Material & Methods

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

12 mai, 2022

Contents

1 Dataset and data selection

The RivFishTime database (Comte et al. 2021) describes the time series of riverine fish communities at large scale, mainly covering western and north of Europe, North America, Japan and Australia. It currently holds 781169 occurrence of 1093 freshwater fishes species over 119962 sampling events distributed over 13549 sites. The minimum time span of a timeseries in the dataset is 10 years.

We selected data to guarantee the quality and the internal homogeneity of the timeseries. In a given site, we selected the sampling events that used the most frequently used protocol (i.e. the mode) and the most reported measurement unit of abundances. We selected the samplings that reported the year, but also the month or the quarter of the year (i.e., January to march, April to June, etc.). For each site,
we selected only the samples that were realized within 1 month around the most frequently sampled month, i.e. one month before or after. When there were several samplings the same year, we selected the sampling that took place the closest of the mode month/quarter of the site. Finally, we kept sites where the number of samplings was equal or superior to five and that had a time span superior of equal 10 years.

The resulting selection of data contains 361093 occurrence of 806 freshwater fishes species over 59904 sampling events distributed over 5359 sites. The mean time span was of 18 years, the median completeness of the timeseries was of 55%. Finally, the median baseline of the timeseries was 1996. The complete distribution of those metrics are shown in supplementary figures (Fig. SX-X). The sites were mostly located in Palearctic (79.2%), Nearctic (16.7%) and Australasia (4.0%). Four countries gathered 88% of the sites, namely Sweden (32%), Great Britain (24%), France (17%) and the United States (15%).

2 Characterise Biodiversity facets

Biodiversity facets were characterised by dissimilarity metrics, species richness and total abundances.

2.1 Dissimilarity metrics

As changes in community composition can be attributed either to abundant or rare species, we described dissimilarity with both the complement of Jaccard (J) and the Simpson-based dissimilarity index (H_d) (Hillebrand et al. 2018). The former index is based on presence/absence and is simply the sum of immigrant and lost species 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)). Large trends for the Jaccard and Simpson-based dissimilarity index indicates that changes in species composition are mainly due to change in species dominance and identity (Hillebrand et al. 2018).

$$J = \frac{S_{imm} + S_{lost}}{S_{tot}}$$

with S_{imm}, S_{lost}, S_{tot} being the numbers of immigrant, lost 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

In order to characterise the most precisely the changes in community composition, we further partitioned the Jaccard dissimilarity index in two sets of complementary indices. The first set was appearance and disappearance, respectively the number of immigrant species ($\frac{S_{imm}}{S_{tot}}$) and the number of disappearing species ($\frac{S_{lost}}{S_{tot}}$). The second set was Turnover and Nestedness (Baselga and Orme 2012), respectively ($J_t = \frac{2 * min(S_{lost}, S_{imm})}{S_{common} + (2 * min(S_{lost}, S_{imm})}$) and ($J_n = 1 - J_t$). High turnover values indicate that the changes in community composition are hold by species replacement, while high values of Nestedness indicate species gain or losses from a nested community, i.e. that a community is a subset of the other (Baselga, Bonthoux, and Balent 2015).

2.2 Total abundance

Total abundance was reported in count (36.83% of the sampling events), abundance density per 100 m^2 (62.24%), Catch Per Unit Effort (0.82%), and Leslie index (0.11%). 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.

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 (Chao and Jost 2012). The coverage of a sample depends both on the number of individuals and number of singletons. We fixed the coverage of all samples to 98.5%.

Since 62.24% of the abundances were measured as density by 100 m^2 and 0.82% by Catch per Unit of Effort, we did not direct have access to the number of individuals and number of singletons. In this case, we divided each species abundance (x_i) by the minimum values of abundance in the community (i.e. $x'_i = \frac{1}{min(x)}$) and

then rounded the obtained abundances. Then, those communities had at minimum one singleton species. The correlation was very high raw species richness and chao richness ($\rho = 0.97$ for both raw variables and log transformed ones).

