Material & Methods

Danet, A, Giam, X, Olden, J, Comte, L

28 June, 2022

Contents

1 Fish community time series

We used the RivFishTime database (**comte_rivfishtime_2021?**), a compilation of more than 12,000 time series of riverine fish community abundances, by species, mainly covering western and northern Europe, northern America, and southeastern Australia. We completed the database with time series from Japan and United States, following the same criteria than RivFishTime for integration (**comte_rivfishtime_2021?**).

We selected data to ensure that each time series had at least 5 years of data over a 10 year period as well as a consistent sampling protocol through time. As several sites had been sampled using different sampling methods (e.g. electrofishing, seining) and/or over different periods of the year, we selected for each site only the sampling events that were performed using the most frequent protocol (i.e. the mode) and within 1.5 month of the most frequently sampled month (i.e. 45 days before or after). When there were several sampling events the same year, we selected the sampling that took place at the closest date from the mode

sampled date of the site. We further checked that the reported unit of abundance was homogeneous for each time series.

The data selection resulted in 4476 fish community time series, totalling 46932 sampling events, 326717 species abundance records, and 806 freshwater fish species. The median time span was of 17 years, the median baseline of the time series was 1997, and the median completeness of the time series was of 55%. Finally, . The distribution of those metrics are shown in supplementary figures (Fig. SX-X). The sites were mostly located in Palearctic (75%), Nearctic (20%) and Australasia (5%). Four countries gathered 85% of the sites, namely Great Britain (29%), France (21%), Sweden (18%), and the United States (18%).

2 Biodiversity facets

To evaluate temporal changes in freshwater fish communities, we assessed changes in several biodiversity facets related to species composition, species richness and total abundances.

2.1 Dissimilarity metrics

As changes in community composition can be attributed either to changes of dominance or of rare species, we described temporal dissimilarity with both the complement of the Jaccard index (J) and the Simpson-based dissimilarity index (H_d) (**hillebrand_biodiversity_2018?**). The former index is based on presence/absence and is simply the sum of species gains and losses over the total number of species across two samples (eq. (2.1)). The latter is based on species relative abundances and their variation across two samples (eq. (2.1)). The Jaccard index quantifies the extent of immigration / extinction processes while the Simpson-based dissimilarity index quantifies the extent of change in species dominance (**hillebrand_biodiversity_2018?**).

$$J = \frac{S_{gain} + S_{loss}}{S_{tot}}$$

with S_{gain}, S_{loss}, S_{tot} being the numbers of immigrant, extirpated and total species respectively.

$$H = 1 - H$$
$$H = 1 - \frac{\sum_i (p_i - p'_i)^2}{\sum_i p_i^2 + \sum_i p_i'^2 - \sum_i p_i p'_i}$$

with i: species i, p: relative abundance and $\prime$: the focal community

We further partitioned the Jaccard dissimilarity index into two sets of complementary indices. The first set was Appearance and Disappearance, respectively the proportion of immigrant species (S_{gain}/S_{tot}) and the proportion of extirpated species (S_{loss}/S_{tot}). The second set was Turnover (J_t) and Nestedness (J_n), respectively $J_t = (2 * min(S_{loss}, S_{gain}))/(S_{common} + (2 * min(S_{loss}, S_{gain})))$ and $J_n = 1 - J_t$ (**baselga_betapart_2012?**), S_{common} being the number of species present in both communities. High Turnover values indicate that the changes in community composition result from species replacement, whereas high values of Nestedness indicate species gains or losses from a nested community, i.e. that a community is a subset of the other (**baselga_temporal_2015?**). We took the first year of sampling of a community as the reference community.

2.2 Total abundance

Total abundance was reported in number of individuals (47.00% of the sampling events), density of individuals per 100 m^2 (51.81%), Catch Per Unit Effort (1.05%), and Leslie index (0.14%). Although we

selected for strict protocol consistency, 70% or more of the sampling events by unit of abundance did not reported sampling effort, preventing us to harmonize count, abundance density and Catch Per Unit Effort (**comte_rivfishtime_2021?**).

2.3 Species richness

As sampled species richness is a negatively biased estimator of the "true" species richness, we corrected sampled species richness with the coverage-based rarefaction and extrapolation methodology (**chao_coverage-based_2012?**). The estimated coverage of a sample is respectively positively and negatively related to the number of individuals and number of singletons. For species richness to be comparable across samples, we fixed the coverage of all samples at 98.5%.

As 51.81% of the abundances were measured as density by 100 m^2 and 1.05% as Catch per Unit Effort, we did not have direct access to the number of individuals and number of singletons to perform the computation. In this case, we first divided each species abundance (x_i) by the minimum values of abundance in the community (i.e. $x'_i = 1/min(x_i)$), which we further rounded so that each community had at least one singleton species, i.e. a species with one individual. The correlation was very high between raw species richness and chao richness ($\rho = 0.97$ for both raw variables and log transformed ones, Fig. SXX).

