Methodology of the dissertation

PhD candidate: François Leroy

Programme: Environmental Earth Sciences

Department: Spatial Sciences

"Mapping biodiversity changes across spatio-temporal scales"

Advisor: doc. Ing. Petra Šímová, Ph.D.
Consultant: Mgr. Petr Keil, PhD
Beginning of study: October 2020

Contents

1. title: “PhD Methodology” # if you indicate the tile, it’ll create a new cover page

Placeholder

1.1 This line is to use when you want several outputs (specified in output.yml)

1.2 The default class

1.3 to colorize the biblio ref and the URLs

2. Introduction

Placeholder

3. Aims

Placeholder

4. Methodological approach

4.1 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, French or North American 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). Species richness has been computed from 1973 to 2020 for areas ranging from less than 1 Km^2 to more than 2 000 Km^2. 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.

4.2 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; Levins, 1966). 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.

4.3 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). 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, 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.

5. Schedule

Table 5.1: 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 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 even more insights about the main trend of the avian biodiversity changes at worldwide scale.

6. Benefits of outcome

My PhD aim at several goals:

1. Better understand the link between spatial and temporal features (*e.g.* latitude, longitude, area, elongation, 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…).

7. Cooperations

This project takes part in the broader research project of Dr. Petr Keil who has been working on the problem of scale-dependent biodiversity change and integration of heterogeneous data for a decade now, and who has published several high-profile publications on these topics. He currently is my PhD supervisor. Petr's expertise will be particularly relevant for tasks requiring advanced statistical modelling, interpretation of the models, and putting the results in a broader macroecological context.

Cooperation is already ongoing with Vladimír Bejček and Karel Šťastný who furnished us time series of avian biodiversity from the Česka Společnost Ornithologiká which were used in the publications of several atlases (see Bejček & Stastný, 2016). On the other hand, Dr. Jiří Reif forwarded me local time series from the Jednotný Program Sčítání Ptáků (JPSP). Their expertise on bird ecologie will be helpful in order to interpret and enhance the outputs of my models.

Finally, discussions with Dr. Marta Jarzyna (University of Colombus, Ohio) are ongoing in order to work together on applying the methods that I use on Czech Republic and Europe to some American states.As a matter of fact, my results will **1)** allow to differentiate biodiversity dynamics on the North American and the European continents and **2)** help to better understand the link between spatio-temporal scales and biodiveristy dynamic by enlarging the databases that I use.

8. Expected outcomes

Placeholder

9. References

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

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

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

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., & 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

Levins, R. (1966). The strategy of model building in population biology [Publisher: Sigma Xi, The Scientific Research Society]. *American Scientist*, *54*(4), 421–431. Retrieved October 13, 2020, from https://www.jstor.org/stable/27836590

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

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