.4 Lithophilic Fish (Ni.LITHO)

For the metric based on fish number of lihtophilic species, the setpwise procedure provides a model based on four variables: natural sediment type, slope, themral amplitude (difference between the July and January temperature) and the second geomorphological variable (Table 29). The quality of this model is globally acceptable. The quasi-normality of residuals is not respected much and the adjustment between the observed and expected values is correct. Additional tests showed that the modification of the observation scale doesn't affect the relationship between expected and observed values and bootstrap representation confirms the good stability of the model. In the representation of residuals and the fitted values (eta, fitted values in the 'link space'), we don't observe relationship between these two components. As with other models, we only observe some potential leverage points. This metrics could be an interesting candidate for the final aggregation.

Table 29. Deviance Analysis of the generalized linear model. The terms 'Df', 'Deviance', 'Resi. Df', 'Resid. Dev' and 'P(>\|Chi\|)'correspond to the degree of freedom, the deviance, residuals degree of freedom, residuals deviance and p-value associated with Chi-squared test.

variable	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			532	458.08	
natsed	2	45.47	530	412.62	< 0.001
poly(lslope, 2)	2	21.86	528	390.76	< 0.001
Tdif	1	6.14	527	384.62	< 0.001
syngeomorph2	1	3.84	526	380.78	0.002

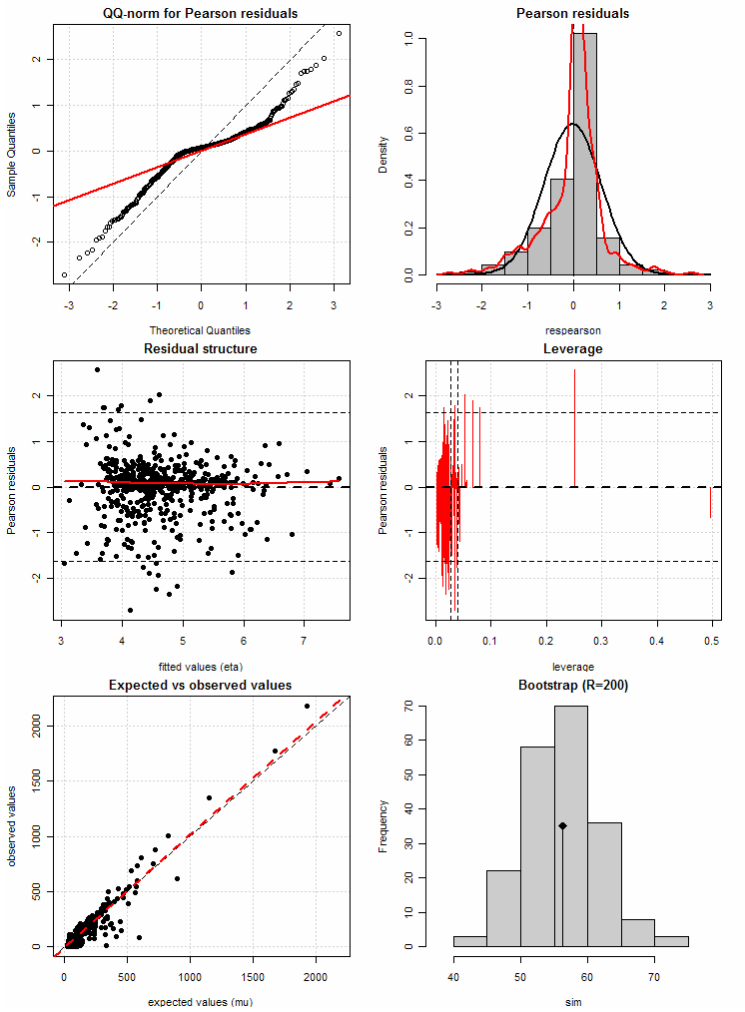


Figure 22. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley 1997).

3.4 Metrics based on Fish Length

Multimetric indices based on fish assemblages, like the European Fish Index (Pont *et al.*, 2006; Pont et al., 2007), have already been demonstrated to be useful tools to assess the "ecological status" of water bodies. Nevertheless, those indices based on aggregation of individuals metrics, seems to be less efficient when applied on low species rivers and especially to headwater systems. In Europe, low local richness mostly concerned the upper part of streams. Headwater systems exhibits particular environment: high slope, cold and well oxygenated water and coarse substrate. At the European scale, species inhabiting low richness rivers shared very close ecological niche, thus belonging to similar guilds. This explains, at least in part, the low

variability of individuals metrics observed in headwater assemblages. Nevertheless; rivers of the Mediterranean and Scandinavian part of Europe present low local richness also in their downstream part.

One of the objectives of the EFI+ project was to develop new metrics, specific to low species rivers, to try to improve the assessment of those particular areas. In addition to the metrics based on guilds, we focused on age and size class of fish assemblages. By working on size or age classes we would take into account on other characteristic of fish assemblages very seldom used in bioassment. Moreover, the use of size or age classes is a requirement of the Water Framework Directive.

We have oriented our research along two axes. First we wanted to develop a metric based on age classes of one species. At the European scale, brown trout (*Salmo trutta fario*) appeared to be the most appropriated species for this experiment. Indeed brown trout is a widespread and abundant species, occurring in low species rivers (especially in upper part of streams) and present in all member countries of the EFI+ project. It's also the species for which we have the greatest number of lengths recorded. Our objective was to separate, for each fishing occasions, the young of the year from the older fishes and then to compute new metrics based on those age classes. For each age class, we tested metric based on absolute and relative abundances.

Secondly, we wanted to develop new metrics based on the combination of guilds and size class. The main objective was to distinguish small fishes from larger ones in all assemblages. To separate fishes in those two categories we used a threshold length. All fishes with an individual length lower than this cut-off were consider as "small" fishes and the others were considered as "large" fishes. New metrics were developed based on the abundance of fishes in each size class. Among this experiment we have also considered two possible ways to compute those metrics. In the first approach, we wanted to focus on the part represented by large or small individuals sharing a specific trait in whole fish assemblages. In the second approach, by removing individuals which are obligatory smalls due to the limited size that they can reach, we wanted to focus on the parts represented by early or late life stages of large species. Here, individuals are small because their growth were limited by their life time (at least partially) and not by a life history traits associated of their species.

As for all metrics tested, we firstly tried to rely the natural variability of the new metrics with environment, and then if the goodness of fit of models were enough satisfying we tested the sensitivity of those metrics to human pressure and we retained the metrics presenting the highest sensitivities.

3.4.1 Environmental variables and data sets definition

To link the field variability of the metrics to environment we had only considered six environmental variables:

- Slope: the river slope (always transformed in natural logarithm);
- July mean air temperature;
- Difference between July and January mean air temperature;
- Syngeomorph1, the coordinates of sites on the first axes of an Hill & Smith analysis, inversely related with longitudinal gradient;
- Syngeomorph2, the coordinates of sites on the second axes of an Hill & Smith analysis, mostly influenced by the geomorphology and by stream water source type;
- The size of sediment naturally occurring in sampling sites (small, medium and large).

More details concerning all the environmental variables retained for the modelling of metrics are available in the section 2.2 . All metrics which were developed were defined for the salmonid sites. This because the main objective of this experiment was to develop metrics for low species rivers, which are mostly located in salmonid reaches or streams excepted in Mediterranean areas.

3.4.1.1 Experiment on the brown trout, Salmo trutta fario.

3.4.1.1.1 Developing a tool to estimate the cut-off between young of the year and older fishes

As ages of fish were not directly available, we had to estimate them from length distributions. For fishing occasions where the brown trout is abundant, the distributions of the young of the year are often well separated from the distribution of older fishes. We supposed lengths distribution of brown trout to be a mixture of normal law. The threshold length was computed as a quantile of the normal law corresponding to the YOY. The parameters were computed with an EM algorithm (Young *et al.*, 2008). For this step we used both reference and calibration sites to increase the number of data. We have only used sites with more than 50 individuals sized to have enough data to estimate the parameters. Finally we had a data set composed of 105 sites.

To be able to estimate a cut-off for each fishing-occasion, we have fitted a model relating the thresholds previously estimated with environment. Then for each fishing occasion we fitted a threshold based on their environment and we counted the number of YOY and OLD.

3.4.1.1.2 Definition of metrics

From the previous estimations we have computed six metrics:

- Abundance of YOY
- Abundance of OLD
- Density of YOY (abundance divided by the fished area)
- Density of OLD
- Proportion of YOY
- Proportion of OLD

3.4.1.1.3 Selection of calibration data set

From the calibration data set we selected fishing occasions sampled between August and November in order to reduce as far as possible a potential temporal effect. We considered only fishing occasions where the brown trout was dominant to reduce a potential bias in length distribution due to biotic interactions. We removed all fishing occasions with to high intervals between size classes. Finally the calibration data set available was composed of 189 fishing occasions. We only took into consideration fishes caught during the first run.

3.4.1.1.4 Modelling of metrics

Depending of the nature of the metrics (abundance or density) we used different kind of models to explain the field variability of the metrics by the environment.

For the metrics based on abundance we fitted Generalized Linear Models (GLM, (Mccullagh & Nelder, 1989) assuming a Poisson distribution or a Negative Binomial distribution if the metrics were overdispersed (Cameron & Trivedi, 1998; Venables & Ripley, 2002). We have also tested to fit rate models of the abundance of the size class considered on the number of trout. Thus, we have integrated the logarithm of the number of trout caught as offset in the previous models (Cameron & Trivedi, 1998).

For the metrics based on densities, due to the skewness of the distributions we modelled the natural logarithm and the fourth square root of the densities as linear combinations of environmental variables. With the logarithm transformations we have also tried to fit models with an offset on the natural logarithm of the number of captures.

For the metrics based on proportion we used logistic regressions (Hosmer & Lemeshow, 2000) and multiple linear regressions on the arcsines square root transformations (Table 30).

To counter balance the disequilibria between ecoregions in the models, we used weigths (excepted for the logistic regressions) such as the sum of weights per region were similar. We also used a stepwise procedure based on Akaike's information criteria (AIC) to select the best combination of environmental variables to explain each metric, whatever the model considered.

Table 30. Summary of the transformations and of the models used to explain metrics variability.

Metric based on	Transformation	Model	Probability law	Offset
Abundance	None	GLM	Poisson	No
	None	GLM	Poisson	Logarithm Number of trout
	None	GLM	Negative Binomial	No
	None	GLM	Negative Binomial	Logarithm Number of trout
Density	4th square root	MLR	Gaussian	No
	Logarithm	MLR	Gaussian	No
	Logarithm	MLR	Gaussian	Logarithm Number of trout
Proportions	Logit	GLM	Binomial	No
	Arcsines square root	MLR	Gaussian	No

3.4.1.2 Crossing metrics based on guilds and size classes

3.4.1.2.1 Definition of metrics

As species inhabiting the "salmonid" areas most often presented very close biological or ecological traits, we focused only on eight metrics: general intolerant (INTOL), oxygen intolerant (O2INTOL), habitat intolerant (HINTOL), rheophilous (RH), insectivorous (INSV), potamodromous (POTAD), lithophilic (LITH) and single reproduction (SIN.B).

For each metrics we have considered two subsets of species: all species and large species. Species were considered as large if their maximal length was greater than 300 mm. For those two subsets we have computed the number of small individuals and the number of large individuals. Conversely to the experiment on the brown trout, we used defined cut-offs to distinguish small and large individuals: 100 mm, 150 mm and 200 mm. Thus for each guilds we have computed twelve news metrics all based on the abundance of individuals from a given size class (Table 31). Finally we have developed and tested 96 different metrics.

Table 31. Summary of the new metrics developed for one a specific guild.

Species subset	Threshold	Size class
All	100	Small
All	100	Large
All	150	Small
All	150	Large
All	200	Small
All	200	Large
Large	100	Small
Large	100	Large
Large	150	Small
Large	150	Large
Large	200	Small
Large	200	Large

3.4.1.2.2 Selection of calibration data set

From the 528 fishing occasions composing the calibration data set we have only retained the fishing occasions located in "salmonid" reaches. We have also selected fishing occasions depending of the sampling date to limit a potential temporal effect in the analysis and also to have enough data. We retained only the fishing occasions sampled between August and November. All fishing occasions without all lengths available for all individuals from a given metrics were removed. Thus the data sets were slightly different between metrics. Nevertheless, the number of fishing occasions was fairly close two 200.

3.4.1.2.3 Specific Modelling of metrics

For the metrics computed on the whole set of species, rather than linking the variation of abundance of small or large fishes with environment, we wanted to explain the variation of their ratio in fish assemblages by the environment. We used GLM with Negative Binomial law due to the overdispersion of data. To model the rate of small or large fishes of a given metrics, we added for each model the logarithm of the number of fish caught in offset. In GLM, the mean (μ) is related to independent variables threw a link function. The link function is expressed as a linear combination of environmental variables. Consequently by using a Negative Distribution and an offset on the number of captures, the logarithm of the ratio of small or large fishes in assemblage was supposed to be a linear combination of environmental variables, i.e the ratio of small rheophilous fishes in the assemblages.

$$\log(Ni) = \log(captures) + f(environment) \quad (1)$$

$$\log(Ni) - \log(captures) = f(environment) \quad (2)$$

$$\log\left(\frac{Ni}{captures}\right) = f(environment) \quad (3)$$

With, Ni the number of small or large individuals; captures, the total number of individuals sampled.

For the metrics computed with large species only, we followed the same methodology excepted for the offset. When we considered all species, small individuals were a mixture of individuals of small species and of the young individuals of large species. Whereas by considering only large species, the small individuals could be associated to young individuals of and large individuals to the adults. Thus we were interested here by the ratio of small or large fishes among the total number of fishes of the considered metrics, i.e. the ratio of small rheophilous fishes among the rheophilous individuals (of large species). Consequently we added the logarithm of the total number of fishes of the metrics of interest as offset in the models.

The diagnostic of the models were done, using the same procedure than for the general metrics. We checked: the normality of standardised residuals, the heteroskedasticity of residuals, leverage points, goodness of fit and stability of models. For more details see the paragraph 'diagnostic and goodness of fit'. Only metrics with models respecting sufficiently those criteria were considered as potential metrics. We then tested the sensitivity of those selected metrics to human pressure.

3.4.1.3 Brown trout experiment

Determining age of fishes

The limits between the length distribution of young of the year and older fishes, were determined with the mixture of normal distributions for 105 fishing occasions. The multiple linear regression between the cut-offs and environmental variables explained 54% of the variance (F= 16.51, p-

value<0.001) and the root mean square error was of 11.6 mm (computed on the fishing occasions of the model). The cut-offs was explained by five environmental variables each heavily significant (Table 32). With this model we were able to compute the lengths determining the limit between YOY and older fishes for all fishing occasions.

Table 32. Summary of the coefficients of the model linking the cut-offs and the environmental variables with: the logarithm of the size of catchments (lctach) and its squared (lcatch²), the mean annual temperature (temp.ann) and its squared (temp.ann²), the julian day (ranging from 1 to 366), the siliceous geology (Geological.typology) and the difference of temperature between July and January. All coefficients were provide with their standard deviation (Sd), the t value statistic associated to the student test of the coefficient and the p-value of the test.

Independant variable	Coefficient	Sd	t	p-value
Intercept	40.873	16.502	2.477	<0.05
lcatch	64.518	12.439	5.187	<0.001
lcatch²	25.996	12.366	2.102	<0.05
temp.ann	99.146	25.161	3.940	<0.001
temp.ann²	-55.324	12.549	-4.409	<0.001
julian day	0.130	0.050	2.591	<0.05
Geological.typology : Siliceous	-10.323	2.890	-3.572	<0.001
Tdif	1.680	0.585	2.873	<0.01

3.4.1.3.1 Environment of the calibration data set for the metrics

Among the calibration data set (528 sites), 189 sites were retained for the computation of the models. Those sites were distributed across eleven of the fifteen member countries (Figure 23).Sites were mostly situated in streams presenting high slope, positive values of syngeomorph1 suggesting low distance from source and low size of catchments and natural medium to large natural sediments characteristic of headstream systems. In addition some sites were located in coastal streams, with lower slope and positive value of syngeomorph2 suggesting a pluvial hydrological regime (Table 33).

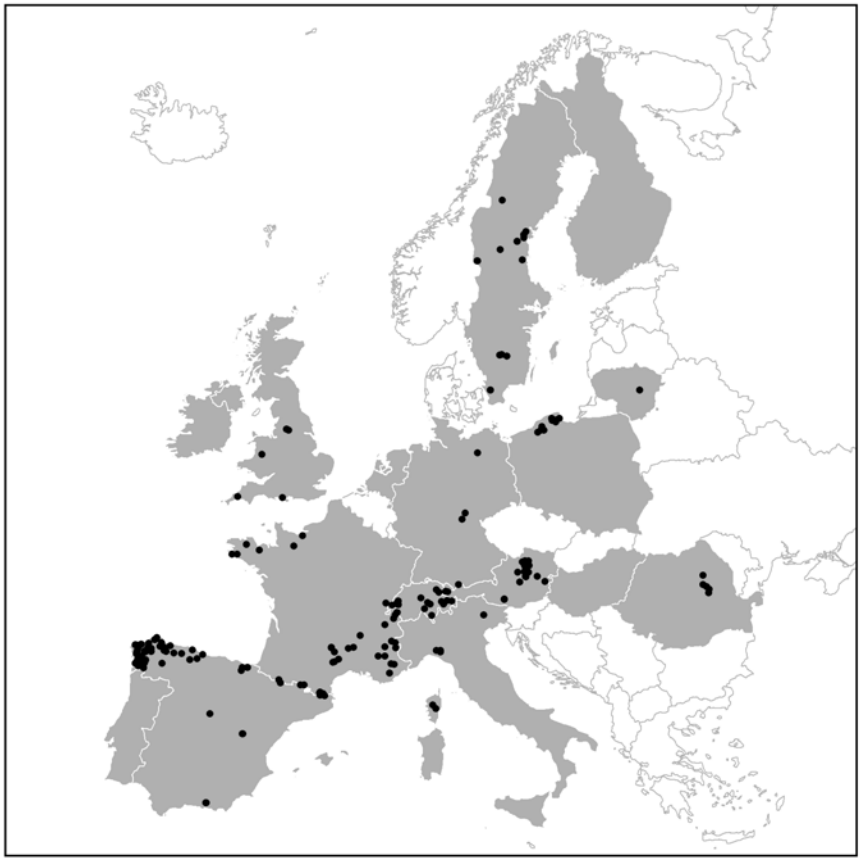


Figure 23. Distribution of the 189 calibration sites. Member countries of the EFI+ project are coloured in grey.

Table 33: Summary of the distribution of environmental variables of the calibration sites employed as independent variables in models.

Environmental variables	Statistics	
Slope	range	1 - 294.657
	mean (sd)	27.414 (35.345)
Tjul	range	8.6 - 21.6
	mean (sd)	17.039 (2.087)
Tdif	range	8.6 - 24.5
	mean (sd)	16.068 (4.301)
syngeomorph1	range	-0.446 - 3.406
	mean (sd)	1.622 (0.857)
syngeomorph2	range	-1.567 - 2.871
	mean (sd)	0.227 (1.177)
natsed	large	38
	medium	142
	small	9

3.4.1.3.2 Metrics selected

On the nine metrics tested (Table 31 and Table 34) based either on the abundance of YOY, either on the density of YOY or on the proportion of YOY in brown trout populations, none of the metrics satisfied sufficiently the statistical criteria. Thus, we didn't consider any of those metrics as a potential new metric for the index. Consequently we didn't test the sensitivity of those metrics to human pressure. The adequacy of the models fitted was always very low; whatever the metric considered (Table 34). Only one metric should have been retained, the abundance of brown trout, but this metric was too much depending of the offset.

Table 34: Summary of the criteria checked on the nine models for the metrics based on brown trout young of the year (YOY) and on older fishes (OLD). Ni, indicated that the metric is based on abundance, dens that the metric is based on density (number of individuals per area), and proportion that the metric is based on the proportion of YOY in brown trout populations.

Metric	Model	Distribution	Transformation	Offset	Normality	Structure residuals	Leverage	Adequacy	Stability
Ni YOY	GLM	Poisson	No	No	No	Lot of values	Few	Poor	Yes
Ni YOY	GLM	Poisson	No	Number of b trout	Yes	Lot of values	Few	Weak	No
Ni YOY	GLM	Negative Binomial	No	No	No	High structu	Few	Poor	Yes
Ni YOY	GLM	Negative Binomial	No	Number of b trout	Yes	No	Few	Poor	No
Dens YOY	MLR	Gaussian	Log	No	No	High structu	Few	Poor	Yes
Dens YOY	MLR	Gaussian	Log	Brown density	Yes	High structu	Few	Poor	No
Dens YOY	MLR	Gaussian	4^{th} square root	No	Yes	High structu	Few	Weak	Yes
Proportion	GLM	Binomial	Logit	No	Yes	High structu	Few	Weak	No
Proportion	MLR	Gaussian	Arcsines square	No	Yes	No	Few	Poor	No
Ni OLD	GLM	Poisson	No	No	No	Lot of values	Few	Poor	No
Ni OLD	GLM	Poisson	No	Number of b trout	Yes	Lot of values	Few	Average	No
Ni OLD	GLM	Negative Binomial	No	No	No	High structu	Few	No	No
Ni OLD	GLM	Negative Binomial	No	Number of b trout	Yes	No	Few	Weak	No
Dens OLD	MLR	Gaussian	Log	No	Yes	No	No	Weak	Yes
Dens OLD	MLR	Gaussian	Log	Brown density	Yes	High structu	Few	Poor	No
Dens OLD	MLR	Gaussian	4th square root	No	Yes	No	Few	Poor	Yes

3.4.1.4 Crossing metrics and size class

3.4.1.4.1 Metrics selected

On the 96 metrics tested, only four metrics based on all species displayed satisfying models: number of oxygen intolerant fishes lower than 150 mm (Ni.O2INTOL.lw.150), number of habitat intolerant fishes lower than 150 mm (Ni.HINTOL.lw.150), number of rheophilous fishes lower than 150 mm (Ni.RH.lw.150) and number of insectivorous fishes lower than 150 mm (Ni.INSV.lw.150). None of the metrics based on large species were retained. Moreover, for a given cut-off, metrics of the same size classes were highly correlated, i.e. metrics based on fish lower than 100 mm. None of the metrics based on large individuals and none of the metrics computed with other cut-offs (100 or 200 mm) were judged satisfying. We will present only the results obtained for the four metrics retained.

3.4.1.4.2 Environment of the calibration data set

Among the 528 calibration sites, only 218 sites were retained to calibrate the models. Those sites are distributed among 10 countries (Figure 24): Austria, Deutschland, France, Italy, Lithuania, Romania, Spain, Sweden, Switzerland and United Kingdom. Finland, Hungary, Netherlands, Poland and Portugal were not represented in calibration sites for this experiment.

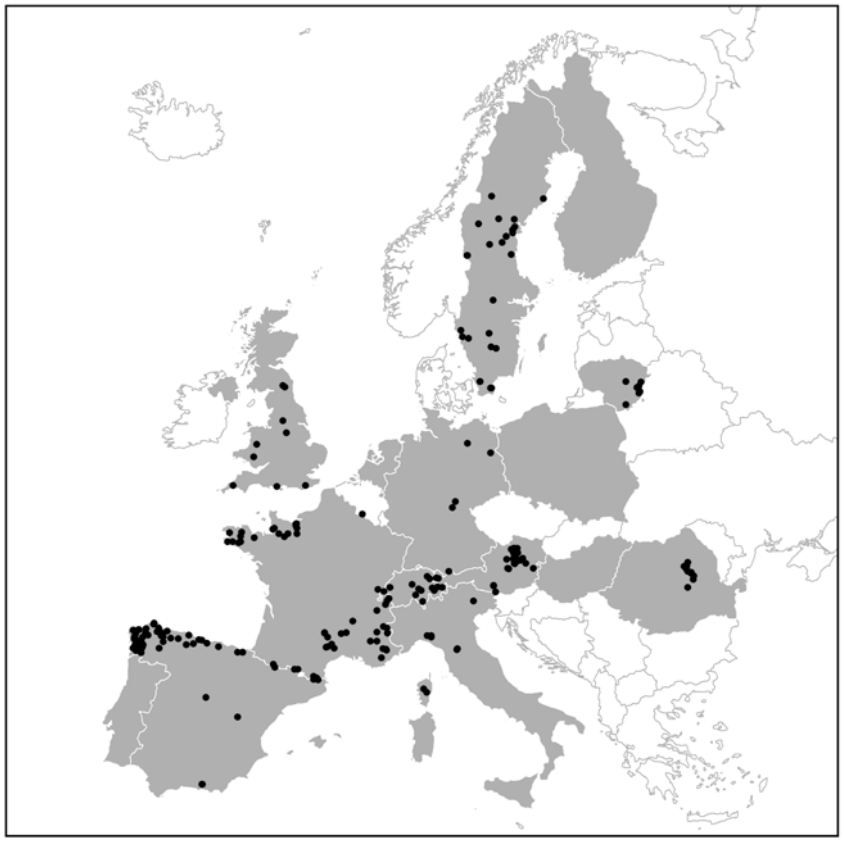


Figure 24. Distribution of the 218 calibration sites. Member countries of the EFI+ project are coloured in grey.

Those 218 sites were mostly situated in the upper part of streams, displaying high values of slope, positive value of syngeomorph1 which is inversely related to distance from source and size of

catchment, medium to large size of sediment (Table 35). Moreover, more than ninety percent of the sites occurred in streams with a size of catchment lower than two hundred square kilometres. An important part of sites were also located in costal streams.

Table 35. Summary of the distribution of environmental variables of the calibration sites employed as independent variables in models.

Environmental variables	Statistics	
Slope	range	0.1 - 294.657
	mean (sd)	25.448 (33.984)
Tjul	range	8.6 - 21.6
	mean (sd)	16.926 (2.069)
Tdif	range	8.6 - 25.4
	mean (sd)	16.255 (4.371)
syngeomorph1	range	-0.81 - 3.406
	mean (sd)	1.567 (0.896)
syngeomorph2	range	-1.658 - 2.915
	mean (sd)	0.251 (1.237)
natsed	large	47
	medium	163
	small	8

3.4.1.4.3 Modelling of the four selected variables

- Number of small oxygen intolerant fishes (Ni.O2INTOL.lw.150v)

To model the abundance of oxygen intolerant fishes lower than 150 mm (all species considered), 214 sites were used and five environmental variables were retained by the stepwise procedure: lslope, Tdif, syngeomorph1, syngeomorph2 and natsed (Table 36). One the five criteria systematically checked, none of them exhibited abnormal values. Pearson residuals were approximately following a normal law with a slight deviation of the distribution toward positive values. No structure in the residuals were observed when related to fitted values on the link (eta), suggesting homoskedasticity in the residuals. Only one site could be potentially influent in the model with a hatvalue of 0.6. The adequacy of the model to observed values was judged enough satisfying, since only few sites displayed high departure from the theoretical model (especially sites with low abundances). The distribution of the RMSE obtained by bootstrap was well balanced suggesting a quite stable model.

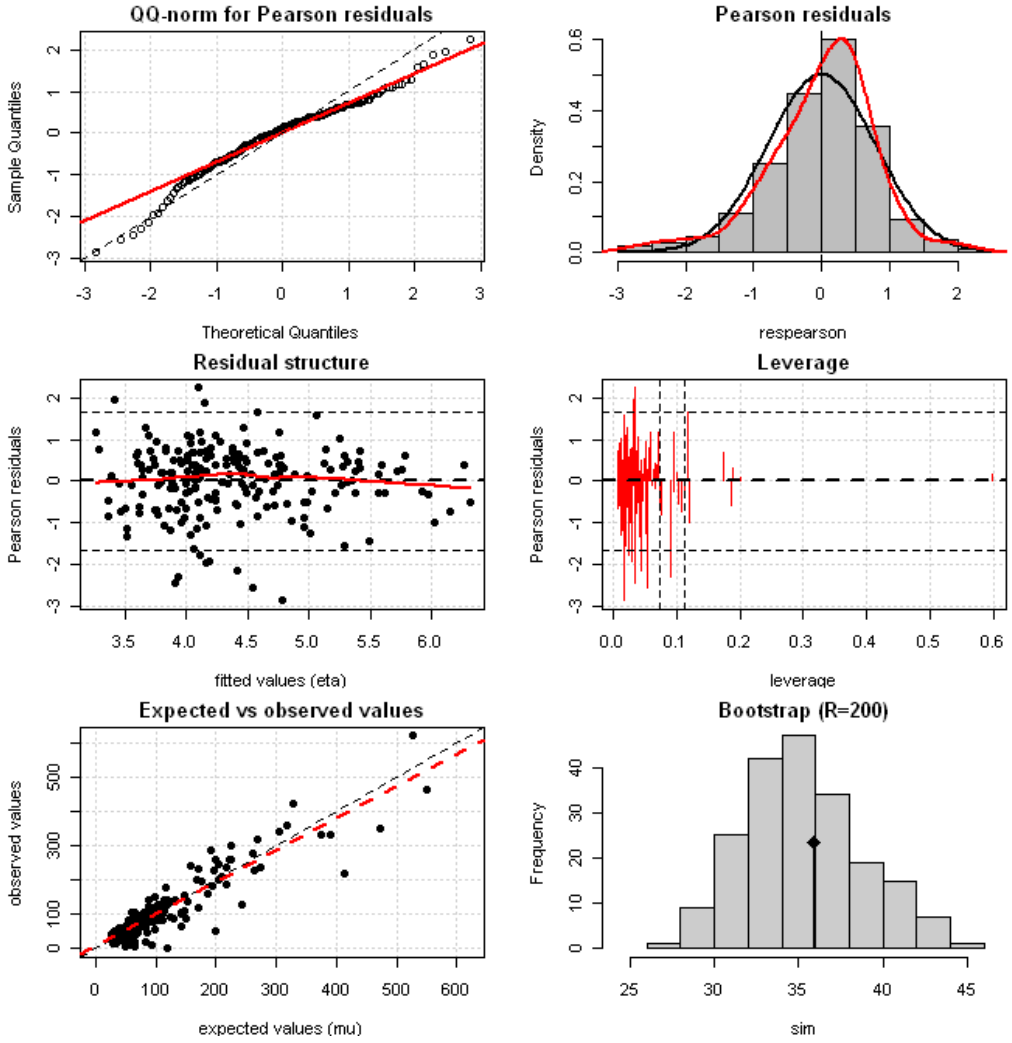


Figure 25.Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley, 1997).

Table 36. Analysis of deviance of the GLM with Negative Binomial distribution selected by stepwise procedure.

Environmental variables	Degree of free	Difference Deviance	P-value
lslope	2	1.544	0.462
Tdif	1	2.640	0.104
syngeomorph1	1	5.449	0.02
syngeomorph2	1	7.724	0.005
natsed	2	12.046	0.002
Residuals	206	238.068	

- Number of small habitat intolerant fishes (Ni.HINTOL.lw.150)

To model the abundance of habitat intolerant fishes lower than 150 mm (all species considered), 214 sites were used and five environmental variables were retained by the stepwise procedure: lslope, Tdif, syngeomorph1, syngeomorph2 and natsed (Table 37). None of the five criteria checked presented abnormal situations. Pearson residuals were approximately following a normal law with a slight deviation of the distribution toward positive values. No structure in the residuals were observed when related to fitted values on the link (eta), suggesting homoskedasticity in the residuals. Only one site could be potentially influent in the model with a hatvalue of 0.6.

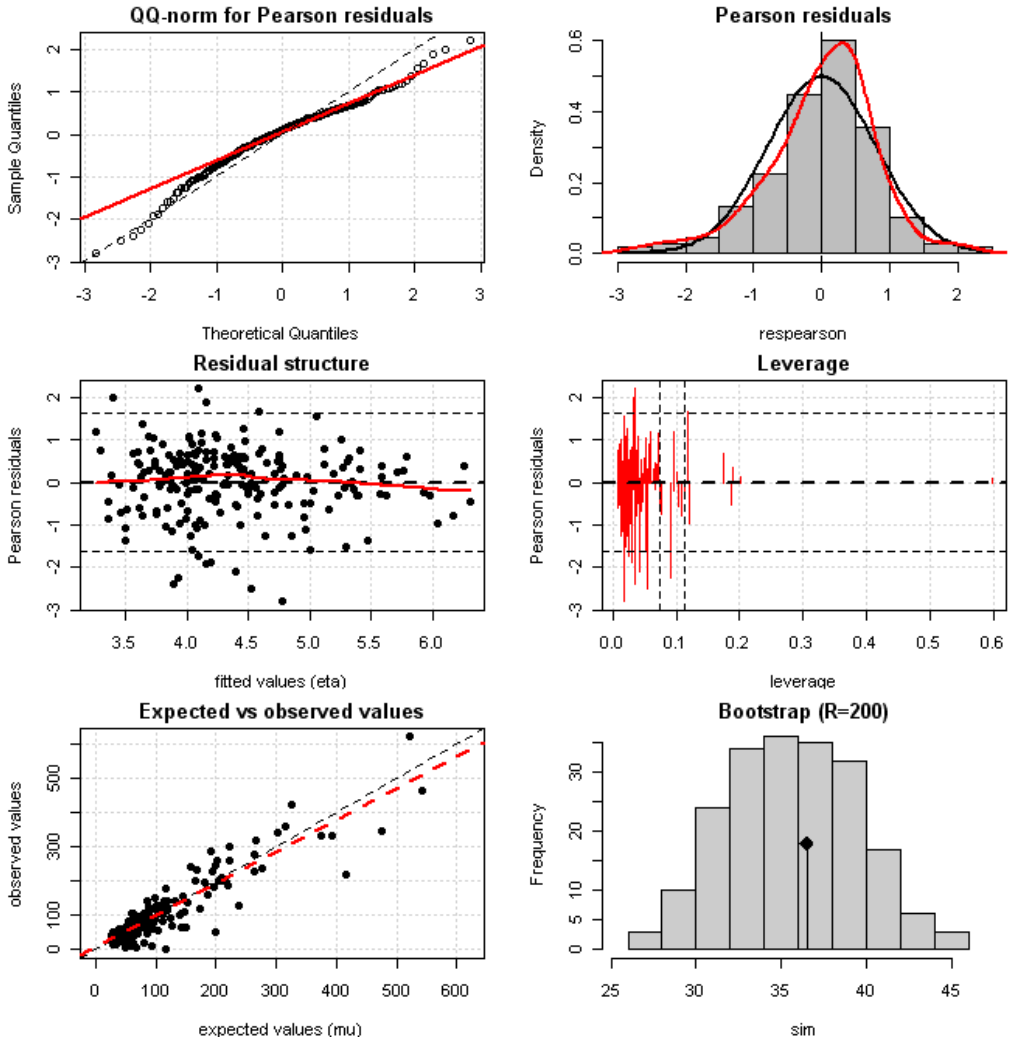


Figure 26. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley, 1997).

The adequacy of the model to observed values was judged enough satisfying, since only few sites displayed high departure from the theoretical model (especially sites with low abundances). However, the simulated distribution of RMSE was slightly to spread in the central part (around the mode), the distribution didn't seem to be skewed and the model seemed to be fairly stable.

Table 37. Analysis of deviance of the GLM with Negative Binomial distribution selected by stepwise procedure.

Environmental variables	Degree of free	Difference Deviance	P-value
lslope	2	1.237	0.539
Tdif	1	1.681	0.195
syngeomorph1	1	4.834	0.028
syngeomorph2	1	6.905	0.009
natsed	2	13.156	0.001
Residuals	206	237.351	

- Number of small rheophilous fishes (Ni.RH.lw.150)

To model the abundance of rheophilous fishes lower than 150 mm (all species considered), 212 sites were used and five environmental variables were retained by the stepwise procedure: lslope, Tdif, syngeomorph1, syngeomorph2 and natsed (Table 38). One the five criteria systematically checked, none of them exhibited abnormal values. Pearson residuals were approximately following a normal law with a slight deviation of the distribution toward positive values. No structure in the residuals were observed when related to fitted values on the link (eta), suggesting homoscedasticity. Only one site could be potentially influent in the model with a hatvalue equal to 0.6. The adequacy of the model to observed values was satisfying. Nevertheless, some sites with few (close to 0) or many rheophilous fishes (more than 300) had been weakly explicated by the model. Bootstrap distribution of RMSE was fairly close to a Gaussian distribution suggesting a model with a high stability.

Table 38. Analysis of deviance of the GLM with Negative Binomial distribution selected by stepwise procedure.

Environmental variables	Degree of free	Difference Deviance	P-value
lslope	2	3.463	0.177
Tdif	1	2.873	0.09
syngeomorph1	1	2.059	0.151
syngeomorph2	1	12.105	0.001
natsed	2	12.141	0.002
Residuals	204	237.339	

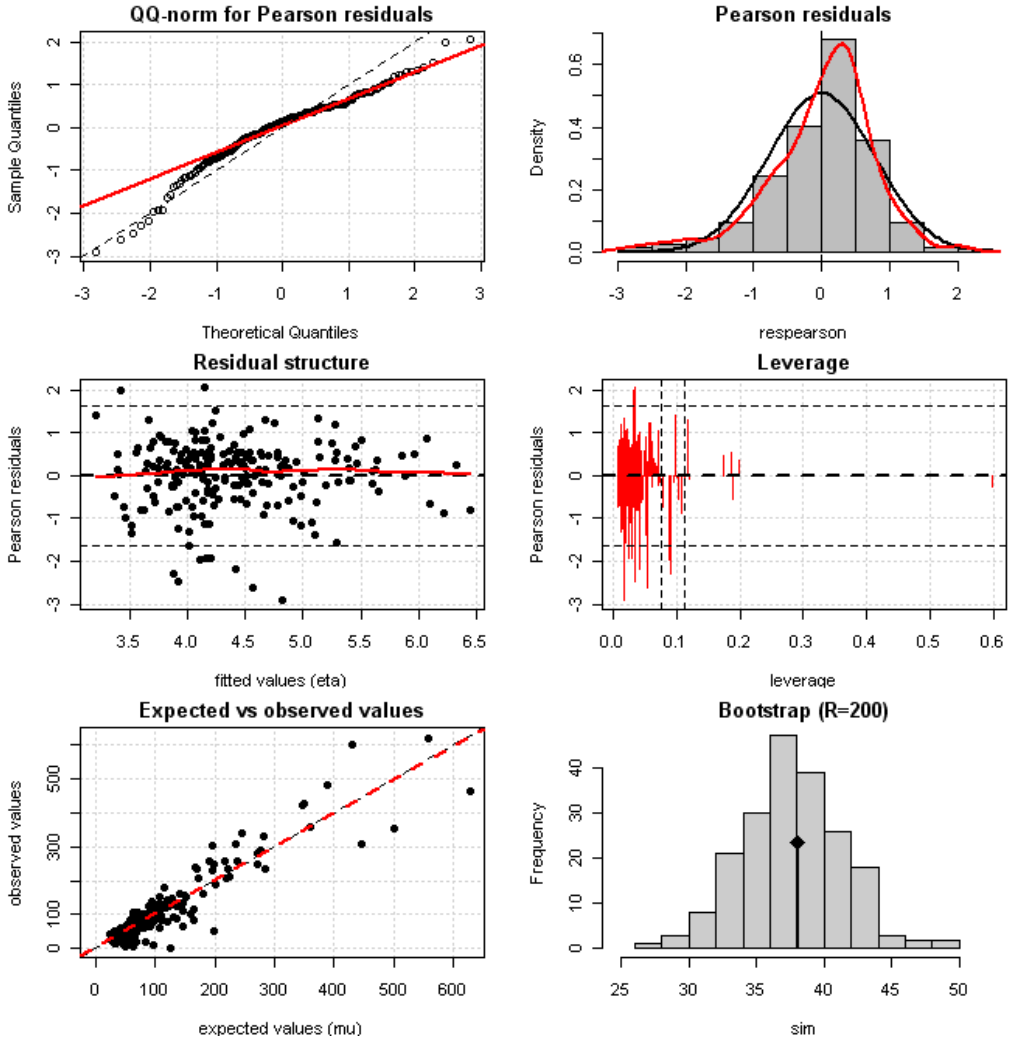


Figure 27. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley, 1997).

- Number of small insectivorous fishes (Ni.INSV.lw.150)

To model the abundance of insectivorous fishes lower than 150 mm (all species considered), 212 sites were used and five environmental variables were retained by the stepwise procedure: lslope, Tdif, syngeomorph1, syngeomorph2 and natsed (Table 39). Compared to the previous models, the skewness of the distribution of Pearson residuals was more marked. A low structure in the residuals could be observed. The smoothing red line of the third graph was always over zero. No heteroscedasticity was suspected in the residuals. Only one site could be potentially influent in the model with a hatvalue equal to 0.6. The adequacy of the model to observed values was satisfying.

Bootstrap distribution of RMSE was fairly close to a Gaussian distribution suggesting a model with a high stability.

