BIODIVERSITY BUILDING BLOCKS FOR POLICY

M17 - Deep learning algorithms for deep learning using B3 and associated open data

31/07/2024

Author(s): Alexis Joly, Diego Marcos, Maxime Ryckewaert

Funded by the European Union

This project receives funding from the European Union's Horizon Europe Research and Innovation Programme (ID No 101059592). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

Table of contents

Summary

This document explores the application of deep learning methods within the B-CUBED project, focusing on the integration of citizen science and opportunistic data for species distribution modelling (SDM). It begins by presenting the general context surrounding the collection of opportunistic data tainted by observation bias. These biases are central to the results obtained from Species Distribution Modelling (SDM) algorithms. This is even more the case with advanced deep learning techniques, collectively known as Deep-SDM. Yet the potential of these methods is to be able to efficiently process and analyse vast datasets. This is followed by a review of the technical aspects and specific deep learning algorithms used in the B-CUBED project, focusing on their application in point process contexts that are crucial for modelling species occurrence distributions. In addition, it discusses the incorporation of various types of data into biodiversity data cubes in order to enrich the analysis. The practical application of these methodologies is illustrated by case studies involving both simulated and real data, including a detailed examination of species classification in Belgium (2010) using B-CUBED data and insights from the NCGEAS project. This overview is supported by an extensive list of references.

List of abbreviations

EU	European Union
SDM	Species Distribution Modelling
Deep-SDM	Deep Learning Species Distribution Modelling
NCGEAS	National Centre for Ecological Analysis and Synthesis
GBIF	Global Biodiversity Information Facility
PO	Presence Only
PA	Presence Absence

1. Introduction

1.1. Citizen Science and Opportunistic Data

The Global Biodiversity Information Facility (GBIF) database is enriched by a combination of probabilistic and opportunistic samples, known as preferred samples. Probabilistic samples are selected at random using statistical methods, providing an impartial and generalisable representation of biodiversity in a given region. Opportunistic samples, on the other hand, often come from unsystematic collections by researchers or amateurs via citizen science applications (Bonnet et al., 2020; Callaghan et al., 2022). Opportunistic data may be influenced by the accessibility of sites, the season, or species of particular interest. The combination of these two types of sampling enables GBIF to maximise the quantity and diversity of the data collected, while mitigating the biases inherent in each method taken in isolation. In this way, this integrative approach offers a more complete and nuanced view of the world's biodiversity.

Despite their potential, presence-only (PO) data have limitations because they only indicate where a species has been observed, without providing information about where the species is absent. These data are typically derived from opportunistic observations or occurrence records. However, using such observation data introduces several inherent challenges. One major issue is the bias arising from imperfect detection; not all individuals of a species present in an area are observed or recorded. Additionally, variations in sampling efforts across different regions and times can further skew the data. The subjective perspectives of individual observers also contribute to inconsistencies, as some species may be more likely to be reported than others. These factors collectively impact the reliability of species distribution models (SDM) that are trained using presence-only data (Fithian et al., 2015; Komori et al., 2020; Phillips et al., 2009).

To overcome the limitations of presence-only (PO) data, researchers have devised various methodologies centred around the concept of pseudo-absences. Pseudo-absences, often referred to as background or pseudo-negative points, involve designating certain geographic locations as negative samples to compensate for the absence data. One common approach involves sampling these pseudo-absences uniformly across the geographic space, creating random background points. Another strategy selects pseudo-absences from locations where other species, which are subject to similar sampling biases, have been observed, known as target-group background points. These techniques aim to provide a more balanced dataset, thereby enhancing the accuracy and reliability of species distribution models (SDMs) that are trained with these augmented datasets

1.2. Deep Learning Methods for SDM (Deep-SDM)

On the other hand, deep learning models have become increasingly prominent in the field of species distribution modelling. These models are capable of processing vast amounts of biodiversity data, effectively capturing the intricate, non-linear relationships between various environmental factors and the presence or absence of species (Deneu et al., 2022; Estopinan et al., 2024; Seo et al., 2021). By leveraging environmental and remote sensing variables, deep learning techniques can uncover patterns that traditional methods might miss.

However, this adaptability also means that deep learning models can inadvertently incorporate and magnify existing biases in the data. When working with datasets that are biased or unbalanced in terms of species representation, the models might produce skewed predictions. This issue underscores the importance of improving the robustness of deep learning methodologies in species distribution modelling.

To address these challenges, researchers are exploring advanced techniques and strategies to mitigate biases and enhance model reliability. Efforts are focused on developing more sophisticated approaches to handle imbalanced data, ensuring that the predictions are more accurate and generalizable. As the field evolves, the integration of robust deep learning models promises to significantly advance our understanding of species distribution and support more effective conservation efforts (Beery et al., 2021).

