PhD Methodology

François Leroy

2021-04-07

Contents

Annotation

Biodiversity, at the basis of many essentials ecosystem services, is in the process of facing its sixth mass extinction. Although global extinction is unprecedented, there is so far no reason to expect that biodiversity dynamic at lower spatial and temporal scales follow this trend. Thus, links between spatio-temporal scales and facets of biodiversity (*i.e.* species richness, species diversity, colonization, extinction, species turnover, etc) need to be fully understood if we want to address this worldwide crisis. So far, attempts to describe biodiversity changes have been limited mainly by heterogeneity in spatial and temporal scales that was hardly taken into account by the statistical modelling frameworks.

My PhD project propose to address this flaws in order to understand in more details biodiversity changes across spatial and temporal scales. Especially, we aim at developing and testing nonparametric tree-based modelling methods allowing to study the non-linear and interacting effects of space and time-span on different aspects of biodiversity.

The specific objectives of my PhD project are:

1. Modelling and mapping avian species richness changes over Czech Republic across space and time scales.
2. Decompose the modelled biodiversity to colonization, extinction, species turnover, across spatiotemporal scales.
3. Estimate the strength of the link between environmental drivers of biodiversity change across spatio-temporal scales.
4. Apply the previously developed method to other European regions (e.g. UK, Switzerland, France)

PhD candidate: François Leroy

Programme: Environmental Earth Sciences

Department: Applied Geoinformatic and Spatial Planning

Advisor: doc. Ing. Petra Šímová, Ph.D.

Consultant: Mgr. Petr Keil, PhD

Beginning of study: October 2020

1. Introduction

Human life quality is intrinsically linked to ecosystems state that he is living in. Indeed, ecosystems services extend in a large spectrum of mechanisms including nutrient cycle, food production, or climate and water cycle regulation (Pereira et al., 2012). Some of those ecosystem functions are managed by bird populations such as seed dispersal, controls pests or pollinate plant. Unfortunately, anthropogenic stressors like habitat loss, over exploitation, pollution or introduction of invasive species could lead biodiversity to its sixth mass extinction (Barnosky et al., 2011).

While the loss of global biodiversity is unprecedented, current scientific literature has also shown that temporal trends in local changes of biodiversity can be opposite to trends at larger scales (Chase et al., 2019). Thus, current changes in biodiversity is far more complex than a simple global decrease: most of the ecosystems undergo alterations of their communities with changes in species composition (Blowes et al., 2019; Dornelas et al., 2014).

Typically, biodiversity is considered for a particular taxon (*e.g.* birds, amphibians, reptiles…), but also according to the spatial scale it is defined by. Here, the term scale refers to the area in which the biodiversity in considered, also referred hereafter as grain size. So far, it has been assumed that holding the spatial scale constant when studying biodiversity is mandatory (Whittaker et al., 2001). As a matter of fact, it is known that species richness increases with the area considered (Arrhenius, 1921) and this relationship is approximately linear on a log-log scale (Species-Area Relationship, SAR). However, this assumption restricts the data accessibility as sampling plans widely differ according to the species studied, the resources available or, the field conditions. Thus, developing a method capable of dealing with biodiversity across varying grain size could increase significantly the data availability. Moreover, it would allow to model biodiversity at different spatial scales than the ones used in the data. Modelling biodiversity indexes at

finer spatial grain size that the data used to learn the model is referred as *downscaling* biodiversity whilst extrapolating at coarser grain-size is called *upscaling*.

So far, there are indications that such method can be used. For instance, Keil and Jetz (2014) and Keil et al. (2013) showed promising downscaling biodiversity models using biodiversity data with different spatial scales, whilst Kunin et al. (2018) showed that upscaling biodiversity is also possible. Thus, all the constituents of cross-scales models are known but still need to be gathered and tested. For instance, Jarzyna et al. (2015) used a Bayesian framework to study temporal changes of avian biodiversity (colonization, extinction, temporal turnover) across scales. However, other approaches such as parametric Generalized Linear Models (GLM), Generalized Additive Models (GAM) and Generalized Linear Mixed Model (GLMM) or non-parametric tree based machine learning methods need to be tested.