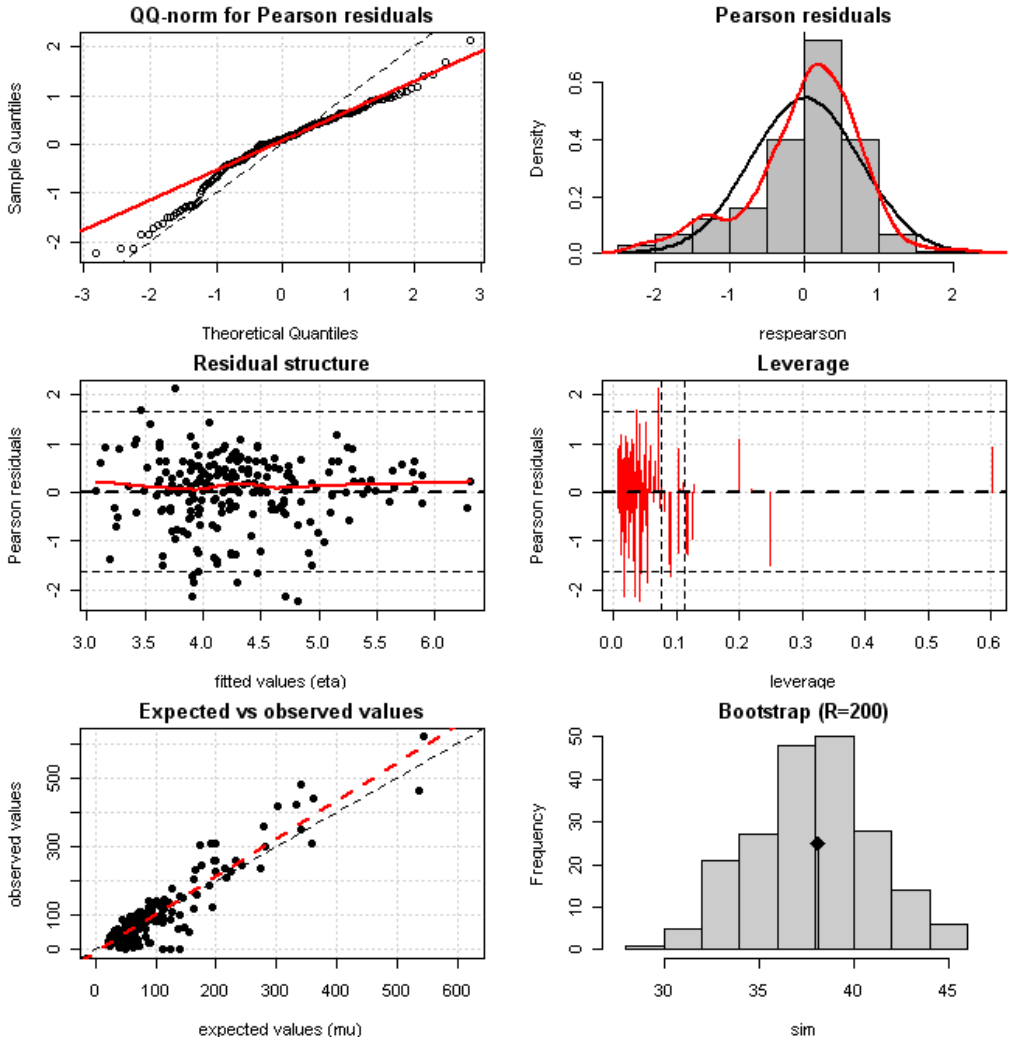


Figure 28. Graphical representation of model diagnostic. The first graphic shows the QQ-plot representation of standardized residuals against normal theoretical quantiles. The second corresponds to histogram of Pearson residuals (with in red, density estimation curves and in black, theoretical normal curves). The third corresponds to residuals in function of the fitted values (eta, fitted values in the 'link space'). The fourth graphic plots the leverage values (hat values) against the standardized residuals to detect the potential influent points. The fifth graphic shows the observed values against expected values from generalized linear model. The last graphic corresponds to the histogram of RMSE obtained by bootstrap (Davidson & Hinkley, 1997).

Table 39. Analysis of deviance of the GLM with Negative Binomial distribution selected by stepwise procedure.

Environmental variables	Degree of free	Difference Deviance	P-value
lslope	2	2.404	0.301
Tdif	1	0.75	0.386
syngeomorph1	1	0.983	0.321
syngeomorph2	1	17.387	<0.001
natsed	2	9.43	0.009
Residuals	204	236.949	

3.4.2 Discussion

None of the sixteen metrics developed on the age classes of brown trout (*Salmo trutta fario*) could have been enough explained by environment to be tested for their sensitivity to human pressure. Consequently it was not possible to retain one of those metric for the development of the new index.

A possible bias could have arisen from the estimation of the abundance of each age classes in the populations. However, the multiple linear regression linking the cut-off length and the environment explained more than fifty percent of the variance, two-third of the sites used in this model was located in the Atlantic part of Spain and in France. The influence of those two areas in the model could have had a negative effect on the estimation for the other regions. Moreover, the estimation of the number of individuals belonging to each age class could have been biased when the lengths distributions were strongly overlapped.

The stocking of brown trout individuals either for maintaining populations or for recreational purpose, was also a possible source of bias. Very often, stocked fishes are larger than indigenous individuals and could reach specific lengths younger than corresponding wild fishes(L'abée-Lund & Saegrov, 1991). As ages of fishes (due to absence of information) were estimated from their sizes, several stocked young of the year should have been identified as older fishes. Moreover, stocking should have artificially inflated the abundance of each cohort compared with indigenous brown trout populations without manipulations.

The high temporal variability in recruitment (Elliott, 1994; Freeman *et al.*, 2001) must have played a significant role in the relative absence of relationship between abundance of each cohort and environmental variables included in the models. Recruitment is dependant of local environmental conditions (Lobon-Cervia, 2003; Lobon-Cervia & Rincon, 2004), random environmental events such as floods and density-dependant feed-back mechanisms (Elliott, 1994; Jenkins Jr *et al.*, 1999; Vollestad, Olsen & Forseth, 2002). The amount of discharge during the emergence of fries affects directly the recruitment threw the availability of habitat for fishes to establish their territory (Lobon-Cervia, 2003; Lobon-Cervia & Rincon, 2004). The timing and intensity of flood could be dramatic for the survival of early life stage and thus on the recruitment (Lobon-Cervia & Rincon, 2004). A peak flow event during the emergence or during the under gravel life should markedly increase the mortality of fry and eggs and thus reduce the recruitment (Jowett & Richardson, 1989). The high temporal variability of recruitment is related at least in part to the random natural variability of environment.

Biological factors threw feed-back mechanisms also controlled the recruitment. The survival and the production of brown trout young of the year is dependant to spawned eggs density a belt shaped relationship (Elliott, 1994). Consequently, beyond a certain density of egg, the abundance of 0+ decreases. Density dependant mechanisms will also affect the growth rates (Jenkins Jr *et al.*, 1999; Vollestad, Olsen & Forseth, 2002) and thus the winter survival probability of young of the year which is directly related to their sizes (Hurst, 2007).

The combination of both abiotic and biotic factors (Milner *et al.*, 2003)controlling the recruitment could explain the relative difficulties to relate statistically the abundances of the two cohorts that we have considered and environment.

Among the ninety six new metrics, based on the interaction of functional metrics and fish size classes, only four cloud be related to environmental gradients. Nevertheless, the abundance of small fishes of each remaining metric was not directly linked with environment but the ratio of their abundance among the total number of fishes caught. Using ratio, instead of abundance, should also reduce the dependency of the results to the number of fish caught. If two sites should present the same ratio of small individuals, but if for sampling raison the abundances were different, scores computed will still be identical. Nevertheless, it implies that sampling efforts in both cases were sufficient to have a good estimation of the fish assemblage structures. Using ration should also accentuate the sensitivity of some metrics to human pressures. The deviation between observed and theoretical values should arise either form both sides of the ratio. A pressure could affect the small individuals but also the total number of fish occurring in a site and thus modifying the theoretical ratio threw two different ways.

3.5 Conclusion

After a few conservative and strict modelling process, the number of candidate metrics is relatively low. We only conserve 13 metrics: five metrics based on species number (Ric.O2.Intol, Ric.Hab.Intol, Ric.Hab.RH, Ric.INSV and Ric.RH.Par), four metrics based on fish number (Ni.O2.Intol, Ni.hab.Intol, Ni.INSV and Ni.LITHO) and four metrics based on individual number with constraints on the fish length (Ni.O2.Intol.150, Ni.Hab.Intol.150, Ni.RH.150 and Ni.INSV.150).

Table 40. Summary of the criteria checked on metrics models. Terms 'Adequacy' and stability correspond to the evaluation of the good adjustment between expected and observed values and model stability measured by a resampling procedure (bootstrap and RMSE).

Metric	Model	Distribution	link	Offset	Residual Quasi-Normality	Residual Structure	Leverage	Adequacy	Stability
Ric.O2.Intol	glm	poisson	log	total richness	yes	low	Few	average	average
Ric.Hab.Intol	glm	poisson	log	total richness	yes	low	Few	average	average
Ric.Hab.RH	glm	poisson	log	total richness	yes	no	Few	average	good
Ric.INSV	glm	poisson	log	total richness	yes	low	Few	average	poor
Ric.RH.Par	glm	poisson	log	total richness	yes	low	Few	average	average
Ni.O2.Intol	glm	negative binomial	log	total captures	average	low	Few	average	average
Ni.hab.Intol	glm	negative binomial	log	total captures	average	low	Few	average	average
Ni.INSV	glm	negative binomial	log	total captures	average	No	Few	average	good
Ni.LITHO	glm	negative binomial	log	total captures	average	low	Few	average	good
Ni.O2.Intol.150	glm	negative binomial	log	total captures	yes	low	Few	average	good
Ni.Hab.Intol.150	glm	negative binomial	log	total captures	yes	low	Few	average	average
Ni.RH.150	glm	negative binomial	log	total captures	average	low	Few	average	good
Ni.INSV.150	glm	negative binomial	log	total captures	average	low	Few	average	good

These 13 selected models are characterized by a satisfactory stability, satisfactory adequacy between expected and observed values, low residuals structure and quasi-normal residuals distribution (Table 40). The consideration of these criteria is strongly required to increase the extrapolation capacity of models and to limit bias of predictions based on environmental conditions in outside the calibration environment.

4 Metric selection

4.1 Introduction

13 metrics are considered as having been correctly modelled (see below and previous section). Five of them are expressed in species richness, 4 in abundance of individuals and 4 in abundance of individuals for a given size class;

The particular case of metric based on historical presence of diadromous species will be considered elsewhere.

Metric names	Detailed names
Metrics expressed in species richness	
Ric.Hab.RH	Rheophilous habitat species richness
Ric.INSV	Insectivorous feeding species richness
Ric.RH.Par	Rheophilic reproduction habitat species richness
Ric.O2.Intol	Oxygen depletion intolerant species richness
Ric.Hab.Intol	Habitat alteration intolerant species density
Metrics expressed in individual abundance	
Ni.O2.Intol	Oxygen depletion intolerant species abundance (Nb. individuals)
Ni.Hab.Intol	Habitat alteration intolerant species density (Nb. individuals)
Ni.INSV	Insectivorous feeding species (Nb. Individuals)
Ni.LITHO	Lithophilic reproduction habitat species abundance (Nb. Individuals)
Metrics expressed in individual abundance in a given size class	
Ni.O2.Intol.150	Abundance of individuals < 15 cm of O2 depletion intolerant species
Ni.Hab.Intol.150	Abundance of individuals < 15 cm of Habitat intolerant species
Ni.RH.150	Abundance of individuals < 15 cm of Rheophilic species
Ni.INSV.150	Abundance of individuals < 15 cm of Insectivorous species

The final selection of metrics is based on three main criteria:

- the correlation between metrics, which must not be too high to avoid redundancy between metrics
- the sensitivity of metrics to pressures
- their representativeness in the different ecoregions

The responses of metrics are examined separately for the two main river zones, the salmonid zone and the cyprinid zone, which have been defined previously (section 2.1.3);

At this step, all the metrics are expressed as the difference between the observed and the expected (predicted) values. They have first to be standardized and rescaled (range from 0 to 1) before analysing their sensitivity.

4.2 Metric computation

4.2.1 Standardization per ecoregion and river zone

In previous work (FAME project), all the metrics were expressed as standardized residuals after modelling are rescaled (from 0 to 1) using a normal transformation. This last transformation is acceptable when the standardized residuals are normally distributed. This was not always strictly the case even if the normality of residual distributions was checked in a rough manner.

Another difference between our present procedure and the FAME procedure is that we did not included any regional classification in the list of explanatory variables used during the stepwise-based modelling phase. The reason is that the combined use of local and regional variables in a model is efficient when the calibration dataset is equilibrate and representative for all regions, i.e. a part of the variability explained by local variables cannot be expressed by a regional parameter if this one is selected previously in the model.

Such "interaction" effect appears when the range and the distribution of each of the considered local variables are not comparable between regions. Then, a part of variability explained by regions in the selected model could be linked to the fact that one or several local environmental variables have a particular range in this region compared to others. This is typically the case when, for example, some regions are more mountainous and others mainly characterized by plains. The effect of river slope (which is for a part dependant of the general physiography of an area, even if it is a variable defined at the local scale) will tend to be underestimated in the model.

Moreover, this non-independence between regional and local variables would have more pronounced effects when the calibration dataset is not enough representative of all the local environmental situations in each region, which is in general the case. In the case of a mountainous region, most of river segments have a high river slope. Nevertheless, slow-flowing section could also exist in valleys, even if it is not the most common type of river segment. If the effect of river slope is mainly expressed by the regional variables in this mountainous region, the model would not be able to correctly predict the fish fauna of such local situation.

For these reason, we have chosen to standardize the residuals per ecoregion and per river zone in a second step, after the modelling procedure. The residual standardization is realized using a larger dataset than the calibration dataset, the undisturbed dataset (N=2526). In this last dataset, the number of sites per ecoregion and river zone is larger enough in comparison with the calibration dataset (N=528). The distance (residuals R_i) among expected (E_i) and observed (O_i) values is given by the following equation:

$$R_i = \log(O_i + 1) - \log(E_i + 1)$$

Nevertheless, as the number of sites per ecoregion and river zone was too limited in some cases, some ecoregions were gathered. In the same way, some river zones are not considered for some region due to the too number of site or the lack of the characteristic fish fauna of the given river zone in the considered ecoregion (see previous section for detailed explanations).

The metric score (M_{iq}) of each of the 13 selected metrics in a given river zone q (salmonid zone or cyprinid zone) is obtained by standardizing the residuals of the model in the following manner in each ecoregion i:

$$M_{iq} = \frac{(R_i - M_{jq})}{S_q}$$

- R_i : Residual value (difference between observed and expected metric) from sites belonging to the ecoregion j and the river zone q.
- M_{jq} : Median value of the residuals in the ecoregion j and the river zone q

- S_q : Standard deviation of the residuals in the whole undisturbed dataset for a given river zone (salmonid or cyprinid)

The value of the median is chosen because it is less sensitive to extreme values than the mean. For the same reason, the variance of residuals of the whole dataset is used instead of the variance of the distribution of residuals corresponding to each ecoregion.

The standardized values of residuals obtained with metrics expressed in abundance of individuals for a given class size are only computed for the salmonid river zone, as the corresponding models are mainly calibrated on sites belonging to only this river zone (only 37 sites in the cyprinid zone).

4.2.2 Rescaling between 0 and 1

Standardized residuals vary from -∞ to + ∞. A requirement is that each metric varies within a finite interval and in addition from 0 to 1. Such result could be obtained using two transformations.

First, for a given river zone, all the values over a maximum and below a minimum have to be replaced by this maximum (Max_j) and this minimum (Min_j). Then the following transformation is applied to each metric score (e.g. Legendre & Legendre 1998, Hann & Kamber 2000):

$$\frac{\left(M_i - Min\right)}{\left(Max_q - Min_q\right)}$$

After several step, the Max_q value has been defined as the quantile 0.95 of the distribution of standardized residuals M_i in the considered river zone q.

An additional requirement is that, after transformation, each metric must have the same median value in the absence of any disturbance (i.e. in the undisturbed dataset). Such result is obtained by computing with an algorithm, for each metric in each river zone the Min_q value corresponding to a median value of 0.80 for the scores in undisturbed sites.

Depending of the considered metric and river zone, the Min_q values vary from quantile values 0.0001 to 0.20, all median values being equal to 0.80. The 0.25 quantile values vary from 0.602 to 0.752.

4.3 Metric selection

4.3.1 Correlations between candidate metrics

The highest correlations between candidate metrics (Pearson coefficient > 0.70)) are presented below. Correlations are computed for the 2 river zones separately.

All the 4 metrics based on oxygen intolerant species and habitat intolerant species guild and expressed in richness or abundance are highly correlated in both salmonid and cyprinid zone.

All the 4 four metric expressed in term of abundance of individuals smaller than 15 cm are also correlated.

Table 41. Rheophilic species and Rheophilic reproductive habitat species are also highly correlated.

	Salmonid Zone	Cyprinid zone
Ric.O2.Intol - Ric.Hab.Intol	0.814	0.753

Ric.O2.Intol - Dens.Hab.Intol	0.749	0.808
Ric.Hab.Intol - Ric.Hab.RH	0.728	
Ric.Hab.Intol - Ni.Hab.Intol.150	0.722	0.803
Ric.Hab.RH - Ric.RH.Par	0.719	0.777
Ni.O2.Intol - Ni.Hab.Intol.150	0.869	0.819
Ni.O2.Intol.150 - Ni.Hab.Intol.150	0.951	
Ni.O2.Intol.150 - Ni.RH.150	0.788	
Ni.O2.Intol.150 - Ni.INSV.150	0.716	
Ni.Hab.Intol.150 - Ni.RH.150	0.826	
Ni.Hab.Intol.150 - Ni.INSV.150	0.713	
Ni.RH.150 - Ni.INSV.150	0.816	

4.3.2 Sensitivity to pressures

The sensitivity of the candidate metrics to pressure are evaluated using the 2 global pressure indices (Global pressure index A and Global pressure index B) for the whole dataset and for each ecoregion separately (see previous section).

Table 42. **Responses of candidate metrics to pressure in the 2 river zones. Comparison of class 1 and class 5 sites using a Kruskall-Wallis test for the 2 pressure indices (p-Press.Index.A and p-Press.Index.B). Difference between the median values of the 2 groups (class 1 - class 5). A very low (negative) value indicates a strong response.**

Metriques	Nb.sites	*p*-Press.Index.B	Diff-Press.Index	p-Press.Index.A	Diff-Press.Index
Salmonid river zone					
Ric.O2.Intol	1731	< 0.0001	-0.161	< 0.0001	0.007
Ric.Hab.Intol	1731	< 0.0001	-0.206	< 0.0001	-0.018
Ric.Hab.RH	1731	0.0469	-0.132	0.0002	0.018
Ric.INSV	1731	0.54	-0.006	< 0.0001	0.03
Ric.RH.Par	1731	< 0.0001	-0.21	< 0.0001	-0.012
Ni.O2.Intol	1731	< 0.0001	-0.467	< 0.0001	0.029
Ni.Hab.Intol	1731	< 0.0001	-0.601	< 0.0001	0.042
Ni.INSV	1731	0.0009	-0.054	< 0.0001	0.057
Ni.LITHO	1731	0.0058	-0.184	< 0.0001	0.027
Ni.O2.Intol.150	1033	< 0.0001	-0.618	< 0.0001	-0.18
Ni.Hab.Intol.150	1054	< 0.0001	-0.656	< 0.0001	-0.158
Ni.RH.150	1036	< 0.0001	-0.273	< 0.0001	-0.116
Ni.INSV.150	1034	0.0518	-0.114	0.3794	-0.087
Cyprinid river zone					
Ric.O2.Intol	795	< 0.0001	-0.306	< 0.0001	-0.293
Ric.Hab.Intol	795	< 0.0001	-0.234	< 0.0001	-0.253
Ric.Hab.RH	795	< 0.0001	-0.149	< 0.0001	-0.138
Ric.INSV	795	< 0.0001	0.009	< 0.0001	-0.04
Ric.RH.Par	795	< 0.0001	-0.11	< 0.0001	-0.122
Ni.O2.Intol	795	< 0.0001	-0.53	< 0.0001	-0.38
Ni.Hab.Intol	795	< 0.0001	-0.446	< 0.0001	-0.357
Ni.INSV	795	< 0.0001	-0.05	< 0.0001	-0.038
Ni.LITHO	795	< 0.0001	-0.282	< 0.0001	-0.208

All the sites are ranked in 5 classes by the 2 pressure indices. The significance of responses are evaluated by comparing the metric values between class 1 sites (low to few disturbance) and

class 5 sites (highest level of disturbance), using a Kruskall-Wallis test and the difference between median values of the 2 group of sites (table below and figures).

In the salmonid zone, 3 metrics do not differ significantly between class 1 and class 5 sites: Ric.INSV, Ric.Hab.RH and Ni.INSV.150. The 2 pressure index demonstrate different reaction, the pressure index A being not very sensitive in general. The differences between median values of the 2 groups are always very low (less than 0.16) and sometimes even positive. Considering the metric responses to the pressure index B, the most sensitive metrics (differences between median values of the two groups > 0.4) are Ni.O2.Intol, Ni.Hab.Intol, Ni.O2.Intol.150 and Ni.Hab.Intol.150. By comparison, the metrics based on the same guilds but expressed in richness are less reacting. A similar result is observed for the metric based on Rheophilic, reproductive-rheophilic and lithophilic species.

In the cyprinid zone, the responses of the two pressure indices are quite similar. Metrics based on O2 intolerant and habitat intolerant species guilds are also the most sensitive. The two metrics based on insectivorous guild species are not sensitive. The others show a more similar level of responses (mean decrease between 0.10 and 0.30).

It is necessary to keep in mind than the intensity of the metric responses to high values of the pressure index cannot be directly compared between the salmonid and the cyprinid zone: the standardization of the residuals is different.

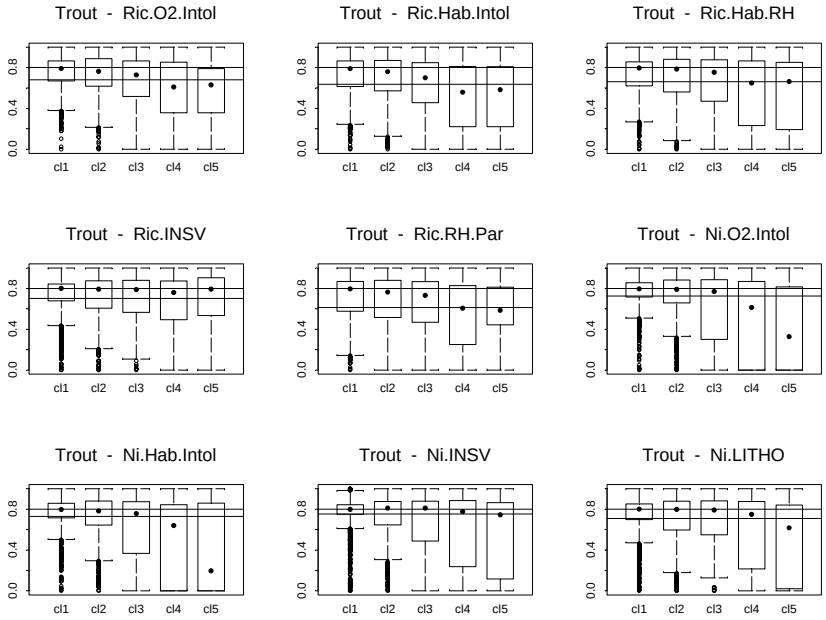


Figure 29. Responses of metrics based on the number of species to the pressures index (Press.Index.B) in the salmonid river zone.

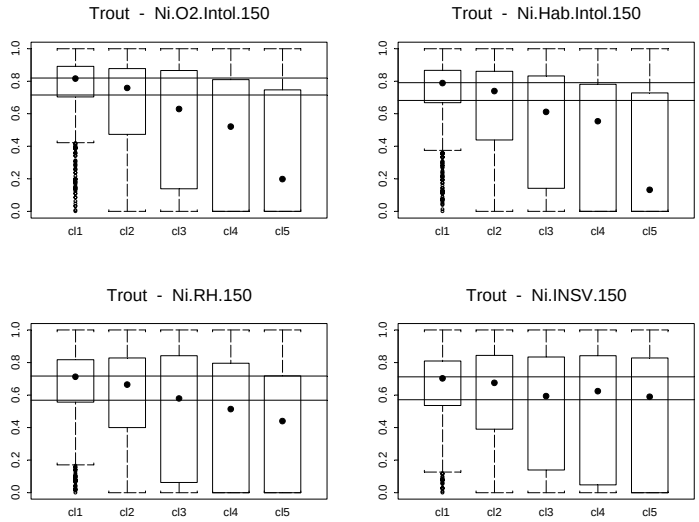


Figure 30. Responses of metrics based on the abundance of individuals smaller than 15 cm to the pressures index (Press.Index.B) in the salmonid river zone.

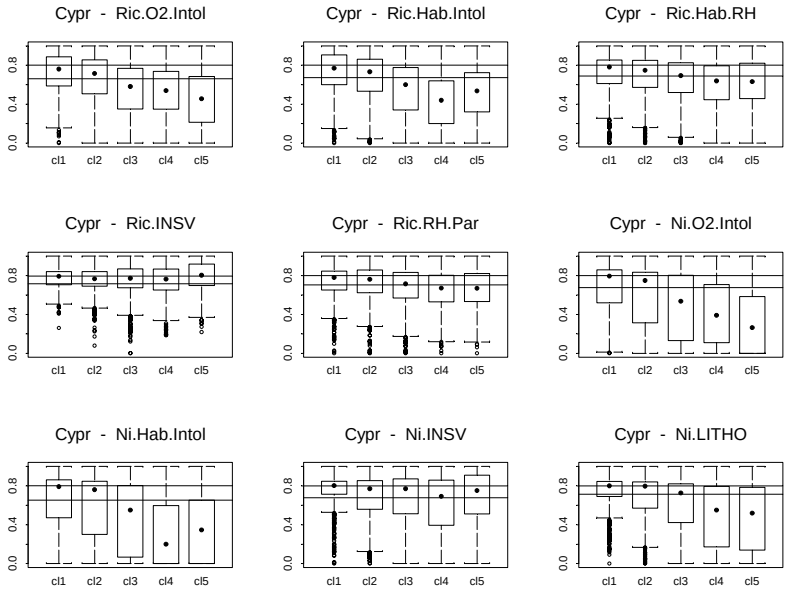


Figure 31. Responses of metrics based on the number of fish to the pressures index (Press.Index.B) in the salmonid river zone.

4.3.3 Representativeness of guilds and metrics

One important point which became obvious during the preliminary phase of the selection of metrics is that selected metrics must be correctly represented in the different fish communities, in the absence of any disturbance.

In some particular ecoregions and/or countries, species belonging to sum of the candidate guild are never abundant, even in undisturbed sites. This is in particular the case for the cyprinid zone (see table below).

Table 43. Relative Abundance of the different candidate guilds per ecoregion in the cyprinid zone (undisturbed sites)

CYPRINID ZONE	C.p	Est	Med	Ibe	W.p	Bal	Car	Eng
Nb.Sites	123	112	83	134	32	54	68	189
O2.INTOL	0.38	0.142	0.294	0.798	0.539	0.498	0.696	0.808
Hab.INTOL	0.391	0.194	0.311	0.798	0.644	0.523	0.772	0.811
Rheophilic	0.691	0.464	0.587	0.923	0.866	0.744	0.999	0.889
RH.PAR	0.581	0.322	0.816	0.911	0.519	0.423	0.473	0.771
Lithophilic	0.483	0.343	0.819	0.925	0.694	0.547	0.869	0.784
Insectivorous	0.58	0.307	0.579	0.817	0.728	0.675	0.934	0.528

Alp (Alps, Pyrenees), Est (Hungarian Lowlands, Eastern Plains, Pontic Province), Nor (Fenno-Scandian Shield, Borealic Uplands, Baltic Province), Med (Mediterranean region), Car (The Carpathians), Eng Great Britain, Ibe (Iberian Penisula), Ita (Italy, Corsica and Malte), W.p (Western Plains), W.h (Western Highlands), C.h (Central Highlands), C.p. (Central Plains).

In the Est group, and in the Mediterranean region, the relative abundance of Oxygen depletion and Habitat intolerant species are much lower than in most of the other ecoregions: respectively less than 20% of individuals and around 30% against 50% to 80%. In some countries, these guilds are close to be absent from the fish community. It is for example the case in Portugal, but also in Hungary and in Netherland.

Several tests and previous analysis demonstrated that in such situation, the score is always underestimated for sites belonging to the lowest pressure group (class 1): the median value of sites is not close to 0.80, as for other metrics, but below 0.50 (Mediterranean area, Hungary).

This aspect could not be correctly considered during the modelling process due to the very few number of calibration sites belonging to these ecoregions and in the corresponding river zone (35 sites among 533).

At the contrary, all the candidate guilds are correctly represented in each of the considered ecoregion in the salmonid zone

Table 44. Relative Abundance of the different candidate guilds per ecoregion in the salmonid zone (undisturbed sites)

SALMONID ZONE	C.h	Alp	W.h	Ita	C.p	Ibe	Nor	W.p	Car	Eng
Nb.Sites	87	160	64	121	117	701	260	72	103	46
O2.INTOL	0.875	0.884	0.896	0.56	0.835	0.874	0.96	0.825	0.83	0.955
Hab.INTOL	0.742	0.84	0.896	0.618	0.833	0.882	0.945	0.859	0.833	0.955
Rheophilic	0.975	0.997	1.001	0.815	0.948	0.926	0.974	0.984	0.99	0.965
RH.PAR	0.902	0.983	0.707	0.981	0.75	0.851	0.748	0.57	0.483	0.945
Lithophilic	0.838	0.918	0.86	0.794	0.75	0.945	0.76	0.838	0.822	0.931
Insectivorous	0.919	0.906	0.986	0.798	0.622	0.923	0.615	0.963	0.941	0.625

4.3.4 Final metric selection

Considering the three criteria used to select metrics (correlation between metrics, sensitivity to the index of pressure, representativeness in the different ecoregions), 2 metrics are finally selected per river zone.

In all case, the metrics based on the guild of insectivorous species are insensitive to pressure.

In the salmonid river zone, the most sensitive metrics are:
- based on oxygen depletion and habitat intolerant guild species.
- and expressed in "relative" abundance of individuals

The 2 corresponding metrics considering all the size class are highly correlated between them (Ni.O2.Intol and Ni.Hab.Intol.150). Among the metrics expressed in term of abundance of small-sized individuals, the 2 based on these species guilds are also highly correlated (Ni.O2.Intol.150 - Ni.Hab.Intol.150).

In order to not use the same guilds with 2 different metrics, and following complementary evaluation of metrics response, the 2 following metrics are selected:

SALMONID ZONE: Ni.O2.Intol and Ni.Hab.Intol.150

In the cyprinid zone, the metrics based on oxygen depletion and habitat intolerance cannot be used due to their lack of representativeness in several ecoregion. Among the others and considering the high correlation between Ric.Hab.RH and Ric.RH.Par, we selected two metrics.

CYPRINID ZONE: Ric.RH.PAR and Ni.LITHO

Ric.RH.PAR has been preferred to Ric.Hab.RH in relation with its higher relative abundance in undisturbed Mediterranean sites.

5 Metric Aggregation, Scoring and Performance analyses

5.1 Index definitions

5.1.1 Indices definition per river zone

As explained in the previous section, 2 metrics were selected per river zone. When only considering undisturbed sites, these 4 metrics have comparable distributions with a same median value of 0.80. The values of the first and the third quartile (1st and 3rd Qu.) are also close.

Table 45. Summary of the values of the 4 selected metrics for undisturbed sites

	Zone	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Ni.Hab.Intol.150	Salm.	0.000	0.691	0.798	0.744	0.870	1.000
Ni.02.Intol	Salm.	0.000	0.727	0.800	0.766	0.859	1.000
Ric.RH.Par	Cypr.	0.000	0.703	0.800	0.770	0.859	1.000
Ni.LITHO	Cypr.	0.000	0.714	0.800	0.726	0.832	1.000

The distributions of the 4 metrics differ a little, mainly in relation with the left tail of distribution for low values.

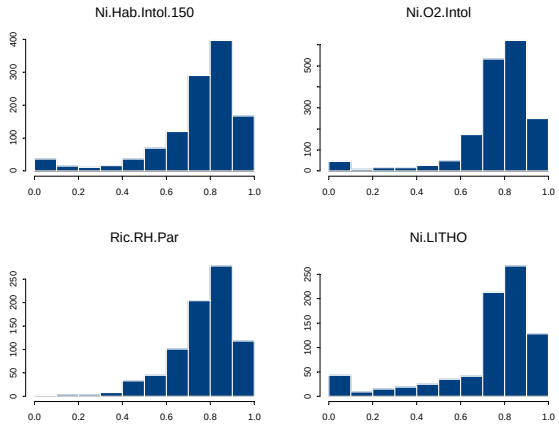


Figure 32. Distribution of the values of the 4 selected metrics (undisturbed sites)

Nevertheless, it has sense to consider the mean of the 2 metrics to define one index per river zone. Equations are defined as follow:

Salmonid zone Index = (Ni.Hab.150 + Ni.O2.Intol) / 2
Cyrpinid zone Index = (Ric.RH.Par + Ni.LITHO) / 2

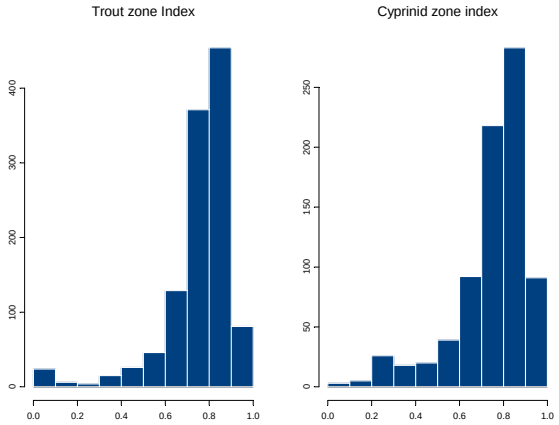


Figure 33. Distribution of the values of the 2 indices (undisturbed sites)

Table 46. Summary of the values of the 2 indices for undisturbed sites

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Salmonid zone index	0.0000	0.7164	0.7902	0.7524	0.8454	1.0000
Cyprinid zone index	0.0000	0.6968	0.7935	0.7481	0.8428	1.0000

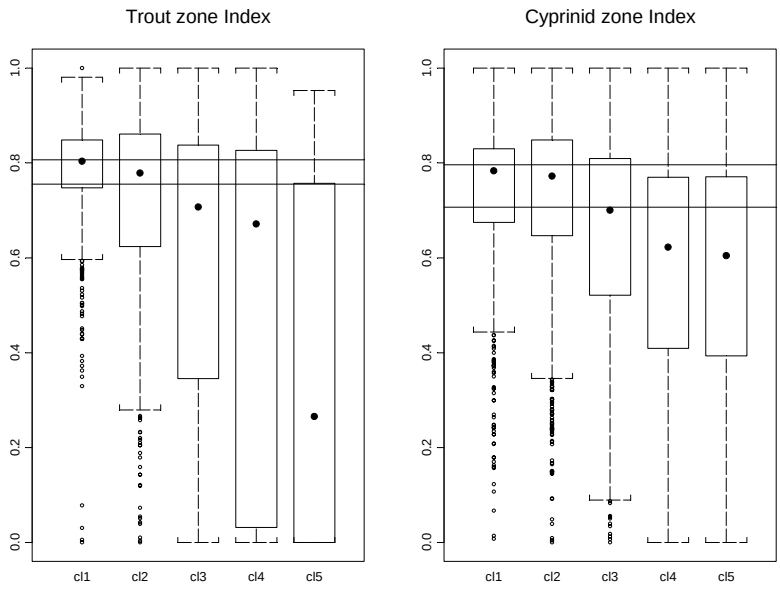


Figure 34. Responses of the two indices to the pressure index A (in 5 class)

5.1.2 Efficiency of the river type classification

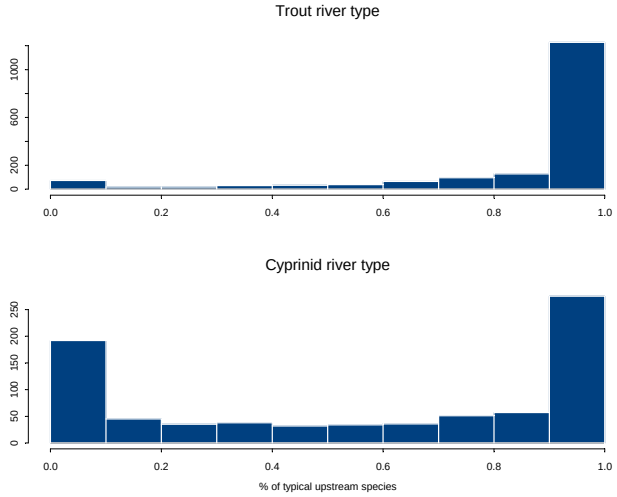
The efficiency of the river type classification used to split sites between "salmonid river type" and "cyprinid river type" is tested using the relative abundance of species typically occurring in the trout zone. As disturbed sites cannot be considered to test this classification (relative abundance of species is potentially modified by the different types of disturbances), only the "undisturbed sites" are considered.

A list of 19 species typically occurring in the upstream part of rivers has been selected after discussion with the partners of the project:

Alburnoides.bipunctatus	*Cobitis.calderoni*	*Coregonus.lavaretus*
Cottus.gobio	*Cottus.poecilopus*	*Eudontomyzon.mariae*
Hucho.hucho	*Lampetra.planeri*	*Phoxinus.phoxinus*
Salmo.salar	*Salmo.trutta.fario*	*Salmo.trutta.lacustris*
Salmo.trutta.macrostigma	*Salmo.trutta.trutta*	*Salmo.trutta.marmoratus*
Salvelinus.fontinalis	*Salvelinus.namaycush*	*Salvelinus.umbla*
Thymallus.thymallus		

These species are considered as typical of the salmonid zone. There all classified as oxygen depletion and habitat intolerant, and most of them as rheophilic.

The relative abundance of these 19 species is examined in both river types.



The classification is more efficient to identify the salmonid river type than the cyprinid one.

Concerning the salmonid river type, only a small number of sites can be considered as misclassified, with a very low relative abundance of "Upstream species". At the opposite, the classification used is not very efficiency concerning the cyprinid river type. A large amount of sites classified as "cyprinid river type" are dominated by typical upstream species.

The results obtained are quite similar when no cyprinid species are included in our upstream species list (*Alburnoides.bipunctatus, Phoxinus.phoxinus*).

To which point the evaluation of an "undisturbed site" could be influenced when the site is misclassified?

The undisturbed sites classified as "salmonid river type" are considered as misclassified when the relative abundance of typically upstream species is below 80%. At the opposite, a site classified as "Cyprinid river type" is misclassified when this relative abundance is over or equal to 80%. The confusion matrix is presented below:

river type	< 80%	>=80%
cyprynid river type	461	334
salmonid river type	371	1360

The distributions of the two indices values put in evidence a quite clear difference between the two types of river (see figure below)

For "undisturbed" salmonid river sites, the distribution of sites with a high relative abundance has a bell shape form (from 0.33 to 1, mean: 0.79, median: 0.807) and most of values are over 0.60.

At the opposite, all the low salmonid index values are related to sites characterized by a relative abundance of typical upstream river species less than 80%. These sites have a median value quite different from 0.80 (from 0 to 1, mean: 0.60, median: 0.67).

Moreover, when examining the values of the cyprinid index for this salmonid sites with a low proportion of typical upstream species, the median index value is again close to 0.80 (from 0 to 1, mean: 0.73, median: 0.77).

For “undisturbed” cyprinid river sites, the cyprinid index values for sites with a high proportion of typical upstream species remain quite high and with a median close to 0.80 ((from 0.36 to 1, mean: 0.81, median: 0.81). When applying the salmonid index on the same sites, the distribution is not very different (from 0.50 to 1, mean: 0.82, median: 0.83).

Conclusion: It is clear that the consequences of a misclassification are quite different, depending of the river type.

For sites misclassified as salmonid river sites (i.e. with a low relative abundance of typical upstream species), and in the absence of any disturbance, the salmonid index cannot be used, and has to be replaced by the cyprinid index.

For undisturbed cyprinid sites with a high relative abundance of typical upstream species, the values given by the cyprinid index are quite close to the one given by the salmonid index. And the misclassification seems not to have important consequences.

Nevertheless, in case of sites heavily disturbed, the relative abundance of typical upstream species cannot be used to classify the site, as human disturbance can significantly alter the abundance of these species. And in any way, this is the main reason we are using a typology based on invariant environmental parameters.

Considering the risk of misclassification and the associated consequences on the evaluation of sites (especially in the salmonid zone), the best solution is to propose systematically to the user the classification of the site (cyprinid or salmonid zone), the relative abundance of typical upstream species and the value of both indices (salmonid index and cyprinid index). The user, as an expert, would have to consider the situation and to make a choice.