3 Environment and anthropogenic pressures

Stream characteristics such as water flow, size of the streams, and position in the dentritic network constitute tremendous constraints on fishes, selecting the body shape of the fishes [ref], their life history traits [ref], immigration rates [ref] of community and so ultimately community composition [ref]. Such stream characteristics are often collinear and can be coined as "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 et al. 2019; Robinson, Regetz, and Guralnick 2014). We did so by snapping the sites to the closest stream segment at the condition that distance of the site to the closest stream segment of the HydroAtlas database was inferior or equal to one kilometer (99% of the sites). We performed a PCA over the site stream characteristics after log transformation (added absolute minimum values plus one to avoid values ≤ 0, $x'_i = x_i + min(x_i) + 1$) and standardization, i.e. centering and scaling and log transformation of the variables. We rotated the two first principal components with the varimax algorithm to increase the quality of variable representation on the two first principal components. The first rotated component was related to average annual discharge, distance from source and strahler order (Fig. S**??**), the second rotated component was related to stream altitude and slope. Then the first rotated component was used as a composite variable characterising the stream gradient from upstream to downstream.

Anthropogenic pressure was quantified with the Human footprint index (Venter et al. 2016a, 2016b). The human footprint aggregates human pressures namely as the extent of build environments, population density, electric infrastructure, croplands, pasture lands, roads, railways and navigatable pathways. It does so by combining remote sensing data and systematic surveys and modelling from the bottom up, making it less prone to errors (Venter et al. 2016a). The human footprint range between 0 and 50, values exceding four being considered as a degraded state (Williams et al. 2020). A study showed that the human footprint is related to the extinction risk of mammals at world scale (Di Marco et al. 2018). Two measures of human footprint exists, in 1993 and in 2009 (i.e. 16 years span), enabling to characterise both the legacy of Anthropogenic pressure and its recent changes. We took the human footprint of 1993 as a measure of the legacy of anthropogenic pressure and the ratio between the human footprint of 2009 and 1993. Only 9% of the samplings took place before 1993, while 58% took place between 1993 and 2009 and 58% after 2009. The human footprint values were extracted from the HydroAtlas database (Linke et al. 2019).

The spatial resolution of human footprint and of the variables characterising the stream gradient data was respectively 1 km and 3 arc degree. We extracted HydroAtlas values at the reach level, which have a length of 4.2 km in average (Linke et al. 2019).

3.1 Exotic species data

The origin status of the fish species were characterised using a global database at the basin scale (Tedesco et al. 2017), for 94.8% of the species occurrence. The sites that were not perfectly included in a basin, such as the sites located close to the shore, were matched to the closest basin in the country. For species missing of the database, we determined the origin status of the species at the country scale using fishbase (Froese, Pauly, and others 2021) (5.0% of species occurrence.). Given the extent of the United States, we specifically completed the global database with the Nonindigenous Aquatic Species (NAS) database of the USGS https://nas.er.usgs.gov/, at the state scale (0.04% of the species occurrence). Finally, we completed the remaining species status by hand (0.1% of the species occurrence) at the country scale, such as *Piaractus brachypomus* and *Rutilus rutilus* in the United States for example.

We then characterised the composition in term of exotic species by the percentage of species richness being exotic and the percentage of community abundance attributed to exotic species.

4 Statistical analysis

4.1 Temporal trends assessment

We modelled the temporal trends of biodiversity facets with hierarchical linear models. All the response variables were modelled with a Gaussian distribution as previous similar studies modelling temporal trends of community composition, species richness and total abundance at global scale (Dornelas et al. 2014; Blowes et al. 2019; Klink et al. 2020). Particularly for the variables bounded between 0 and 1, namely here dissimilarity metrics and percentages of exotic species, Blowes et al. (2019) 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. Furthermore, the response variables are modelled on logit scale (i.e. $\log \frac{p}{1-p}$) when assuming a beta error distribution and are therefore interpretable at logit scale, i.e. temporal trends are not interpretable as rates of change (Blowes et al. 2019).