2. Aims

The main aim of my PhD will be to test different statistical modeling methods allowing to integrate cross-scales biodiversity data. This method will allow to model biodiversity facets at various spatial and temporal grain sizes.

This principal objective can be divided into four sub-objectives:

1. **I will model temporal dynamics of avian species richness in the Czech Republic across spatio-temporal scales** (*i.e.* grain size ranging from less than 1 Km² to more than 2 000 Km²). This will allow to map biodiversity over Czech republic at any desired spatial and temporal scales. Given the well known Species-Area Relationship (SAR, Arrhenius, 1921; Storch & Gaston, 2004) and Species-Time Relationship (STR, White et al., 2006) we expect to see higher species richness at coarser spatial and temporal scales than at finer scales. Moreover, we will be able to look at the effect of grain size over biodiversity trends. So far, Chase et al. (2019) showed that North American avian biodiversity is largely stable at fine scales, but that it tends to increase with spatial scale. Thus, we can expect to observe the same trends for the Czech Republic.

2. **I will decompose the modelled biodiversity to colonization, extinction, and species turnover, across spatio-temporal scales.** Indeed, biodiversity dynamic is underlined by those ecological processes whose trends can be opposite to the species richness' one (Dornelas et al., 2014). Thus, understanding how they fluctuate according to the spatio-temporal scales they are considered will help understand the global biodiversity dynamic. So far, Jarzyna et al. (2015) showed that those facets of biodiversity was in general declining with increasing spatial grain size for the state of New-York. However the colonization undergoes a slower steep. Thus, those trends are also expected to occur with our model for the Czech Republic.

3. **I will estimate the strength of the link between environmental drivers of biodiversity change across spatio-temporal scales.** So far, land use change, habitat loss, or changes in climatic conditions are good candidates for such drivers. Jarzyna et al. (2015) showed that environmental parameters, such as climate change variables (*e.g.* temperatures, pluviometry) and landscape variables (*e.g.* elevation, sampling effort), were dependent on the spatial scale on the one hand, but also on the metric considered on the other hand (*i.e.* colonization, extinction...). Thus, the

expectations are more difficult to assess here. What we know is that at larger scales (*i.e.* biogeographic or continental), evolutionary processes tend to drive the biodiversity patterns (Keil & Jetz, 2014). Climatic and land cover parameters, for their part, intervene at scales ranging from tens to hundreds of kilometers. At even finer scales, biotic and population dynamics processes are driving.

4. **Once the methods will be validated for the Czech Republic, the logical continuation will be to use it over other European regions (*i.e.* Switzerland, United-Kingdom, Brittany, North America).** The main advantage of looking at other countries is to look at the trend of biodiversity facets at larger scales. As for objective 1, we can also expect that the species richness trends will tend to increase with spatial scale (Chase et al., 2019) whilst, as for objective 2, other biodiversity dynamic indexes will decrease (Jarzyna et al., 2015).

3. Methodological approach

3.1   Data

A significant part of this project will consist in **1)** harvesting, gathering and managing biodiversity datasets of the aimed taxon (i.e. birds here) in order to **2)** use them to model biodiversity facets across spatial and temporal scales.

Birds represent a key taxon for this problematic as they are various in morphology and colors, allowing to easily identify and list them. We already have access to high quality avian biodiversity time series over Czech Republic from the Česka Společnost Ornithologiká (Bejček & Stastný, 2016) and the Jednotný Program Sčítání Ptáků (JPSP, Reif et al., 2006, objective 1 and 2). Avian biodiversity will also be studied in other European countries (objective 4) and data for those regions are needed. I already contacted the Bretagne Vivante association, which handle biodiversity data for Brittany (*i.e.* French region). It will allow us to access avian biodiversity data for oceanic climate in order to contrast with the continental climate of the Czech Republic. Other datasets are aimed such as Swiss, British, Cataloninan or other French biodiversity data. In order to achieve the third objective of this project, environmental datasets are needed. For instance, the CORINE and HYDE (Goldewijk et al., 2011) datasets are aimed to access landcover and land use data, respectively. Climatology timeseries can also be found with Chelsa (Karger et al., 2018; Karger et al., 2017) and WorldClim datasets (Fick & Hijmans, 2017). Data management represent a significant time consuming part of a modelling project. So far, the beginning of my PhD consisted mainly into gathering and shaping datasets in order to be able to analyse and use them to train my models (objective 1). So far, species richness has been computed from 1973 to 2020 and for areas ranging from less than 1 Km² to more than 2 000 Km². For this, I used the avian biodiversity atlas data from the Česka Společnost Ornithologiká available at one grainsize that I aggregated into coarser 2 by 2 and 4 by 4 grain size (Fig. 1). On the other hand, I managed the JPSP dataset in a singular way, allowing me to extract species richness from censuses of local points to censuses of entire transects. Thanks to those, I was able to train my first random forests (see Pilot results part below). Thus, I am already able to shape any biodiversity dataset to use them into the machine learning framework desired. The next step will be to compute dynamic biodiversity indexes such as colonization, extinction, temporal turnover and community dissimilarity for objective 2.