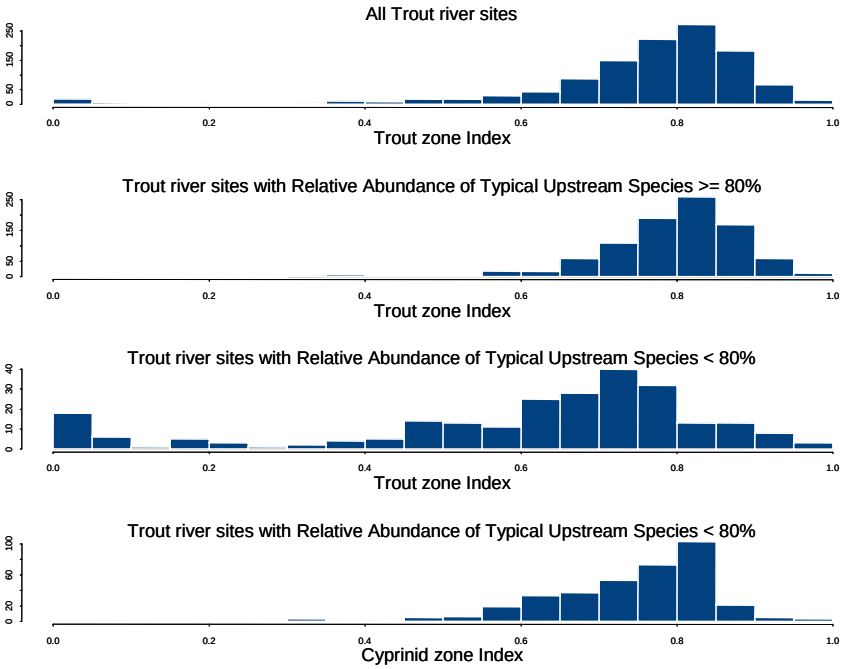


Figure 35. Frequency of the fish index in the salmonid zone

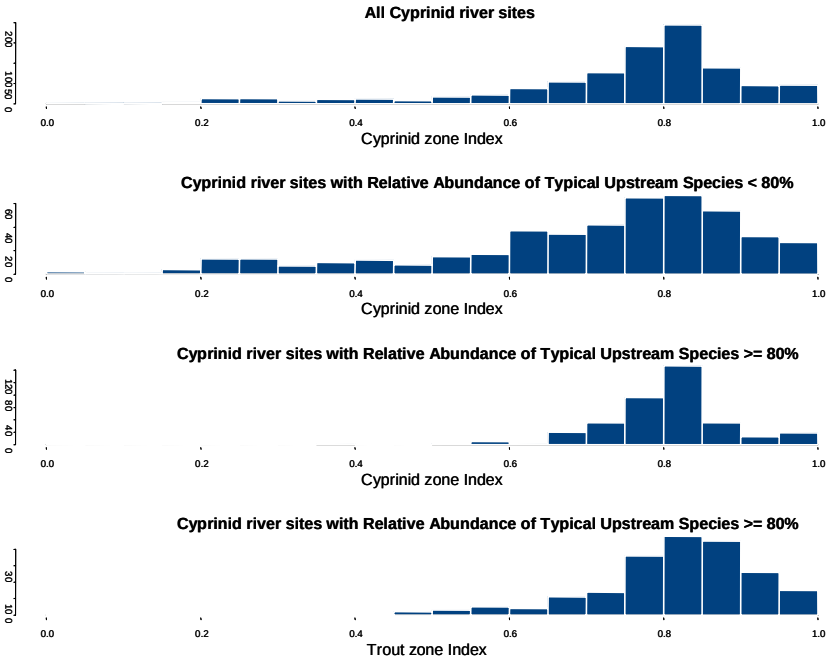


Figure 36. Frequency of the fish index in the Cyprinid zone

5.1.3 Limitations of the index

5.1.3.1 Sensitivity of the indices to the sampling method

The sampling method and the sampling strategy is, for a part a function of the river size and of the river depth. In case of deep river, the river is sampled by boating and in general, the sampling area is limited to the area close to the shore line. Thus, the fish sample is not really representative of the whole river section and the error associated to the evaluation of the fished area could be important and, in any case much higher than for sites sampled by wading on the whole river section.

As written previously, most of the sites used to calibrate the metric models are wadeable sites. Only 19 of the 533 calibration sites were sampled by boating. Due to this too low frequency, it was not possible to consider this variable as one of the potential explanatory variables in the modelling approach. Nevertheless, these sites were not excluded and it is necessary to evaluate the sensitivity of the selected metrics to this parameter. The hypothesis is that, for both salmonid and cyprinid zone, the distributions of the corresponding index must be similar.

We only consider "undisturbed" sites which are correctly classified (salmonid vs cyprinid zone).

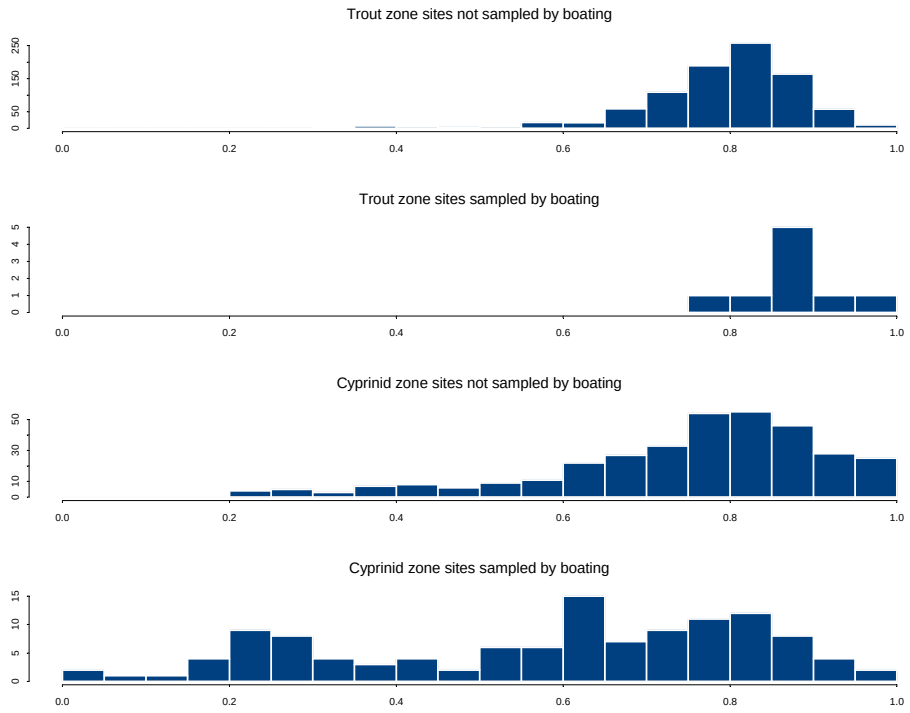


Figure 37. Distribution of the two index values in the corresponding zones as a function of the sampling method (only "undisturbed sites").

The number of sites sampled by boating in the salmonid zone is limited. But their range is not too different from the range sites sampled by wading.

At the opposite, there is a clear effect of the sampling method on the index values for the cyprinid zone. Most of low index values are related to boating sites. These low value boating sites are not belonging to a particular region or country.

As a first conclusion, it seems that the fish index, at the present state, could be used only with caution when sites have been sampled by boating, especially in the cyprinid zone, i.e. for larger and deeper rivers.

5.1.3.2 Sensitivity of the index to the sampling strategy

The same methodology is used to assess the effect of the sampling strategy (Whole or Partial) in both salmonid and cyprinid zone. Only correctly classified undisturbed sites not sampled by boating are considered (salmonid zone: 1701 sites, cyprinid zone: 661 sites).

For the trout zone, the distribution of the salmonid index values differs (Kruskall-Wallis test, p=0.0001) and the median value for sites with a Partial sampling strategy is higher (0.876 against 0.806). But the 2 distributions remain relatively similar and with no low value (i.e. no risk of site underscoring due to the sampling method).

For the cyprinid zone, the distributions of the cyprinid index values do not differ (Kruskall-Wallis test, p=0.126).

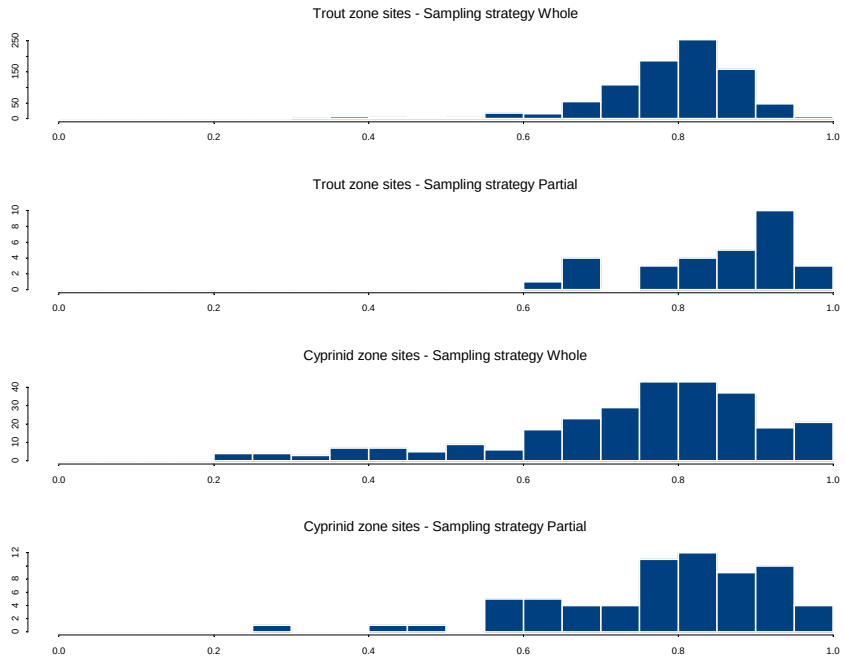


Figure 38. Influence of sampling strategy on index values for "undisturbed sites".

At the opposite of the sampling method, the sampling seems not too influent on the assessment method.

In addition, we also examined the case of fishing occasion where the lateral water bodies from the floodplain were sampled with or without the main channel (N=13). In such case, the indices values are significantly (Kruskall-Wallis test, $p<0.0001$) and clearly lower in comparison with sites where only the main channel is sampled (median of 0.29 against 0.77).

5.1.3.3 Sensitivity of the index to the number of fish caught

The same methodology is used to assess the effect of the number of fish caught in both salmonid and cyprinid zone. Only correctly classified undisturbed sites not sampled by boating are considered (salmonid zone: 902 sites, cyprinid zone: 661 sites).

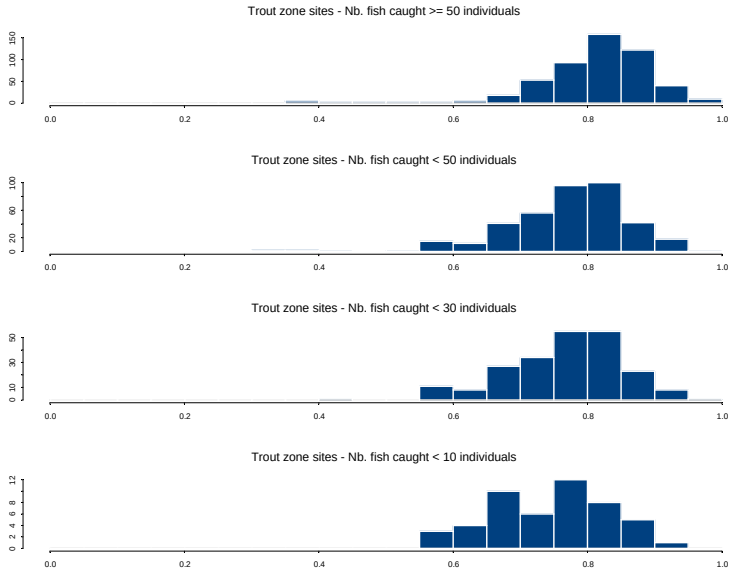


Figure 39. Influence of number of fish caught on index values for "undisturbed sites" (salmonid zone).

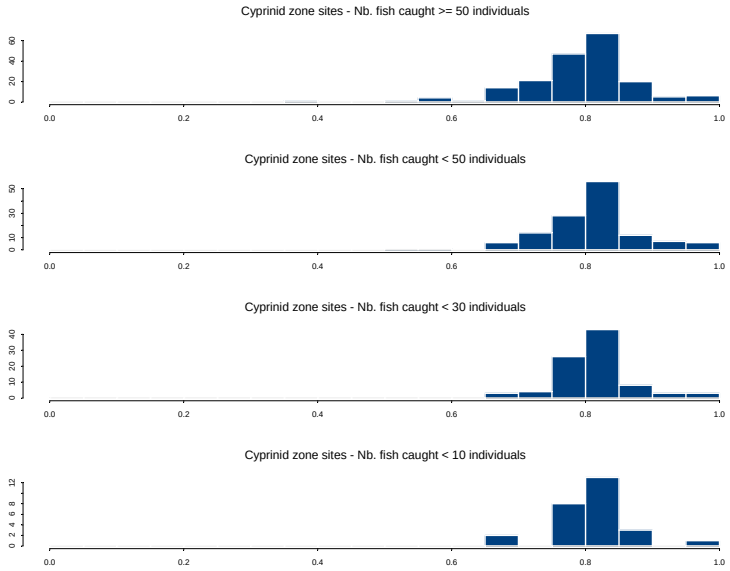


Figure 40. Influence of number of fish caught on index values for "undisturbed sites" (Cyprinid zone).

In both salmonid and cyprinid zones, the index distributions remain relatively similar whatever the number of fish caught.

For the cyprinid zone, there is no significant effect of the sampling effort on the cyprinid index value (nb. fish caught >=50 or <50, Kruskall-Wallis test, p=0.062). The same result is obtained with a threshold between the 2 groups corresponding to 30 and 10 fish caught.

For the salmonid zone, the difference between the group is significant (Nb. fish caught >=50 or <50, Kruskall-Wallis test, p< 0.0001). But the two median values are close (respectively 0.82 and 0.78), and most of the values are over 0.5. The difference between the 2 median values does not increase when considering a threshold between the 2 groups corresponding to 10 fish caught (respectively 0.81 and 0.77)

5.1.3.4 Sensitivity of the index to specific environmental situations

The same methodology is used to assess the effect of several environmental variables in both salmonid and cyprinid zones. Only correctly classified undisturbed sites not sampled by boating are considered (salmonid zone: 902 sites, cyprinid zone: 661 sites).

Environmental Variables	**River Zone**	**Kruskall-Walis Test (p value)**	**Median values of the correspond index per modality**
Flow regime	Salm.	K=0.791,df=3, p-value=0.852	
	Cypr.	K=1.173,df=2, p-value=0.556	
Geomorphological River type	Salm.	K=88.154,df=4, p-value <0.00001	Braided (0.869), Meand regular (0.82), Meand tortous(0.848), Naturally constraint(0.848),

			No mobility (0.781) Sinuous (0.830)
	Cypr	K=27.114,df=4, p-value<0.00001	Braided (0.887), Meand regular (0.802), Meand tortous(0.767), No mobility (0.710) Sinuous (0.786)
Geological typology	Salm.	K=1.2778,df=1, p-value=0.258	
	Cypr.	K=0.4019,df=1, p-value=0.526	
Water.source type	Salm.	K=16.3382,df=3, p-value=0.001	Glacial (O.73) Groundwater (0.70) Nival (0.784) Pluvial (0.812)
	Cypr.	K=3.0817,df=2, p-value=0.214	
Floodplain site	Salm.	K=21.109,df=1, p-value<0.0001	No (0.805) Yes (0.854)
	Cypr.	K=11.308,df=1, p-value=0.0008	No (0.774) Yes (0.804)
Valley form	Salm.	K=27.890, df=3, value<0.0001	Gorges (0.803) Plains (0.794) U-shape (0.847) V-shape (0.804)
	Cypr	K=9.7198,df=3, p-value=0.0211	
Natural sediment	Salm.	K=37.8226,df=2, p-value<0.0001	Boulder/Rock (0.779) Gravel/Pebble/Cobble (0.816) Sand (0.866)
	Cypr	K=6.320, df=4, p-value=0.177	
Lakes.upstream	Salm.	K=7.1106,df=1, p-value=0.0077	No (0.807) Yes (0.75)
	Cypr	K=9.0839,df=1, p-value=0.0026	No (0.789) Yes (0.570)

Flow regime: the four regimes did not differ significantly in the 2 river zones. But the range of the Index value for winter dry condition in the salmonid zone is very large, compared to the others (from 0.37 to 0.94, n=13).

Gemorphological river type: There are some significant differences between modalities. But all median values are over 0.70. In both river zones, the rivers characterized by an absence of mobility have the lowest indices median values. At the opposite, braided rivers have the highest.

Geological typology: The indices values do not differ depending of the dominant geogical substrate in the upstream watershed.

Water source type: In the salmonid zone, the sites characterized by a pluvial regime have a higher median value than other regime. The range of index values in glacial and groundwater influenced regime is larger, with the presence of low values (range from 0.36 to 0.90).

Presence of a floodplain: In both river types, the index values are significantly higher in the presence of a flooplain.

Valley form: the form of valley is mainly influential in the salmonid zone where U form valley has a higher score probably in relation with a braided-dominant fluvial dynamic.

Natural sediment: Significant differences only occur in the salmonid zone and are related to lower values for sites dominated by coarse sediment (boulder and rock).

Presence of a natural lake upstream: The presence of a lake upstream from the site have a negative influence on the score, in particular in the cyprinid zone with a low median value (0.57) and a large range of values (from 0.23 to 1.00).

Additional particular situation: Case of polish organic rivers: When only considering polish rivers classified as organic river by our polish partner and undisturbed, it is also clear that the response of the cyprinid zone index is not appropriate with a median value of only 0.39, i.e. much lower than the expected 0.80.

Conclusion

In conclusion, some environmental situations are not correctly handled by the two indices. We will mainly consider case where the range and/or the median value of the index values for undisturbed sites are clearly altered, i.e. much lower than 0.80. These situations are:

- presence of a natural lake upstream from the site
- presence of a winter dry period
- case of "organic" rivers

Several others parameters influence significantly the index values in undisturbed sites but the effect is not a clear underestimation of the site status, i.e. the indices values remain close to 0.80. These effects, which are not completely controlled by the models seems to have sense from the ecological point of view:

- negative effect in relation to an absence of river mobility
- positive effect in relation with braided-type fluvial dynamic
- positive effect for pluvial regime rivers
- positive effect in the presence of a floodplain
- negative effect for very coarse sediment dominated rivers

5.1.3.5 Case of large rivers

As previously, we only consider here sites not sampled by boating (mainly wading) and without natural lakes upstream. Due to the very low number of sites belonging to the salmonid zone and characterized by a large upstream drainage area (> 500km2), we only consider cyprinid type sites.

The size of the drainage area do not have a significant effect on the index values when considering sites with a drainage area below or over 500 km2 (Kruskal-Wallis chi-square = 2.841, df = 1, p-value = 0.092), and below or over 1,500 km2 (Kruskal-Wallis chi-square = 3.7711, df = 1, p-value = 0.0521).

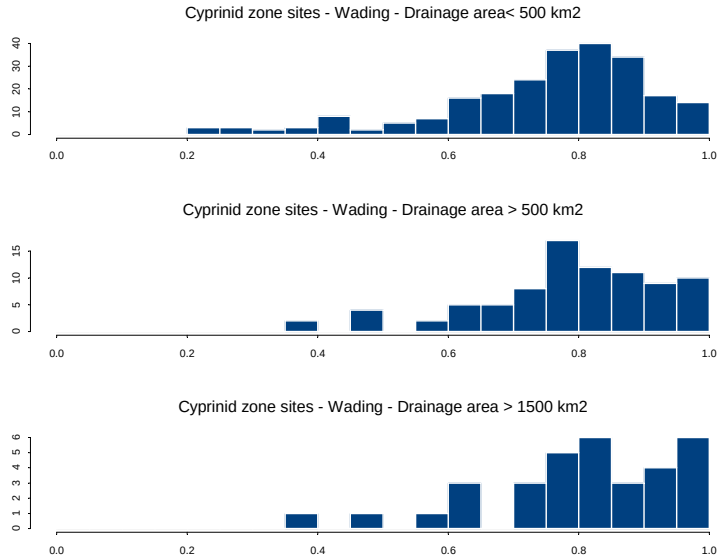


Figure 41. Influence of the drainage area on index values for "undisturbed sites" (Cyprinid zone) not sampled by boating (mainly wading)

Nevertheless, it must be mentioned that very large rivers (> 10,000 km2) are not considered if this test: 35 sites between 1500 and 10,000 km2, only one site over (16,825 km2).

Case of sites sampled by boating

Only 14 sites sampled by boating and classified undisturbed have an upstream drainage area > 10,000 km2.

When only considering undisturbed sites sampled by boating in the cyprinid zone, no significant differences appear between sites with a drainage area <500 km2, between 500 and 1500 km2 and > 1500 km2 (Kruskall-Wallis chi-square = 0.8633, df = 2, p-value = 0.6494).

At the opposite, the three categories of sites are characterized by low median values in comparison with sites not sampled by boating (mainly wading): respectively 0.627 0.718 and 0.691).

This result confirms the influence of the sampling method on the index values, in particular in the cyprinid zone.

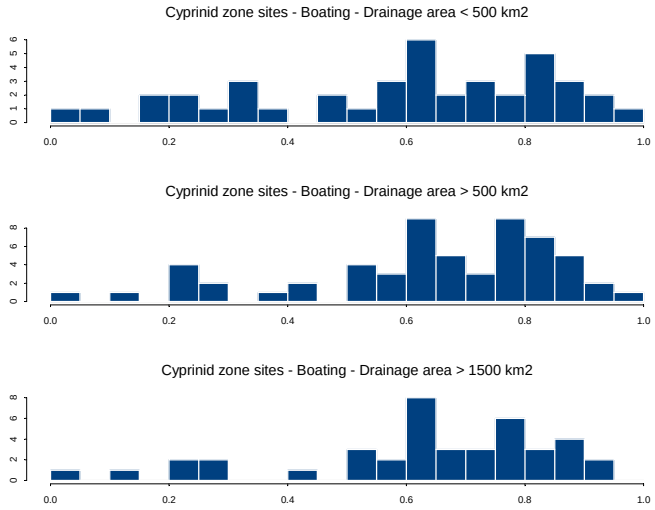


Figure 42. Influence of the drainage area on index values for "undisturbed sites" (Cyprinid zone) sampled by boating

5.1.3.6 Sensitivity of the index to the species richness

The influence of the species richness on indices values is tested separately in the 2 river zones.

Salmonid river type:

In the salmonid zone, the species richness has a slight influence on the index values (Kruskal-Wallis chi-square = 8.8915, df = 3, p-value = 0.0308). But the median values do not differ a lot (range from 0.802 to 0.832), the main effect being related to a higher median value for river with a species richness >=5.

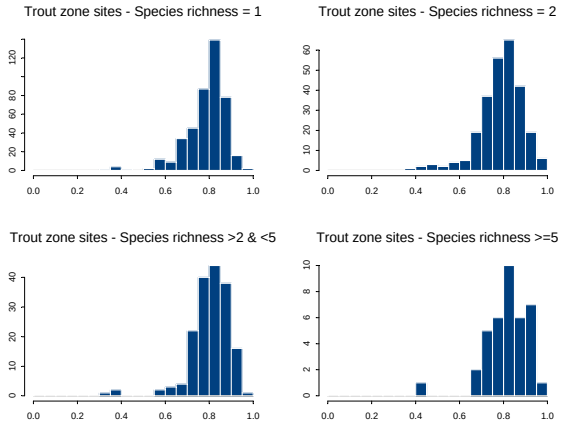


Figure 43. Influence of total species richness on index values for "undisturbed sites" in the salmonid zone

Cyprinid river type:

In the cyprinid zone, the species richness has no significant influence on the index values (Kruskal-Wallis chi-square = 2.5906, df = 3, p-value = 0.459), and the median values for the 4 richness classes are close to 0.80.

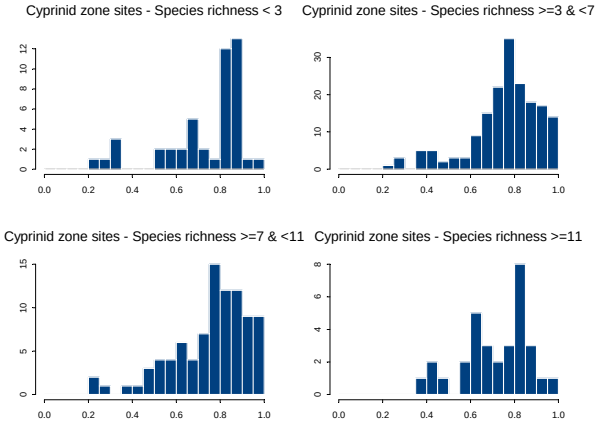


Figure 44. Influence of total species richness on index values for "undisturbed sites" in the cyprinid zone.

In both river zones, our modelling approach is able to correctly handle the influence of species richness in undisturbed sites. Nevertheless, these results do not demonstrate that the responses of the two indices to pressures are independent from the species richness.

5.1.4 Conclusion and recommendation

Two indices, each composed of 2 different metrics, are computed for each site, depending on the river type classification.

Salm.Fish.Index = (Ni.Hab.150 + Ni.O2.Intol) / 2
Cypr.Fish.Index = (Ric.RH.Par + Ni.LITHO) / 2

Indices values vary between 0 and 1. As for each metric, an undisturbed site would have an index value close to 0.80, and a highly disturbed site a value lower than the 25% quantile of the index distribution for undisturbed sites.

Ric.RH.Par	Rheophilic reproduction habitat species richness
Ni.O2.Intol	Oxygen depletion intolerant species abundance (Nb. individuals)
Ni.LITHO	Lithophilic reproduction habitat species abundance (Nb. Individuals)
Ni.Hab.Intol.150	Abundance of individuals < 15 cm of Habitat intolerant species

5.1.4.1 River zonation classification

A critical point to use the method is the classification of sites in one of the two river zone (salmonid river zone versus cyprinid river zone). From our definition, in the absence of any human disturbance, a salmonid river zone site is characterized by a very high proportion of the intolerant ST-species (most of them with more than 80% of individuals belonging to

these species). At the opposite, a typical cyprinid site is characterized by a relative low abundance of these species (lower than 20%, in most of cases 10%).

The classification is more efficient to identify the salmonid river type than the cyprinid one. Concerning the salmonid river type, only a small number of sites can be considered as misclassified (i.e. with a very low relative abundance of ST-species). At the opposite, a larger amount of sites classified as "cyprinid river type" are dominated by ST- species.

It is clear that the consequences of a misclassification are quite different, depending of the river type.

- For sites misclassified as salmonid river sites (i.e. with a low relative abundance of ST-species), and in the absence of any disturbance, the salmonid fish index cannot be used, and has to be replaced by the cyprinid fish index.
- For undisturbed sites misclassified as cyprinid sites with a high relative abundance of ST- species, the values given by the cyprinid index are quite close to the one given by the salmonid index when the site is not disturbed. However, in case of disturbance, the impact would not be correctly evaluated if the cyprinid index is used instead of the salmonid index.

Considering the risk of misclassification and the associated consequences on the evaluation of sites the best solution is to give systematically to the user the initial classification of the site (cyprinid or salmonid river zone), the relative abundance of ST-species and the value of both indices (salmonid fish index and cyprinid fish index) when they can be computed.

Very often, the proposed river zone type is correct and the user has to consider the corresponding index. In other cases, the users, as expert, will have to evaluate the situation and to confirm the proposed classification or will have to make their own choice between the two fish indices.

There are several possibilities and associated recommendations:

<u>Sites classified by the EFT classification as Salmonid river zone site</u>

The site is classified as a "Salmonid" site and the % of ST- species is high (i.e. > 80%). The classification is correct and the Salmonid fish index has to be used.

The site is classified as a "Salmonid" sites and the % of ST-species is relatively high (from 50 to 80%). The reduction of the relative abundance of ST-species could be due to a human disturbance of the river ecosystem. The risk of misclassification is relatively low but the user has to check the proposed typology.

The site is classified as a "Salmonid" sites and the % of ST-species is relatively low (from 20 to 50%) to very low (less than 20%). The reduction of the relative abundance of these intolerant species can only be due to a very severe human disturbance (i.e. heavy impoundment, high level of water quality degradation …). The risk of misclassification is important and the user has to evaluate the proposed typology and to confirm or reject the choice of the adapted fish index. A warning is included in the output of the software.

<u>Sites classified by the EFT classification as a Cyprinid river zone site</u>

The site is classified as a "Cyprinid" site and the % of ST-species is very low (less than 20 %). The classification is correct and the Cyprinid fish index has to be used.

The site is classified as a "Cyprinid" sites and the % of ST-species is relatively high (from 20 to 50%). The increase of the relative abundance of these intolerant species can be due to some particular human disturbance of the river ecosystem (extreme channelization and huge increase of the water velocity, water cooling downstream from a dam …). Nevertheless, in most of cases, a misclassification of the site is possible. The software proposes to classify the site as a salmonid river zone type and to use the Salmonid.Fish.index. The user has to evaluate the proposed typology and to confirm or reject the choice of the adapted fish index. A warning is included in the output of the software.

The site is classified as a "Cyprinid" sites and the % of ST-species is quite high (from 50 to 80%) or very high (more than 80%). The increase of the relative abundance of these intolerant species can also be due to particular severe human disturbances (see upper § for examples) but the risk of misclassification is very important. A correction for the river zone is included in the output of the software (site reclassified as a Salmonid river type site) and the value of the Samonid fish index is proposed. The software proposes to classify the site as a salmonid river zone type and to use the Salmonid.Fish.index. The user has to evaluate the proposed typology and to confirm or reject the choice of the adapted fish index. A warning is included in the output of the software.

The different options are summarized in Table 47.

Table 47. **Summary of the different options to select the appropriate fish index.**

	% of ST-species (intolerant salmonid type species)			
Initial site classification	[0% – 20%]	]20% - 50%]	]50% - 80%]	]80% - 100%]
Salmonid zone	Risk of misclassification **Salmonid index proposed** User has to confirm the river zone and the index choice	Risk of misclassification **Salmonid index proposed** User has to confirm the river zone and the index choice	**Salmonid Index recommended** User has to check the classification	Correct classification **Salmonid Index should be used**
Cyprinid zone	Correct classification **Cyprinid Index should be used**	Increase of % of intolerant species can be linked to a human disturbance **Salmonid Index proposed** User has to confirm the river zone and the index choice	Increase of % of intolerant species can be linked to particular extreme disturbance **Salmonid Index proposed** User has to confirm the river zone and the index choice	High risk of misclassification **Salmonid Index proposed** User has to confirm the river zone and the index choice

In particular ecoregions, the possibilities for a site to be a salmonid river zone site are very low (see section 1.1.1). This is the case for Hungarian lowlands, Eastern plains, Pontic province, Baltic province and Mediterranean region.

In particular ecoregions, the possibility for a site to be a cyprinid site is very low (see section 1.1.1). This is the case for Alps, Pyrenees, Fenno-Scandian shield and Borealic uplands.

5.1.4.2 Limitation of the Index in relation with the environment

The statistical models that are used for the EFI reflect the average response of fish communities to environmental conditions. The application of the EFI for particular environmental situations might cause problems.

This index has been developed for sites located in the ecoregions presented in annex. Therefore, the index should not be applied in areas with a fish fauna deviating from those of the tested ecoregions.

The model was developed using data from sites with environmental characteristics ranging between specific limits. These values are given in Table 4 and Table 5. Your site should have characteristics within these ranges in order to obtain a confident EFI.

Some environmental situations are not correctly handled by the two indices. These situations are:

- presence of a natural lake upstream from the site
- presence of a winter dry period
- case of "organic" rivers

Even if no clear effect have been observed, the indices must be used with caution for intermittent/ summer dry rivers due to the low number of undisturbed sites used to test the index.

River size: The metrics have been mainly calibrated for rivers with an upstream drainage area less than 10,000 km2. Independently from the sampling method, the river size seems not to significantly influence the index values for undisturbed sites when the upstream drainage area is less than 10,000 km2.

The index should be used with caution in the lowland reaches of very large rivers as no reference sites from these reaches have been used for the calibration of the index. In those cases the index uses only extrapolated predictions based on the trends observed in the models.

5.1.4.3 Limitations in the use of the Index due to the number of fish caught

When few specimens were caught the software still allows you to calculate the index, but the results must be considered with care. The same applies when the sampled area is smaller than 100 m². Consequently, when no fish occur at a site, this method is not applicable.

The index seems relatively independent from the number of fish caught. This could be directly related to our modelling methods. All the 4 selected metrics are modelled after taking into account the sampling effort (i.e. the total richness or the number of fish caught depending

of the metric). Nevertheless, a too low number of fish caught would alter the capacity of the index to respond correctly to a pressure. The user has to be careful when the number of fish caught is less than 30 individuals and a warning has to be included in the output of the software in such a situation.

Two cases could be problematic and the EFI should be used with care:

(2) undisturbed rivers with naturally low fish density and (2) heavily disturbed sites where fish are nearly extinct. In the first case, fish are close to the natural limits of occurrence and therefore might not be good indicators for human impacts. The occurrence of fish in those rivers is highly coincidental and therefore not predictable. If the very low density is caused by severe human impacts more simple methods or even expert judgement are sufficient to assess the ecological status of the river.

5.1.4.4 Limitations in the use of the Index due to the sampling method

Only fish data obtained with single-pass electric fishing may be used to calculate the EFI. If data from multiple passes are used (i.e. same site fished several times and catches cumulated) the EFI produces erroneous results.

The sampling method (boating or wading) has a strong impact on the index values. Most of our calibration sites were sampled by wading and it was not possible to include the variable describing the sampling method as a potential explanatory variable.

The number of sites sampled by boating in the salmonid river zone is limited. But their range is not too different from the range sites sampled by wading. At the opposite, there is a clear effect of the sampling method on the index values for the cyprinid zone. Most of low index values are related to boating sites. These low value boating sites are not belonging to any particular region or country.

As a first conclusion, it seems that the fish index, at the present state, could be used only with caution when sites have been sampled by boating, especially in the cyprinid zone, i.e. for larger and deeper rivers. The boating effect is not only to reduce the mean value of the cyprinid index but to increase its variability.

Nevertheless, as additional information, we propose to the user a classification of sites sampled by boating in 5 specific classes, defined in a different way than for wadeable sites (see next section). This specific scoring has just to be considered as a preliminary one and a more specific work is needed in the future if enough undisturbed or slightly disturbed sites sampled by boating are available.

5.1.5 Scoring in 5 classes

Ecological class boundaries are only based on the distributions of indices values for undisturbed sites in the two river types (Table 48).

As the sampling method greatly influences the score value especially in the cyprinid zone, class boundaries have been computed separately for sites sampled by boating and wading in the cyprinid zone (see Indices limitations section below).

The limits between class 1 and 2 correspond to the value of the 95% quantile of the index distribution for undisturbed sites.

The limits between class 2 and 3 correspond to the value of the 25% quantile of the index distribution for undisturbed sites.

The limits between classes 3-4 and 4-5 are defined in a way that the ranges between classes 3, 4 and 5 are similar.

The specific scoring for cyprinid zone sites sampled by boating has to be considered as a preliminary one. A more specific work is needed in the future, by using enough undisturbed or slightly disturbed boating sites and being able to correctly handle these parameters in the different models.

Table 48. Ecological class boundaries for the 2 indices.

	Salmonid Zone index	Cyprinid Zone Index	
		Wading	Boating
Class 1	[0.911 -1]	[0.939 -1]	[0.917 - 1]
Class 2	[0.755- 0.911[	[0.655- 0.939[	[0.562 - 0.917[
Class 3	[0.503 -0.755[	[0.437 -0.655[	[0.375 - 0.562[
Class 4	[0.252 -0.503[	[0.218 -0.437[	[0.187 - 0.375[
Class 5	[0 - 0.252[	[0 - 0.218[	[0 - 0.187[

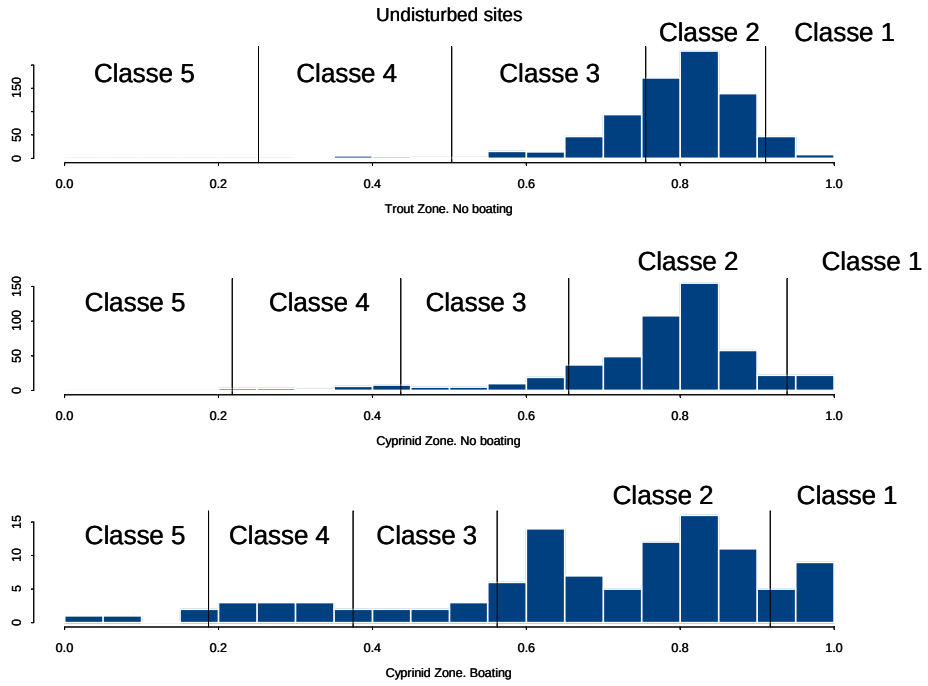


Figure 45. Distribution of salmonid index values and cyprinid index values (wading and boating sites). Vertical lines represent the boundaries between classes for each of the three cases.

5.2 Performance analyses

5.2.1 Tools and concepts

To quantify the performance of altered site detection and to determinate an optimal threshold, we used the slightly disturbed sites as unexposed sites, and the sites classified in the classes 4 and 5 by the restricted pressure index (Press.Index.B described in section 2.1.4) as exposed sites. However, it's clear that the pressure variables are not exact measures of the site status and they only correspond to the specific situation with a high potential risk of alteration. The exposures to pressures don't systematically improve an effect on the biological and functional structure of fish assemblage. The pressure indices can't be considered as a TRUE gold standard. Consequently, measures of performance won't provide the true capacity of the fish index to detect altered sites. Nevertheless, they give information on the coherence (or association) on the both indices. The evaluation of index "performance" requires some adaptations of classical epidemiological concepts (e.g. Fletcher & Fletcher 2005, Spitalnic 2004a, 2004b, Freeman & Moisen 2007, Table 49):

- **Sensitivity a/(a+c):** the probability that a test will be positive given a sites with pressures (index indicate that the site is altered).
- **Specificity d/(d+b):** the probability that a test will be negative given a site without pressures (index indicate that the site is unaltered).
- **Positive predictive value (PPV, a/(a+b)):** the probability that a site will have pressures given a positive test result (index indicate that the site is altered).
- **Negative predictive value (NPV, b/(c+d)):** the probability that a site will not have pressures given a negative test result (index indicate that the site is unaltered).

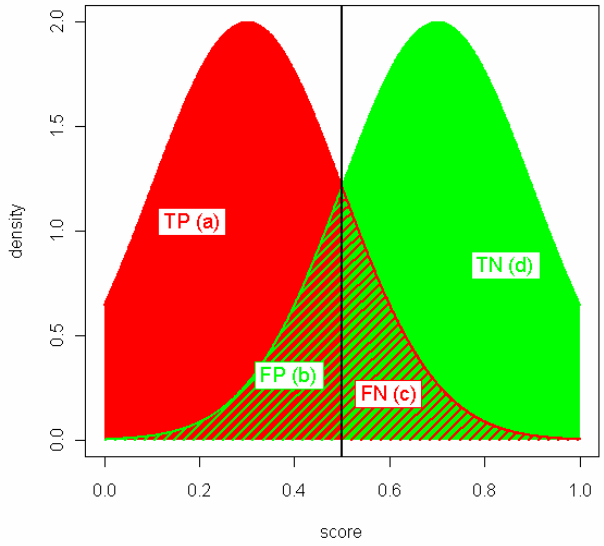


Figure 46. Graphical representation of the status of sites in function of the score (altered/unaltered) and the exposure level (exposed/unexposed): a: exposed sites detected as altered (True positive); b: unexposed sites detected as altered (False positive); c: exposed sites detected as unaltered (False negative); d: unexposed sites detected as unaltered (True negative). The black vertical defines the cut-off values.

Table 49. Relationship between the pressure index and ecological status defined by the multi-metric index.

Ecological status	**Pressure Index B**	
	Affected sites (1)	Unaffected sites (1)
Altered sites (+)	a (TP)	b (FP)
Unaltered sites (-)	c (FN)	d (TN)

All of these parameters are not intrinsic to the performance test and are determined by the ecological and management context in which the index is employed. As results, it's essential that diagnostic accuracy studies address how well the index identifies the target condition of interest (Fletcher & Fletcher 2005). In a similar way, Florwski (2008) stress the need to define the action range of the diagnostic tools and writes that "*a high sensitivity corresponds to high negative predictive value and is the ideal property of a "rule-out" test. In contrast, high specificity corresponds to high positive predictive value and is the ideal property of a "rule-in" test*".

The index performance estimation requires a specific dataset characterized by the exclusion of the following sites:

- Sites sampling by boating
- Sites with flow regime classified 'winter dry'
- ~~Sites with ST-species proportion inferior to 80% for the sites localised in salmonid zone~~
- Sites with ST-species proportion superior to 20% for the sites localised in the cyprinid zone (A verifier)
- Sites with lake upstream
- Sites with sampling location coding in "Backwaters" or "Mixed"
- Sites classified in organic
- Sites without the information on the seven pressures used to compute the pressure index (Press.Index.B)

Supprimé : in

The definition of this dataset reduces the potential bias induced by particular sites and it integrates the limitation of the index use presented in the previous section 5.1.4.

5.2.2 Evaluation classification

To quantify the performance of the classification (see in section 5.1.5), we used only the slightly disturbed sites and exposed sites (pressure index equal to classes 4 or 5). This manner, we can evaluate the capacity of the index to discriminate the exposed and unexposed sites. The confusion matrices associated with pressure index (Press.Index.B) and the both fish index are available in the Table 50.

Table 50. Confusion matrices computed with the pressure index (Press.Index.B) and Fish indices for the both zone (cyprinid and salmonid zone).

salmonid zone				**cyprinid zone**	
Fish index	**Press.Index.B**			**Pressure.Index.B**	
Inverse dess ???	Exposed (1)	Unexposed (0)		Exposed (1)	Unexposed (0)
Altered (+)	64	54		186	52
Unaltered (-)	72	666		185	416

The design of the cyprinid dataset is relatively balanced. We count 839 sites which contain 468 unexposed sites and 371 exposed ones and the pressure exposure prevalence (sites in class 4 or 5) is equal to 0.44 (Table 51). The sensitivity of the Cyprinid index is relatively low, but the specificity (0.89) and positive predictive values are relatively strong (0.78). Thus, if we suppose

that the pressure index provides true information on the ecological status of a given site, the index recognizes undegraded sites in most cases and the detection of an altered situation efficiently confirms the high degraded level of this site. As results, the cyprinids index appears to be a typical *"rule-in"* test (e.g. Fletcher & Fletcher 2005, Florwski 2008). From the economical point of view, this objective corresponds to the idea that the risk for managers to invest in restoration measures for undegraded sites is low.