The value of each biodiversity facet (Y) at a site i and time (t) was modelled as a function of time ($\beta_0 Time_t$), the covariates ($\sum_{k=1} \beta_k X_k$), the interaction between time and covariates ($\sum_{k=0,n\neq k} \beta_{kn} X_k X_n$), and the double interaction among time and the covariates ($\sum_{k=0,m\neq n\neq k} \beta_{kmn} X_k X_m X_n$) plus an error term ($\epsilon_{it}$). We added random effects on the intercept (α) and the slope of the temporal trends (β_0) such as they had random variations added to a main value (α_0 and μ respectively). Random variations were added as dependant on the basin identity (n) and on the site identity (o), nested in basin. The random effects and the error terms were modelled as a Normal distribution of mean 0 and variance (σ^2).

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

- $\alpha = \alpha_0 + a_n + a_o$
- $\beta_0 = \mu + b_n + b_o$
- $a_n, a_o, b_n, b_o, \epsilon_{it} \sim \mathcal{N}(0,\, \sigma^2)$
- n: basin, o: site within basin
- i: site i, t: time t

Little cleaning to do in the notation: $Y_i|n,t and i|n instead of o ?

The general model includes time and the ecological drivers, namely the PCA axis related to the stream gradient, the Human footprint 1993, and the ratio of the Human footprint between 1993 and 2009 as main effects. We modelled the interactions among time and the ecological drivers to test how the latter covariates affect the temporal trends of biodiversity facets. Finally, we included the triple interactions among time and the pairs of other ecological drivers to test for the presence synergistic or antagonistic effects of the ecological drivers on the temporal trends of biodiversity facets.

In all the models, time was quantified as the number of years since the begining of the sampling at each site with $t_0 = 0$. We log-transford the number of years as $\log(year + 1)$ to signicantly improved the quality of the fit of the model on the data. We indeed found that log-transformed number of year increased
the AIC by 29% in average from 0% for total abundance to 124% for the abundance based dissimilarity index (Table S**??**) [missing chao richness and percentage of exotic species in the comparison + refaire avec INLA?]. It indicates the presense of non-linearity in the temporal trends, which is particularly expected in the case of bounded variables such as the dissimilarity metrics. In this situation, logging the predictor is a convenient way to linearise the relationship. Then the rate of change of a given biodiversity facet for a decade is found for a time log-transformed value of 2.4 approximatively (i.e. $\log(10 + 1)$) [Point to the model prediction plots with double abscisse scale (log + year)]. In order to obtain interpretable coefficients of recent changes in human footprint, we log-transformed the ratio of human footprint with a base 2 (i.e. $\log_2(\frac{HFT_{2009}}{HFT_{1993}})$). 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.

4.1.1 Statistical Models

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

We logged transformed total abundance and chao species richness in order to get temporal trends expressed in percentage by unit of time (Klink et al. 2020). Furthermore, the general statistical model (eq. (**???**)(eq:gen)) 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 (Klink et al. 2020). The temporal trends in the main text and supplementary are presented for raw count, i.e. the reference factor level in modelling was raw count.

The value of dissimilarity metrics at each site was 0 are the begining of the temporal trends and the values were all bounded between 0 and 1. Then, we fixed the intercept was fixed to zero. Because the values of dissimilarity metrics are relative to the site, we do not expect average difference in dissimilarity which are not due to differences in the temporal trends. Then, we kept only the ecological driver effects as interactions with time, i.e. we removed the $\sum_{k=1} \beta_k X_k$ term from eq.(4.1).

4.1.2 Variable standardization

In order to compare the magnitude of the temporal trends among biodiversity facets and to compare the magnitude of the effects of the ecological drivers, we scaled of the variables by their standard deviation prior to the model evaluation.

As our models contains interactions, the individuals slope coefficients can be difficult to interpret without centering explicative variables around ecological relevant values (Gelman 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 values of PCA axis related to the stream gradient, 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 two human footprint ($\log_2(\frac{HFT_{2009}}{HFT_{1993}})$), 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.