3.2 Modelling methods

Non-parametric tree-based machine learning methods uses variance partitioning to iteratively split the feature space (features can also be named predictors, covariates, independent or, input variables) in order to obtain a tree in which one just need to follow the splits to predict an output (*i.e.* the response variable or dependent variable) such as species richness, colonization, extinction…

In order to make a model both understandable and predictive, a balance must be found between complexity and explicative power (Houlahan et al., 2017). Thus, using as few covariates as possible to predict biodiversity is necessary if we want to make the forecasts conveniently and if we want to discuss our models. We aim to start by using very few covariates such as latitude, longitude, area, time and time span in order to then add environmental parameters step by step. Tree-based machine learning methods such as random forests or boosted regression tree **1)** allows to study the interacting effect of drivers on the output variable and **2)** also represent a convenient way of dealing with nonlinear relationships between the response variable and the covariates. Indeed, Keil and Chase (2019) showed that **1)** area does have an interacting impact with other environmental and spatial drivers of biodiversity and **2)** that this relationship is non linear. Moreover, Viana et al. (2019) showed that boosted regression trees and random forests predicted ecological indexes more accurately than other methods. Thus, tree-based modelling methods are totally suited for our purpose. Other parametric methods such as generalized linear, additive or mixed models (GLM, GAM, GLMM) have already shown to give satisfying results (Keil & Chase, 2019) and could be used in some of my analysis.

It is important to point out that the proposed methodology here can be applied to any other taxa (**e.g.** lepidoptera, large mammals) and any other spatial range (**e.g.** Europe, North America, South Africa), which represent the next steps of this project.

3.3 Pilot results

So far, I have already been able to produce random forests using only latitude, longitude, area, date, time span and elongation as covariates that explain around 90% of the species richness variance over the Czech Republic, which is encouraging (see Fig. 2, https://github.com/FrsLry/IGA_figures/blob/main/pred_vs_obs.pdf). An other advantage of the complex nonparametric models is that you can

represent the dependence between the outcome and a predictor of interest called a marginal plot or partial plot. For instance, in Fig.3 (https://github.com/FrsLry/IGA_figures/blob/main/marginal_plot.pdf), I represented the influence of the interacting area and time factors on the species richness for one of my model. In order to validate and enhance the models performance, the next steps will be to **1)** perform cross-validation to avoid overfitting, **2)** add the adequate environmental parameters and **3)** connect the local time series from the JPSP data to the atlas data by using both of them in a single model.

4. Schedule

Table 4.1: Planning of the methodological steps of my PhD

Start date	End date	Description of the tasks
01/10/2020	01/06/2021	1. Gathering and managing avian biodiversity datasets over Czech Republic (objective 1, 2 and 3). 2. Computing biodiversity indexes (species richness, colonization, extinction, temporal turnover, community similarity) from the data (objective 2 and 3).
01/06/2021	01/09/2021	Building models for Czech Republic: machine learning models are powerful but need to be parametrized in order to increase predicting power and avoid overfitting. Moreover, I will build my models with significant databases and expect to run long time consuming computing jobs.
01/01/2022	01/05/2022	1. Testing addition of environmental variables to the random forests in order to increase predictive power (Objective 3). 2. Harvesting and managing avian biodiversity datasets across Europe: objective 4 of this project will be to study biodiversity changes across scales at the European extent. I will have more datasets and management is expected to take more time than for Czech Republic.
01/05/2022	01/08/2022	Building models for Europe (objective 4): I will build my models with significant databases and expect to run long computing jobs.

