PhD Proposal:
Imputation Models for Biodiversity and Ecological Science

Keywords. imputation models, biodiversity, data scarcity, latent variable models, variational inference

Summary

In biodiversity monitoring, large datasets are becoming widely available and are used globally to estimate species trends and conservation status. However, analyzing them is challenging, as they often simultaneously display missing values, zero-inflation, overdispersion and detection bias. Existing statistical methods used to address these issues usually fail to tackle them simultaneously, leading up to bias in trends estimations. This problem is particularly blatant in low-income countries, where temporal and spatial gaps in biodiversity datasets are major concerns. The proposed project is a multidisciplinary initiative gathering statisticians and ecologists to develop a new method to analyze large-scale monitoring data in the context of data scarcity. The developed method should account for missing values, zero-inflation, overdispersion and possibly detection bias in a unified model. This new framework will allow us to analyze a dataset monitoring 173 waterbirds species across 30 years (1990-2019) and 785 sites in North Africa, a region of strategic importance for the conservation of waterbirds suffering from many spatial and temporal gaps in survey data. We aim at estimating the long-term trends of rare or threatened species as well as the impact of several meteorological and anthropogenic factors, for the first time at the North African scale.

A first method based on advances in low-rank matrix completion has been proposed by Robin et al. [2019]. However, this method is not suited for rare species for the reasons previously cited. The methodological objective of the thesis is to propose an extension of this model adapted to overdispersion and excess of zeros. Latent-space species distribution models have proved to be a convenient framework to account for both abundance predictors and, through the latent layer, for over-dispersion and excess of zeros. The latent layer of these models also provides a natural framework to deal with missing data. The inference of these models benefit from recent advances in variational approximations. Various modeling options will be tested on simulated and real-life biodiversity data, and implemented in an open source R package. We believe this new approach opens promising perspectives to increase the accuracy of species trend estimation in large-scale surveys, as well as citizen-science monitoring programs.

Description of the proposal

Context

The proposal emerged following a decision from France as a contracting party to the African-Eurasian Waterbird Agreement to support the AEWA plan of Action for Africa and in particular waterbird surveys in Africa by setting up a Technical Support Unit to the AEWA African initiative and the Mediterranean Waterbirds network. This Network achieved an in-depth review, update and completion of the waterbird database for North Africa. This database has numerous gaps in space and time which prevent comprehensive analysis. Upgrading imputation tools and collecting predictor covariates was identified as one strategy allowing use of most count data collected in Africa. An early collaboration was initiated between École Polytechnique, Tour du Valat (TdV), Office Français de la Biodiversité (OFB) as well as North African governments and NGOs to test a new modeling tool for North-African waterbird species. Publications by Sayoud et al. [2017] and Robin et al. [2019] were both preliminary studies to the present proposal.

General goals and objectives

Biodiversity monitoring datasets, emerging in particular from citizen-science monitoring programmes, are becoming more and more complex and high dimensional. They bring hope of answering many important ecological or

conservation issues [Pereira et al., 2013, Stephenson et al., 2017]. However, they are challenging to analyze when they display missing values, zero-inflation, overdispersion, large amounts of predictor covariates and detection bias.

Statistical methods have been developed to deal with these issues. For instance, missing values through imputation methods (Van Strien et al. 2004, van Buuren and Groothuis-Oudshoorn 2011, Stekhoven and Bühlmann 2012), zero excess with zero-inflated models (Cunningham and Lindenmayer 2005, Dénes et al. 2015), estimation of the detection bias (Coron et al. 2016, Johnston et al. 2020). However, to the best of our knowledge, there is currently no available method to address these issues simultaneously. This gap is particularly problematic in the monitoring of rare or threatened species, and in regions where data surveys are difficult to carry out for financial, political or logistic reasons. Indeed, in such cases, large spatial and temporal gaps lead to data scarcity, while the rare occurrence of species increases the risk of zero-inflation and overdispersion. In addition, the availability of many covariates leads to complex statistical models. This context pushes available methods outside the boundaries of their statistical guarantees and yields inaccurate trends estimation.

The updated International Waterbird Census (IWC) dataset for North Africa is a striking example of such issues. North Africa (Morocco, Algeria, Tunisia, Libya and Egypt) is of considerable importance for the conservation of waterbirds migrating along the African-Eurasian flyway (Sayoud et al. 2017). However, for financial and political reasons, coverage of North African wetlands has been highly irregular (Dakki et al. [2001], EGA - RAC/SPA waterbird census team 2012), leading to large amounts of missing values. To analyze this data set and estimate species trends in spite of the large amount of missing values, additional factors assumed to be relevant predictors of species counts [Amano et al., 2018] were gathered by the data management team.