4.1.3 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. We used default priors and checked that we obtain the same estimation of slope coefficients, random effects and the same estimation of temporal trends by basin and site (obtained by Best Linear Unbiased Prediction) than with a implementation in frequentist with glmmTMB. When estimating conjointly the temporal trends at multiple location, a advantage of the bayesian approach over the frequentist one is that it allows for the estimation of confidence intervals around the temporal trends estimated at each location. A second advantage is that it allows the computation of credible intervals at 80%, 90% and 95%. 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 (Klink et al. 2020). We specifically computed the credible intervals using the Highest Posterior Density method, which fix the interval in the presence of asymmetric posterior distribution of the parameters [ref].

4.1.4 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 ($\sigma = \frac{1}{\sqrt{\tau}}$). As the estimation of the main effects and the random effects are of the same with INLA than in frequentist in glmmTMB [show the table!], we concluded that the quality of parameter inference did not suffer from the informative priors.

4.1.5 Model validity

We checked the model validity visualy by plotting the fitted values versus the observed [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 (Klink et al. 2020)

4.2 PCA and clustering of the temporal trends of biodiversity facets

We performed a PCA over the temporal trends of biodiversity facets at the site level in order to characterise the dimentionality of community changes. In complement of this analysis, we performed a cluster analysis to isolate and illustrate different groups of community changes. The temporal trends of biodiversity facets at site scale 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. Furthermore, we did not include the temporal trends of the variables describing composition of exotic species.

We performed the clustering using k-means and data trimming (Fritz, García-Escudero, and Mayo-Iscar 2012), this method is part of the family of robust clustering methods. The method consists on trimming the α most outlying data while taking in account is multidimentional structure. To choose a relevant number of clusters, we plotted the trimmed log-likelihood of the function in function of the proportion of the most outlying data trimmed [ref plot]. We finally selected six
clusters. We did not contraints the algorithm for the size the clusters, as we had no apriori expectation about them and we used a restricting factor. We started with 100 iterations and set the maximum to 125. To further control for the quality of the site assignment on a cluster, we set cluster assignment as NA if another cluster assignment would have been more 50% better. The clustering was performed using `tclust` R package.

5 Reproducibility, ROBITT assessment and open science statement

All the code to reproduce the analysis from the raw data to the manuscript is available on github. We further implement a pipeline using the `targets` R package to ensure that all the code, the data, and the results are up to date.

- todo: Robitt assessment

References

Baselga, Andrés, Sébastien Bonthoux, and Gérard Balent. 2015. "Temporal Beta Diversity of Bird Assemblages in Agricultural Landscapes: Land Cover Change Vs. Stochastic Processes." *PLOS ONE* 10 (5): e0127913. https://doi.org/10.1371/journal.pone.0127913.

Baselga, Andrés, and C. David L. Orme. 2012. "Betapart : An R Package for the Study of Beta Diversity: *Betapart Package*." *Methods in Ecology and Evolution* 3 (5): 808–12. https://doi.org/10.1111/j.2041-210X.2012.00224.x.

Blowes, Shane A., Sarah R. Supp, Laura H. Antão, Amanda Bates, Helge Bruelheide, Jonathan M. Chase, Faye Moyes, et al. 2019. "The Geography of Biodiversity Change in Marine and Terrestrial Assemblages." *Science* 366 (6463): 339–45. https://doi.org/10.1126/science.aaw1620.

Chao, Anne, and Lou Jost. 2012. "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather Than Size." *Ecology* 93 (12): 2533–47. https://doi.org/10.1890/11-1952.1.

Comte, Lise, Juan Carvajal-Quintero, Pablo A. Tedesco, Xingli Giam, Ulrich Brose, Tibor Erős, Ana F. Filipe, et al. 2021. "RivFishTIME: A Global Database of Fish Time-Series to Study Global Change Ecology in Riverine Systems." *Global Ecology and Biogeography* 30 (1): 38–50. https://doi.org/10.1111/geb.13210.