The low sensitivity and high number of False Negative can be partly explained by the systematically low index values observed in the sites sampling by boating (see previous section on the limitation of the indices). Consequently, these results trend to justify the necessity to correct the classification for these particular sites.

Table 51. Quantification of the Index performances computed from the Press.Index.B for the cyprinid zone. In these results, we excluded the intermediate sites (class 3) and we compare slightly disturbed sites against disturbed sites (class 4 and 5). The interval estimation of good classification was computed by weighting bootstrap procedure and the parameters 'est', 'lower' and 'upper' correspond to the mean and percentiles of the simulated distribution. For the other values, we used classical 95% confidence intervals.

cyprinid zone	measure	estimation	lower	upper
prevalence	true	0.4422	0.4089	0.476
performance	sens	0.5013	0.4507	0.552
	spec	0.8889	0.8572	0.9143
	Good classification	0.6955	0.6651	0.7259
	Kappa	0.4053	0.3412	0.4694
predictive.value	Positive	0.7815	0.7248	0.8293
	Negative	0.6922	0.6542	0.7278

In the salmonid zone, the performance information must be considered with caution, because the design of the dataset is strongly unbalanced. Our working dataset (N=856) contains 720 unexposed sites and 136 exposed sites. The prevalence of disturbance (class 4or 5) is only equal to 0.16. As with the Cyprinid zone, we observe that the specificity (0.93) is largely superior to sensitivity (0.47) and it appears to be a *"rule-in"* test (e.g. Fletcher & Fletcher 2005, Florwski 2008). However, the low positive predictive value (0.54) casts some doubt on the validity of our test. A more balanced dataset with more degraded sites should improve the pertinence of this one.

Supprimé : y

Table 52. Quantification of the Index performances computed from the Press.Index.B for the salmonid zone. In these results, we excluded the intermediate sites (class 3) and we compare slightly disturbed sites against disturbed sites (class 4 and 5). The interval estimation of good classification was computed by bootstrap procedure and the parameters 'estimation', 'lower' and 'upper' correspond to the mean and percentiles of the simulated distribution. For the other values, we used classical 95% confidence intervals.

salmonid zone	measure	estimation	lower	upper
prevalence	true	0.1589	0.1359	0.1849
performance	sens	0.4706	0.3887	0.5541
	spec	0.9250	0.9034	0.9421
	Good classification	0.6981	0,6659	0.7278
	Kappa	0.4180	0.3242	0.5119
predictive.value	Positive	0.5424	0.4526	0.6295
	Negative	0.9024	0.8789	0.9218

The low sensitivities could be partly explained by the misclassification associated with the river type (cyprinids and salmonid, see the previous section on the limitation of the use of the indices).

In addition, the sites characterized by pressures without effect on the fish assemblage also reduce the index sensitivity.

5.2.3 Classification and optimisation

In the precedent section, class limits was only defined on the slightly disturbed dataset and we evaluated the performance of this classification against the high disturbed dataset (class 4 and class 5). However, other strategies can be performed a categorization of indices based on the optimisation of statistical criteria adapted to binary data such as kappa index (Conger 1980, Altman et al. 2000) or the sum of sensitivity and specificity (e.g. Hosmer & Lemeshow 2000, Bardos 2001, Collett 2003, Saporta 2006 , Freeman & Moisen 2008). As results, we propose to reduce our working dataset in searching the index limits between the both main groups formed by slightly disturbed sites and by high disturbed sites (classes 4 and 5). In the following list, we present two methods to define an optimal threshold:

Supprimé : y

- maximization of Kappa index defined as follow:

$$\kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i.}x_{.i}\right)}{N^2 - \sum_{i=1}^{r}\left(x_{i.}x_{.i}\right)}$$

The terms x_{ii}, $x_{.i}$, $x_{i.}$ and N represent respectively the i^{th} values of the diagonal of confusion matrix, the margin values of the columns and rows and the number of observation (Conger 1980, Altman et al. 2000).

- maximization of the sum of the specificity (spec) and sensitivity (sens)

Other criteria based on percentage of good classification or minimization of $(1-sens)^2+(spec-1)^2$ could be used to defined the threshold.

Table 53. Quantification of the cyprinid index performances for the both optimal thresholds. The cyprinid dataset (N=951) integrates slightly disturbed sites and disturbed sites (class 4 and 5).

Criteria	measure	estimation	lower	upper
Kappa				
	cut-off	0.2665568		
	goodclassification (raw)	0.7087277		
	goodclassification (boot)	0.7054562	0.6761	0.7361
	prevalence	0.4637	0.4322	0.4955
	sensitivity	0.6757	0.6307	0.7177
	specificity	0.7373	0.6974	0.7736
	positive predictive value	0.6898	0.6447	0.7316
	negative predictive value	0.7245	0.6845	0.7612
	kappa	0.4136	0.3554	0.4717
Sens+Spec				
	cut-off	0.2619492		
	goodclassification (raw)	0.7087277		
	goodclassification (boot)	0.7068562	0.6761	0.735
	prevalence	0.4637	0.4322	0.4955
	sensitivity	0.6757	0.6307	0.7177
	specificity	0.7373	0.6974	0.7736
	positive predictive value	0.6898	0.6447	0.7316

	negative predictive value	0.7245	0.6845	0.7612
	kappa	0.4136	0.3554	0.4717

The cyprinid zone dataset (N=951) is relatively balanced and it includes 510 unexposed and 441 exposed sites. The optimal thresholds obtained by the both methods are relatively low 0.267 and 0.262. The procedure seems increase the sensitivity (sens=0.68) and reduce the specificity (spec=0.74). The balance between these two measures provides a balance of classification error and it increases the index capacity to detect the degraded sites. At the opposite, it reduces the index capacity to detect the undegraded sites. As results, the index appears to be more polyvalent and less efficient to confirm a high degraded level.

The salmonid zone dataset (N=923) include 779 unexposed and 144 exposed sites. This unbalanced design (prevalence=0.16) involves the necessity to consider the results with caution. The optimal thresholds are relatively similar to the threshold proposed in the precedent section: the cut-off is equal to 0.45 with kappa criterion and it equal to 0.38 with the criterion based on the sum of specificity and sensitivity against 0.503 for the previous classification (section 5.1.5). The categorization procedure based on optimal criteria increases the specificity (spec=0.98 and spec=0.96, Table 54) and positive predictive values. In the case of the threshold based on kappa criterion, the high specificity (spec=0.98) associated with high positive predictive values (ppv=0.81) provides an interesting *"rule-in"* test (e.g. Fletcher & Fletcher 2005, Florwski 2008). For the second method based on the sum of specificity and sensitivity, the increase of the positive predictive value is more moderate.

Table 54. Quantification of the salmonid index performances for the both optimal threshold. The cyprinid dataset (N=923) integrates slightly disturbed sites and disturbed sites (class 4 and 5).

Criteria	measure	est	lower	upper
Kappa				
	cut-off	0.4500		
	goodclassification (raw)	0.8906		
	goodclassification (boot)	0.6867	0.6587	0.7183
	prevalence	0.1560	0.1340	0.1808
	sensitivity	0.3889	0.3131	0.4704
	specificity	0.9833	0.9717	0.9902
	positive predictive value	0.8116	0.7039	0.8865
	negative predictive value	0.8970	0.8748	0.9156
	kappa	0.4725	0.3754	0.5696
Sens+Spec				
	cut-off	0.3806		
	goodclassification (raw)	0.8776		
	goodclassification (boot)	0.6944	0.6652	0.7226
	prevalence	0.1560	0.1340	0.1808
	sensitivity	0.4306	0.3525	0.5122
	specificity	0.9602	0.9441	0.9718
	positive predictive value	0.6667	0.5659	0.7542
	negative predictive value	0.9012	0.8790	0.9197
	kappa	0.4567	0.3628	0.5505

5.2.4 Conclusion

As mentioned earlier, the idea of fish index is that the metrics quantify the distance between the predicted fish community and the observed one and allow us to evaluate the risk (« probability ») for a site to be an undegraded site. This risk for a site to be an undegraded site will decrease when the distance increase. Eventually, the Water Framework Directive (WFD) considers three main objectives in relation with the assessment of water bodies:

- The ecological status of one site cannot be worst in the future. This is especially true for reference sites (class1)
- Sites having an ecological status clearly altered (i.e. the bio-indicator values deviate significantly from the « reference condition value ») have to be restored.
-

For the first point and in particular in the case of detection of class1 sites, our method is probably not the best. Our models are calibrated with sites no or slightly disturbed, which means that our capacity to distinguish between this two categories of sites « reference (class1) and slightly disturbed (class2) is limited. At the opposite, our main task is linked to the detection of sites which need restoration measures, i.e. sites characterized by a very low chance to be an undegraded site. From the economical point of view, this objective corresponds to the idea that the risk for managers to invest in restoration measures for undegraded sites is low. The cyprinid index typically comes up to these expectations. For the salmonid index, the results of performance analyses are less clear, because our working dataset is particularly unbalanced (excess of undegraded sites compared with degraded sites).

6 How estimate the error of multi-metric index based on modelling step: an proposition

In this chapter, we propose a general procedure to compute the predictive error induced by the modeling process. To estimate the tolerance interval associated with individual and global scores, we use a hybrid approach based on three main elements: i) theoretical knowledge on generalized linear model (GLM), ii) simulation procedure and iii) principle of the error propagation. Thus, we propose to compute limit values based on tolerance interval for each model and under the assumptions of error propagation, the final tolerance interval for the scores are obtained by the transformation of the limit values. For example, in the case of symmetrical interval and if we consider the function $f(\ )$, the interval $[y-\Delta y; y+\Delta y]$ provides the interval $[f(y-\Delta y); f(y+\Delta y)]$, where Δy is the distance between the value y and limit values. This approach is used at each step of the index computation.

6.1.1 Confidence and Prediction Intervals

The computation of the confidence or prediction intervals (or tolerance intervals) associated with the expected values is based on solid theoretical knowledges (e.g. Atlman et al. 2000, Fox 2002). Thus, the limits values are obtained by the classical formula defined as follow:

$$\hat{y}_x \pm \hat{\sigma}(\hat{y}_x) \cdot t_{1-\alpha,n-p}$$

Where $\hat{y}_x$, $\hat{\sigma}(\hat{y}_x)$ and $t_{1-\alpha,n-p}$ correspond to the estimated values, standard deviation associated with estimated values and the theoretical values from Student distribution (α = error type I; $n-p$ = degree of freedom).

The confidence interval estimates the error associated with the expected values used in model construction. On the other hand, prediction (or tolerance) interval is used to describe the error associated with an individual and new observation independent to data used in the model. The both intervals differ from the computation of the variance (more mathematical detail in Greene 2002 and Saporta 2006, Chatfield 1993, Altman et al. 2000, Fox 2002). The Estimation of the variance of the estimated values $\hat{y}_x$ for confidence interval ($Var(y_x - E(Y|X_x))$) is given by:

$$\hat{\sigma}^2(\hat{y}_x) = \hat{\sigma}^2 \mathbf{X}_x^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_x \text{ where } \hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and the estimation of the variance of the estimated values $\hat{y}_x$ for prediction interval ($Var(y_x - Y|X_x)$) is given by:

$$\hat{\sigma}^2(\hat{y}_x) = \hat{\sigma}^2 (1 + \mathbf{X}_x^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_x)$$

In the literature, the following formula to compute the estimation of variance can be fund:

$$\hat{\sigma}^2(\hat{y}_x) = \hat{\sigma}^2 \left(\frac{1}{m} + \mathbf{X}_x^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_x \right)$$

In this equation, the term m is the number of observation used to estimate the data (sample size). The value m is generally equal to one or infinite. If m is equal to one, we consider that y_x comes from individual observation (Alvord et Rossio 1993, see section 4.4.1). In contrast, if m is equal to infinite, we consider that y_x comes from infinity of observation (Carroll et al. 1988) and we are only interested in the positioning of curve. The tolerance interval is larger than the confidence interval, but it converges on confidence interval when m indefinitely increases. For this reason, the

increase of the sample size is an efficient method to reduce the variability associated with prediction (e.g. Saporta 2006).

For generalized linear model, the computation of the error interval is more complex. Two methods are in competition to provide confidence interval for response prediction: the first, delta method, provides symmetric confidence intervals. However, the confidence interval can be outside the definition domain of the variable. The variance of the values of the additive function is obtained as follow:

$$\mathrm{var}(\hat{\mu}) = \left(\frac{\delta\mu}{\delta\eta}\right)^2 \mathrm{var}(\hat{\eta}) \textit{ where } \mathrm{var}(\hat{\eta}) = \mathbf{X}\,\mathrm{var}(\hat{\beta})\mathbf{X}^t$$

In the second approach, the error interval is systematically contained in the definition domain of the variable. The computation of the interval is based on the additive function and the inverse transformation (inverse link, g^{-1}) provides the final interval.

$$IC(\mu) = g^{-1}(IC(\eta))$$

For generalized linear model, the covariance matrix is given by Fisher information matrix ($\mathbf{X}^t\mathbf{W}\mathbf{X}$). The dispersion parameter is equal to 1.0 in Binomial and Poisson models. However, if we observed problem of over or under-dispersion, we can take in account the dispersion parameter in the variance computation of expected values. This parameter corresponds to the sum of the squared Pearson residuals (e.g. McCullagh & Nelder 1989, Cameron & Trivedi 1998, Faraway 2006) divided by the residuals degree of freedom and it obtains as follow:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(\frac{w_i\left(y_i-\hat{\mu}_i\right)}{\mathrm{var}\left(\hat{\mu}_i\right)}\right)^2$$

6.1.2 Simulation of tolerance intervals for individual score

For one single metric, the model provides expected values and standard errors. By extending the classical regression propriety, a random sampling procedure based on normal, expected values and standard errors produces an empirical distribution of the expected values in the 'link space' (Algorithm 1). After the inverse link transformation, the computation of the standardized distance between the quantile values (e.g. 0.1 and 0.9) and the observed ones provides a good approximation of the predictive intervals. This strategy is commonly uses in epidemiological studies to estimate error interval of odds-ratio (e.g. Hosmer & Lemeshow 2002).

Algorithm 1. Simulation of tolerance intervals of one given metric. The terms $\hat{\eta}_x$ and $\hat{\sigma}(\hat{\eta}_x)$ correspond to the expected values (in the 'link space') and the standard errors.

For each values

⇨ 99 random samples in a normal distribution with mean equal to expected values ($\hat{\eta}_x$) and standard deviation equal to

$$\hat{\sigma}(\hat{\eta}_x) = \sqrt{\hat{\sigma}^2\left(1+\mathbf{X}_x^t\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}_x\right)}.$$

⇨ Computation of mean, standard error and quantile values (e.g. 0.10 and 0.90)

The tolerance intervals for the both metrics used in the computation of the cyprinid index (Ni.LITHO and Ric.RH.Par) are presented in the Figure 46. For the metric based on species with preference to spawn in running waters (Ric.RH.Par), we observe that the size of 80 % tolerance intervals are close to 0.39 units (+/- 0.11 units). For the one based on Lithophilic Fish, the tolerance interval appears to be larger and it's on average close to 0.42 (+/- 0.14 units).

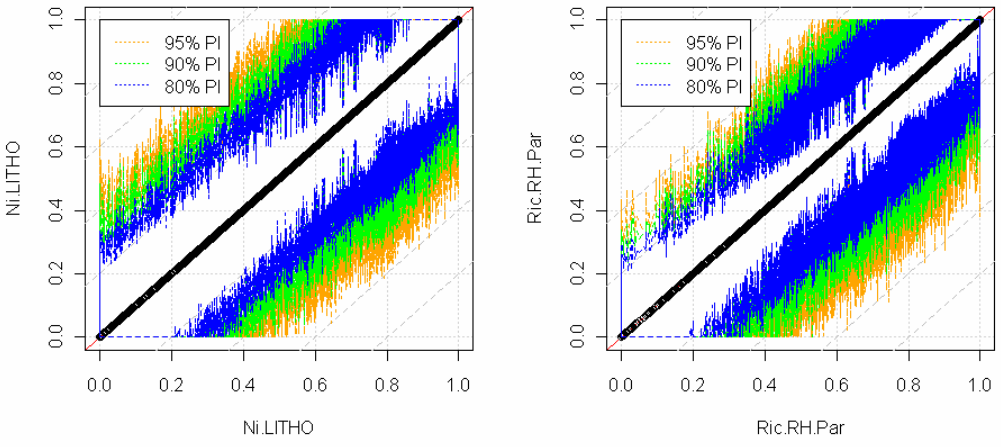


Table 55. Representation of tolerance intervals (PI) of score based on simulation procedure (99 random samples). Red, orange, green and blue lines correspond to the tolerance intervals based on percentiles (80%, blue; 90%, green; 95%, orange).

6.1.3 Predictive error after metric aggregation

As described in previous sections, the salmonid and cyprinid indices are obtained by the mean of the two different metrics. By default and without a priori knowledge on the relationship between the metrics, this aggregation mode is certainly a good pragmatic solution. However, the addition of the non-independent variables involves some statistical difficulties. At the time of metric selection, we have only selected metrics with low correlation levels, but these ones are not null and the metrics are not strictly independent. For example, in the case of the computation of the cyprinid index, the correlation between the metric Ric.RH.Par and Ni.LITHO is equal to 0.51 (R2=0.26). Consequently, the computation of theoretical variances required to compute the predictive error is extremely complicate. The classical elementary formulas on variance computation give us the following equations:

- For two independent variables A and B:
$$Var(A+B)=Var(A)+Var(B)$$

- For two non-independent variables A and B:
$$Var(A+B)=Var(A)+Var(B)+2\cdot Cov(A,B)$$

- For three non-independent variables A, B and C:
$$Var(A+B+C)=Var(A)+Var(B)+Var(C)+2\cdot Cov(A,B)+2\cdot Cov(A,C)+2\cdot Cov(B,C)$$

It's clear that the addition of supplementary non-independent variables strongly increase the variance of the index. For this reason, selection criterion based on a low correlation level is really essential to limit the size of predictive error. To reduce all these difficulties, we generalize the

previous results in adapting a simulation strategy to estimate empirical distribution of aggregated scores. At each step, we compute an empirical value from normal distribution based on expected value and expected variance for each metric and we calculate the new scores. Afterwards, we aggregate these ones to obtain the final indices. The consideration of quantile values (e.g. 0.1 and 0.9) easily completes the construction of the tolerance interval. The algorithm is defined below:

Algorithm 2. Simulation of tolerance intervals

```
For k in 1 : 99
   For each metrics
      For each values
         Random sample in a normal distribution with mean equal to
         expected values ( ŷ_x ) and standard deviation
         equal to σ̂(ŷ_x) = √(σ̂²(1 + X_x^t (X^t X)^-1 X_x)).
      Metric standardization
   Aggregation of metrics
Computation of quantiles values(e.g. 0.10 and 0.90 or 0.05 and 0.95)
```

To illustrate the computation of error associated with fish index, we have tested our algorithm for cyprinid index. The simulation procedure provides relatively encouraging and interesting results. We show that the size of the 80% tolerance interval is on average close to 0.30 units (+/- 0.06 units, Figure 47). In other words, this corresponds more or less to one class. The error estimation is presently in an experimental phase and it requires some additional tests before the implementation in the software.

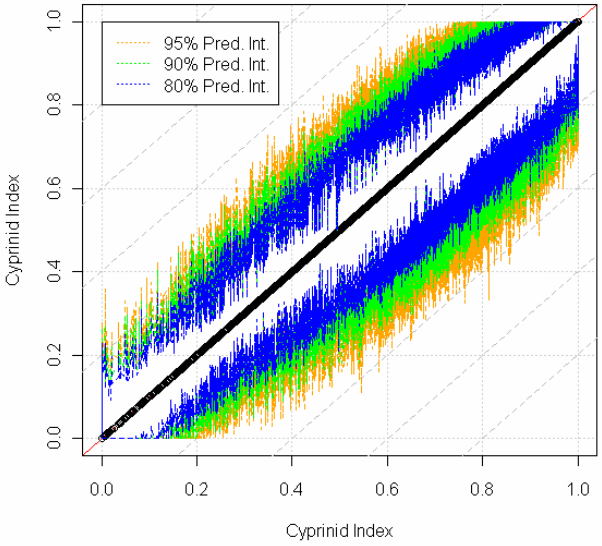


Figure 47. Simulated tolerance error associated with the cyprinid index. Red, orange, green and blue lines correspond to the tolerance intervals based on percentiles (80%, blue; 90%, green; 95%, orange).

7 References

Altman DG, Machin D, Bryant TN, Gardner MJ (2000). Statistics with Confidence, second edition. British Medical Journal, London, pp. 116 - 118.

Alvord G. & Rossio J.L. (1993) Determining confidence limits for drug potency in immunoassay. Journal of Immunological Methods . 157-155.

Austin M.P. (1980) Searching for a model for use in vegetation analysis. Vegetatio, 42, 11-21

Austin M.P., Cunningham R.B. & Fleming P.M. (1984) New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. Vegetatio, 55, 11-27

Austin M.P. (1987) Models for the analysis of species' response to environmental gradients. Vegetatio, 69, 35-45

Austin M.P., (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecological Modelling, 157, 101-118.

Austin M.P., Belbin L., Meyers J.A., Doherty M.D. & Luoto M. (2006) Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory, Ecological Modelling, 199 (2), 197-216.

Austin M.P. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches Ecological Modelling, 200 (1-2), 1-19.

Bailey R.C., Kennedy M.G., Dervish M.Z. & Taylor R.M. (1998) Biological assessment of freshwater ecosystems using a reference condition approach: comparing predicted and actual benthic invertebrate communities in Yukon streams. Freshwater Biology, 39, 765-774.

Bardos, M. (2001) Analyse discriminante. Application au risque et scoring financier. Dunod, Paris.

Ben M.G. & Yohai V.J. (2004) Quantile-quantile plot for deviance residuals in the generalized model. Journal of Computational and Graphical Statistics, 13, 36-47

Belsey D., Kuh E. & Welsch R. (1980) Regressions diagnostics. John Wiley & Sons, New York.

Bonada, N., Prat, N., Resh, V.H. Statzner, B. (2006). Developments in aquatic insect biomonitoring: comparative analysis of recent approaches. Annual Review of Entomology 51:495-523.

Buja A. (2000) Discussion of additive logistic regression: A statistical view of boosting. Annals of Statistics, 28, 387–391.

Cameron A. C. & Trivedi P. K. (1998), Regression Analysis of Count Data, Econometric Society Monograph No.30, Cambridge University Press,.

Candy S.G. (2003) Predicting time to peak occurence of insect life-stages using regression models calibrated from stage-frequency data and ancillary stage-mortality data. Agrcultural and forest Entomology, 5, 43-49

Carroll R.J., Spiegelman C.H. & Sacks, J. (1988) A quick and easy multiple-use calibration-curve procedure. Technometrics. 30(2), 137-141.

Chatfield, C. (1993) "Calculating Interval Forecasts," Journal of Business and Economic Statistics, 11 121–135.

Chatterjee S., Hadi A.S. & Price B. (2000) Regression analysis by example, third edition. 3rd edn. John Wiley & Sons, New York.

Cleveland, W.S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. The American Statistician, 35, 54.

Collett D. (2003). Modelling Binary Data, Second Edition. Texts in Statistical Sciences, Chapman & Hall/CRC, Boca Raton.

Conger, A.J. (1980), Integration and generalisation of Kappas for multiple raters, Psychological Bulletin, 88, 322-328.

Davidson A.C. & Hinkley D.V. (1997) Standard deviation from the multivariate delta method. Booststrad Methods and Their Application, 45-46.

De Leeuw J. & Van Rijckevorsel J. (1980) HOMALS and PRINCALS - Some generalizations of principal components analysis. in E. Diday and Coll., editors. Data Analysis and Informatics II. Elsevier Science Publisher, North Holland, Amsterdam, 231-241.

Dormann C. F., McPherson J. M., Araújo M. B., Bivand R., Bolliger J., Carl G., Davies R. G., Hirzel A., Jetz W., Kissling W. D., Kühn I., Ohlemüller R., Peres-Neto R. P., Reineking B., Schröder B., Schurr F. M. & Wilson R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review, Ecography, **30**, 609-628.

Dray S & Dufour A-B (2007), The ade4 Package: Implementing the Duality Diagram for Ecologists, Journal of Statistical Software, 22, 1-20.

Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman & Hall.

Elliott J.M. (1994) Quantitative ecology and the brown trout. Oxford University Press, Oxford, New York and Tokyo.

Eyre M., Foster G. & Young A. (1993) Relationships between water-beetle distributions and climatic change. Archiv für Hydrobiologie, 127, 437-450

Faraway, J. J. (2006) Extending the linear model wit R: generalized linear, mixed effects and nonparametric regression models, Chapman & Hall/CRC, New-York.

Fletcher R. H. & Fletcher S. W. (2005) Clinical Epidemiology: The Essentials, fourth Edition, Lippincott Williams & Wilkins, Philadelphia.

Florkowski C M (2008) Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests, Clin Biochem Rev, 29 Suppl (i), 83-87.

Florwski C. (2008) Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests. The Clinical Biochemist Reviews, 29(Supp i), 83–S87.

Fox, J. (2002) Applied Regression, Linear Models, and Related Methods. Sage.

Freeman E. A & Moisen G. G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa, ecological modelling 217, 48–58.

Freeman M.C., Bowen Z.H., Bovee K.D. & Irwin E.R. (2001) Flow and habitat effects on juvenile fish abundance in natural and altered flow regimes. Ecological Applications, 11, 179-190.

Freeman S.N., Pomeroy D.E. & Tushabe H. (2003) On the use of timed species counts to estimate avian abundance indices in species-rich communities. African Journal of Ecology, 41, 337-348.

Freeman E. A. & Moisen G. G. (2007) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling,217, 48–58.

Greene (2002) Econometric Analysis, Prentice Hall,US, 5e International Ed, 1026 pp.

Guisan A., Edwards J., Thomas C. & Hastie T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling, 157, 89-100

Hann, J., Kamber, M. (2000). Data Mining: Concepts and Techniques. Morgan Kaufman Publishers.

Hardin, J.W. and Hilbe, J.M. (2007). Generalized Linear Models and Extensions (2nd Edition). Stata Press.

Hartigan, J. A. & Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics, 28, 100-108.

Hastie, T., Tibshirani, R. & Friedman J. (2001). The Elements of Statistical Learning: Data mining, inference and prediction, Springer-Verlag, New-York, 536 pages.

Hellmann J.J. & Fowler G.W. (1999) Bias, precision and accuracy of four measures of species richness. Ecological Applications, 9(3), 824-34.

Hughes R.M., Kaufmann P.R., Herlihy A.T., Kincaid T.M., Reynolds L. & Larsen D.P. (1998) A process for developing and evaluating indices of fish assemblage integrity. Canadian Journal of Fisheries and Aquatic Sciences, 55, 1618-31.

Hughes, R.M., Kaufmann, P.R., Herlihy, A.T., Instelmann, S.S., Corbett, S.C., Arbogast, M.C. & Hjort, R.C. (2002) Electrofishing distance needed to estimated fish species richness in raftable Oregon rivers. North American Journal of Fisheries Management, **22**, 1229-40.

Hughes, R.M. & Herlihy, A.T. (2007) Electrofishing distance needed to estimate consistent index of biotic integrity (IBI) score in raftable Oregon rivers. Transactions of the American Fisheries Society, **136**, 135-41.

Hill M. O. & Smith A. J. E. (1976) Principal component analysis of taxonomic data with multi-state discrete characters. Taxon, 25, 249-255.

Hosmer D.W. & Lemeshow S. (2000) Applied Logistic Regression 2nd Edition. John Wiley & Sons, Inc., New York.

Hurst T.P. (2007) Causes and consequences of winter mortality in fishes. Journal of Fish Biology, 71, 315-345.

Illies J. (1978) Limnofauna Europaea. A check-list of the animal inhabiting European inland waters, with an account of their distribution and ecology. 2nd ed., Gustav Fischer Verlag, Stuttgart.

Jenkins Jr T.M., Diehl S., Kratz K.W. & Cooper S.D. (1999) Effects of population density on individual growth of brown trout in streams. Ecology, 80, 941-956.

Jongman R.H.G., Ter Braak C.J.F. & van Tongeren O.F.R. (1995) Data analysis in community and landscape ecology, 2nd edition. Cambridge University Press, Cambridge.

Jowett I.G. & Richardson J. (1989) Effects of a severe flood on istream habitat and trout populations in seven New Zealand rivers. New Zealand Journal of Marine & Freshwater Research, 23, 11-17.

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

Karr J., Fausch K.D., Angermeier P.L., Yant P.R. & Schlosser I.J. (1986) Assessing biological integrity in running waters. A method and its rationale. Illinois Nat. Hist. Surv. Spec. Publ., 5, 23.

L'abée-Lund J.H. & Saegrov H. (1991) Resource use, growth and effects of stocking in alpine brown trout, Salmo trutta L. Aquaculture and Fisheries Management, 22, 519-526.

Legendre P., Dale M. R. T., Fortin, M.-J., Gurevitch J., Hohn M. & Myers D. (2002) The consequences of spatial structure for the design and analysis of ecological study, Ecography, **25**, 601-615.

Legendre P. & Legendre L. (1998) Numerical Ecology. Second English Edition. Elsevier.

Lobon-Cervia J. & Rincon P.A. (2004) Environmental determinants of recruitment and their influence on the population dynamics of stream-living brown trout Salmo trutta. Oikos, 105, 641-646.

Lobon-Cervia J. (2003) Spatiotemporal dynamics of brown trout production in a Cantabrian stream: Effects of density and habitat quality. Transactions of the American Fisheries Society, 132, 621-637.

Magurran, A.E. (1988) Ecological diversity and its measurement Croom Helm Limited, London.

Marazzi, A. (1993) Algorithms, routines, and S functions for robust statistics. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Martens H. & Naes T. (1989) Multivariate calibration. Chichester: Wiley. 438 pages

McCullagh P. & Nelder J. A. (1989) Generalized Linear Models, second edition edn. Chapman & Hall/CRC, London.

McCulloch C.E. & Searle S.R. (2001) Generalized, Linear, and Mixed Models, second ed. Wiley, New York.

MacKinnon JG, White H (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. Journal of Econometrics, 29, 305–325.

Mease D. & Wyner A. (2008) Evidence Contrary to the Statistical View of Boosting. Journal of Machine Learning Research, 9,131--156, 2008.

Melcher A., Schmutz S., Haidvogl G. & Moder K. (2007). Spatially based methods to assess the ecological status of European fish assemblage types. Fisheries Management and Ecology 14, 453–463.

Milner N.J., Elliott J.M., Armstrong J.D., Gardiner R., Welton J.S. & Ladle M. (2003) The natural control of salmon and trout populations in streams. Fisheries Research, 62, 111-125.

Nelder J. A. & Wedderburn R. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135, 370-384.

Oberdorff T., Pont D., Huggueny B. & Chessel D. (2001) A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. Freshwater Biology, 46, 399-415

Oberdorff T., Pont D., Hugueny B. & Porcher J.-P. (2002) Development and validation of a fish-based index (FBI) for the assessment of 'river health' in France. Freshwater Biology, 47, 1720-1734

Oksanen J. & Minchin P.R. (2002) Continuum theory revisited: what shape are species responses along ecological gradients? Ecological Modelling, 157, 119-129

Pinherio J. C. & Bates D. M. (2000) Mixed-effect models in S and S-plus, Springer Verlag, New York.

Pont D., Hugueny B. & 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.

Pont D., Hugueny B., Beier U., Goffaux D., Melcher A., Noble R., Rogers C., Roset N. & 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., Beier, U., Goffaux, D., Melcher, A., Noble, R., Rogers, C., Roset, N. & 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(1), 70-80.

Potts, J. & Elith, J. (2006) Comparing species' abundance models. Ecological Modelling 199, 153-163.

Quataert P., Breine J. & 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,

http://fame.boku.ac.at/downloads/D16_17_MethodComparison.pdf

Quataert, P., Breine, J. & 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(6), 465-72.

R Development Core Team (2008) R: A language and environment for statistical computing. In R Foundation for Statistical Computing. Vienna, Austria.

Reyjol, Y., Hugueny, B., Pont, D., Bianco, P.G., Beier, U., Caiola, N., Casals, F., Cowx, I., Economou, A., Ferreira, T., Haidvogl, G., Noble, R., de Sostoa, A., Vigneron, T. & Virbickas, T. (2007) Patterns in species richness and endemism of European freshwater fish. Global Ecology and Biogeography, 16(1), 65-75.

Reynolds, L., Herlihy, A.T., Kaufmann, P.R., Gregory, S.V. & Hughes, R.M. (2003) Electrofishing efoort requirements for assessing species richness and biotic integrity in Western Oregon streams. North American Journal of Fisheries Management, **23**, 450-61.

Saporta, G. (2006) Probabilités, analyses de données et statistique, Editions TECHNIP, Paris.

Schabenberger, O. & Gotway, C.A. (2005) Statistical methods: for spatial data analysis Chapman & Hall/CRC, Florida.

Schmutz, S., Cowx, I.G., Haidvogl, G. & Pont, D. (2007) Fish-based methods for assessing European running waters: a synthesis. Fisheries Management and Ecology, 14(6), 369-80.

Snee R.D. (1977) Validation of regression models: methods and examples. Technometrics, 19, 415-428.

Spitalnic S. (2004a) Test Properties I: Sensitivity, Specificity, and Predictive Values, Hospital Physician, 40 (9), 27-31.

Spitalnic S. (2004b) Test Properties 2: Likelihood Ratios, Bayes' Formula, and Receiver Operating Characteristic Curves, Hospital Physician, 40 (10), 53-58.

Statzner B. & Moss B. (2004) Linking ecological function, biodiversity and habitat: a mini-review focusing on older ecological literature. Basic and Applied Ecology, 5, 97-106

Tenenhaus M. (1998) La Régression PLS. Théorie et pratique. Technip, Paris.

Tenenhaus, M. & Young, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. Psychometrika, 50, 1, 91-119.

van Tongeren O. F. R. (1995) Data analysis or simulation model: a critical evaluation of some methods. Ecological Modelling,78 (1-2), 51-60.

Venables W. N. & Ripley B. D. (1999) Modern Apllied Statistics with S-plus, Third edition, Statistics and computing, Springer-Verlag, New York.

Venables W.N. & Ripley B.D. (2002) Modern Applied Statistics with S 4th Edition. Springer, New York.

Vollestad L.A., Olsen E.M. & Forseth T. (2002) Growth-rate variation in brown trout in small neighbouring streams: Evidence for density-dependence? Journal of Fish Biology, 61, 1513-1527.

Young D., Benaglia T., Chauveau D., Elmore R., Hettmansperger T., Hunter D., Thomas H. & Xuan F. (2008) mixtools: Tools for analyzing finite mixture models, pp. A collection of R functions for analyzing finite mixture models.

8 Annexes

8.1 Biological variables descriptions (guilds and traits)

Num	Trait	Modality	Abbrev.	Definition EFI+	Num	Trait	Modality	Abbrev.	Definition EFI+
1	Native	N/A	N/A	If a fishspeci native (N) or ali within a catchment or not.	36	Salinity	freshwater	FRE	Fish that exclusive in freshwater habi
2	Water tolerance gene	tolerant	TOL	In general toleran usual (national) quality parameter	37	Salinity	estuary - brackish	ESTU	Fish that spen periods both freshwater and br habitats.
3	Water tolerance gene	intermediate	IM	In general tole intolerant to (national) water parameters.	38	Salinity	marine	MAR	Fish that spen periods mainly marine habitats
4	Water tolerance gene	intolerant	INTOL	In general intole usual (national) quality parameter	39	Salinity	anadrom - catadr	ANCA	Fish that spen periods in fresh brackish and habitats.(long mig species)
5	Water tolerance O2	tolerant	O2TOL	Tolerant to low C concentration (C mg/l or less	40	Reproductive gui	lithopelagophilic	LIPE	Rock and spawners with p free embryos.
6	Water tolerance O3	intermediate	O2IM	Relatively tolera low C concentration (O2	41	Reproductive gui	lithophilic	LITH	Fish spawn exclu on gravel, stones, rubbles pebbles. hatchlin photophobic.
7	Water tolerance O4	intolerant	O2INTOL	Tolerant to low C concentration more than 6 m water	42	Reproductive gui	ostracophilic	OSTRA	Spawning in she bivalve molluscs.
8	Water tolerance TOX	tolerant	TOXTOL	In general toler toxic contaminati	43	Reproductive gui	pelagophilic	PELA	Fish spawn int pelagic zone.
9	Water tolerance TOX	intermediate	TOXIM	In general tole intolerant to contamination.	44	Reproductive gui	phytophilic	PHYT	Fish deposit eg clear water habit submerged plants
10	Water tolerance TOX	intolerant	TOXINTOL	In general intole toxic contaminati	45	Reproductive gui	phyto-lithophilic	PHLI	Fish deposit eg clear water habit submerged plants other submerged such as logs, and rocks. larva photophobic.
11	Acid tolerance	tolerant	ATOL	Tolerant acidification.	46	Reproductive gui	polyphilic	POLY	Non-specialised spawners.
12	Acid tolerance	intermediate	AIM	Tolerant / intole acidification.	47	Reproductive gui	psamnmophilic	PSAM	Fish spawn on ro grass above bottom or on the itself.
13	Acid tolerance	intolerant	AINTOL	Intolerant acidification.	48	Reproductive gui	speleophilic	SPEL	Fish spawn in inte spaces, crevice caves.
14	Temperature tolerance	eurythermal	EUTHER	Fish capable withstanding a range of tempera	49	Reproductive gui	viviparous	VIVI	Live bearers.

15	Temperature tolerance	stenothermal	STTHER	Fish able to witl only a r temperature rang	50	Reproductive gui	ariadnophilic	ARIAD	Specialised building fish tha exhibit some fo parental care.
16	Habitat degra tolerance	tolerant	HTOL	Tolerant to mentioned metric	**51**	**Habitat spa preferences**	**rheopar**	**RHPAR**	**Preference to s in running waters**
17	Habitat degra tolerance	intermediate	HIM	Tolerant / intoler the mentioned m	**52**	**Habitat spa preferences**	**euryopar**	**EUPAR**	**No clear spa habitat preferenc**
18	Habitat degra tolerance	intolerant	HINTOL	Intolerant to mentioned metric	**53**	**Habitat spa preferences**	**limnopar**	**LIPAR**	**Preference to s in stagnant wate**
19	Habitat	rheophilic	RH	Degree of rhe (habitat). Fish pr live in a habita high flow con and clear water.	54	Reproductive behaviour	single	SIN	spawing event a one time in the se
20	Habitat	eurytopic	EURY	Degree of rhe (habitat). Fish exhibit a tolerance of conditions, alt generally considered to rheophilic.	55	Reproductive behaviour	fractional	FR	Fractional spawn spawning events repeated in a sea different compone the population spa different times.
21	Habitat	limnophilic	LIMNO	Degree of rhe (habitat). Fish pr live, feed reproduce in a l with slow flowi stagnant conditio	56	Reproductive behaviour	protraced	PRO	Protracted spaw spawning occurs long period durir potential season.
22	Feeding Habita	water column	WC	Species that liv feed in the column. They u do not go to the b to search for food	**57**	**Parental care**	**no protection**	**NOP**	**No protection of or larvae (pa care)**
23	Feeding Habita	benthic	B	Fish prefer to liv the bottom from they take food. usually do not go surface for f purpose.	**58**	**Parental care**	**protection**	**PROT**	**Protection of and/or larvae, including some of parental care or eggs hided i manner)**
24	Adult trophic gu	detritivorous	DETR	Adult diet consi high proportio detritus, the dig tract is unspeci Combinations other modalities possible, if you ex	59	Length	numeric	LENG	Maximum fish [mm]. From Fishb others; no exce cases.

25	Adult trophic gu	herbivorous	HERB	Diet of adult consists of more 75% plant materi have terminal subterminal mou bony slashing ja clipping and t aquatic vege Often the dig tract is as lo longer than the length of the indi Combinations other modalities possible, if you e	60	Length relation a	numeric	LWa	Length relationship; Fishbase or Mathematical f for the weight of in terms of its When only o known, the formu be used to det the other. Ty given as w=a*L1 weight in gram length in centimet
26	Adult trophic gu	insectivorous	INSV	Adult diet consi more than insects. indiv have terminal supraterminal take aerial, drift swimming insect invertebrates. largest and diverse trophic Combinations other modalities possible, if you e	61	Length relation b	numeric	LWb	Length relationship; Fishbase or Mathematical f for the weight of in terms of its When only o known, the formu be used to det the other. Ty given as w=a*L^ weight in gram length in centimet
27	Adult trophic gu	omnivorous	OMNI	Adult consists o than 25% material and mor 25% animal ma Generalists. Combinations other modalities possible, if descri	62	Shape factor	numeric	SHAF	Minor ratio of larg smallest part body (in genera body length divi maximum body de
28	Adult trophic gu	parasitic	PARA	Fish that exhi parasitic feeding Combinations other modalities possible, if you e	63	Swimming factor	numeric	SWF	Defined as the minimum depth caudal peduncle maximum cauda depth. Fish hav small ratio are c of strong swi (thunniform f After Poff and 1995 (initially Scarnecchia 198

29	Adult trophic gu	piscivorous	PISC	Other fish rep more than 75 adults diet. indiv have a wide aperture with n like teeth and a jaw with margina palatal bones. pursue a pre stalking, ch ambushing or ly wait app Combinations other modalities possible, if you ex	64	Longevity	numeric	LONG	Maximum longevi species [years] exeptional cases.
30	Adult trophic gu	planktivorous	PLAN	Adult diet consi more than zooplankton phytoplankton. having fine gill- and elon pharyngal teeth have no stomac an elon undifferentiated intestine. Combin with other mod are possible, i explain.	65	Fecundity	numeric	FEC	Maximum numb oocytes, no exce cases.
31	Trohic Index		TROPHIC	See Fishbase.org where you can several values species….	66	Relative fecundit	numeric	RFEC	Maximum numb oocytes per gramm, no exce cases.
32	Migration guild	resident	RESID	Species that only within a particula segment.	67	Egg diameter	numeric	EGG	Average egg dia [mm].
33	Migration guild	potamodrom	POTAD	Species n between river zo more than 5-1 (more than a segment)	68	Age at maturity	numeric	MATU	Average age at m of female fish [yea
34	Migration guild	long catadrom	LONG-LMC	Refers to fish tha lived their early fresh water - fe and growing - a maturity migrate rivers to spawn sea.	69	Incubation	numeric	INCU	Average incubatio / period of eggs for the co temperature
35	Migration guild	long anadrom	LONG-LMA	Refers to fish th as older juvenile sub-adults in th but at maturity m up rivers to spaw					

8.2 Computation of the Habitat Index, water alteration and Channel-Crossection variables.

8.2.1 Habitat Index

The habitat index is based on the aggregation of the value of the three pressure descriptors describing direct habitat alteration at the local scale:

- Instream.habitat
- Riparian.vegetation
- Embankment

These 3 variables have to be fulfilled in the dataset to compute the Habitat.index (no missing value « NA » allowed)

The modalities are first replaced by numerical values.