The resulting statistical problem boils down to modeling high-dimensional, largely incomplete data with complex distributions, in the presence of large predictor sets. Low-rank matrix completion has recently been applied to monitoring count data as a proof of concept [Robin et al., 2019, Dakki et al., 2021]. Tested on a subset of our North African data corresponding to one of the most common waterbird species, it demonstrated competitive performance in comparison to existing methods. We propose to extend this approach, introducing a new methodology based on latent variable models, which provides a sound statistical framework for the prediction of missing values. Latent variable modelling typically allows accounting for over-dispersion (thanks to the addition of random effects) or for an excess of zero (using a two-component mixture model). The inference of complex latent variable models can generally not be carried with the EM algorithm [Dempster et al., 1977] due to a complex dependency structure, but variational approximations [Blei et al., 2017] provide a computationally efficient way to circumvent these issues. These approximations have been proved to be efficient and accurate in the context of joint species distribution models to analyze large datasets [Chiquet et al., 2018, 2019, 2021].

We propose a multidisciplinary project gathering ecologists and statisticians, with the double objective of developing a new methodology for rare species monitoring in the context of data scarcity, and of analyzing the IWC North African data set to estimate species trends and conservation status at a regional scale. The first step will be to extend the existing framework of matrix completion to rare species modeling. Then, the method will be tested thoroughly in several simulated settings, and evaluated on species monitoring data sets. Finally, it will be used to analyze the IWC North African data set in collaboration with local organizations managing the national monitoring programs. We emphasize that, though our project focuses on an application to waterbirds monitoring in North Africa, the proposed statistical methodology could be used on other biodiversity data sets, and have a broader impact on other fields where similar issues occur.

Proposed activities

This project will be carried out in three main phases corresponding to

i) the development of the statistical methodology,

ii) its empirical evaluation based on simulations as well as on species monitoring data, and

iii) its application to the analysis of the North African data set.

The first objective is to extend low-rank matrix completion to rare species modeling. For the moment it is restricted to Poisson models. We aim to extend it to more sophisticated models incorporating zero-inflation, overdispersion, and possibly detection bias. Using latent variables models is a generic and convenient way to tackle this point. The main anticipated difficulty will be the development of an efficient inference algorithm. However, the latent variable framework allows to resort to variational technics of inference, which have proved their efficiency and robustness. Classically, over-parameterization will be tackled through two main assumptions. First, we assume that not all predictors have a non-zero effect on the counts. Second, we assume the existence of a few groups of

similar sites and similar years. To implement these constraints, we will resort to variational penalized estimation procedures. These procedures are known to benefit from strong statistical guarantees in similar problems (Robin et al. 2019, Chiquet et al. [2018].

To evaluate the performance of the method, we will consider several simulation models using different distributions as well as different missing data scenarios. An important aspect will be the comparison to existing methods for biodiversity monitoring data analysis. The scope of this first empirical study will be to understand precisely the range of applicability of the proposed method, in order to release recommendations to potential users. We will pay particular attention to the impact of the size of the data set, the level of zero-inflation and overdispersion, the proportion of missing values, and the impact of measurement errors in survey data. Then, a second empirical study will be carried out on waterbirds monitoring data from public data sets resulting from citizen science programms. The goal of this second study will be to evaluate the robustness of the method on real data, and whether the resulting trends are consistent with known results from existing studies.

Finally, the method will be applied to the analysis of the North African data sets in order to estimate the trends and conservation status of rare or threatened species. This phase will be carried out in close collaboration with field observers. During this last phase conservation recommendations will have to be identified in collaboration with experts of the OFB and TdV and international partners.

Expected results

The anticipated results of this multidisciplinary project consist of methodological developments, an open source, documented software, and trend analysis for a selection of North-African waterbird species of strong conservation concern. The proposed methodology could improve inference and decision processes in biodiversity monitoring by providing a unified statistical package dedicated to incomplete count datasets of any animal or plant. In addition, we emphasize that this method could also have an impact on other fields where similar issues occur, such as genomics data analysis. To make it widely available, we plan to produce a methodological publication on statistical developments and capacities of the method on simulated and real biodiversity data, and to deliver an R package. We also aim at presenting it in international statistics, but also R conferences. The expected trend analyses could also have an impact on waterbirds conservation in North Africa. These results will be delivered in a conservation publication.

Dataset to be used.