Di Marco, Moreno, Oscar Venter, Hugh P. Possingham, and James E. M. Watson. 2018. "Changes in Human Footprint Drive Changes in Species Extinction Risk." *Nature Communications* 9 (1): 4621. https://doi.org/10.1038/s41467-018-07049-5.

Dornelas, Maria, Nicholas J. Gotelli, Brian McGill, Hideyasu Shimadzu, Faye Moyes, Caya Sievers, and Anne E. Magurran. 2014. "Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss." *Science* 344 (6181): 296–99. https://doi.org/10.1126/science.1248484.

Fritz, Heinrich, Luis A. García-Escudero, and Agustín Mayo-Iscar. 2012. "Tclust: An R Package for a Trimming Approach to Cluster Analysis." *Journal of Statistical Software* 47 (May): 1–26. https://doi.org/10.18637/jss.v047.i12.

Froese, Rainer, Daniel Pauly, and others. 2021. *FishBase*. Fisheries Centre, University of British Columbia.

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27 (15): 2865–73. https://doi.org/10.1002/sim.3107.

Hillebrand, Helmut, Bernd Blasius, Elizabeth T. Borer, Jonathan M. Chase, John A. Downing, Britas Klemens Eriksson, Christopher T. Filstrup, et al. 2018. "Biodiversity Change Is Uncoupled from Species Richness Trends: Consequences for Conservation and Monitoring." *Journal of Applied Ecology* 55 (1): 169–84. https://doi.org/https://doi.org/10.1111/1365-2664.12959.

Klink, Roel van, Diana E. Bowler, Konstantin B. Gongalsky, Ann B. Swengel, Alessandro Gentile, and Jonathan M. Chase. 2020. "Meta-Analysis Reveals Declines in Terrestrial but Increases in Freshwater Insect Abundances." *Science* 368 (6489): 417–20. https://doi.org/10.1126/science.aax9931.

Linke, Simon, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, et al. 2019. "Global Hydro-Environmental Sub-Basin and River Reach Characteristics at High Spatial Resolution." *Scientific Data* 6 (1): 283. https://doi.org/10.1038/s41597-019-0300-6.

Robinson, Natalie, James Regetz, and Robert P. Guralnick. 2014. "EarthEnv-DEM90: A Nearly-Global, Void-Free, Multi-Scale Smoothed, 90m Digital Elevation Model from Fused ASTER and SRTM Data." *ISPRS Journal of Photogrammetry and Remote Sensing* 87 (January): 57–67. https://doi.org/10.1016/j.isprsjprs.2013.11.002.

Tedesco, Pablo A., Olivier Beauchard, Rémy Bigorne, Simon Blanchet, Laëtitia Buisson, Lorenza Conti, Jean-François Cornu, et al. 2017. "A Global Database on Freshwater Fish Species Occurrence in Drainage Basins." *Scientific Data* 4 (October): 170141. https://doi.org/10.1038/sdata.2017.141.

Venter, Oscar, Eric W. Sanderson, Ainhoa Magrach, James R. Allan, Jutta Beher, Kendall R. Jones, Hugh P. Possingham, et al. 2016a. "Global Terrestrial Human Footprint Maps for 1993 and 2009." *Scientific Data* 3 (1): 160067. https://doi.org/10.1038/sdata.2016.67.

———. 2016b. "Sixteen Years of Change in the Global Terrestrial Human Footprint and Implications for Biodiversity Conservation." *Nature Communications* 7 (1): 12558. https://doi.org/10.1038/ncomms12558.

Williams, Brooke A., Oscar Venter, James R. Allan, Scott C. Atkinson, Jose A. Rehbein, Michelle Ward, Moreno Di Marco, et al. 2020. "Change in Terrestrial Human Footprint Drives Continued Loss of Intact Ecosystems." *One Earth* 3 (3): 371–82. https://doi.org/10.1016/j.oneear.2020.08.009.