Instream.habitat	No	Intermediate	High
	1	2	3

Riparian.vegetati	No	Slight	Intermediate	High
	1	1.5	2	3

Embankment	No	Local	Continuous permeable	Continuous permeability
	1	1.5	2	3

The numerical values are summed up:
A = Instream.habitat + Riparian.vegetation + Embankment

The habitat index is defined as follow:

Habitat Index Class	No	Slight	Medium	High
A	3	]3 – 4]	]4 – 6]	]6 – 9]

8.2.2 Water alteration Index

The water alteration index is based on the aggregation of the value of the three pressure descriptors describing water alteration at the local scale:

- Eutrophication
- Organic pollution
- Organic pollution

These 3 variables have to be fulfilled in the dataset to compute the Water.alteration.index (no missing value « NA » allowed).

Eutrophication	No	Low	Intermediate	Extreme
	1	1.5	2	3

Organic.pollution	No	Weak	Strong
	1	1.5	3

Organic.siltation	No	Yes
	1	3

The numerical values are summed up:
B = Eutrophication + Organic.siltation + Organic pollution

The water alteration index is defined as follow:

Water alteration Class	No	Medium	High
A	3	]3 – 5[	]5 – 9]

8.2.3 Channel Cross Definition

The two pressure “Channelisation” and “Cross.section” are highly correlated. They can be combined as follow

```
Channel.cross modalities:

"No"
(Cross.section = "No" and Channelisation="No")

"Strong"
(Cross.section = "Techn.U-profile" and Channelisation="Intermediate")
or
(Cross.section = "Intermediate" and Channelisation="Straightened"

"Extreme"
(Cross.section = "Techn.U-profile" and Channelisation="Straightened")

"Moderate"
Others combinations

table(a$Channelisation,a$Cross.sec)
               A.No B.Intermediate C.Techn.U-profile
          A.No 6168            804               131
B.Intermediate  369            720               337
C.Straightened  100            272               970

table(a$Channel.cross)
 A.No B.Slight C.Moderate D.Strong E.Extreme
 6168     1173        951      609       970
```

8.3 Environmental variable distributions

A first classification of the 10063 EFI+ sites gave the following groups: 459 Reference sites (not impacted), 1265 Calibration sites (little impacted) and 8339 Impacted sites. Impacted should be understood in its broad meaning here, say under anthropic pressure and/or really impacted. Models are to be fitted using a subset of this calibration dataset.

This chapter is a presentation of environmental data in the EFI+ database, with a focus on the comparison of the environmental variables within reference, calibration and impacted sites, as well as on patterns within calibration sites.

First, the geographical locations of reference, calibration and impacted sites, as well as the distribution of environmental variables within each subset will be shown. Then, the geographical distribution of these variables within the calibration sites only will be developed, before trying to bring out a few patterns.

8.3.1 Reference, Calibration and Impacted Sites

8.3.1.1 Locations of reference, calibration and impacted sites

The following maps display the location of respectively calibration, impacted – as defined above – and reference sites.

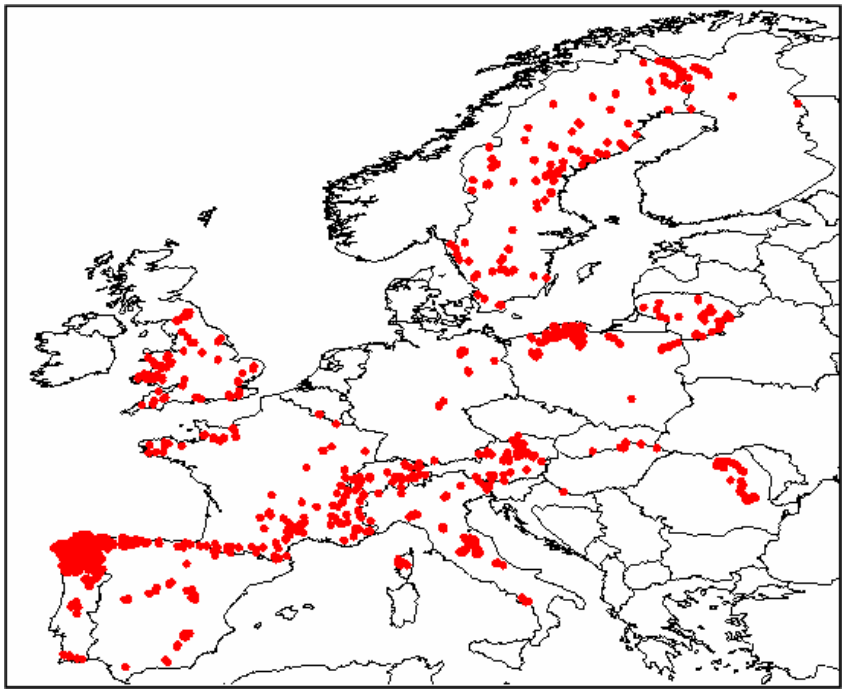


Figure 1a. Locations of Calibration sites

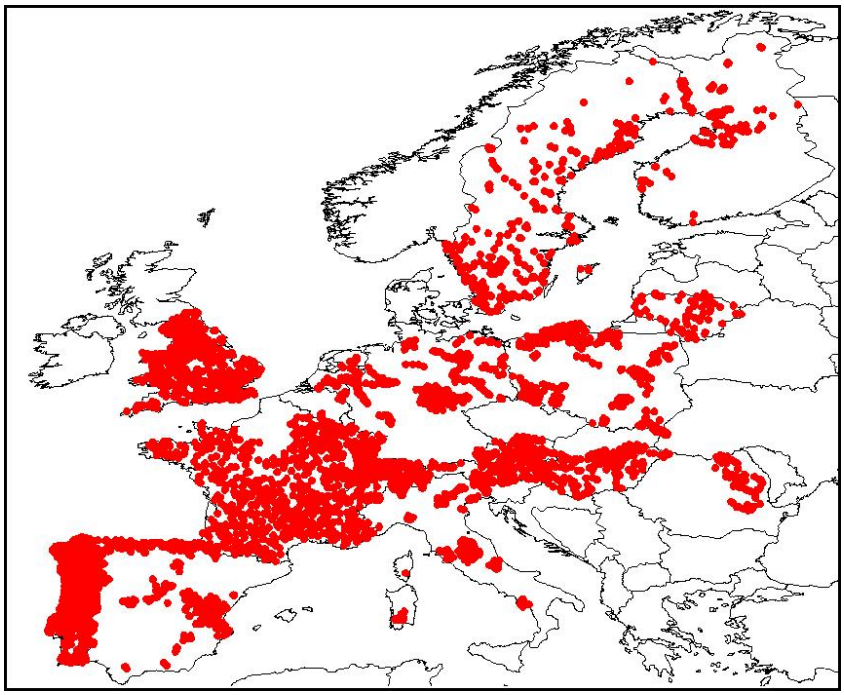


Figure 1b. Locations of Impacted sites

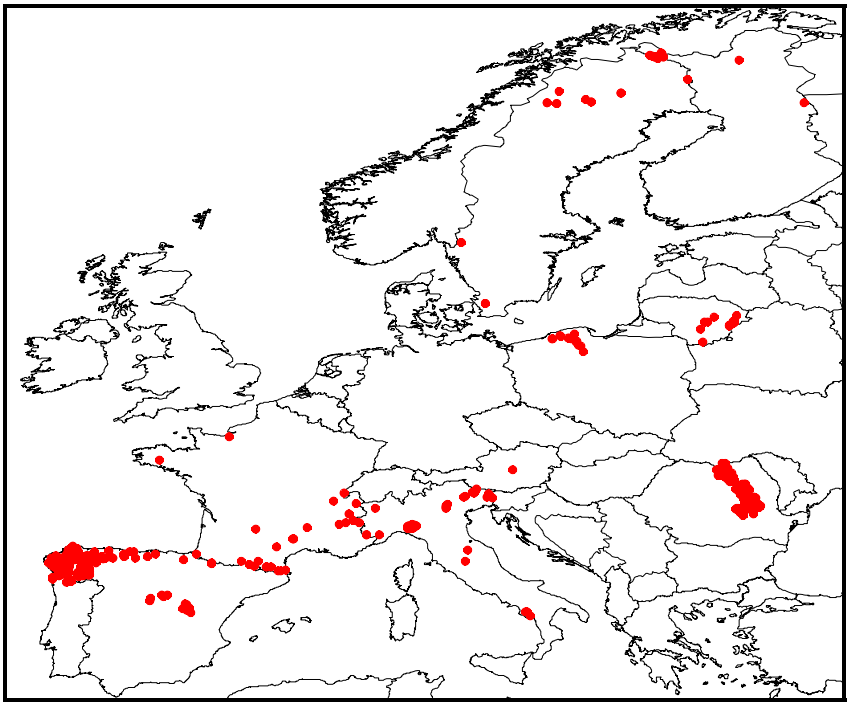


Figure 1c. Locations of Reference sites

The next series of maps shows clusters of sites with a dot size for each cluster function of the number of sites present in the cluster area.

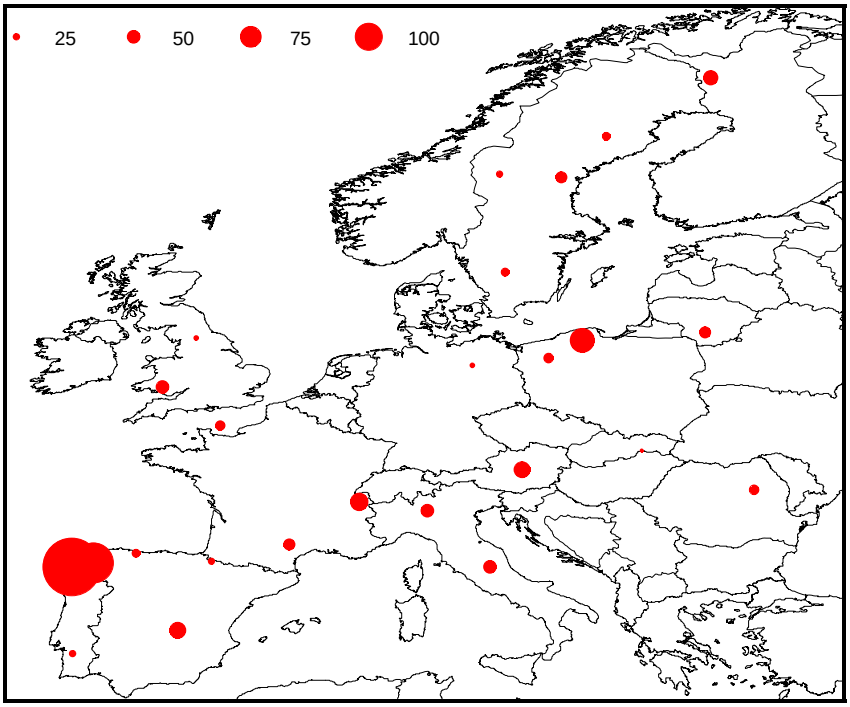


Figure 2a. ***Locations and Number of Calibration sites***

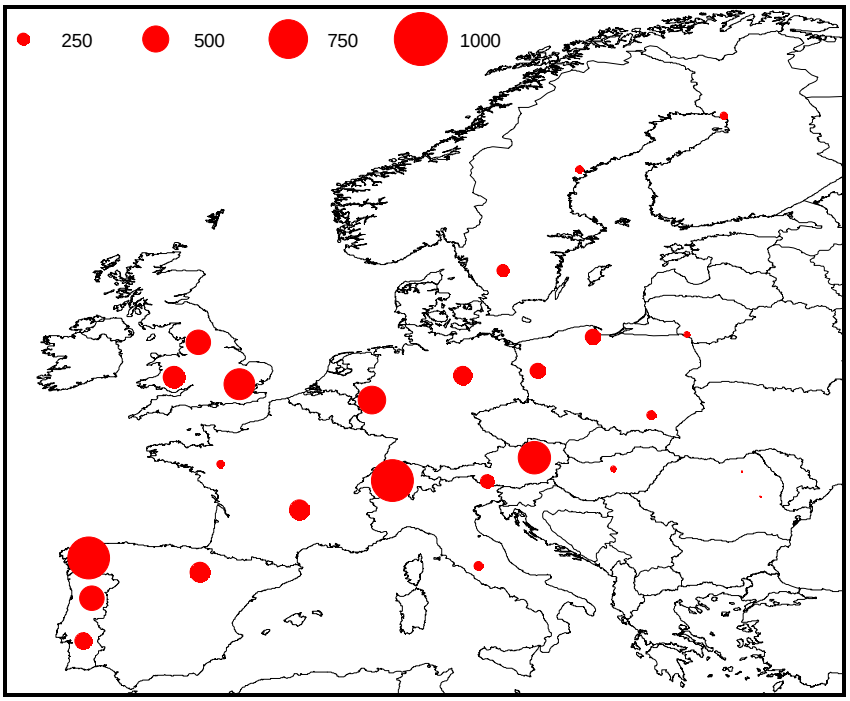


Figure 2b. ***Locations and Number of Impacted sites***

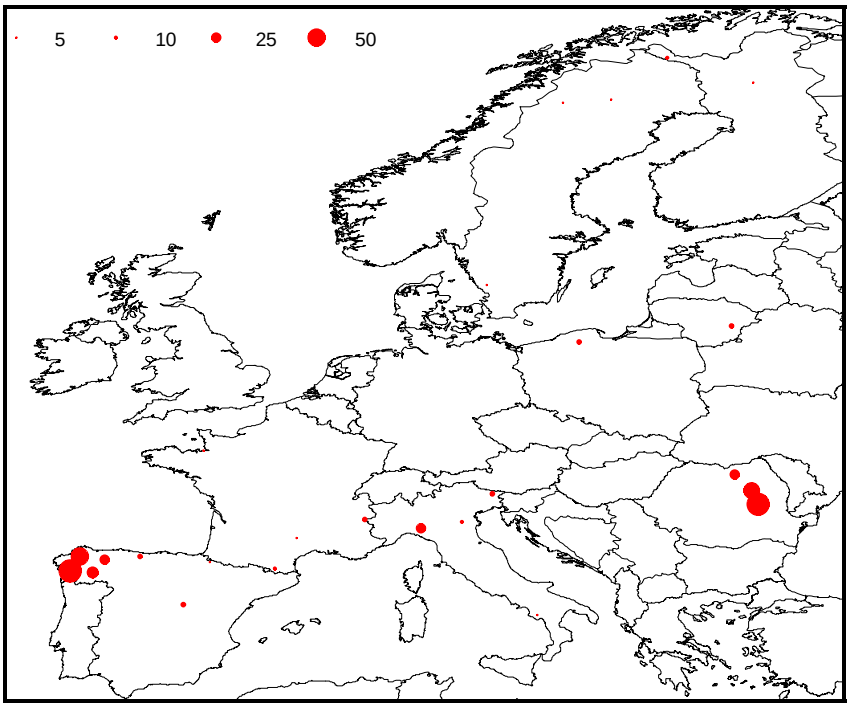


Figure 2c. ***Locations and Number of Reference sites***

These few maps show the uneven distribution of sites, especially as regards calibration and even more as regards reference sites. Reference sites are highly concentrated in Galice and the Asturies, as well as in Romania. Calibration sites are mainly concentrated in Galice and the Asturies, and to a lesser extent in the North of Poland. But, they spread out over almost all regions. Impacted sites are found everywhere with a relatively high concentration in Galice and the Asturies too, but also in the Alps and in Great Britain.

8.3.1.2 Distributions of environment variables

The next bar charts display the distributions of qualitative environmental variables, which are to be included in models, as well as the fishing method.

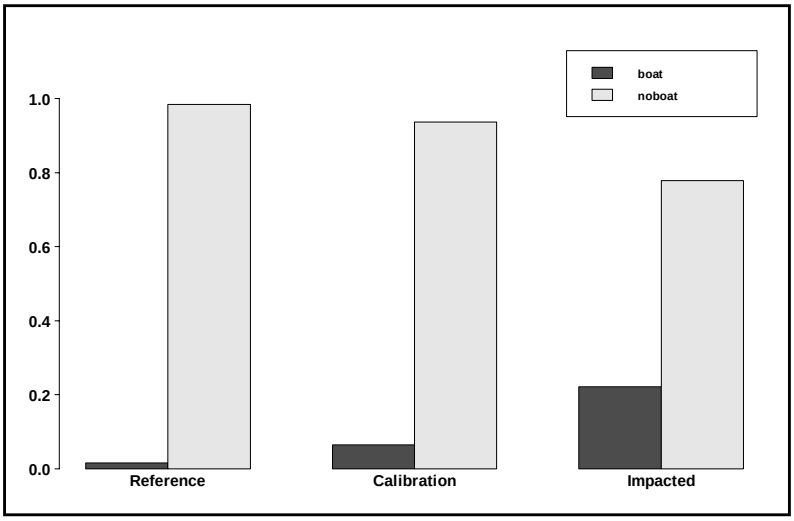


Figure 3a. ***Distribution of sites according to fishing method***

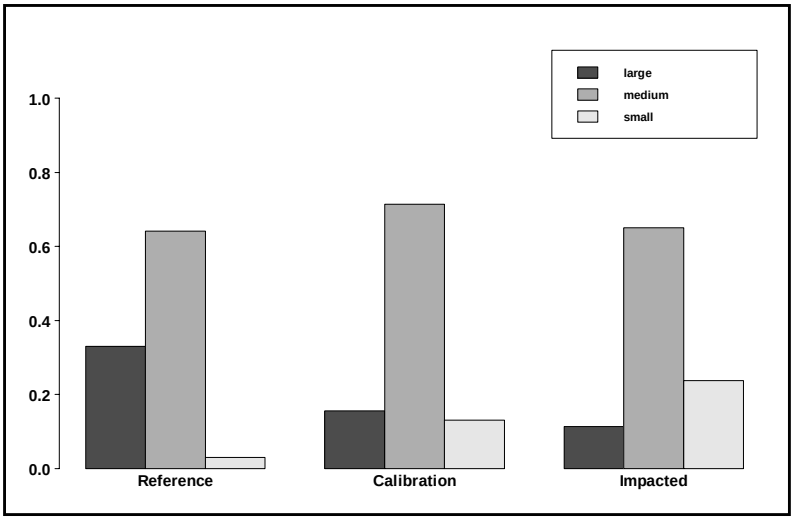


Figure 3b. ***Distribution of sites according to natural sediment size***

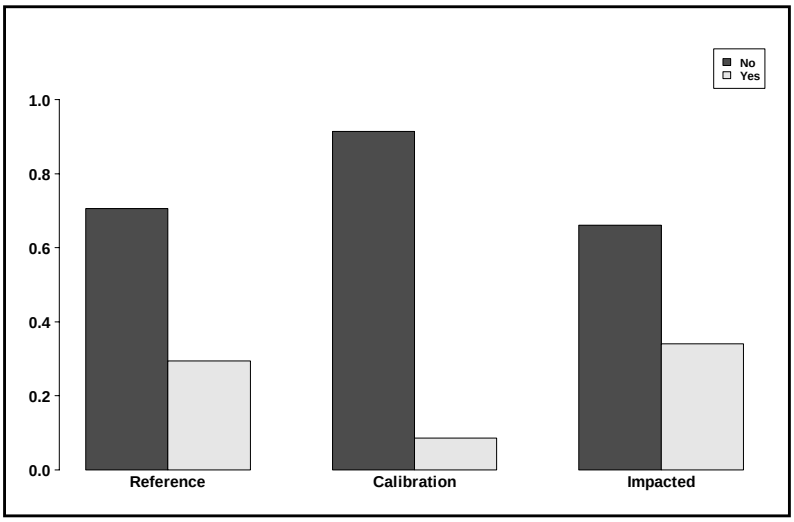


Figure 3c. ***Distribution of sites according to presence/absence of a floodplain***

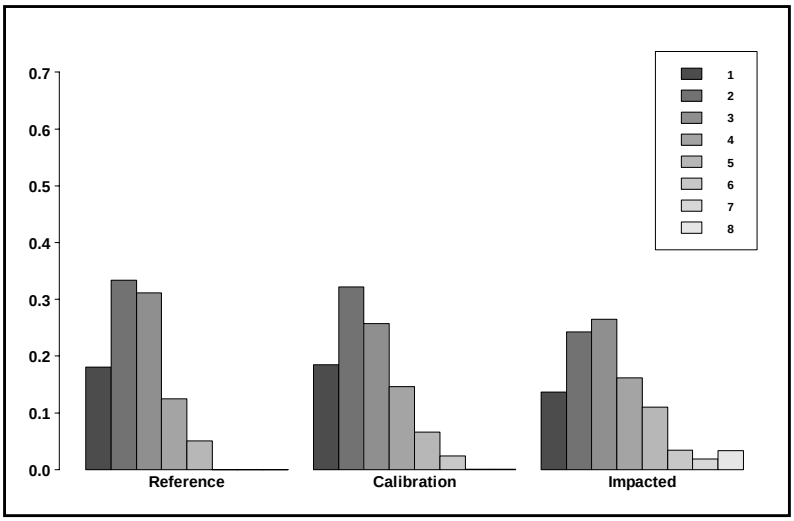


Figure 3d. ***Distribution of sites according to Strahler Index***

Along the gradient Reference-Calibration-Impacted, the following trends can be noticed:

1) An increasing proportion of boat fishing method;
2) A decreasing proportion of large sediment, with a corresponding increase in the proportion of small sediment;
3) And an increase in the number of higher Strahler orders.

However, as regards the presence of floodplain sites, a smaller proportion in calibration in comparison with both reference and impacted sites can be noticed.

The next graph is the boxes and whiskers charts of quantitative environmental variables for each subset of sites.

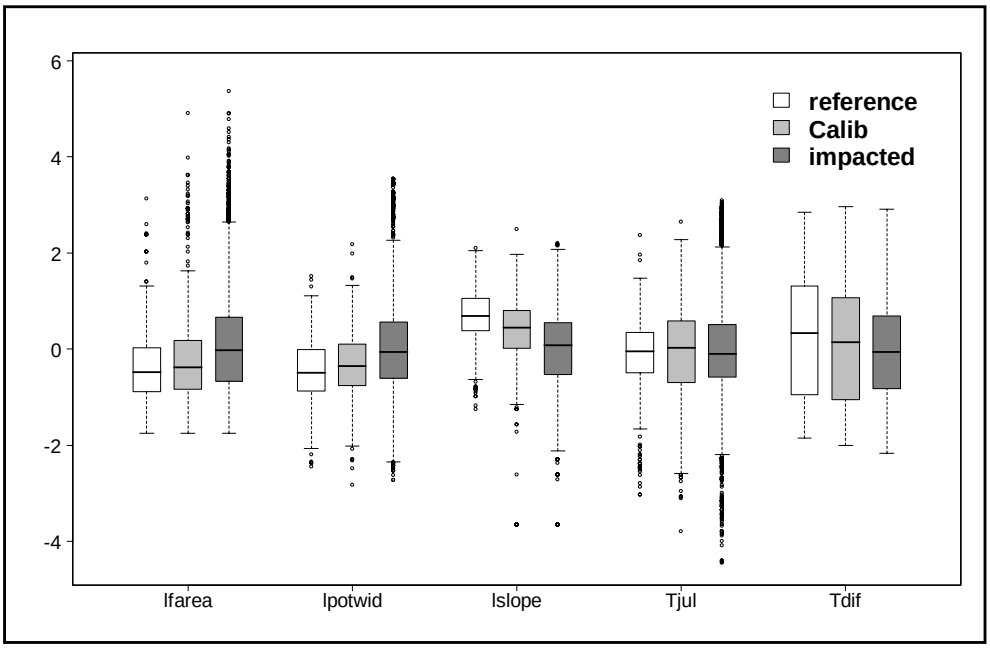


Figure 4. ***Box and whiskers charts of a set of continuous variables (scaled) in each subset: Log-transformed fished area (lfarea), log-transformed potential width (lpotwid), log-transformed slope (lslope), July temperature (Tjul), temperature of July minus temperature of January (Tdif).***

Similar trends can be noticed, say increasing for fished area and potential width; decreasing for river slope and, but to a lesser extent, for the temperature range between January and July. As regards July temperatures, there is no obvious trend.

Which follows is based on a factorial analysis of an extended set of environmental variables when considering all sites using a Multiple Component Analysis (MCA). Environmental variables, which could be included in models, as well as fishing variables – because closely linked to the environment, and since they are to be included in models too, both qualitative and quantitative, were included in MCA. Quantitative variables were previously grouped into classes, after transformation when necessary.

The next graphs are the projection of sites on the first factorial plan, grouped according to variable modalities.

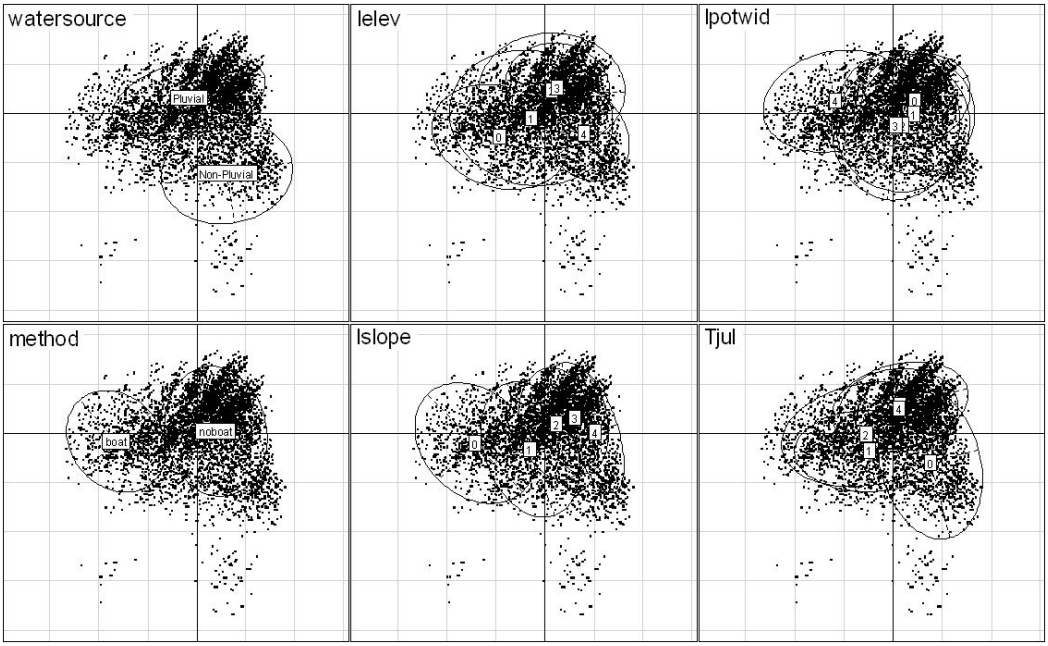


Figure 5a. ***Projection of sites on the first factorial plan grouped by the modalities of variables included in MCA: water source type (watersource), log-transformed elevation (lelev), log-transformed potential width (lpotwid), fishing method (method), log-transformed slope (lslope), July temperature (Tjul).***

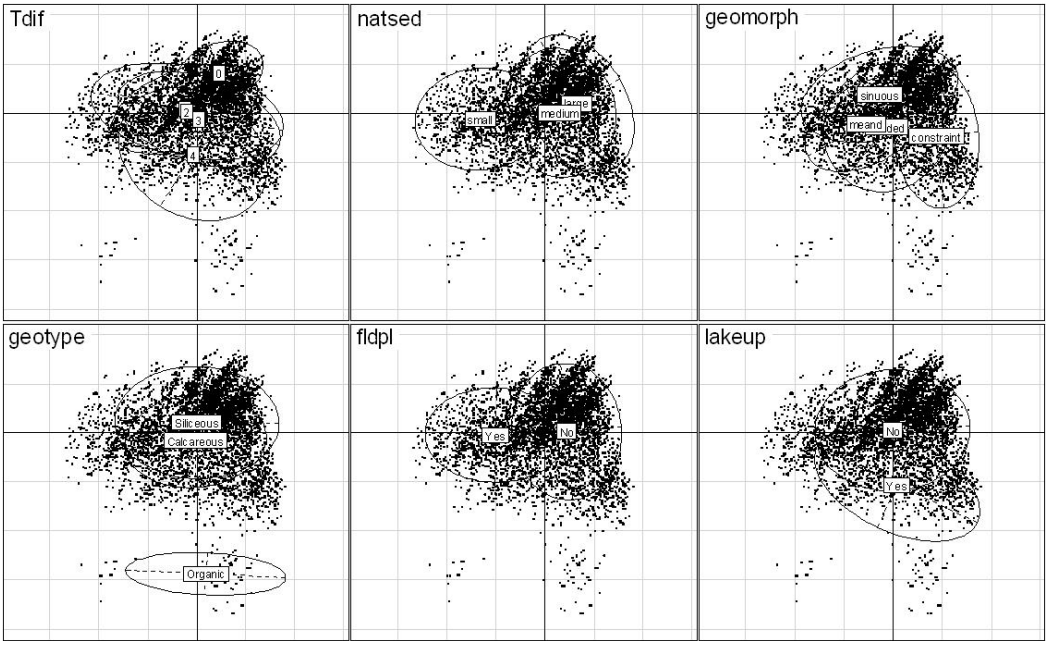


Figure 5b. ***Projection of sites on the first factorial plan grouped by the modalities of variables included in MCA: temperature of July minus temperature of January (Tdif), natural sediment (natsed), geomorphological type (geomorph), geological type (geotype), floodplain site (fldpl) and lake upstream (lakeup) indices.***

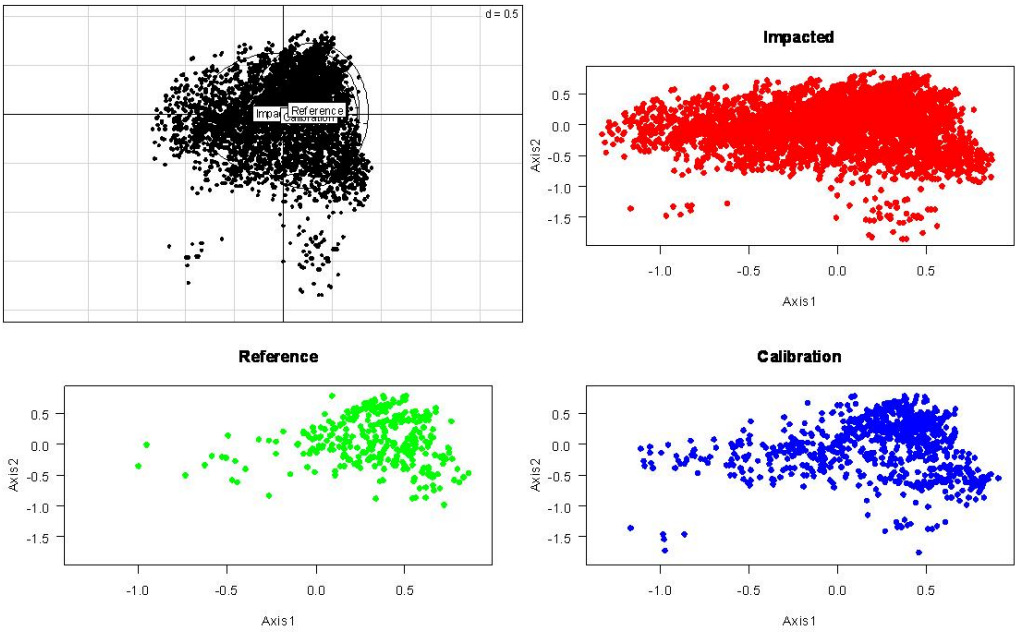


Figure 5c. ***Projection of sites on the first factorial plan grouped into site type: reference, calibration or impacted and projections of each group separately.***

The first axis clearly opposes boat and no-boat fishing method, larger potential widths and others, floodplain sites and others, smaller natural sediments and others. Most of all, it is highly correlated with increasing slopes and, but much less obviously, with increasing elevations.

The second axis opposes organic geo-morphological type and other types, as well as, but to a lesser extent, non-pluvial and pluvial water source types. In addition, the presence of upstream lake seems to be also associated with this axis, as well as the temperature range between January and July (Tdif).

Besides, variables such as elevation and July temperature correlate, but only partly, with these first two axes.

Last graphs show that most of not or little impacted sites are projected in the top right corner of the first factorial plan. However, whereas reference sites are clustered in this corner with only a few sites in the top left side, calibration sites spread almost everywhere, and thus should be more representative of the whole environmental variability.

8.3.2 Geographical distribution of environment variables within the calibration dataset

In the following part, a series of maps are displayed to appreciate geographical distribution of the main environmental variables, both quantitative and qualitative, to be included in models.

To represent the geographical distribution of quantitative variables, sites were first clustered according to their locations. Three statistics are shown for each cluster: the 5% and 95% quantiles, as well as the arithmetic mean.

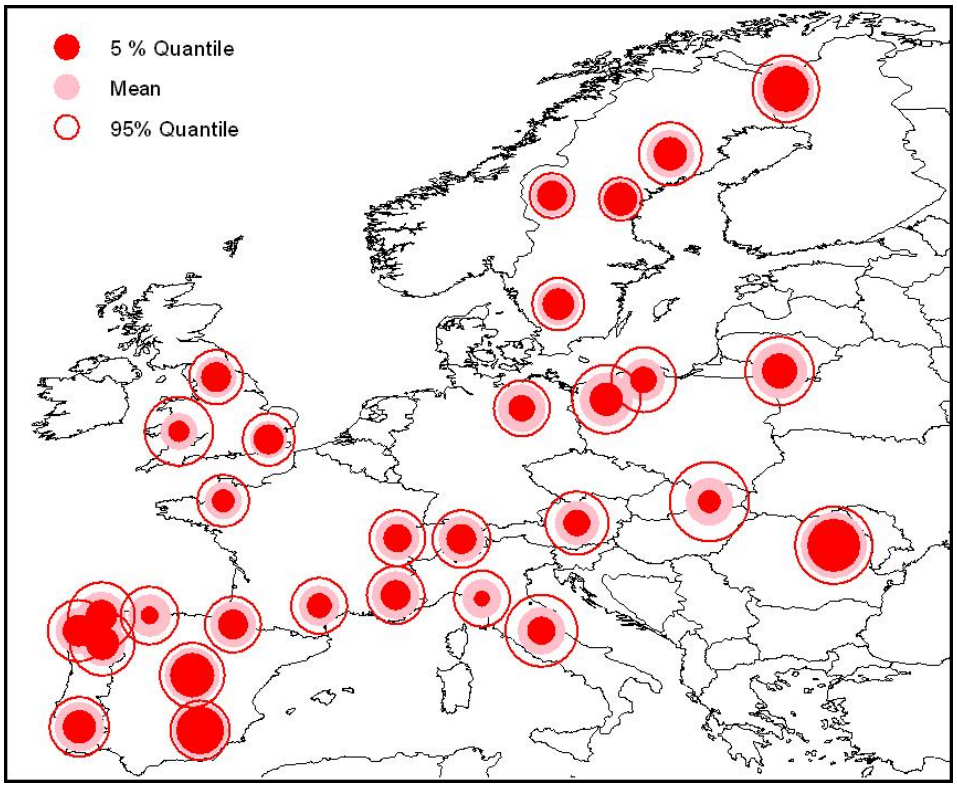


Figure 6a. ***Map for scaled log-transformed potential width.***

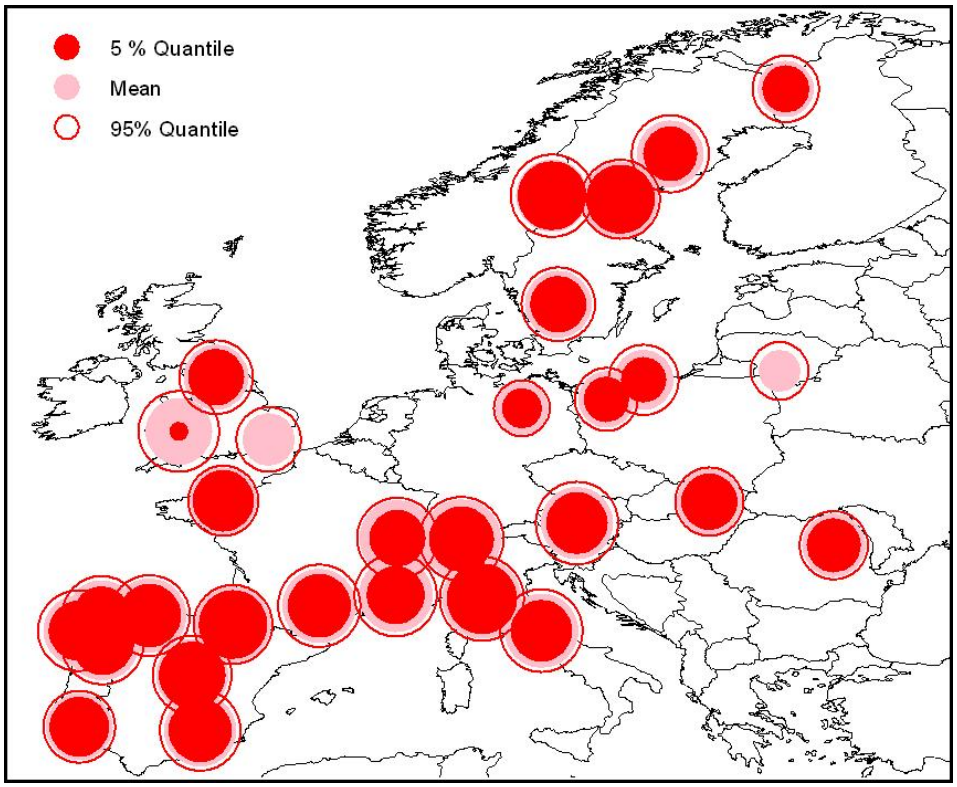


Figure 6b. ***Map for scaled log-transformed slope.***

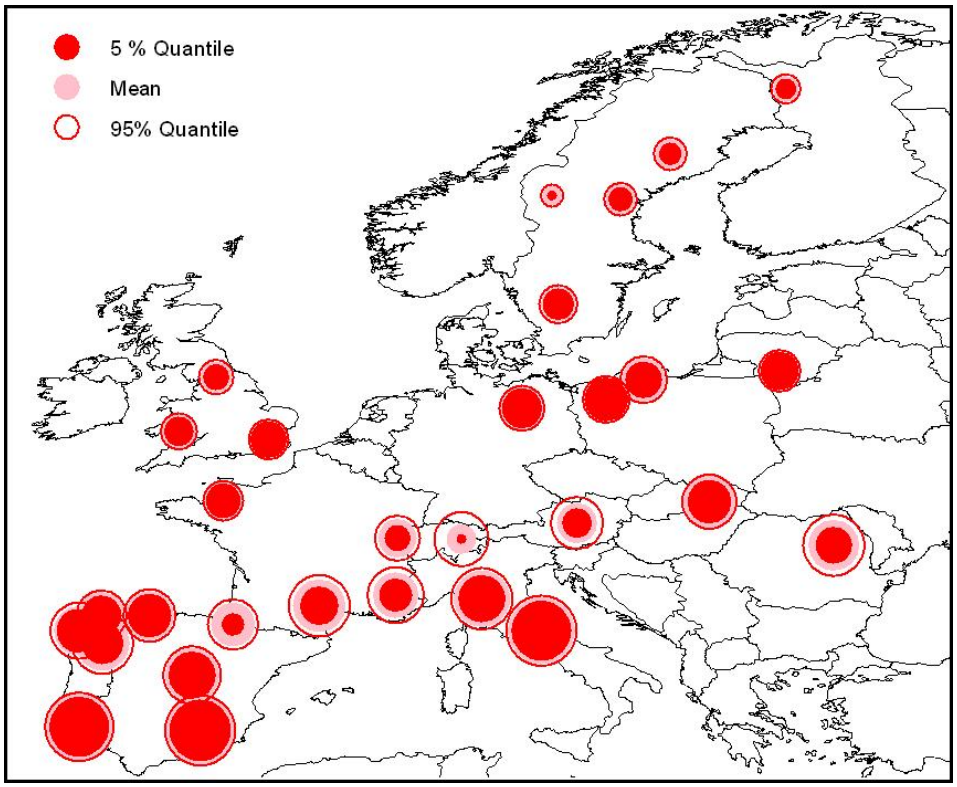


Figure 6c. ***Map for scaled July temperature.***

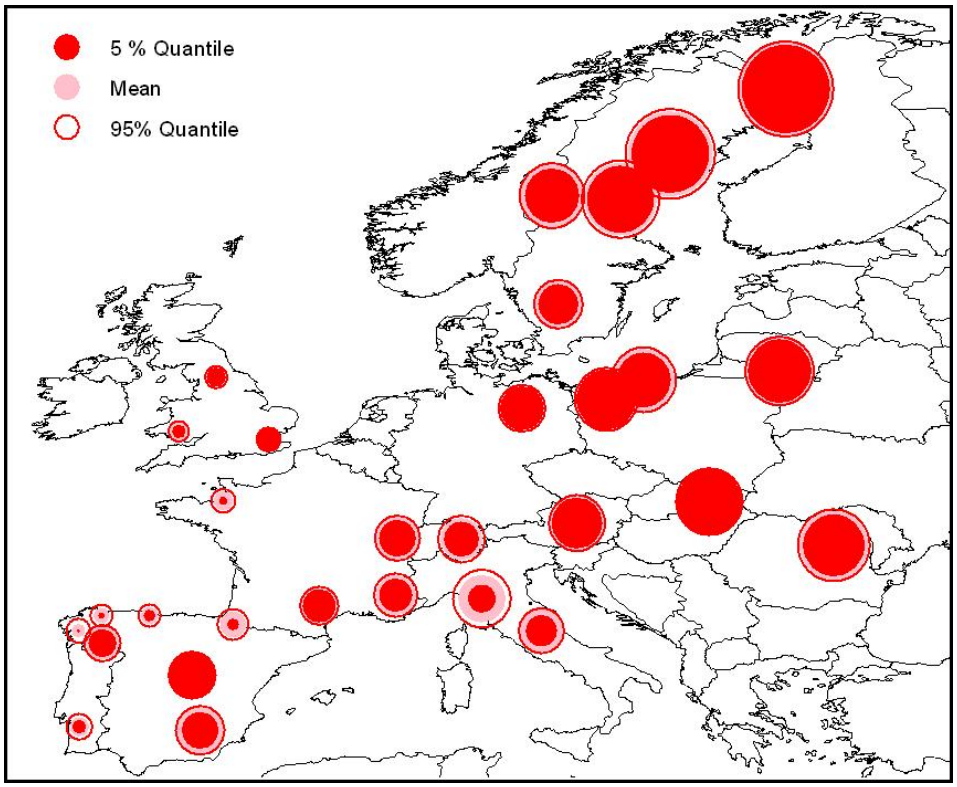


Figure 6d. ***Map for scaled temperature range between January and July***

These series of maps show:

1) A strong gradient from the North to the South of Europe for July temperature, with more variability in the mountains, especially in the Alps;
2) And strong gradient from the West to the East of Europe for temperature range.