3 Stream gradient and anthropogenic pressures

In dendritic networks, the environmental heterogeneity and connectivity along the longitudinal (upstream-downstream) gradient strongly shape species occurrences, immigration rates and community composition [ref]. To capture this stream gradient, we described stream characteristics at each site by the altitude (m), slope (deg.), average annual discharge ($m^3.s^-1$), distance from source (km), and strahler order that we extracted from the HydroAtlas database (**linke_global_2019?**; **robinson_earthenv-dem90_2014?**). We did so by snapping the sites to the closest stream segment using a one kilometer buffer (99% of the sites). We performed a Principal Component Analysis over the site stream characteristics after log transforming (added absolute minimum values plus one to avoid values ≤ 0, $x'_i = x_i + min(x_i) + 1$) and standardizing all the variables, i.e. centering and scaling . We orthogonally rotated the two first principal components, using the varimax criterion (**revelle_psych_2019?**), to increase the quality of the variable representation (i.e. their loadings) on the two first principal components. The first rotated component, related to average annual discharge, distance from source and Strahler order and capturing 56% of the variance (Fig. S**??**), was used as a composite variable describing the stream gradient from upstream to downstream.

The degree of anthropogenic pressure was quantified using the human footprint index (**venter_global_2016?**; **venter_sixteen_2016?**). The human footprint index aggregates an array of human pressures, including the extent of build environments, population density, electric infrastructures, croplands, pasture lands, roads, railways and navigable pathways. It does so by combining remote sensing data, systematic surveys and modelling from ground data, making it less prone to errors (**venter_global_2016?**). The human footprint range between 0 and 50, with values superior to four being considered in a degraded state (**williams_change_2020?**). To capture both the effects of the legacy of past anthropogenic pressures and its recent changes, we considered the human footprint index computed in 1993 and 2009 (i.e. 16 years span). Specifically, we used the human footprint index of 1993 as a measure of the legacy of anthropogenic pressures and the ratio between the human footprint of 2009 and 1993 as a measure of the recent changes in anthropogenic pressures. The human footprint values were extracted from the HydroAtlas database at the reach scale (original resolution of 450meters, **linke_global_2019?**). In order to obtain interpretable coefficients of recent changes in human footprint, we log-transformed the ratio of human footprint with a base 2. Then, a value of minus one and one represent a division by two and a multiplication by two of the human footprint between 1993 and 2009, respectively. Only 7% of the samplings took place before 1993, while 58% took place between 1993 and 2009 and 34% after 2009. The human footprint indexes 1993 and 2009 were not correlated with the stream gradient (Spearman's ρ, 0.08 and 0.1 respectively).

3.1 Non-native species data

The biogeographic origin of the fish species describing whether species were native or introduced to a given drainage basin was retrieved using the global database of (**tedesco_global_2017?**) (94.4% of the species occurrences). For the sites falling outside of the river basins provided in (**tedesco_global_2017?**), such as for the sites located close to the shore, we used the closest basin in the country (mean distance = xxx). For species not included in a given drainage basin, we determined the origin of the species at the country scale using Fishbase (**froese_fishbase_2021?**) (5.5% of species occurrences). Given the extent of the United States, we specifically completed the global database with the Nonindigenous Aquatic Species (NAS) database developed by the U.S. Geological Survey (https://nas.er.usgs.gov/), at the state scale (0.05% of the species occurrences). Finally, we completed the remaining species origins at the country scale, using national atlases and Fishbase data in neighboring countries, such as for *Piaractus brachypomus* and *Rutilus rutilus* in the United States (0.1% of the species occurrences).

We then estimated the percentage of non-native species for each sampling events, both for species richness and abundances.

4 Statistical analysis

4.1 General statistical model

- Lise suggests to present the drivers before the equations
- A proposition:
  - Present the philosophy: response versus drivers: why not all interactions? justify them
  - Transformation of the variables

We modelled the temporal trends of the different biodiversity facets (Y) as dependant of time ($\beta_0 Time_t$, eq. (4.1)) measured as the number of years since the beginning of the sampling at each site with $t_0 = 0$, the stream gradient measured by the rotated PCA axis over stream characteristics, the legacy of past anthropogenic pressures measured by the human footprint index of 1993, and the recent changes in anthropogenic pressures measured by the ratio between the human footprint index of 2009 and 1993. We included all the predictors as main effects ($\sum_{k=1} \beta_k X_k$) to capture the differences in biodiversity facets attributed to spatial effects of the ecological drivers. We further included interactions between time and the ecological drivers ($\sum_{k=0,l\neq k} \beta_{kl} X_k X_l$) to test how stream gradient and anthropogenic pressures affect the temporal trends in biodiversity facets. Finally, we included the triple interactions between time and the pairs of other ecological drivers ($\sum_{k=0,m\neq n\neq k} \beta_{kmn} X_k X_m X_n$) to test for the presence synergistic or antagonistic effects of the stream gradient and anthropogenic pressures on the temporal trends in biodiversity facets.