Start date	End date	Description of the tasks
01/08/2022	End of the PhD	1. Building models for North America: Discussions about collaboration with Dr. Marta Jarzyna is ongoing in order to apply these developed method to North American breeding birds datasets they have access to. 2. Merging the European and North American models will be a good start to look at a the more global biodiversity changes across different continents. 3. Gathering and merging biodiversity datasets from other part of the globe will give us more insights about the main trend of the avian biodiversity changes at worldwide scale.

5. Benefits of outcome

My PhD aim at several goals:

1. Better understand the link between spatial and temporal features (*e.g.* latitude, longitude, area, time span) and biodiversity changes.

2. Developing a modelling framework (tree-based models) allowing to forecast biodiversity indexes across various space and time scales would help to **1)** understand and thus **2)** forecast biodiversity dynamic.

3. This method will allow to integrate heterogeneous biodiversity datasets that could usually not be used together with classical statistical models due to their inconsistencies in space and/or time scales. Thus, it will be possible to link local biodiversity datasets (*e.g.* censuses of ornithological associations, participative sciences) to broader ones (*e.g.* atlas, time-series assemblage such as Biotime database by Dornelas et al., 2018).

4. After focusing on the species richness changes, the link between spatial and temporal scales and the dynamic of other facets of biodiversity (*e.g.* colonization, extinction, β diversity) will be investigated.

5. Produce avian biodiversity maps at different spatial and temporal grain sizes over Czech Republic/Europe/North America.

6. The parametrized models for birds will give insights on how to apply them to other taxa (*e.g.* lepidoptera, odonata, large mammals...).

6. References

Arrhenius, O. (1921). Species and area [Publisher: [Wiley, British Ecological Society]]. *Journal of Ecology*, *9*(1), 95–99. https://doi.org/10.2307/2255763

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., & Ferrer, E. A. (2011). Has the earth's sixth mass extinction already arrived? [Number: 7336 Publisher: Nature Publishing Group]. *Nature*, *471*(7336), 51–57. https://doi.org/10.1038/nature09678

Bejček, V., & Stastný. (2016). Velké ptačí mapování. *Vesmír*. Retrieved December 10, 2020, from https://vesmir.cz/cz/on-line-clanky/2016/04/velke-ptaci-mapovani.html

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., Moyes, F., Magurran, A., McGill, B., Myers-Smith, I. H., Winter, M., Bjorkman, A. D., Bowler, D. E., Byrnes, J. E. K., Gonzalez, A., Hines, J., Isbell, F., Jones, H. P., Navarro, L. M., … Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages [Publisher: American Association for the Advancement of Science Section: Research Article]. *Science*, *366*(6463), 339–345. https://doi.org/10.1126/science.aaw1620

Chase, J. M., McGill, B. J., Thompson, P. L., Antão, L. H., Bates, A. E., Blowes, S. A., Dornelas, M., Gonzalez, A., Magurran, A. E., Supp, S. R., Winter, M., Bjorkman, A. D., Bruelheide, H., Byrnes, J. E. K., Cabral, J. S., Elahi, R., Gomez, C., Guzman, H. M., Isbell, F., … O'Connor, M. (2019). Species richness change across spatial scales. *Oikos*, *128*(8), 1079–1091. https://doi.org/10.1111/oik.05968

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., Akhmetzhanova, A. A., Appeltans, W., Arcos, J. M., Arnold, H., Ayyappan, N., Badihi, G., Baird, A. H., Barbosa, M., Barreto, T. E., Bässler, C., Bellgrove, A., Belmaker, J., Benedetti-Cecchi, L., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time series for the anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. https://doi.org/10.1111/geb.12729

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., & Magurran, A. E. (2014). Assemblage time series reveal biodiversity change but not systematic loss [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *344*(6181), 296–299. https://doi.org/10.1126/science.1248484

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas [_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5086]. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/https://doi.org/10.1002/joc.5086