As regards potential width and slope, there is no obvious pattern. Potential slope levels and variability seems to be much more related to the size of catchment and the location within the basin. As regards slope, only a line from Great Britain to the Northern continental Europe with lower levels is really noticeable.

The next maps display the locations of qualitative variables' modalities.

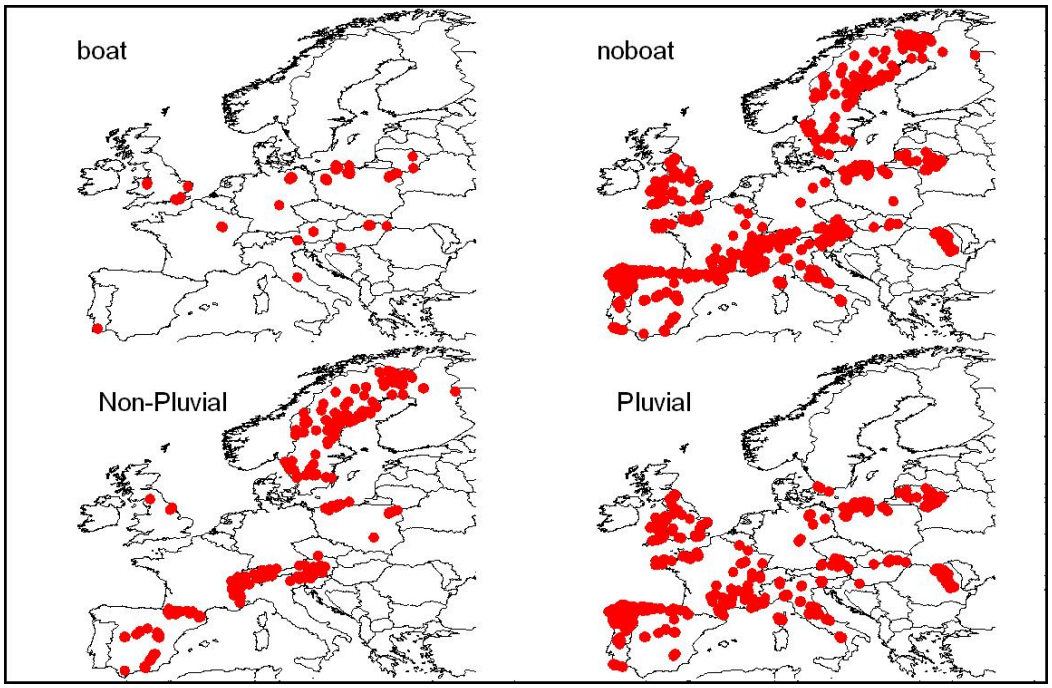


Figure 7a. ***Maps for fishing methods and water source types.***

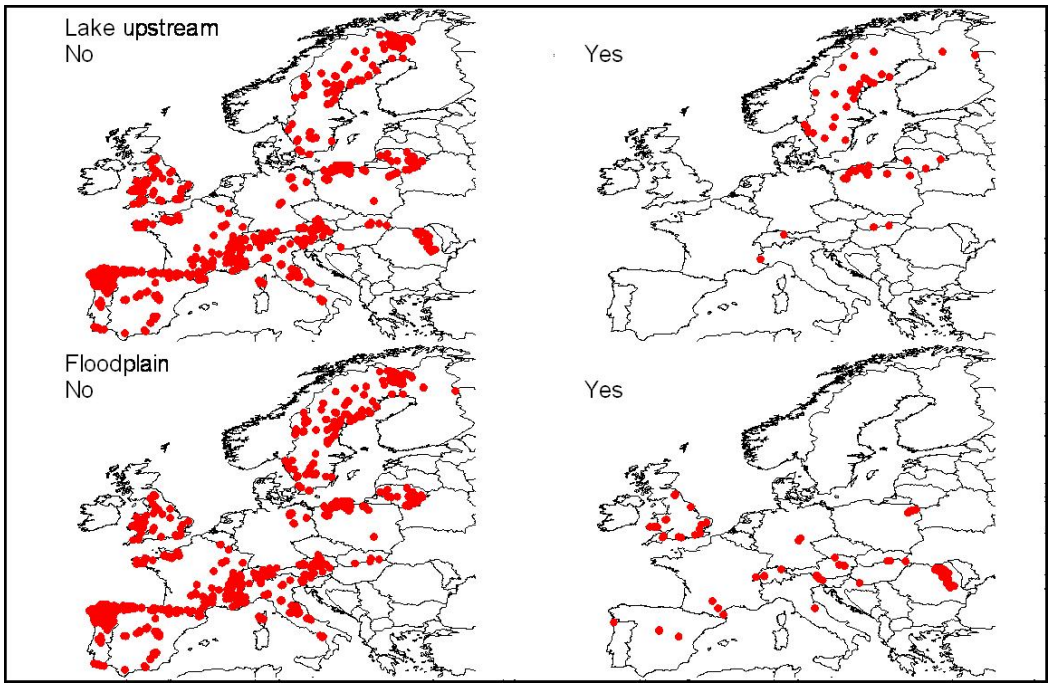


Figure 7b. ***Maps for presence/absence of lake upstream and floodplain.***

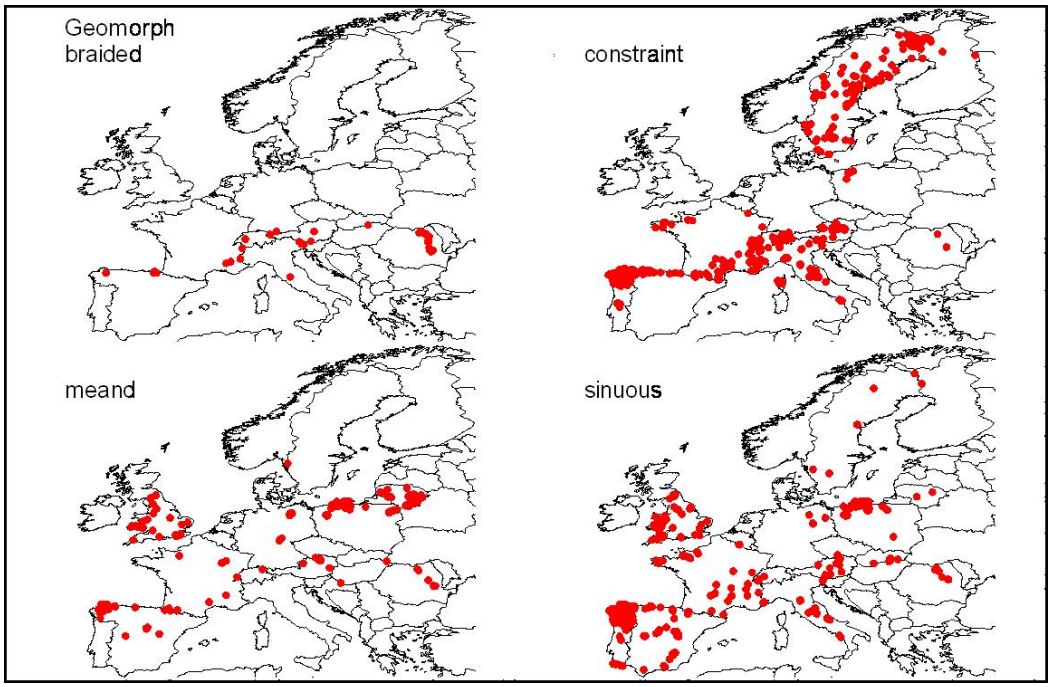


Figure 7c. ***Maps for Geo-morphological types.***

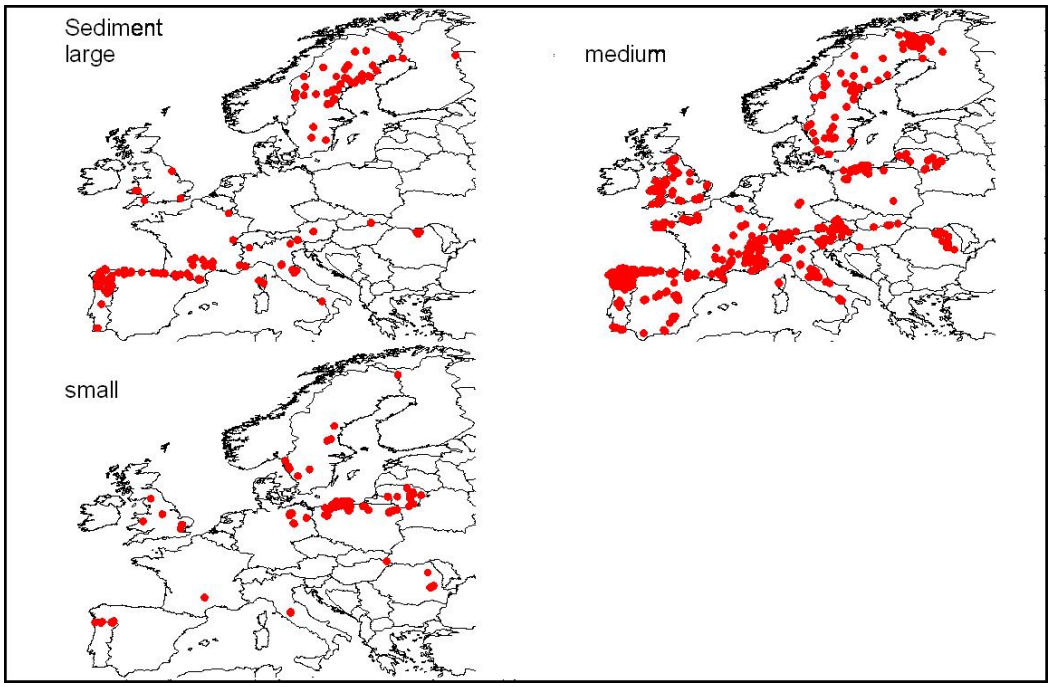


Figure 7c. ***Maps for natural sediment size.***

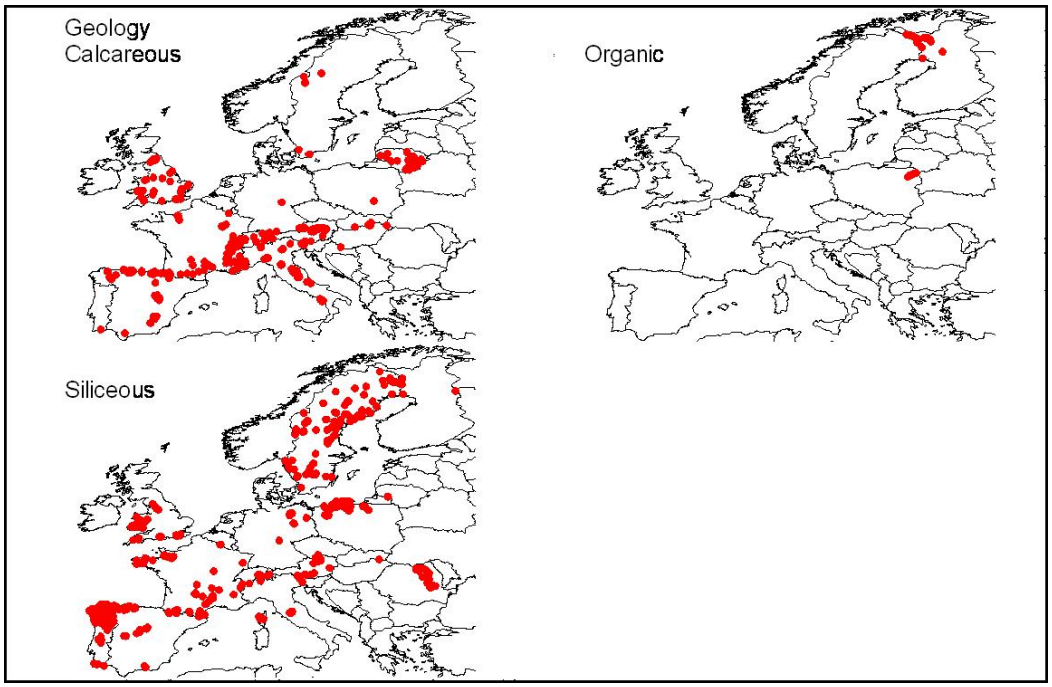


Figure 7d. ***Maps for geological types.***

The main noticeable points are:

1) There is a very small number of boat fishing sites among calibration sites.
2) Non pluvial are mainly distributed in Scandinavian Europe, in mountains (the Pyrenees, the Alps and Spanish sierras), as well as in Poland;
3) Sites with lakes upstream are located in North-Eastern Europe, notably in Poland and Sweden;
4) Sites with floodplain are located in Great Britain and central Europe, particularly in the Danube in Romania;
5) As regards geo-morphological types, braided type is mainly found in the Alps and the Danube in Romania, whereas constraint type is found in the North of Spain, in the Alps and in Sweden; Meander and sinuous types show quite similar locations, almost everywhere, notably all Great Britain calibration sites;
6) Medium natural sediment is distributed quite everywhere; large sediment is mostly found in Sweden and in the North of Spain, whereas small sediment is mainly found in North-Eastern Europe;
7) Siliceous sites are located in Western Great Britain, in Galice and the Asturies, in the Danube in Romania, in the North of Poland and in Sweden; calcareous sites rather elsewhere; organic geological type is only found in the North-East of Poland and in Finland.

8.3.3 Looking for patterns…

In order to display the main patterns in the environment of calibration sites, a multiple component analysis, similar as the previous, one but on the calibration subset, was performed.

The next graphs are the projection of sites on the first and second factorial plans, grouped according to the modalities of some of MCA variables.

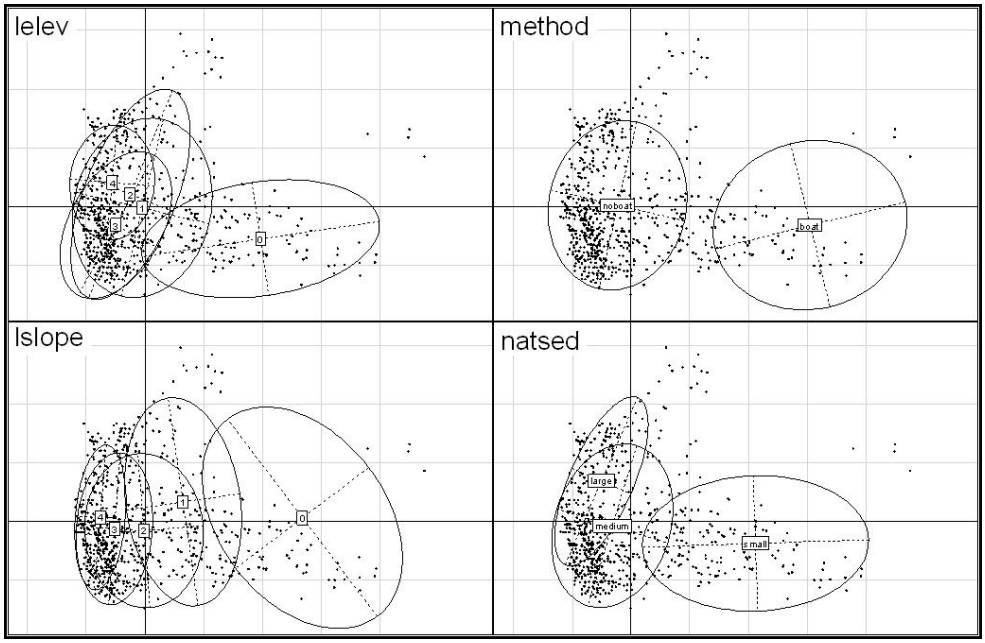


Figure 8a. ***Projection of sites on the first factorial plan grouped by the modalities of variables included in MCA: log-transformed elevation (lelev), fishing method, log-transformed slope (lslope), natural sediment (natsed).***

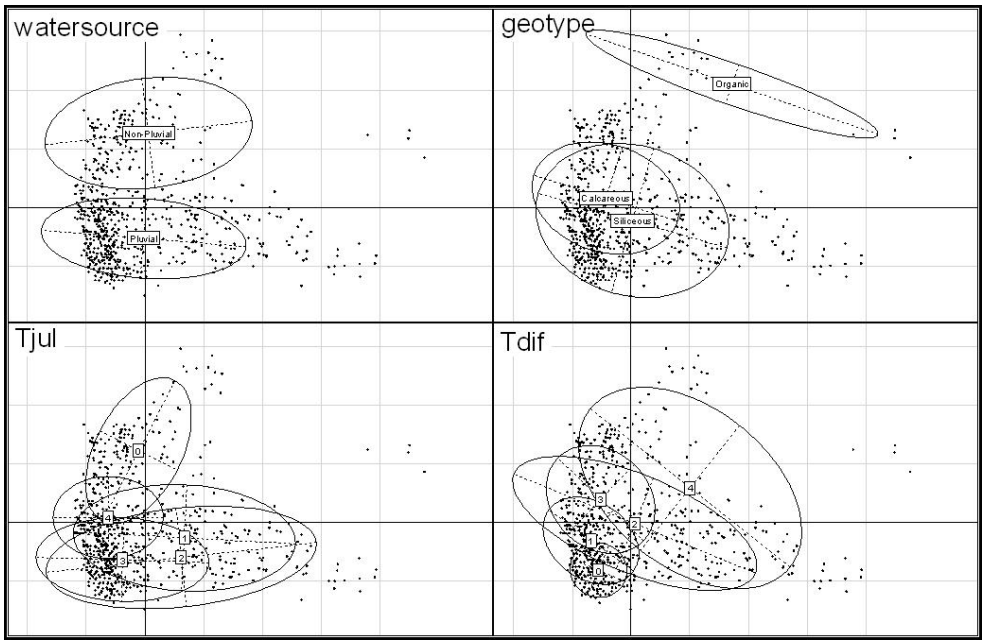


Figure 8b. ***Projection of sites on the first factorial plan grouped by the modalities of variables included in MCA: water source type (watersource), geological type (geotype), July temperature (Tjul), temperature range (Tdif).***

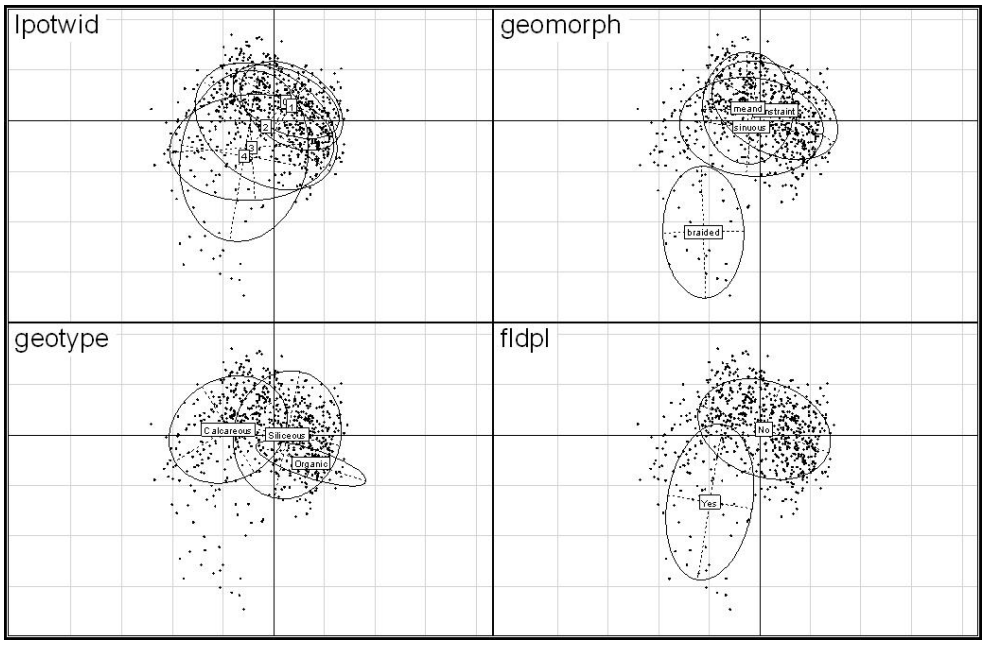


Figure 8c. ***Projection of sites on the second factorial plan grouped by the modalities of variables included in MCA: log-transformed potential width (lpotwid), geo-morphological type (geomorph), geological type (geotype), floodplain site (fldpl).***

The next graphs are the corresponding projection of MCA variables' modalities on the first and second factorial plans.

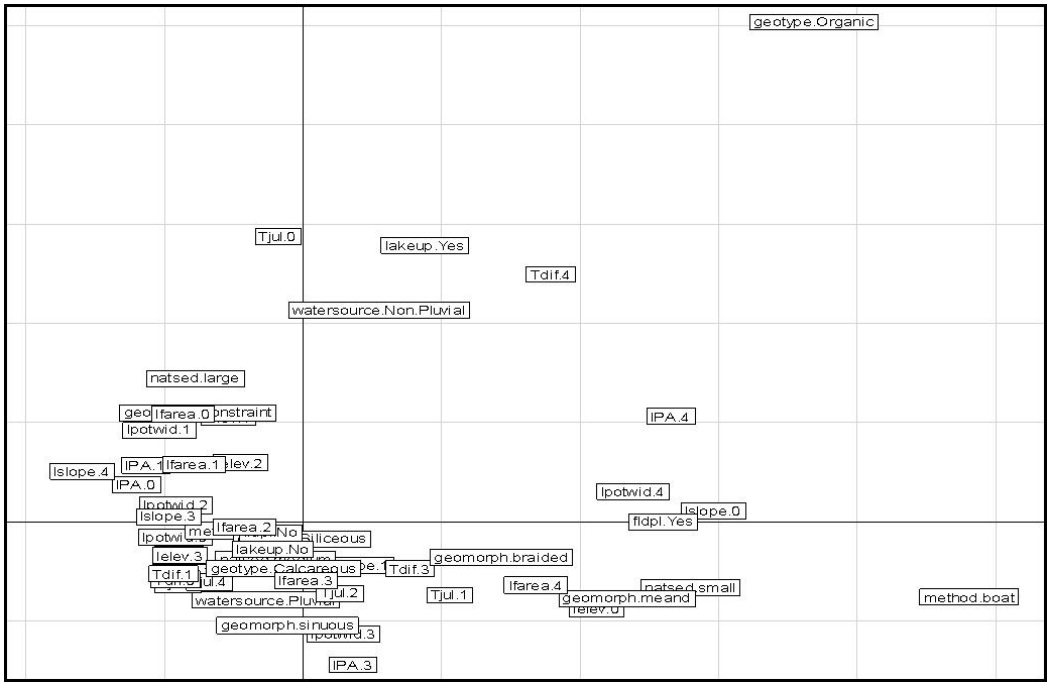


Figure 9a. ***Projection of variables' modalities on the first factorial plan.***

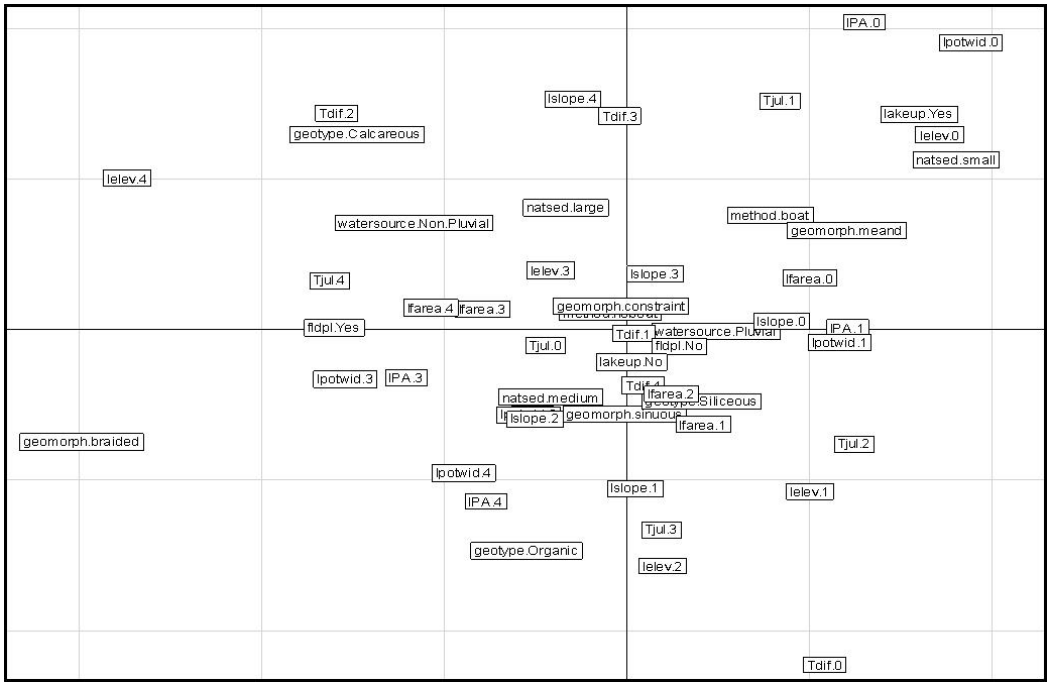


Figure 9b. ***Projection of variables' modalities on the second factorial plan.***

Here again, the first component is highly correlated with slope, and opposes small sediment and boat fishing method to others, whereas the second opposes pluvial and non pluvial water source. Also notice that, if the third axis is difficult to interpret, the fourth clearly opposes braided geo-morphology and sites with a flood plain to others.

The next graphs are the projection of sites on the first and second factorial plans, grouped by eco-regions.

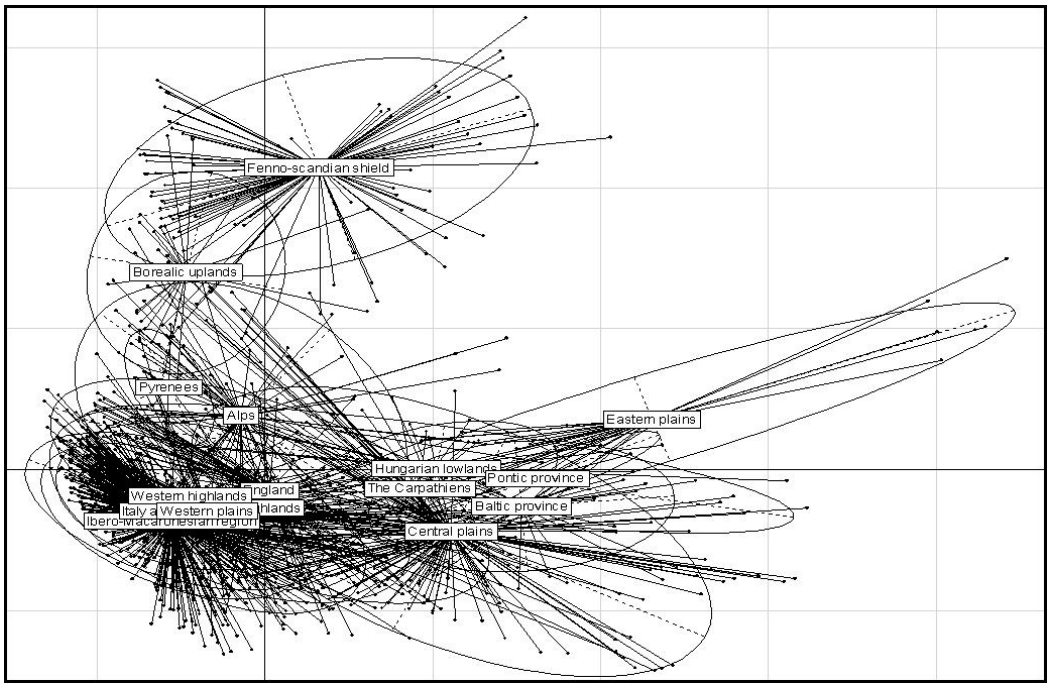


Figure 10a. ***Projection of sites on the first factorial plan grouped by eco-regions.***

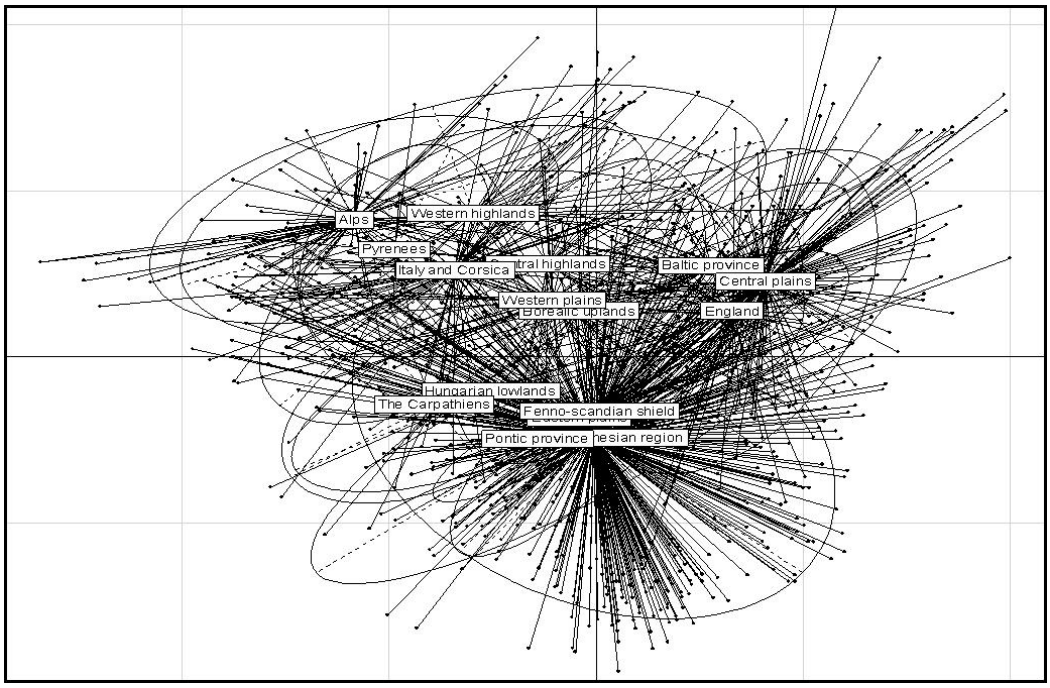


Figure 10b. ***Projection of sites on the second factorial plan grouped by eco-regions.***

As regards the projection of regions, the first axis clearly opposes eastern plains to western plains, whereas the second axis opposes Scandinavian Europe to other regions. Finally, the third axis opposes mountains to lower regions.

Appendix 1 – Models calibration sites (N=533)