Furthermore, the statistical model (eq. (4.1)) was adapted according of the nature of the variables, namely total abundance and dissimilarity metrics. For total abundance, we added the measurement unit of abundance as a categorical variable both as a main effect and in interaction with time (**van_klink_meta-analysis_2020?**). The temporal trends in the main text and supplementary are adjusted for raw count, i.e. the reference factor level in modelling was raw count. Because dissimilarity metrics at each site was 0 at the beginning of the time series, we fixed the intercept was fixed to zero. The dissimilarity metrics were relative to the site and bounded between 0 and 1, we did not expect average difference in dissimilarity which are not due to differences in the temporal trends. Then, we removed the main effect of ecological drivers ($\sum_{k=1} \beta_k X_k$ term from eq. (4.1)), but kept their effects on the temporal trends.

We accounted for the spatial structure of the data by adding random effects to the intercept (α) and the slope of the temporal trends (β_0) on the basin identity (n) and on the site identity, nested in basin ($i|n$). The random effects and the error terms were modelled as a Normal distribution of mean 0 and variance (σ^2).

$$Y_{i|n,t} = \alpha + \beta_0 Time_t + \sum_{k=1} \beta_k X_k + \sum_{k=0,l\neq k} \beta_{kl} X_k X_l + \sum_{k=0,m\neq l\neq k} \beta_{klm} X_k X_l X_m + \epsilon_{i|n,t}$$

- $\alpha = \alpha_0 + a_n + a_{i|n}$
- $\beta_0 = \mu + b_n + b_{i|n}$
- $k, l, m \in [1, 2, 3]$: ecological drivers including stream gradient, legacy of past and recent changes in anthropogenic pressures
- $a_n, a_{i|n}, b_n, b_{i|n}, \epsilon_{i|n,t} \sim \mathcal{N}(0,\, \sigma^2)$
- n: basin, i: site i, t: time t
- suggestion:
  - How model abundance and dissimilarity metrics?
  - Choice of response distribution?
  - Variable transformation

All the response variables were modelled with a Gaussian distribution following previous studies modelling temporal trends of community composition, species richness and total abundance at the global scale (**dornelas_assemblage_2014?**; **blowes_geography_2019?**; **van_klink_meta-analysis_2020?**). Other error structures might be more appropriate to model response variables bounded between 0 and 1 and representing ratio of discrete numbers such as the dissimilarity metrics and the percentage of non-native species, doing so allows to obtain easily interpretable coefficients across all biodiversity facets (e.g. temporal trends are not interpretable as rates of change when modelled using a logit scale such as when using a beta distribution). In addition, (**blowes_geography_2019?**) previously found that slope coefficients estimated with a gaussian error and a beta error had a spearman correlation superior to 0.90 and give qualitatively similar results. We therefore believe that this choice is not likely to alter our conclusions.

4.2 Variable transformations

We log-transformed the number of years as $\log(year + 1)$ as it improved the quality of the model fitting to the data, decreasing the AIC by 29% in average from 0% for total abundance to 124% for the Simpson based dissimilarity index (Table S**??**) [missing chao richness and percentage of exotic species in the comparison + refaire avec INLA?]. It suggests the presence of non-linearity in the temporal trends, which is particularly expected in the case of bounded variables such as the dissimilarity metrics. Then the rate of change per decade of a given biodiversity facet is found for a time value of $\log(10 + 1)$, i.e. 2.4 [Point to the model prediction plots with double abscissa scale (log + year)]. We further log-transformed with a base 2 the recent changes in anthropogenic pressures quantified by the ratio of human footprint index between 2009 and 1993, i.e. $\log_2(HFT_{2009}/HFT_{1993})$. Then, a value of minus one and one represent respectively a division by two and a multiplication by two of the human footprint between 1993 and 2009. We log-transformed total abundance and Chao species richness, then the temporal trends are expressed in percentage by unit of time (**van_klink_meta-analysis_2020?**).

[Add a table in supplementary with all the transformations of response and explicative variables.]

4.2.1 Variable standardization

In order to compare the magnitude of the effects of time, stream gradient, and anthropogenic pressures among biodiversity facets and to compare the magnitude of the effects of the predictors, we scaled both biodiversity facets and the predictors by their standard deviation prior to the model evaluation.