- Dataset name: IWC_North Africa
- Description (type of date, format, size): Postgres/PostGis database containing waterbird data from 1990 to 2019. 2049 sites over 5 countries, and 173 waterbird species.
- Current location/owner: Mediterranean Waterbirds Network
- Accessibility: Medwaterbirds network (`www.medwaterbirds.net/datacounts.php`)

Tour du Valat, with permanent support from OFB, promotes and coordinates waterbird data centralization from North Africa in partnership with the following partners: BirdLife Morocco/GREPOM, Direction Générale des Forêts d'Algérie, AAO/BirdLife Tunisia, Libyan Society for Birds, Egyptian Environmental Affairs Agency (EEAA), Wetlands International. The database is managed by Tour du Valat and OFB and belongs primarily to field data contributors under a charter signed in October 3, 2016 (attached). Because data collection in the field is the result of a strong long-term commitment of all these partners, yet without a noticeable publication impact and recognition, and because data collection in the field is the first fundamental process in biological conservation, all partners in this proposal agree to give full recognition to field data contributors in the publication strategy.

Supervision and environment

The Phd will be supervised by Dr. S. Donnet (INRAE, MIA-Paris-Saclay) and Dr. S. Robin (Sorbonne Université, LPSM) for all the mathematical, statistical and software developments. The Phd student will also be supervised by Mme. Dami from Tour du Valat, Dr. Defos Du Rau from the French Agency for Biodiversity (OFB) and Dr. Galewski Thomas from Tour du Valat (TdV), on all the developments and also on the ecological developments of the study. She/he will collaborate with the project leader, Mme Dami Laura, and an engineer, both based at the TdV, notably on ecological aspects of the interpretation of results and conservation issues for the North African countries. She/he will probably need to collaborate or exchange with the different partners in the countries that

sustained strong long-term efforts to provide the databases used in this study. The Phd student will be based at LPSM (Paris) or at MIA-Paris-Saclay (Palaiseau) most of the time, but will need to spend time at the Tour du Valat, south of France, in order to work with all the staff (OFB and TdV) on the ecological aspects, at least partly of the second and the third year of the PhD.

Contacts

Les candidatures doivent être adressées **avant le 26 août 2022** aux quatre personnes suivantes :

- S. Donnet (INRAE, MIA-Paris-Saclay) : sophie.donnet@inrae.fr,
- S. Robin (Sorbonne Université, LPSM) : stephane.robin@sorbonne-universite.fr,
- L. Dami (Tour du Valat) : dami@tourduvalat.org,
- P. Defos Du Rau (French Agency for Biodiversity) : pierre.defosdurau@ofb.gouv.fr.

References

T. Amano, T. Székely, B. Sandel, S. Nagy, T. Mundkur, T. Langendoen, D. Blanco, C.U Soykan, and W.J Sutherland. Successful conservation of global waterbird populations depends on effective governance. *Nature*, 553(7687): 199–202, 2018.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171, 2019.

J. Chiquet, M. Mariadassou, and S. Robin. The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:188, 2021. doi: 10.3389/fevo.2021.588292.

M Dakki, A Qninba, M-A El Agbani, A Benhoussa, and P-C Beaubrun. Waders wintering in morocco: national population estimates, trends and site-assessments. *Bulletin-Wader Study Group*, 96:47–59, 2001.

M. Dakki, G. Robin, M. Suet, A. Qninba, M. A El Agbani, A. Ouassou, R. El Hamoumi, H. Azafzaf, S. Rebah, C. Feltrup-Azafzaf, et al. Imputation of incomplete large-scale monitoring count data via penalized estimation. *Methods in Ecology and Evolution*, 12(6):1031–1039, 2021.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

H.M Pereira, S. Ferrier, M. Walters, G.N. Geller, R.H.G. Jongman, R.J. Scholes, M.W. Bruford, N. Brummitt, S.H.M. Butchart, A.C. Cardoso, et al. Essential biodiversity variables. *Science*, 339(6117):277–278, 2013.

G. Robin, J. Josse, É. Moulines, and S. Sardy. Low-rank model with covariates for count data with missing values. *Journal of Multivariate Analysis*, 173:416–434, 2019.

M.S. Sayoud, H. Salhi, B. Chalabi, A. Allali, M. Dakki, A. Qninba, M.A. El Agbani, H. Azafzaf, C. Feltrup-Azafzaf, H. Dlensi, et al. The first coordinated trans-north african mid-winter waterbird census: the contribution of the international waterbird census to the conservation of waterbirds and wetlands at a biogeographical level. *Biological Conservation*, 206:11–20, 2017.

P.J. Stephenson, T. Brooks, S. Butchart, E. Fegraus, G. Geller, R. Hoft, J. Hutton, N. Kingston, B. Long, and L. McRae. Priorities for big biodiversity data. *Frontiers in Ecology and the Environment*, 15(3):124–125, 2017.