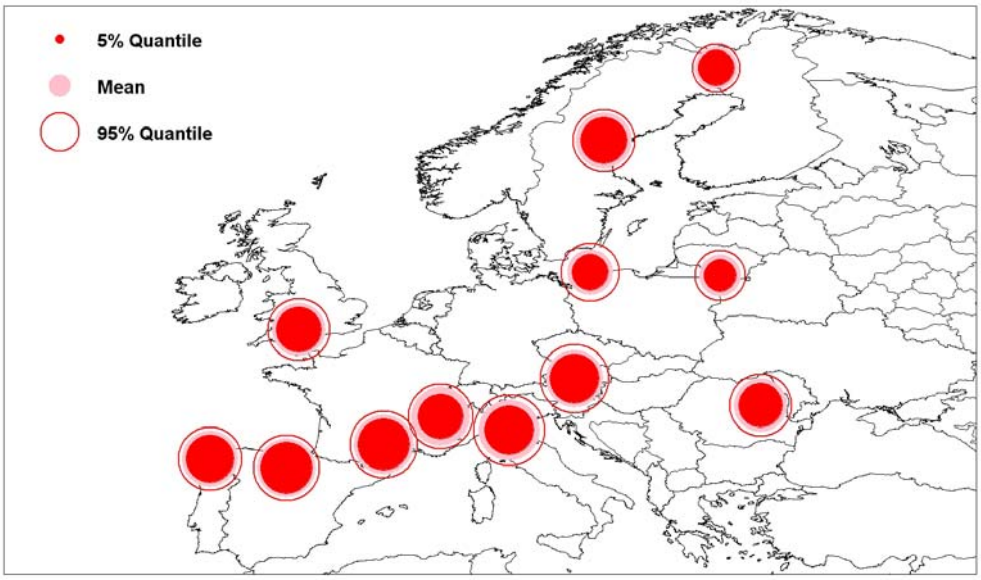


Figure 4. ***Log-transformed slope***

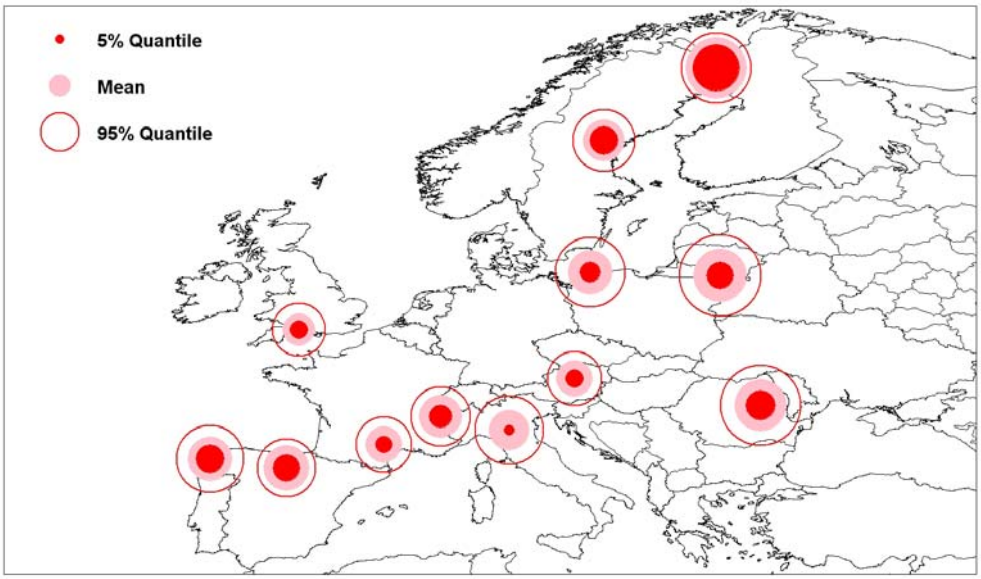


Figure 4. *Log-transformed potential width*

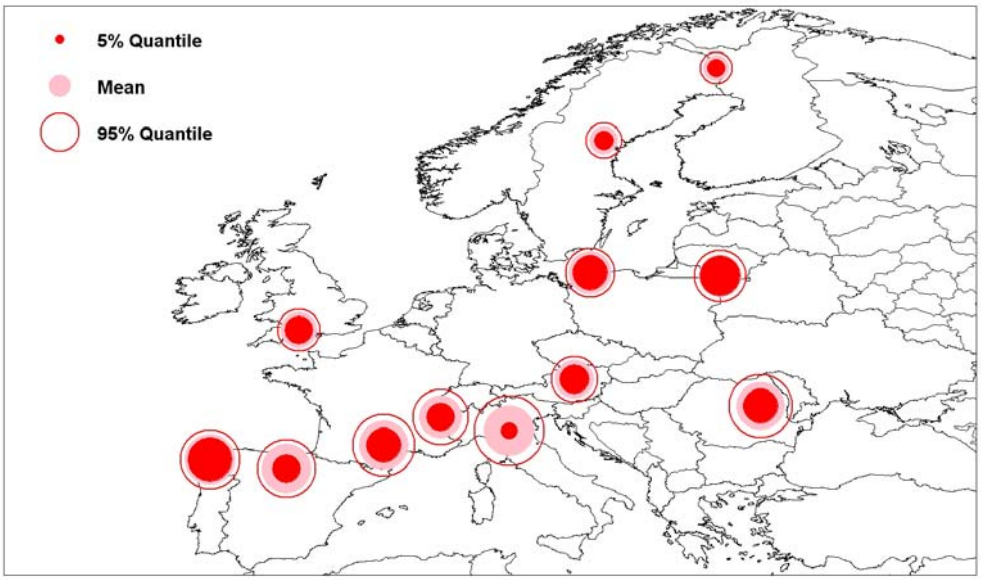


Figure 4. *Temperature of July*

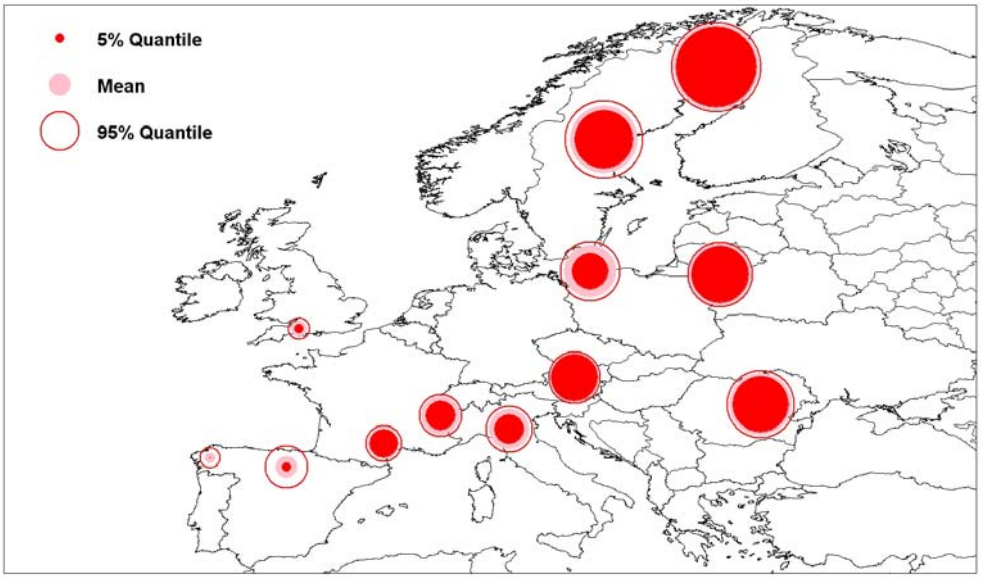


Figure 4. *Temperature of July minus temperature of January*

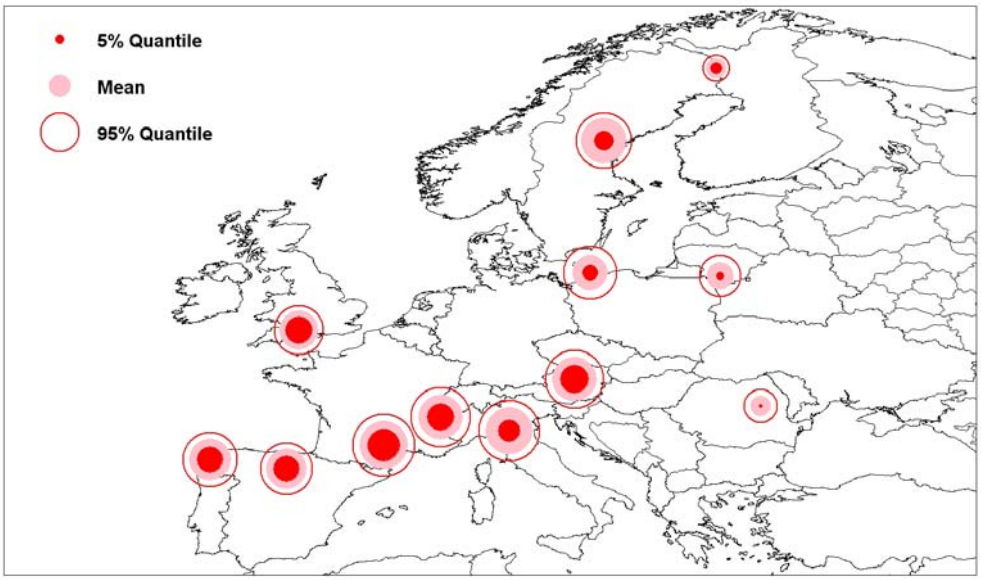


Figure 4. *First morphological factor*

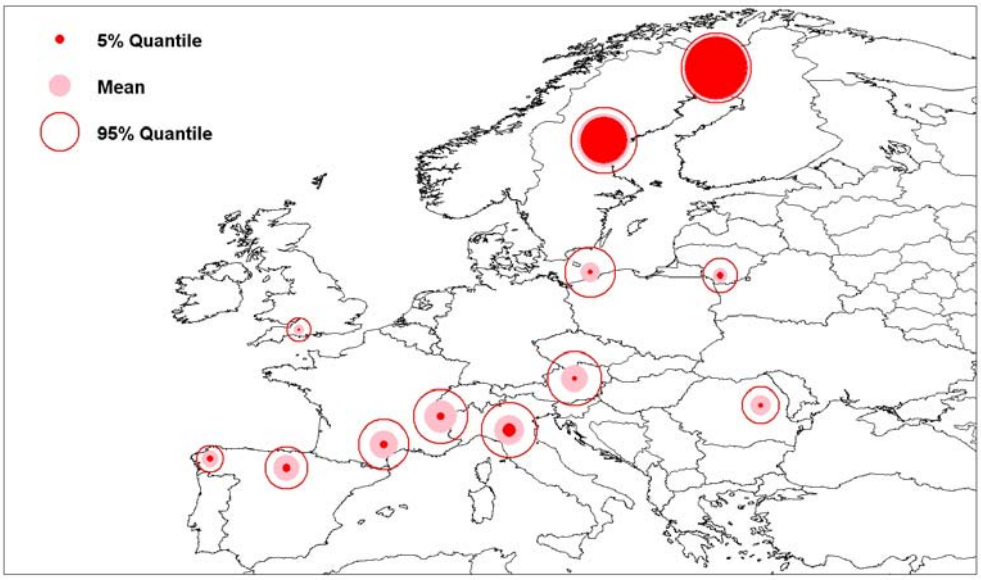


Figure 4. *Second morphological factor*

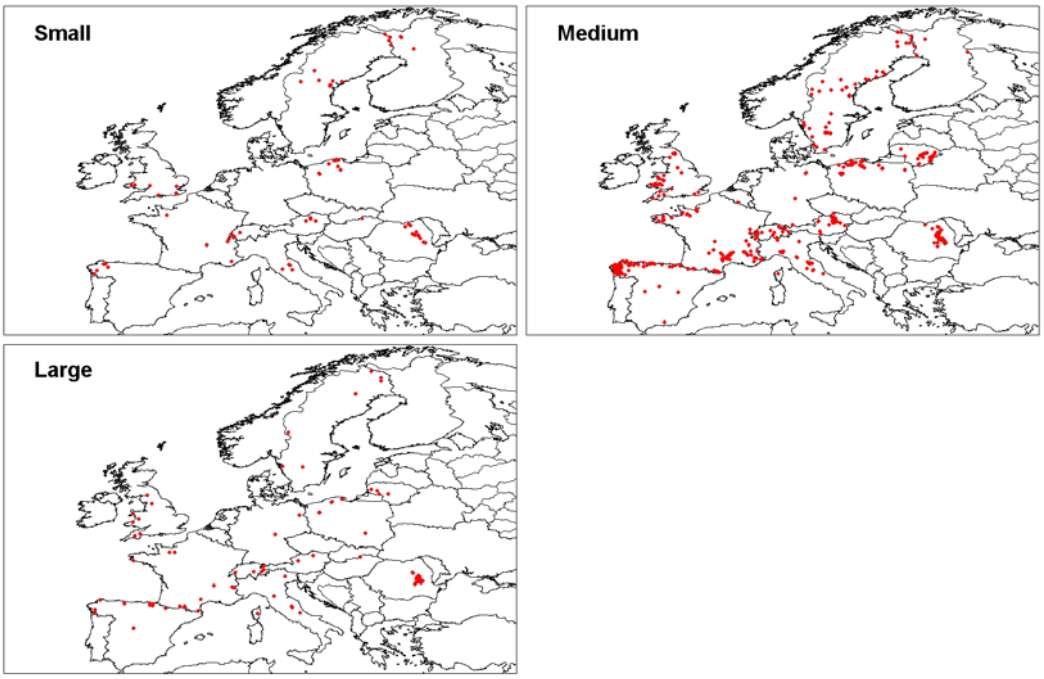


Figure 4. *Natural sediment*

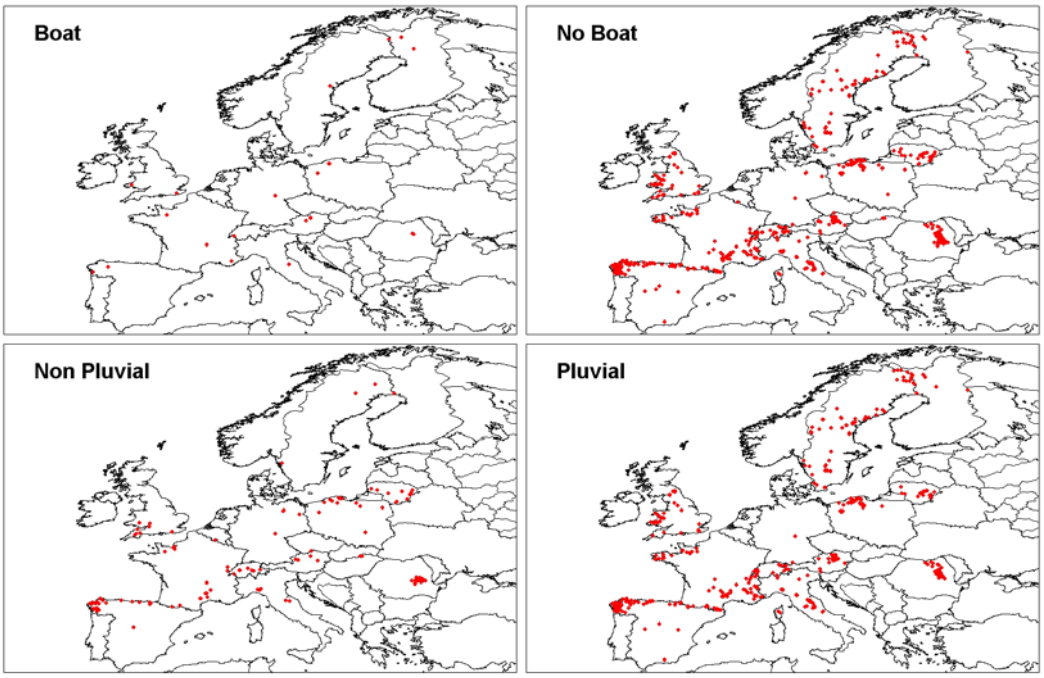


Figure 4. *Fishing method and water source type*

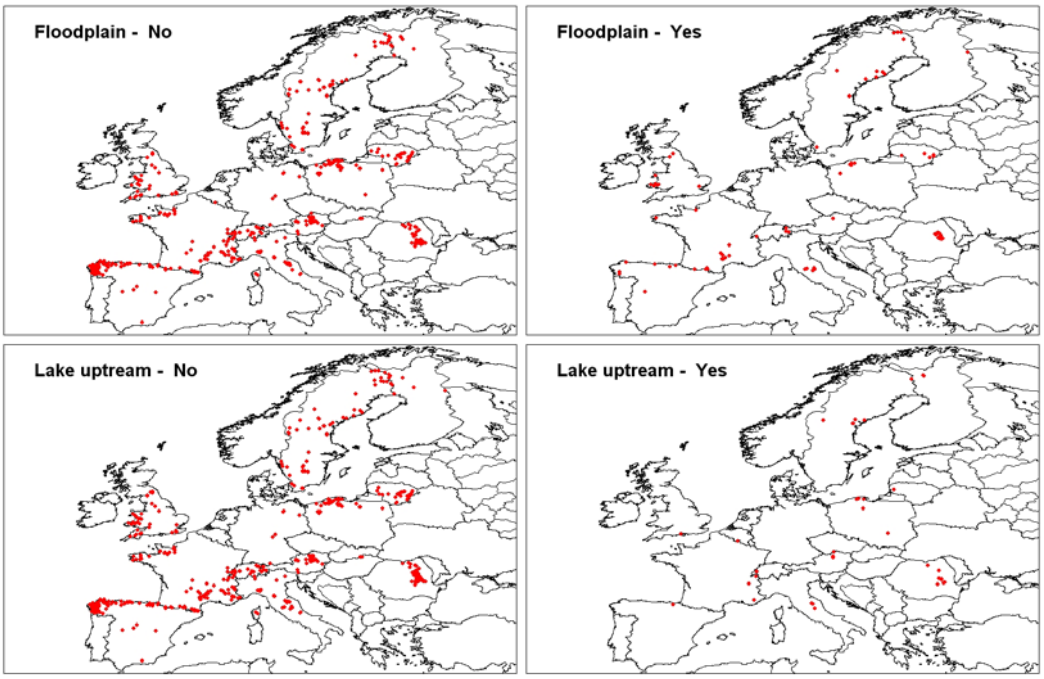


Figure 4. *Floodplain and upstream lake indicators*

Appendix 2 – Clean sites (N=1741)

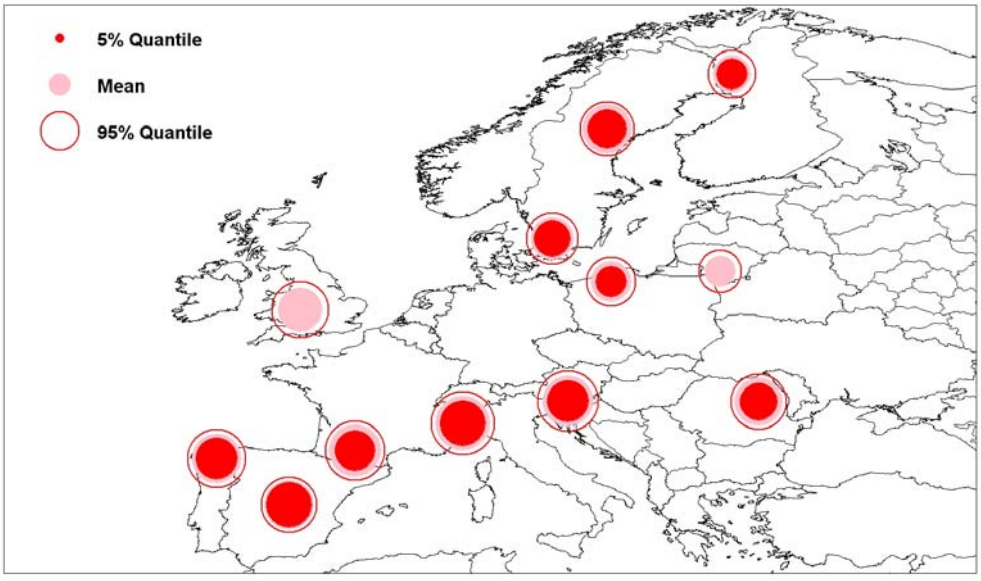


Figure 12a. ***Log-transformed slope***

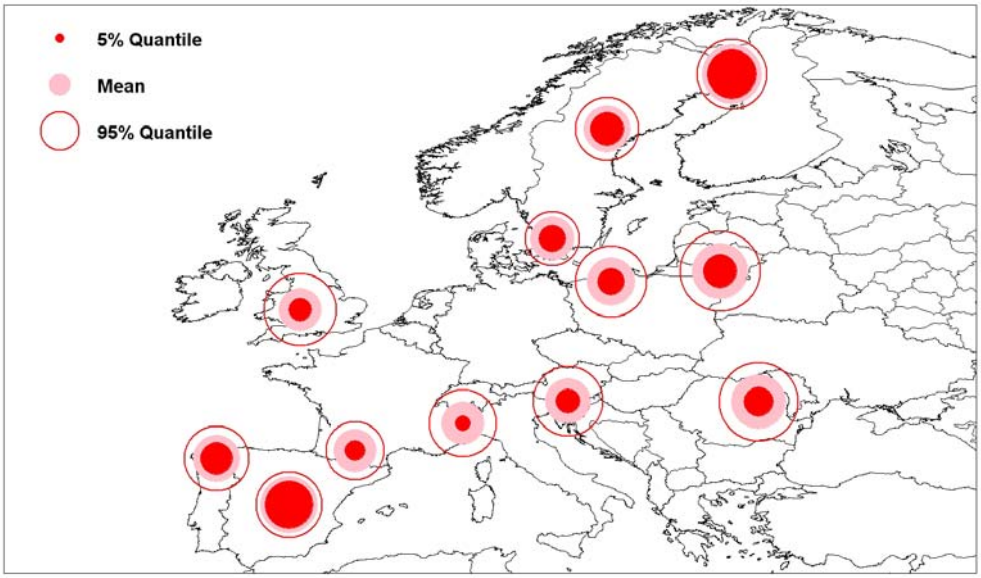


Figure 12b. *Log-transformed potential width*

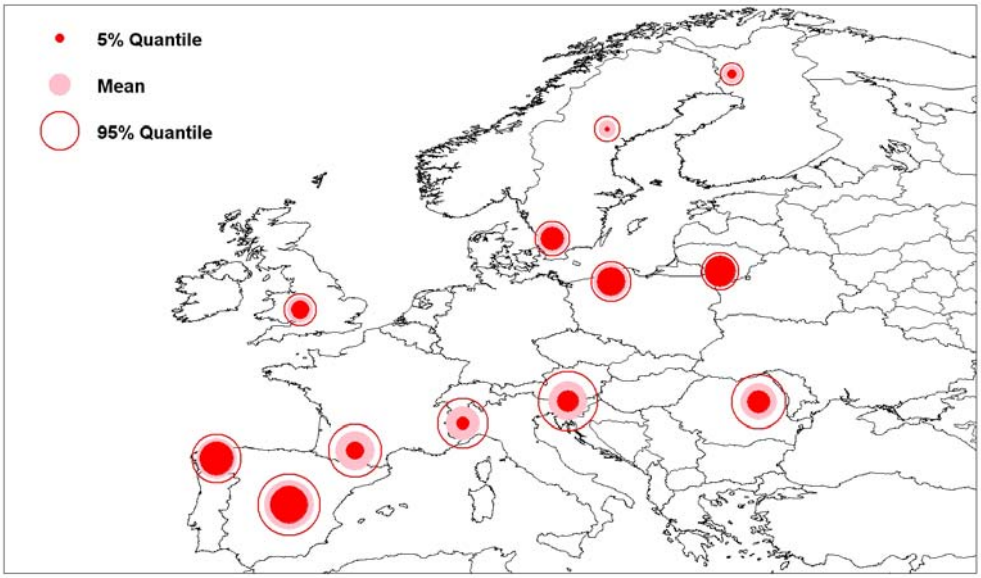


Figure 12c. *Temperature of July*

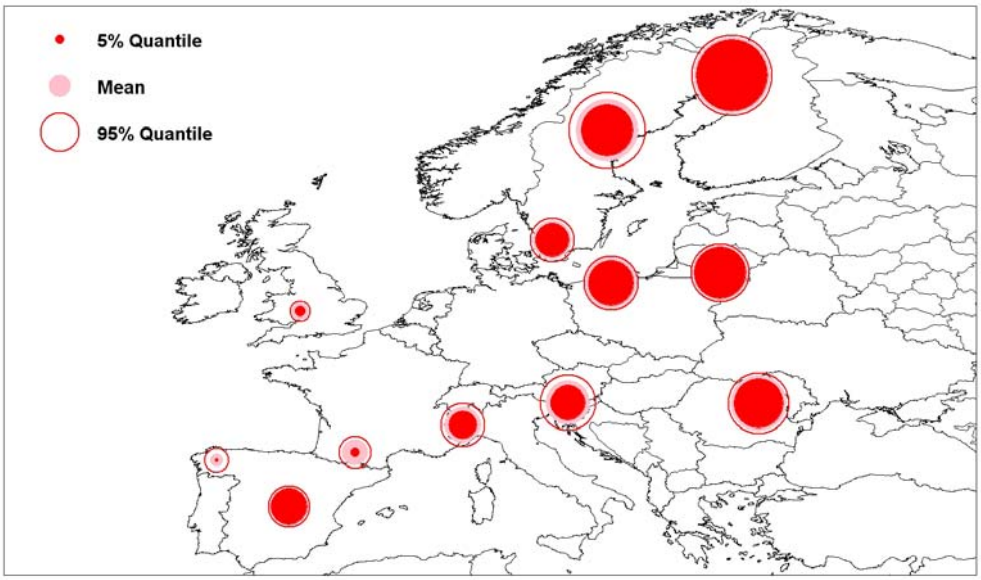


Figure 12d. *Temperature of July minus temperature of January*

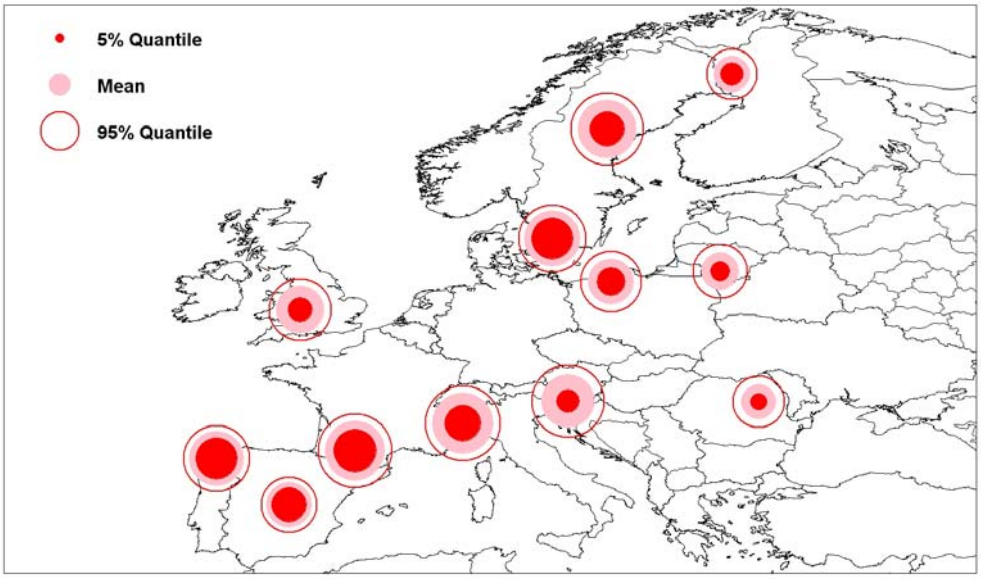


Figure 12e. *First morphological factor*

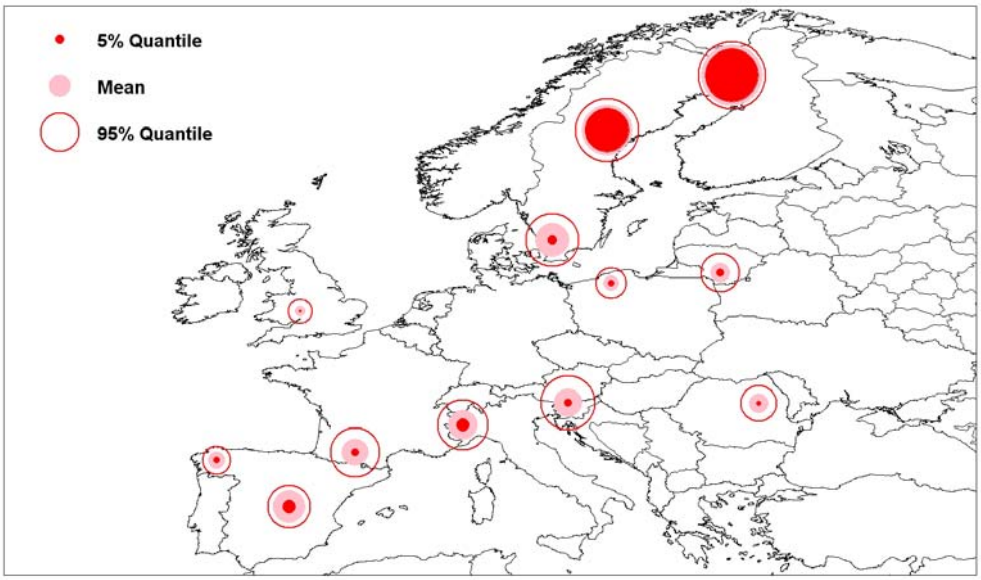


Figure 12f. *Second morphological factor*

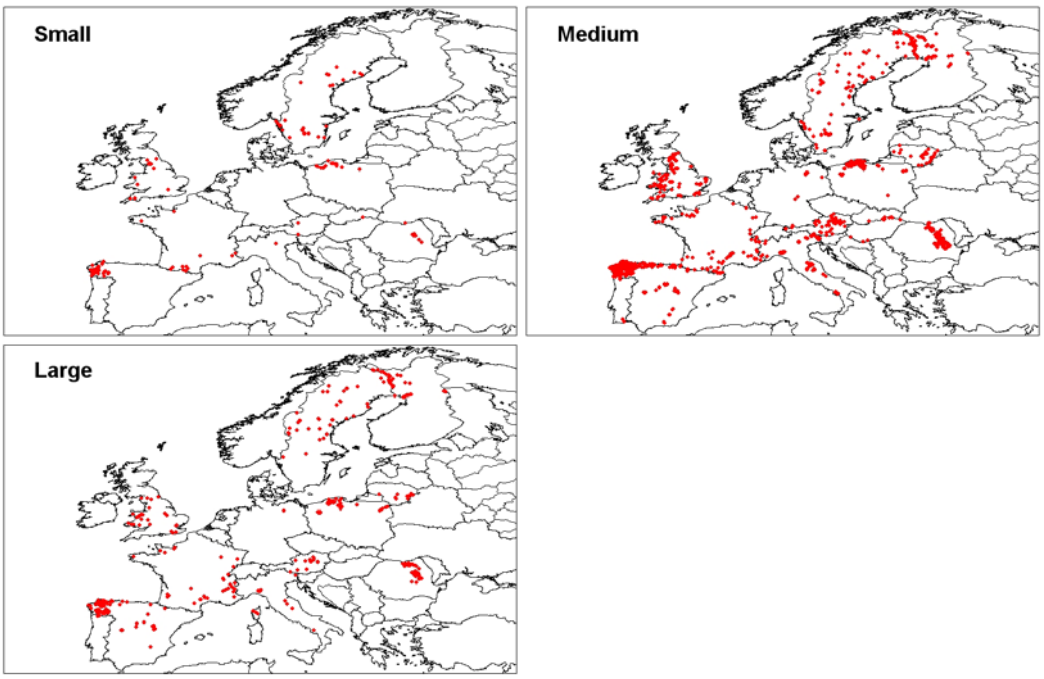


Figure 13a. *Natural sediment*

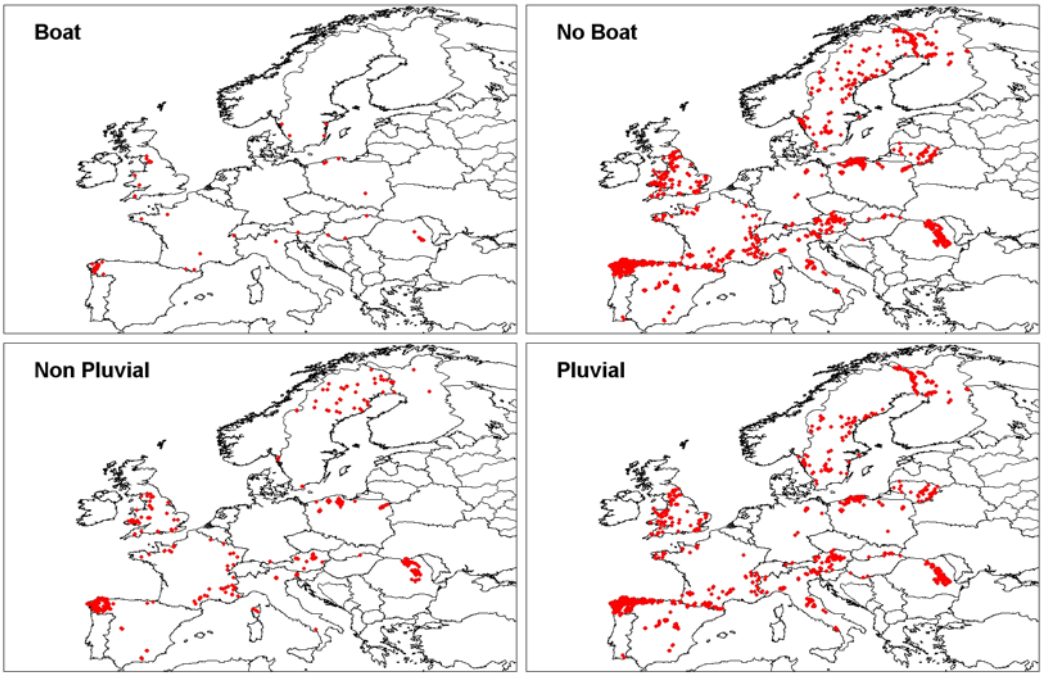


Figure 13b. *Fishing method and water source type*

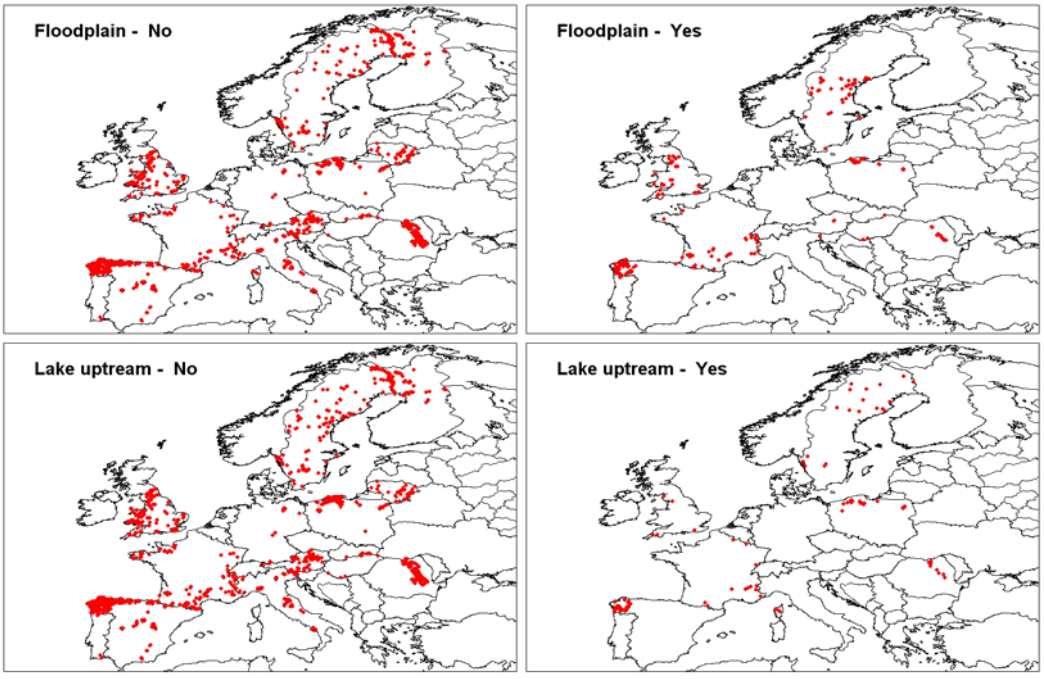


Figure 13c. *Floodplain and upstream lake indicators*

8.4 Modelling metrics using models for count data

8.4.1 Considered strategies for modelling metrics

8.4.1.1 Classical problems and strategies considered

Non-Normality

Richness and density metrics consist respectively in count or count-like data. The distribution of such response variable is almost never Gaussian, except in asymptotic cases, notably because underlying processes involved are based on more or less rare events, whose occurrence numbers are classically distributed according to dissymmetric histograms. Moreover, as for richness metrics, data are discrete. And, finally, such data are in general intrinsically heteroskedastic, i.e. variance is function of the mean.

Correlation between predictors

A second problem consists in the correlation structure within the set of predictors, which could bring to bad estimates of model coefficients and then to bad predictions.

Non-Linearity of predictors' effects

The possible non-linearity of predictors' effects on the response variable (or transformed variable) mean can drive to misspecification.

Robustness, too influential data

The influence of extreme values and/or possible outliers could also strongly impact the model, since this latter can become unstable, notably because of leverage effects.

Extra zeros

Finally, another problem often encountered with count data is zero inflation, which consists in an unexpected number of zeros according to classical probability distributions, which results in difficulties to fit a relevant model to such count data.

8.4.1.2 Considered strategies

To normalize, linearize and stabilize variance

To model such response variable, two main strategies could be considered: First, to transform the response variable, in order to normalize and/or to stabilize variance; Second, to find another probability distribution to model the response variable.

Classically, log, square root and double square root transformations are used to normalize count data. Boxcox power transformations can also be used, but they are uneasy to interpret and less familiar. Finally, one must keep in mind that such transformations mean a rescaling of the initial variable, that is a distortion of the variable. For example, log transformation fixes heteroskedasticity through an artificial increase in the variance of the smaller values and an artificial decrease in the variance for the greater values. Besides it is not only harder to interpret data, but also to do inference on the initial variable and to compare models with and without transformations.

To fit specific variance-mean relationships

A probability distribution commonly used to model count data is the Poisson distribution which is part of the exponential family and then can be modelled in the framework of Generalized Linear Models. A problem often encountered is over- (or under-) dispersion in counts. A solution is to take account of this overdispersion through a parameter which is to estimate, for example by using a Quasi-Poisson in the GLM context.

Another solution is to use Negative Binomial distribution which is a probability distribution suited to count data (positive, dissymmetric) with two parameters (against one for Poisson), and thus more flexible to take account of overdispersion.

To 'decorrelate' predictors

To avoid problems due to multi-collinearity in predictors, a classical solution is to use an orthogonal basis of predictors' space as a new set of predictors.

For this, two classical techniques are:

- PCR regression that uses main factors of a PCA of initial predictors;
- PLS regression that uses orthogonal factors, that are linear combinations of predictors which maximise their covariance with the response variable.

To approach predictors' true effects

A classical technique to include non-linear effect of predictors is provided by the Generalized Additive Models framework. It consists in including functional transformations of predictors; among smoothing functions, B-splines provide new predictors easy to use in this context; these latter are projections of the initial predictor on an orthogonal basis, whose elements are piecewise polynomial functions.

A second approach is neural networks. Such method automatically combines different functional transformations of predictors to maximize a fit criterion.

To stabilize models

Robust regression is based on iterative re-weighted least squares. It consists in giving less weight to influential values in the fitting process.

To model extra zeros

Two approaches to take extra zeros into account are:

- First, zero-inflated models, which use a mixture of distribution functions: a count process and a point mass at zero to 'inflate' zero counts;
- Second, hurdle models – also called zero-altered models – which use truncated distribution functions.

8.4.1.3 Description of modelling strategies

Generalized Linear Models

Most of the time, a functional transformation of the response variable helps and is sometimes enough for recovering symmetry, stabilizing variance, linearizing and normalizing data. But, in some cases, it is necessary to consider other distributions for the response variable.

Generalized Linear Models are the classical framework used for modelling count data. It is based on log-Likelihood maximisation and allows to model distributions among those of the exponential family, e.g. Gaussian, Poisson, binomial, Gamma, as well as so-called quasi-families.

The density function of one of the exponential family is of the following form:

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}$$

where:

$a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$

are known functions characterizing the distribution;

and:

θ_i and ϕ

are, respectively, the location and scale parameters.

The mean of the response variable is related to the location parameter as follows:

$$\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$$

The mean is related to predictors through the link function:

$$\eta_i = g(\mu_i)$$
$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$$

The function $g(\mu_i)$ will be called the *link* function.
The quantity η_i is called the *linear predictor*.

Generalized Additive Models

Linear models can be extended to non-linearity in the framework of additive models, which consists in using functional transformation of predictors in the specification of the model, as follows:

$$Y = \alpha + \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

Generalized Additive Models (GAM) are the extension of Generalized Linear Models to additive models. Classically, the functions of the predictors used here are smoothing functions, such as local polynomials or piecewise polynomials (splines).

B-splines, used here, consist in a functional basis of positive piecewise polynomials, which notably satisfies:

$$0 \leq B_i(x) \leq 1$$

$$\sum_{i=1}^{r} B_i(x) = 1, \text{pour } x \in [a, b]$$

A projection of the initial predictor on this basis is included into the model. It results in an increase of dimensionality, i.e. an increase of the number of the parameters to estimate. However, the effect of the predictor is better taken into account especially when this effect is clearly non linear. In addition, B-splines help reducing the influence of atypical values because of the nullity of a B-spline is non-zero only on certain knots.

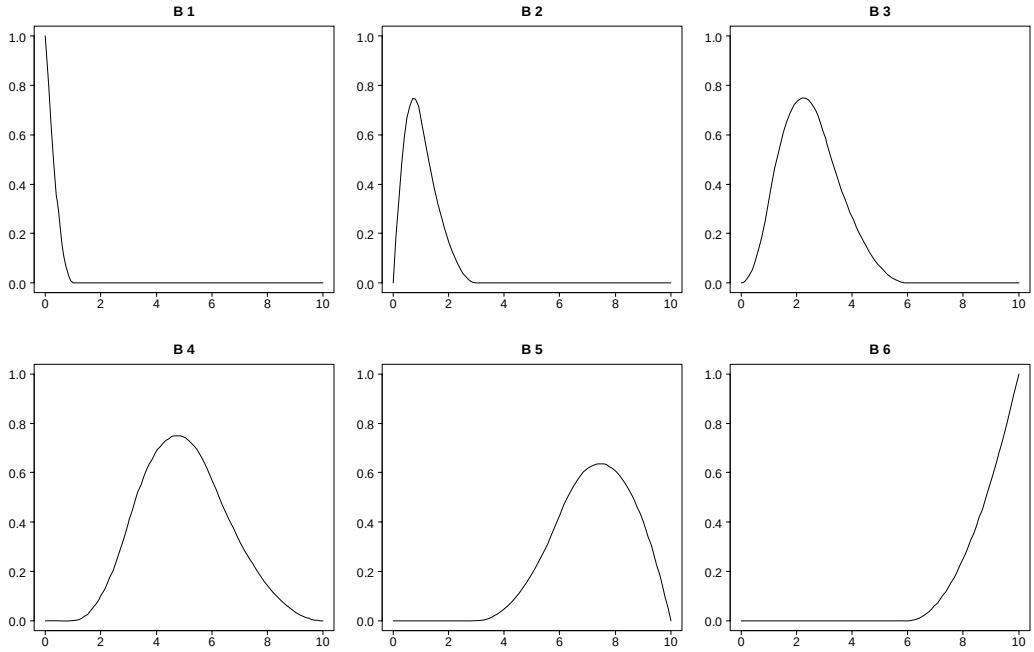


Figure 1. *Example of a B-spline basis*

Negative Binomial Distribution

The Poisson model is a model suited for data showing equi-dispersion, that is a constant mean/variance ratio. The link function used here is the log function (canonical link). The scale parameter is always 1.

The Poisson model can be considered as an extreme case of the negative binomial model. This latter can be written as:

$$Y_i \sim \text{Poisson}\ (\lambda_i)$$

With:

$$\lambda_i = \exp\ (x_{i\beta} + u_i)$$
$$\exp(u_i) \sim \text{Gamma}(1/\alpha, \alpha)$$

where :

α is the over-dispersion parameter (Note: In Poisson, $\alpha = 0$).

$$\mu_i = \lambda_i \qquad \text{(as in Poisson)}$$
$$\omega_i = \mu_i + \alpha\, \mu_i^p$$

p is a constant (usually 1 or 2).

Note that Negative Binomial is not part of the exponential family.

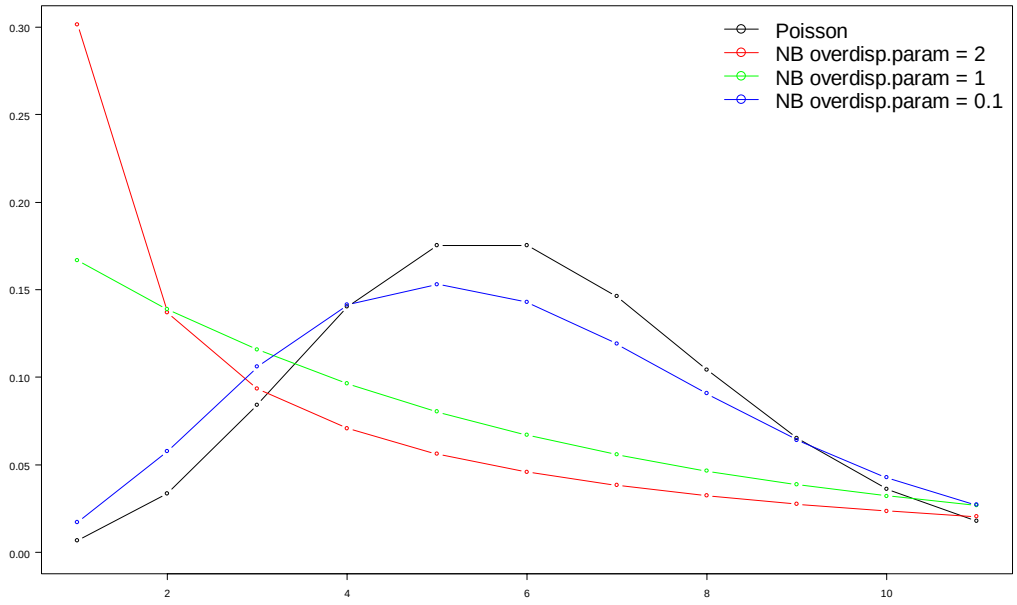
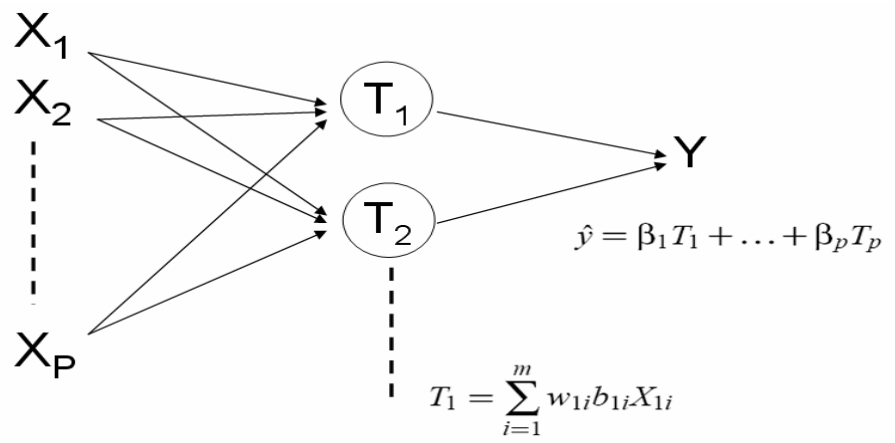


Figure 2. ***Probability functions for Poisson (lambda=2) andNegative Binomial with the same mean as the Poisson and 3 different values for overdispersion parameter (2, 1, 0.1)***

Orthogonal Regression

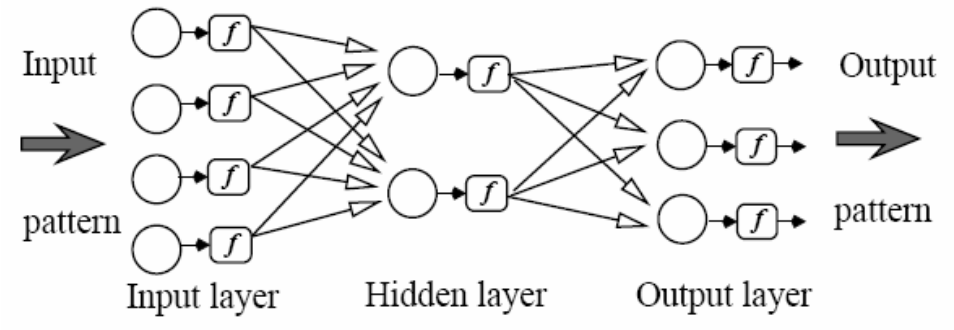
Partial Least Square and Principal Component Regression consist in a Least Square Regression on an orthogonal basis of predictors' space. There is then no correlation left between these new regressors, which are linear combinations of the initial predictors. PCR Factors are the components of the Principal Component Analysis of predictors, thus built with no a priori on their capacity to explain the response variable. PLS Factors are built so that they maximize their covariance with the response variable. The remaining question is how many factors should be kept in the regression.

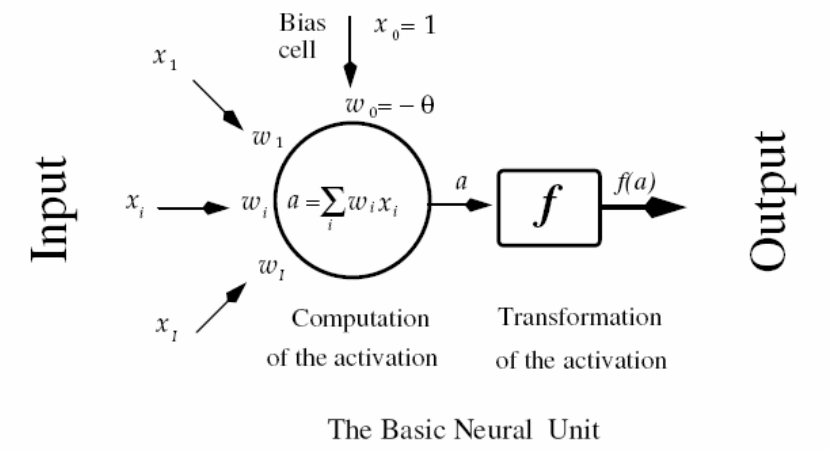


Neural networks

Neural network models consist in a series of computations based on the functional transformation (activation) of linear combinations of inputs. In the input layer, inputs are linear combinations of predictors. In the following layers, all inputs and outputs can be combined before activation, using functions among: linear, logistic, indicator or Gaussian.

$$y_k = \phi_0(\alpha_k + \sum_{i \to k} w_{ik} x_i + \sum_{j \to k} w_{jk} \phi_j(\alpha_j + \sum_{i \to j} w_{ij} x_i)$$





The Basic Neural Unit

Zero-inflated and hurdle models

Zero-inflated models are true mixture models with two sources for zeros: that of a count process (e.g. Poisson or Negative Binomial), and that of a point mass process. The weight of the mixture is fitted using a Bernoulli model with similar predictors.

$$f_{\text{zeroinfl}}(y; x, y, \beta, \gamma) \quad =$$

$$\pi \cdot I_{\{0\}}(y) \; + \; (1 - \pi) \cdot f_{\text{count}}(y; x, \beta)$$

Point mass at 0

Binomial GLM
for weights

Count distribution
(Poisson, NB…)

Hurdle Models consist in truncated distributions, a truncated count model (e.g. Poisson or Negative Binomial) for positive counts and a Binomial or a censored count model for zero counts. Hurdle models are not true mixture models, and there is only one source of zeros in such models.

$$f_{\text{hurdle}}(y; x, z, \beta, \gamma) =$$

$$\begin{cases} f_{\text{zero}}(0; z, \gamma) & \text{if } y > 0 \\ (1 - f_{\text{zero}}(0; z, \gamma)) \cdot f_{\text{count}}(y; x, \beta) / (1 - f_{\text{count}}(0; x, \beta)) & \text{if } y > 0 \end{cases}$$

Presence/Absence distribution for zeros

Truncated distribution for counts

8.4.2 Considered models and retained criteria for comparing them

8.4.2.1 Predictors

The predictors used for every models are :

- The fishing method (method), with two modalities: boat/noboat;
- The log-transformed slope (lslope), and its squared term;
- The log-transformed potential width (lpotwid);
- The July temperature (Tjul), and its squared term;
- The temperature of July minus the temperature of January (Tdif);
- The Natural sediment size (natsed), with three modalities: small/medium/large.