As our models contain interactions, the individuals slope coefficients can be difficult to interpret without centering the predictors around ecological relevant values (**gelman_scaling_2008?**). As an example, the average temporal trends estimated by β_0 in eq.(4.1) can only be interpreted when all the $X_k = 0$. Without centering, it means that β_0 is interpretable when the legacy of anthropogenic pressures, measured by human

footprint at the year 1993, and the stream gradient, measured by the values of the rotated PCA axis over stream characteristics, are equal to 0. Hence, we centered the human footprint at the year 1993 and the coordinates along the PCA axis around their average values. Recent changes in anthropogenic pressures, measured by the ratio between the human footprint index of 2009 and 1993, was not centered as 0 values indicate no recent changes in anthropogenic pressures. Time variable was not centered either because then the main effects of the ecological drivers ($\sum_{k=1} \beta_k$) are then interpreted as a baseline effect, i.e. when time is equal to 0.

4.2.2 Model evaluation and confidence intervals

The models were evaluated in a Bayesian framework using Integrated Nested Laplacian Approximation (INLA), which approximates the posterior distribution of the parameters and then do not rely on Markov chains and Monte Carlo simulations, and then is a computationally efficient method to evaluate Bayesian models (**rue__approximate__2009?**; **rue__bayesian__2017?**). When estimating conjointly the temporal trends at multiple location, a advantage of the Bayesian approach is the estimation of credible intervals around the temporal trends estimated at each location. A second advantage is that it allows the computation of credible intervals, which translates in the probability that an unobserved parameter falls in a given interval, to the difference with frequentist approach (**greenland__statistical__2016?**). We computed the credible intervals at 80%, 90% and 95% using Highest Posterior Density method (**hyndman__computing__1996?**). When 0 is outside the credible intervals of the coefficients at 80%, 90% and 95%, they can respectively be interpreted as weak, moderate and strong evidence of an effect (**van__klink__meta-analysis__2020?**; **mastrandrea__guidance__2010?**).

4.2.3 Bayesian priors

INLA models were evaluated with defaults uninformative priors. The prior distribution of fixed coefficients followed a flat zero centered normal distribution ($\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(0, 1000)$). The prior distribution of the random effects and the gaussian error (ϵ_{it}, eq. (4.1)) followed a log gamma distribution with shape and inverse scale parameters ($\mathcal{G}(s, \tau) = \mathcal{G}(1, 5.10^{-}5)$). We there had to back-transformed the estimated coefficients in order to obtained the standard deviations attributed to the random effects and the gaussian error ($\sigma = 1/\sqrt{\tau}$). We checked that the slope coefficients, random effects and the temporal trends by basin and site were the same than with a implementation in frequentist. Then, we concluded that the quality of parameter inference did not suffer from the uninformative priors.

4.2.4 Model validity

We checked the model validity visually by plotting the fitted versus the observed values [ref plot].

- Add plots of posterior marginal distribution of the parameters? (it would be like 8*10 plots for the fixed effects only)
- Add PIT and CPO (**van__klink__meta-analysis__2020?**)

4.3 Assessing the dimensionality of temporal community changes

We performed a PCA over the temporal trends of biodiversity facets at the site level. A PCA finds the linear combination of variables that explained the most variance and separates linearly uncorrelated variables. In complement, we performed a clustering analysis to group together sites that displayed similar temporal community changes. The temporal trends for each biodiversity facet and site were extracted by the Best Linear Unbiased Prediction method. Prior to the analysis, we adjusted the temporal trends of total abundance for the unit of measurement with the main effect coefficients and the one in interaction with temporal trends. We did not include the temporal trends of the variables describing composition of non-native species as the

they are part of biodiversity (**schlaepfer_non-native_2018?**) and site temporal trends in percentage of non-native species displayed very little variation, then it was of little interest to include them in the dimensionality analysis.

We performed the trimmed k-means clustering method (**fritz_tclust_2012?**), a robust clustering method because it avoids the identification of spurious clusters. The method consists of trimming the α most outlying data while taking in account the multidimentional structure, the number of dimension being the number of the biodiversity facets. To choose a relevant number of clusters, we plotted the trimmed log-likelihood of the function as a function of the proportion of the most outlying data trimmed (*alpha*) [ref plot]. We thus selected a partition of temporal community changes in six clusters with $\alpha = 5\%$. We did not constraint the algorithm for the relative size of shape of the clusters, as we had no apriori expectation about them. The clustering algorithm was run for a minimum number of one hundred iterations and up to 125. To further control for the quality of each fish community changes assignment to a given cluster, we discarded any fish community for which the second best cluster assignment is 50% better, we did so by comparing the degree of affiliation to the clusters (**fritz_tclust_2012?**). The clustering was performed using `tclust` R package.

5 Reproducibility, ROBITT assessment and open science statement

The manuscript and the supplementary materials are written in Rmarkdown, i.e. combining code and text, and are available on github. We further implement a code pipeline using the `targets` R package to ensure that all the code, the data, the figures, the manuscript and the results are up to date.

- todo: Robitt assessment

References