Goldewijk, K. K., Beusen, A., Drecht, G. v., & Vos, M. d. (2011). The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-8238.2010.00587.x]. *Global Ecology and Biogeography*, *20*(1), 73–86. https://doi.org/https://doi.org/10.1111/j.1466-8238.2010.00587.x

Houlahan, J. E., McKinney, S. T., Anderson, T. M., & McGill, B. J. (2017). The priority of prediction in ecological understanding. *Oikos*, *126*(1), 1–7. https://doi.org/10.1111/oik.03726

Jarzyna, M. A., Zuckerberg, B., Porter, W. F., Finley, A. O., & Maurer, B. A. (2015). Spatial scaling of temporal changes in avian communities: Scale dependence of community turnover. *Global Ecology and Biogeography*, *24*(11), 1236–1248. https://doi.org/10.1111/geb.12361

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2018). Data from: Climatologies at high resolution for the earth's land surface areas [Artwork Size: 7266827510 bytes Pages: 7266827510 bytes Version Number: 1 type: dataset]. https://doi.org/10.5061/DRYAD.KD1D4

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas [Number:

1 Publisher: Nature Publishing Group]. *Scientific Data*, *4*(1), 170122. https://doi.org/10.1038/sdata.2017.122

Keil, P., Belmaker, J., Wilson, A. M., Unitt, P., & Jetz, W. (2013). Downscaling of species distribution models: A hierarchical approach [_eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2041-210x.2012.00264.x]. *Methods in Ecology and Evolution*, *4*(1), 82–94. https://doi.org/10.1111/j.2041-210x.2012.00264.x

Keil, P., & Chase, J. M. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution*, *3*(3), 390–399. https://doi.org/10.1038/s41559-019-0799-0

Keil, P., & Jetz, W. (2014). Downscaling the environmental associations and spatial patterns of species richness. *Ecological Applications: A Publication of the Ecological Society of America*, *24*(4), 823–831. https://doi.org/10.1890/13-0805.1

Kunin, W. E., Harte, J., He, F., Hui, C., Jobe, R. T., Ostling, A., Polce, C., Šizling, A., Smith, A. B., Smith, K., Smart, S. M., Storch, D., Tjørve, E., Ugland, K.-I., Ulrich, W., & Varma, V. (2018). Upscaling biodiversity: Estimating the species-area relationship from small samples. *Ecological Monographs*, *88*(2), 170–187. https://doi.org/10.1002/ecm.1284

Pereira, H. M., Navarro, L. M., & Martins, I. S. (2012). Global biodiversity change: The bad, the good, and the unknown. *Annual Review of Environment and Resources*, *37*(1), 25–50. https://doi.org/10.1146/annurev-environ-042911-093511

Reif, J., Voříšek, P., & Šťastný, K. (2006). Population trends of birds in the czech republic during 1982–2005, 16.

Storch, D., & Gaston, K. J. (2004). Untangling ecological complexity on different scales of space and time. *Basic and Applied Ecology*, *5*(5), 389–400. https://doi.org/10.1016/j.baae.2004.08.001

Viana, D. S., Keil, P., & Jeliazkov, A. (2019). Partitioning environment and space in species-by-site matrices: A comparison of methods for community ecology and macroecology [Publisher: Cold Spring Harbor Laboratory Section: New Results]. *bioRxiv*, 871251. https://doi.org/10.1101/871251

White, E. P., Adler, P. B., Lauenroth, W. K., Gill, R. A., Greenberg, D., Kaufman, D. M., Rassweiler, A., Rusak, J. A., Smith, M. D., Steinbeck, J. R., Waide, R. B., & Yao, J. (2006). A comparison of the species–time relationship across ecosystems and taxonomic groups [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0030-1299.2006.14223.x]. *Oikos*, *112*(1), 185–195. https://doi.org/https://doi.org/10.1111/j.0030-1299.2006.14223.x

Whittaker, R. J., Willis, K. J., & Field, R. (2001). Scale and species richness: Towards a general, hierarchical theory of species diversity [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2699.2001.00563.x]. *Journal of Biogeography*, *28*(4), 453–470. https://doi.org/https://doi.org/10.1046/j.1365-2699.2001.00563.x