For additive models, non-linear effects of three predictors were considered: log-transformed slope, log-transformed potential width and July temperature.

Finally, for zeros and extra-zeros in hurdle and zero-inflated models, the same predictors were considered, but squared terms.

8.4.2.2 Fitted models

For Gaussian regression, the following models were fitted:

- An Ordinary Least Square model;
- An Additive Gaussian model, using cubic B-splines;
- Robust Regression using Iterated re-Weighted Least Squares (IWLS);
- A Partial Least Squares model (PLS);
- A Principal Component model (PCR);
- Two versions of two-layer neural network model using different levels for decay parameter.

For each of the above models, 3 transformations of the initial response variable were used:

- Natural logarithm, $log(Y+1)$;
- Square root, $\sqrt{Y+1}$;
- Double square root, $\sqrt[4]{Y+1}$.

For other statistical models, the following models were fitted:

- A Poisson Generalized Linear Model;
- A Poisson Generalized Additive Model, using cubic B-splines;
- A Negative Binomial model;
- Two versions of Zero-Inflated models, both with a Bernoulli model for the point mass at zero; the first one with a Poisson distribution for counts, the second with a negative binomial distribution;
- Two versions of Hurdle models, both with a Binomial model for zero counts; the first one with a Poisson distribution for positive counts, the second with a negative binomial distribution.

For density metrics, abundances were used with fishing area as an offset.

Finally, to get a benchmark, rate models were also fitted, using all models mentioned before, except orthogonal regression, zero-inflation and hurdle models. An offset were included in these models with total richness and total captures, for, respectively, richness and density metrics.

8.4.2.3 Criteria

Akaike Information Criterion (AIC) – or a Taylor-series approximation for models with a transformed response variable, is shown for direct comparison of models, even if its use is controversial when models are not nested or are not part of the exponential family.

A series of criteria was considered:

Goodness of Fit

For Goodness of Fit, two criteria were used:

- Pearson Correlation coefficient between observed and predicted values;
- Spearman Correlation coefficient between observed and predicted values.

The goodness of fit were also assessed within a subset of calibration set:

- In Mediterranean and non-mediterranen regions;
- In smaller Strahler order sites (<4) and in greater Strahler order sites (>3).

Prediction

For prediction error estimation:

- An internal validation process using cross-validation (bootstrap samples; Residual Mean Square Errors as a cost function);
- An external validation, using an extra set of calibration sites (extended set of calibration sites without retained calibration sites), using Pearson Correlation coefficient between observed and predicted values.

Normality

As regards normality of residuals, skewness and kurtosis were computed as well as a not too conservative test of normality, Jarque-Bera test.

8.4.3 Comparison of models

Next graphs are the distributions of the 26 ranks – corresponding to the 26 richness metrics fitted - of each of the 20 model families considered, according to the retained criteria, mentioned above. For a given criterion, a model family is good when its rank tends to 20, and is bad when it tends to 1. The size of the dot is function of the number of occurrences of this rank value. For computing ranks, the best output is retained for a given model family; thus, these outputs cannot be considered additively.

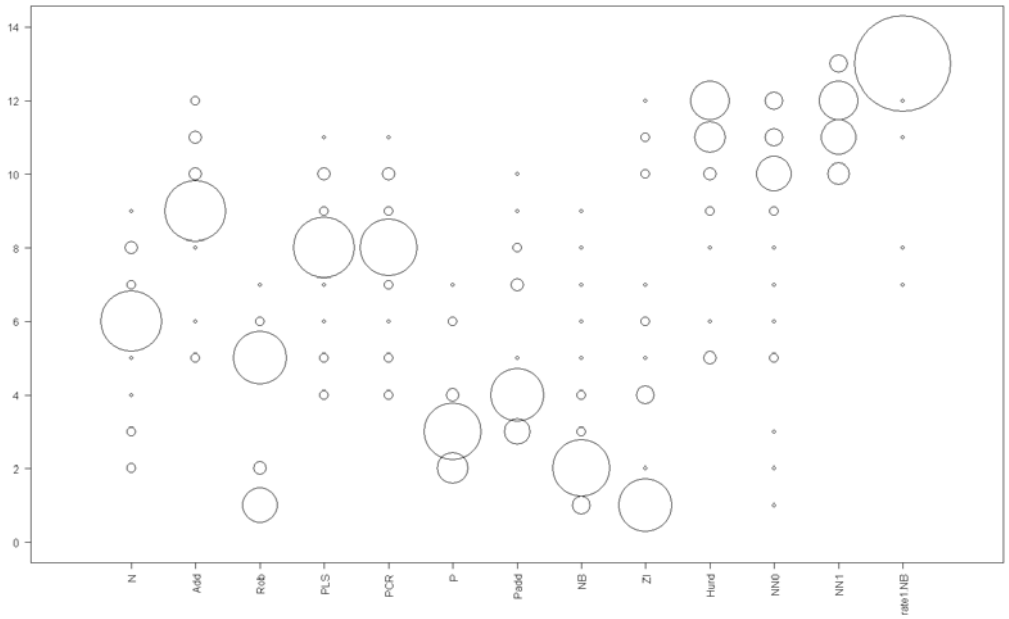


Figure 3a. ***=Corrected AIC of models for richness metrics (AICc) – The graphe shows ranks for each type of model considered; the best the model is, higher is the level; the dot size corresponds to the number of times the level is reached.***

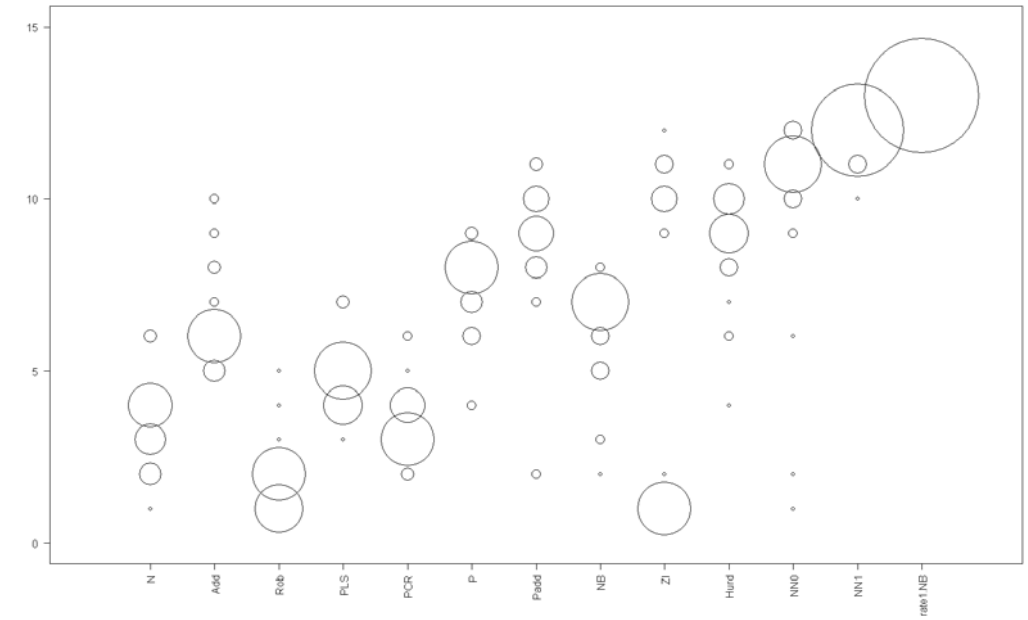


Figure 3b. ***Pearson Correlation coefficient between observed and fitted values for richness metrics (R21)***

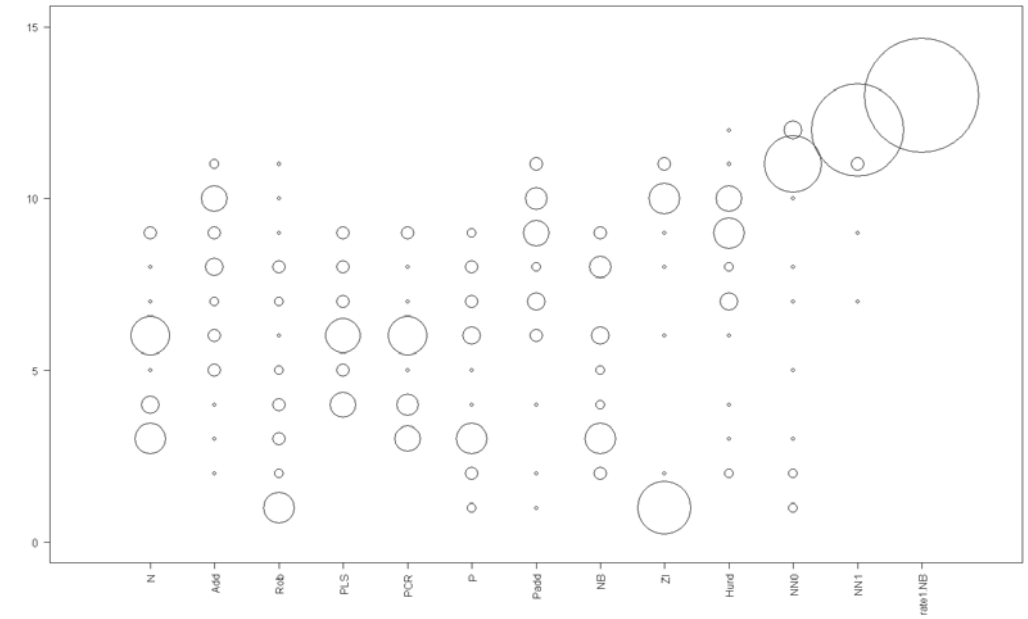


Figure 3c. ***Spearman Correlation coefficient between observed and fitted values for richness metrics (R22)***

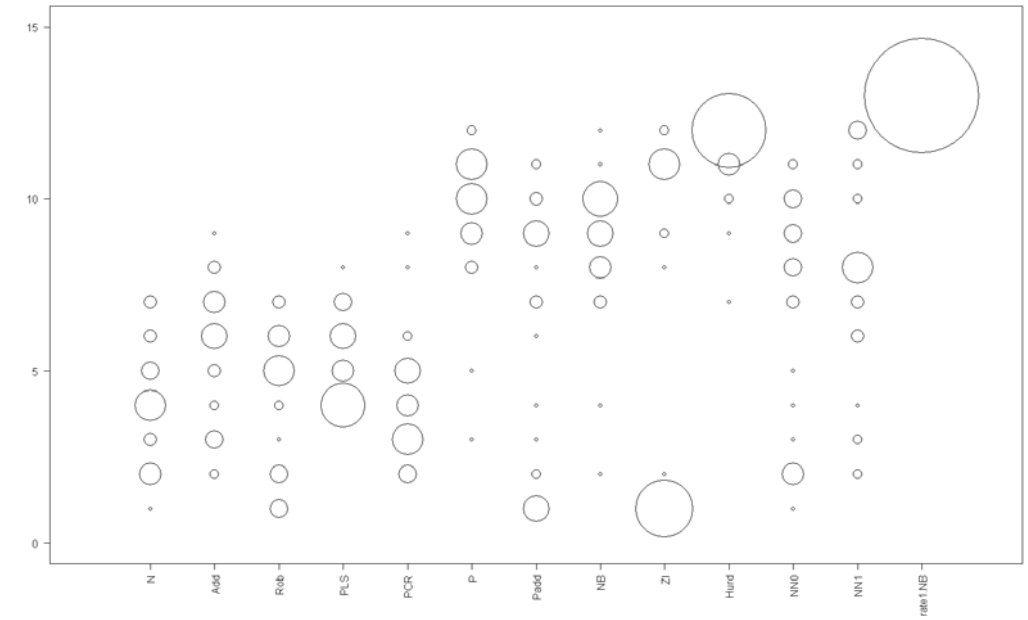


Figure 3d. ***Pearson Correlation coefficient between observed and predicted values for richness metrics (R23)***

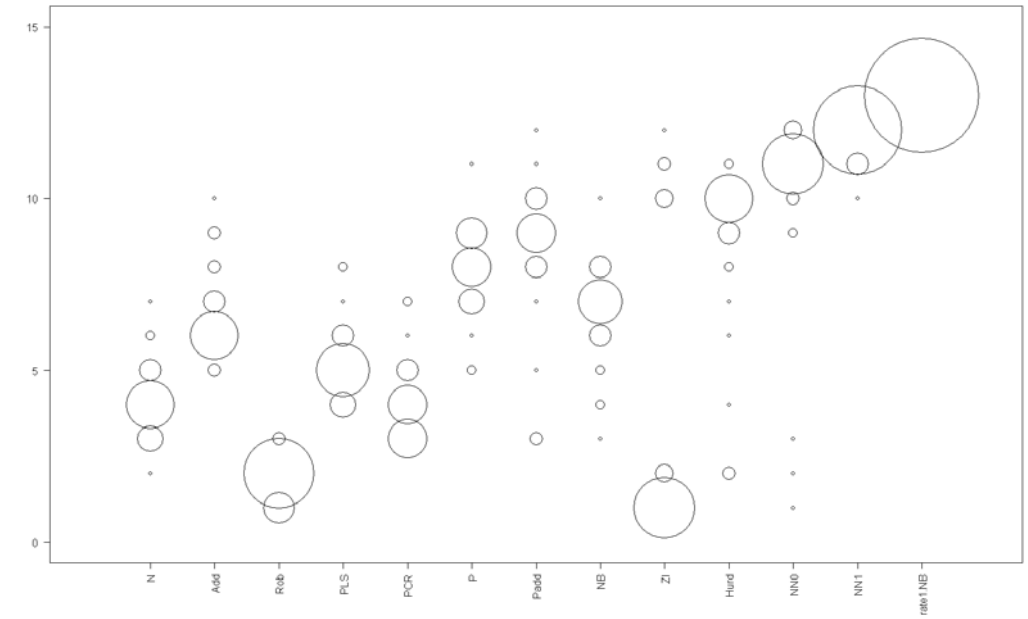


Figure 3e. ***Root Mean Square Errors of Cross Validation for richness metrics (CV)***

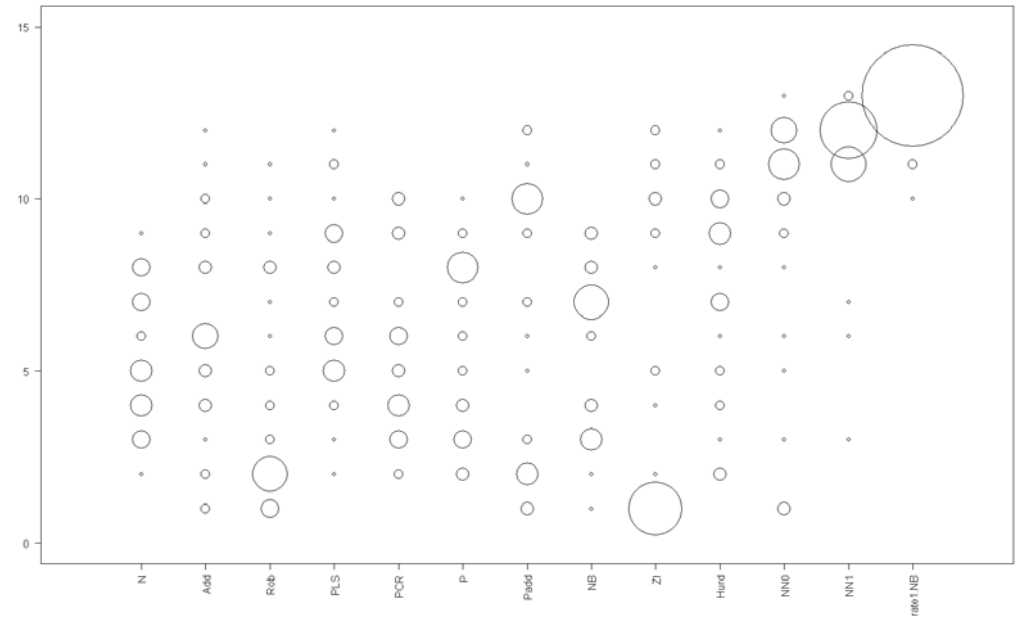


Figure 3f. ***Minimum of Pearson correlation coefficients within bio-ecological regions.***

First, it is noticeable that higher ranks are those of rate models, meaning that these models fit clearly better, but one must of course keep in mind that this approach does not model the same thing.

As regards count models, some points are noticeable:

First, neural networks show good results as regards fit criteria since they have best ranks among count model families; but, as for internal and external validation, they are not always as good. Finally, the difference in the ranks for fit and prediction criteria of the 2 versions of these models are not clear; NN1 models should be better in term of fit whereas NN0 should be less sensible to over-specification. So, the problem with neural network will notably be to find and be sure that parameters – the number of layers, the connection between layers, the decay parameter, etc. – have been well tuned.

Second, zero-inflation and hurdle models' ranks are good, both in term of fit and in term of prediction. If some metrics, those with few zeros, have not been well fitted by zero-inflation models, hurdle models remain rather good, notably as regards external validation. Hurdle models seem then to be a nice alternative to model richness metrics.

Third, Poisson and Negative Binomial families are almost always better than Gaussian models, except when looking at AIC, but their ranks are not as good as those of zero-inflation and hurdle models.

Four, orthogonal regression appears rather better than linear regression on predictors, especially as regards prediction, and PLS is in general better than PCR. For both techniques,

however, the number of components to keep remains difficult to choose, notably to find an optimum in the balance between fit and prediction.

Finally, for both Gaussian or Poisson families, generalized additive models improve often clearly the fit, but is not always as good when looking at prediction since the inclusion of smooth functions can bring to over-specification. Robust regression provides better predictions but only for a few metrics.

Next graphs are the true-value and corresponding ranks for R21 and R23 criteria computed on the same – for both criteria - selected model in each family. The selection was performed in order to have relatively good R21 and R23 in the same time.

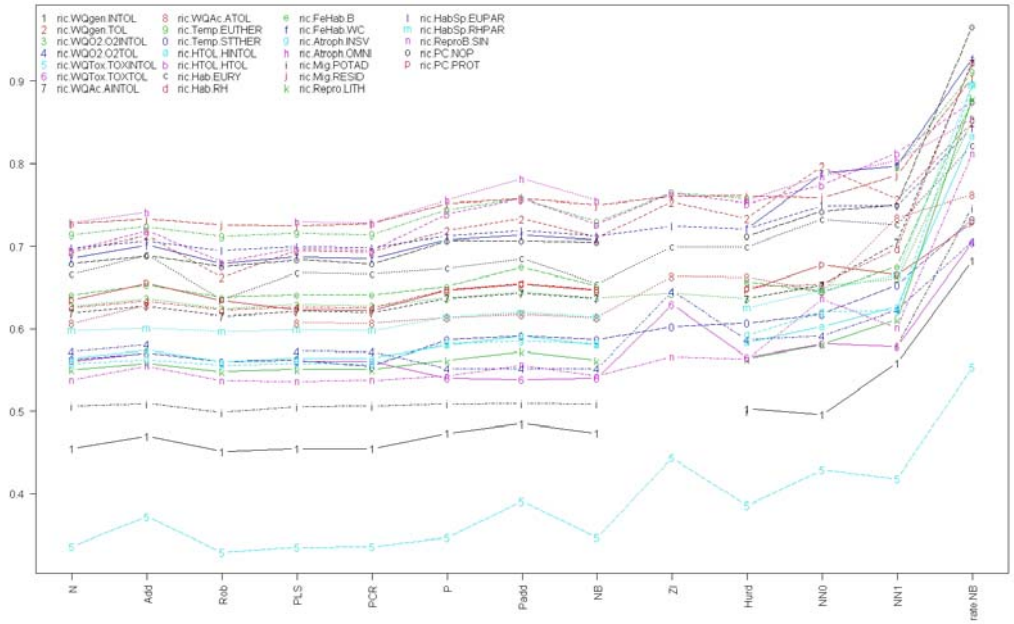


Figure 4a. ***Pearson Correlation coefficient between observed and fitted values for richness metrics (R21)***

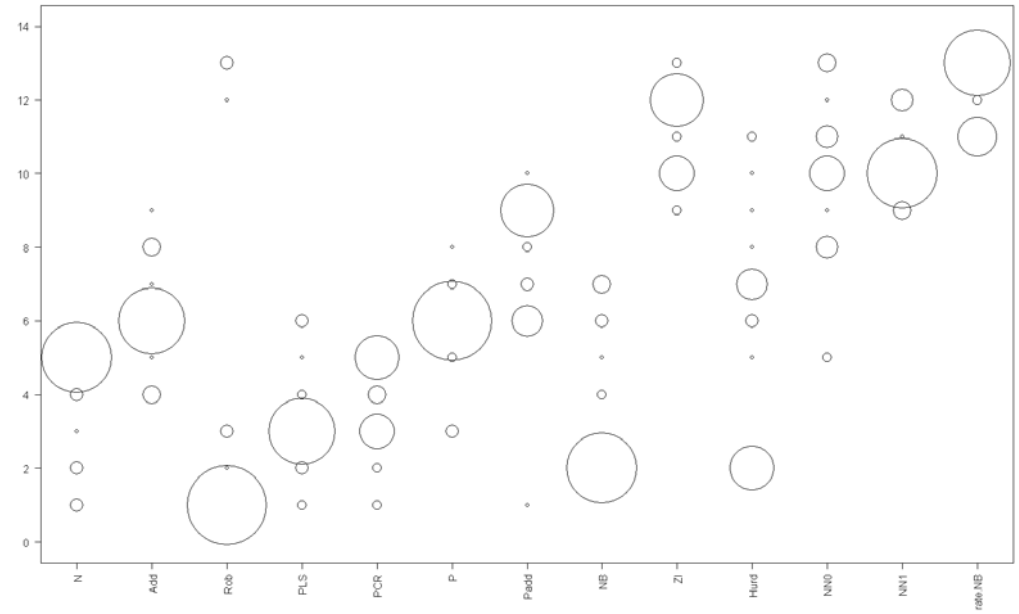


Figure 4b. ***Corresponding ranks for R21.***

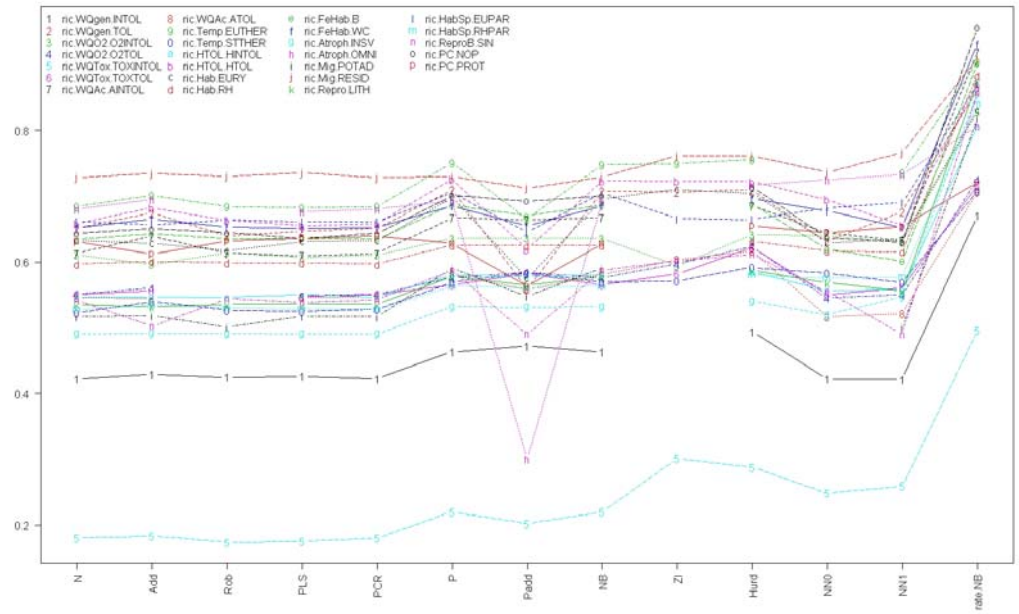


Figure 5a. ***Pearson Correlation coefficient between observed and predicted values for richness metrics (R23)***

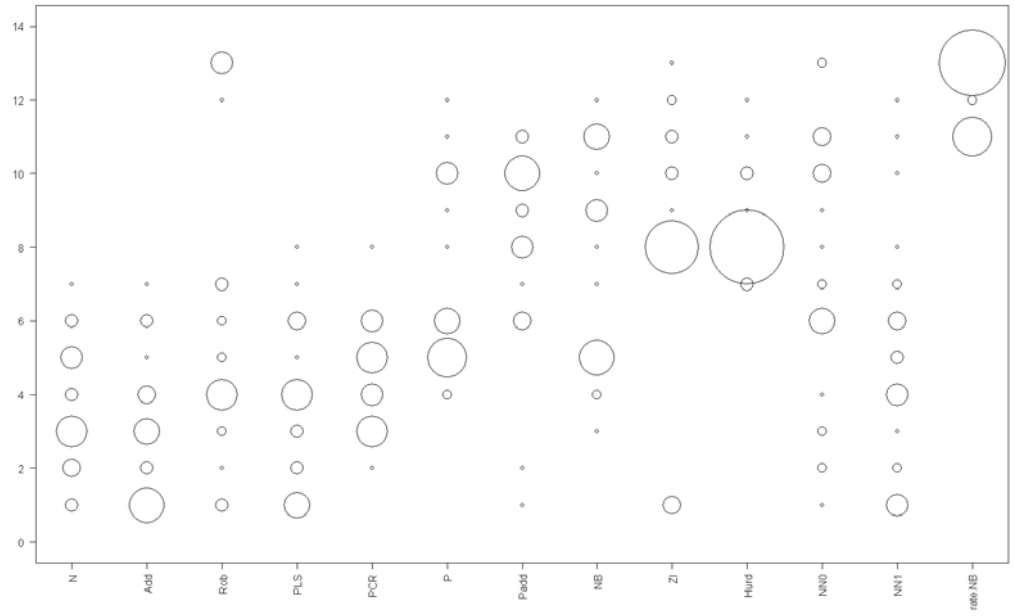


Figure 5b. ***Corresponding ranks for R23.***

8.4.4 Concluding remarks

Of course, rate models fit and predict better, but such models consist in modelling a proportion, which is other information.

This review of some alternative techniques to model count metrics provides a little information on the use of these techniques in this context, that is with zero inflation for some metrics – but not for all, multi-collinearity, atypical points or extremes, etc.

Neural networks show good results but a fine tuning is necessary, notably to avoid over-specification, and this could remain a difficult task, especially when automated procedure are needed.

Zero-inflation and hurdle models show also good outputs in term of fit and in term of prediction, better than those of classical Poisson or negative binomial models. However, for a few metrics, zero-inflation models used to fail during fitting process. Conversely, hurdle models appeared much more robust and may be a good alternative, especially to model richness metrics.

Generalized linear models appear almost ever better than linear models on transformed response. But negative binomial models are rarely better than Poisson models.

Orthogonal regression could be a relevant alternative if a Gaussian model is preferred, since it clearly improves validation outputs but, in this case too, a fine tuning is needed to

choose the number of components to include, which is not easy in an automated script. Using PCR with no a priori, on a fixed set of components could be an option.

Finally, if additive models sometimes improve predictions, these latter can also be worsened in the case of over-specification.

Another criterion which has not been really considered here concerns the variance of predictions, but Akaike criterion has something to do with the precision of model parameters, and then on predictions.

Other approach, such as mixed models or Generalized Estimation Equation (GEE), which was not considered here, could be an interesting way to improve models, notably to take into account possible patterns in scale parameters. But this supposes strong hypothesis, notably because the environment features are themselves highly correlated.

Appendix 1 – Metrics based on densities

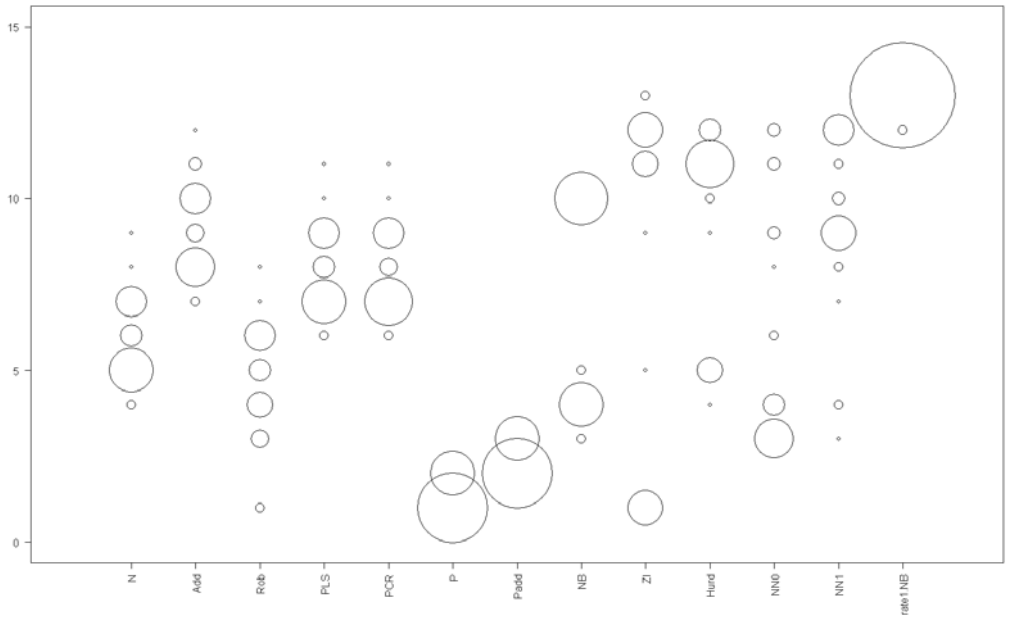


Figure 6a. ***Corrected AIC of models for density metrics (AICc) – The graph shows ranks for each type of model considered; the best the model is, higher is the level; the dot size corresponds to the number of times the level is reached.***

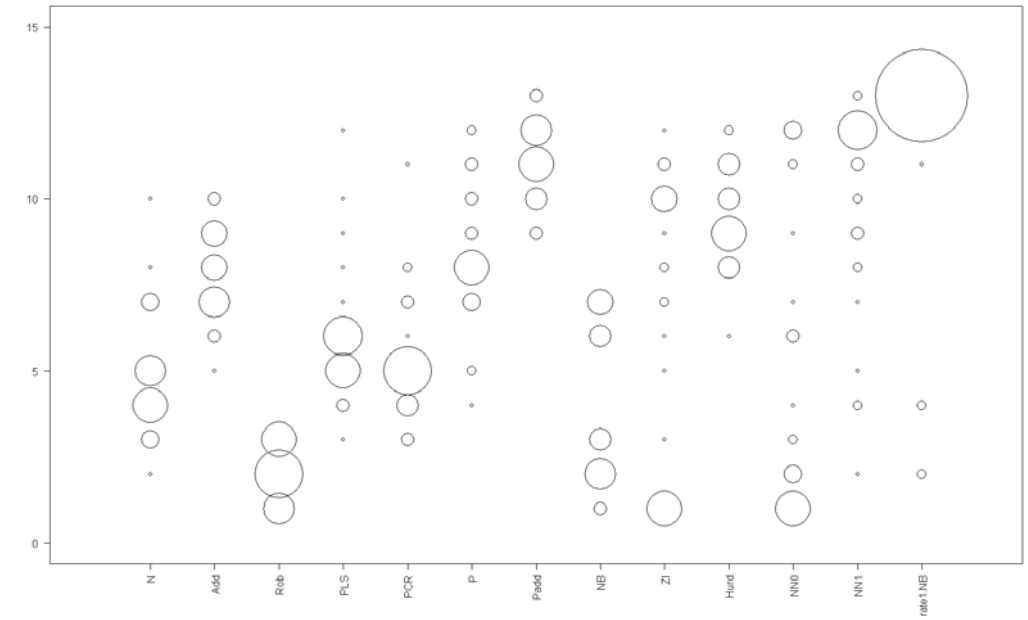


Figure 6b. ***Pearson Correlation coefficient between observed and fitted values for density metrics (R21)***

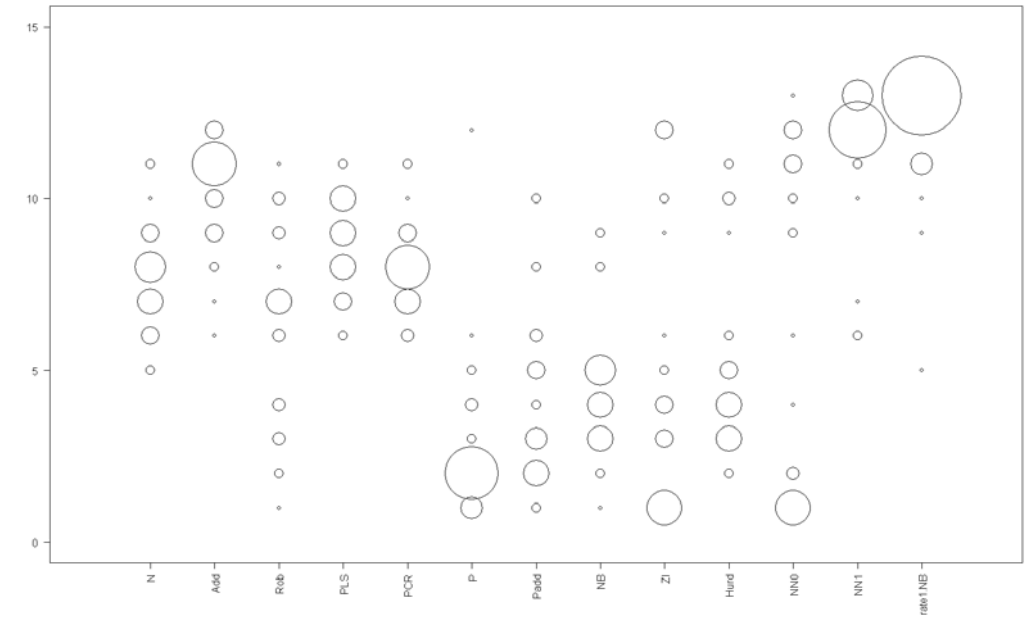


Figure 6c. ***Spearman Correlation coefficient between observed and fitted values for density metrics (R22)***

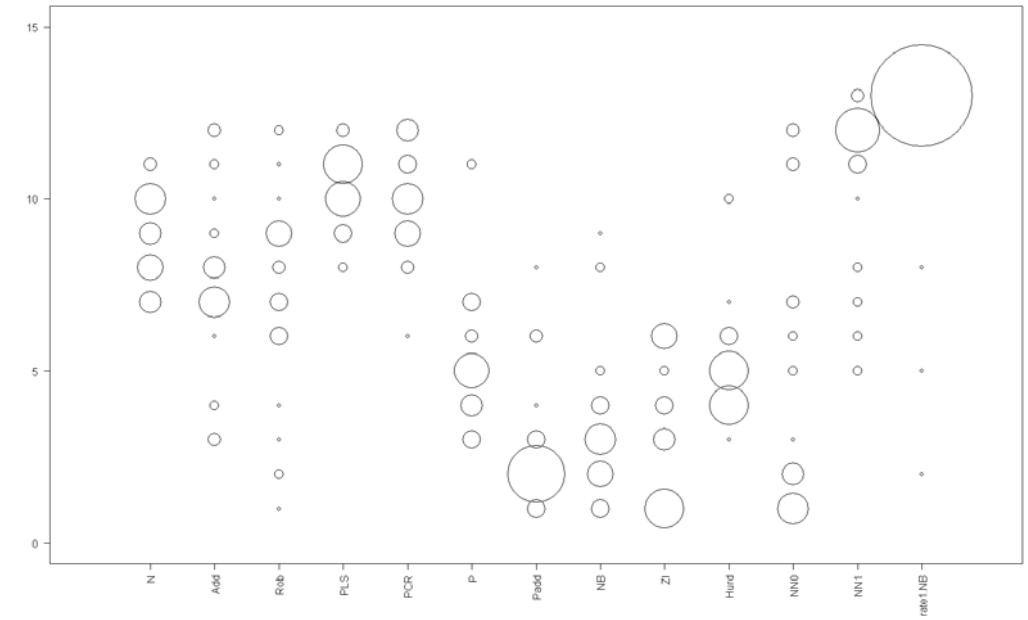


Figure 6d. ***Pearson Correlation coefficient between observed and predicted values for density metrics (R23)***

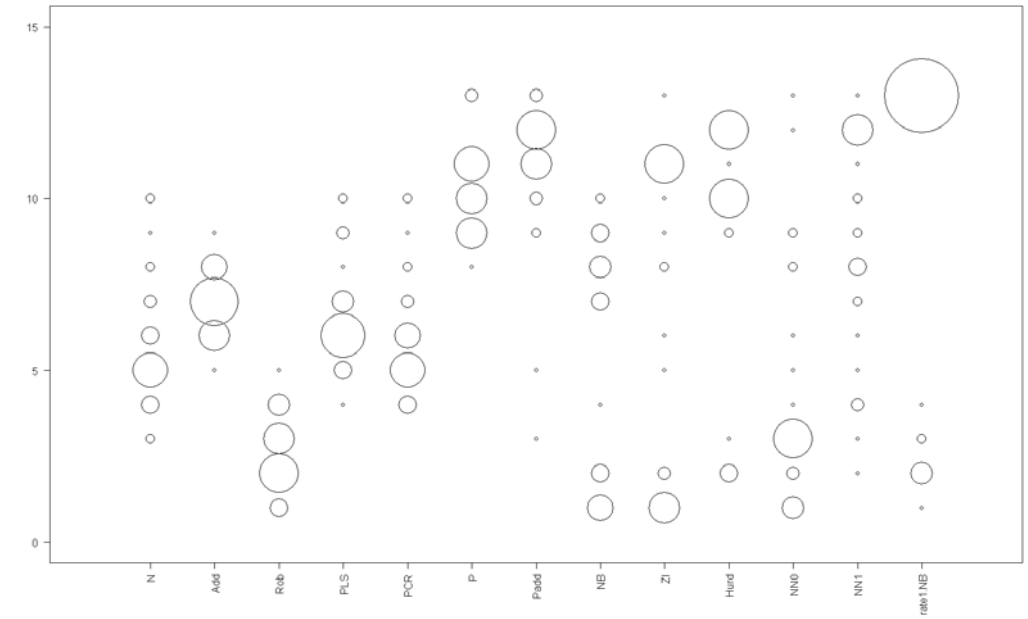


Figure 6e. ***Root Mean Square Errors of Cross Validation for density metrics (CV)***

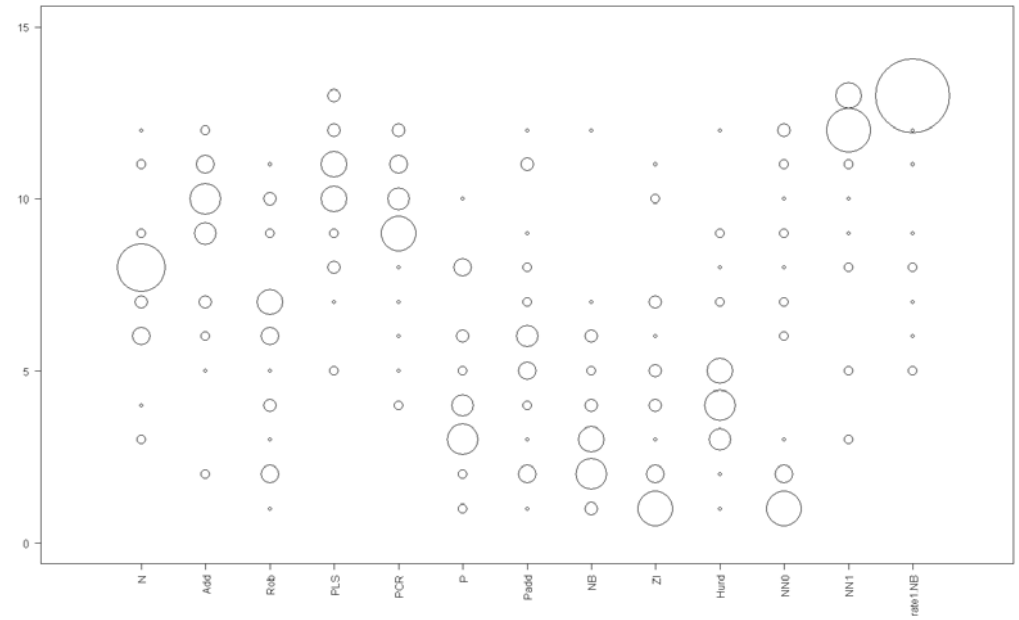


Figure 6f. ***Minimum of Pearson correlation coefficients within bio-ecological regions.***