2. Deep learning algorithms for B-CUBED project

2.1. Technical aspects

A repository can be found at the following address on GitHub: https://github.com/RYCKEWAERT/b-cubed_deep-sdm. The algorithms in the repository have been developed exclusively using the Python programming language. The main library for deep learning used is PyTorch, an open-source machine learning framework that offers great flexibility and efficiency for the development of deep learning models. PyTorch is particularly appreciated for its ability to perform calculations on GPUs, making it easier to train complex neural networks.

2.2. Deep Learning Methods

In the broad families of deep learning and artificial intelligence, there are several key architectures, including Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs). MLPs are one of the simplest forms of artificial neural networks. They consist of several layers of neurons, where each neuron in one layer is connected to all the neurons in the next layer. This dense architecture enables MLPs to capture complex relationships between input variables and expected responses, such as the presence or absence of a species in a given geographical area. MLPs are particularly useful when the input data are feature vectors extracted from environmental data such as temperature, humidity or altitude. CNNs, and in particular deep architectures such as ResNet, are better suited to processing spatial data and images thanks to their ability to capture local features through convolution operations. ResNet, or residual networks, introduces residual connections that allow information to pass directly

between layers, making it easier to train very deep networks without the problem of vanishing gradients arising. For species distribution modelling, ResNet can be used to analyse satellite images or environmental maps, identifying complex patterns and spatial features relevant to predicting the presence of species in specific regions.

2.3. Deep Learning Methods in a context of point process

Assuming that the generation of observations follows a Poisson distribution, a loss function must be defined to match the underlying Poisson distribution. The development of deep learning methods has focused on a Poisson loss function $\mathcal{L}_{\mathcal{P}}$ which is formulated as the negative log-likelihood for a Poisson model. Based on this loss, several methods have been developed to resolve problems related to bias in the data, and these are still being evaluated.

2.4. Others data associated with biodiversity data cubes

To build a species distribution model, additional datasets known as covariates are also required. Covariates include various environmental variables such as temperature, precipitation, soil type, and vegetation cover, which influence species distributions. Terrestrial observation data and satellite missions, like those providing remote sensing data, are invaluable in this context. They supply detailed and up-to-date information on environmental conditions across vast areas.

- Observation Data: Information collected in the field or from databases on the presence and abundance of species.
- Habitat Mapping: Spatial data on the distribution of habitats used by different species.
- Climatic Data: Climatic variables such as temperature, rainfall and sunshine that influence the distribution of species.
- Geographical data: Information on topography, altitude and geology that may affect the distribution of species.
- Anthropogenic data: Human factors such as land use, urbanisation and agricultural activities that modify natural habitats.

3. Case studies

3.1. Simulated data

A simulation framework has been developed to build and evaluate deep learning algorithms. This solution has been proposed in order to obtain ground truth. To this end, a database was used to generate virtual species. The means and standard deviations of the real species are used to generate the virtual species that have relationships between the input variables. For a virtual species n, we construct a virtual ``ground truth'' intensity function:

$$\lambda_n(\mathbf{x}) = f(\mathbf{x}; \mu_n, \Sigma_n)$$

where μ_n and Σ_n are estimated by sampling from one randomly selected real species. To do so, species intensity is built as a multivariate Gaussian of climatic variables.

$$f(\mathbf{x};\mu_n,\Sigma_n) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma_n)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_n)^T\Sigma^{-1}(\mathbf{x}-\mu_n)\right)$$

where $\mathbf{x}$ are the bioclimatic variables used, μ a vector containing means of all variables and Σ is the covariance matrix containing variances and covariances between variables.

3.2. Real data

3.2.1. Using B-CUBED data for species classification (Belgium, 2010)

This dataset is a typical biodiversity dataset from the B-CUBED project in Belgium. It represents a subset from the year 2010, extracted from a more comprehensive dataset. The data is organised into spatial cubes to facilitate detailed biodiversity analysis for that year. For more information and access to the full dataset, please refer to the following resources.[https://www.gbif.org/occurrence/download/0096919-240321170329656] (https://doi.org/10.15468/dl.e3j5kv).

The covariate dataset contains 19 bioclimatic rasters obtained from the WorldClim and CHELSA databases. The rasters represent various environmental factors such as temperature, precipitation, and altitude. The full dataset is available : https://chelsa-climate.org/bioclim/ https://doi.org/10.1038/sdata.2017.122

3.2.2.NCGEAS

Data from the National Centre for Ecological Analysis and Synthesis (NCEAS) have been openly released recently (Elith et al., 2020). This dataset includes presence-only and presence-absence data from six global regions: Australian Wet Tropics (AWT), Canada (CAN), New South Wales (NSW), New Zealand (NZ), South America (SA), and Switzerland (SWI). It comprises data for 226 anonymized species from different biological groups. The dataset contains different environmental predictive variables for each region, including climatic, soil variables or location information (more details in Elith et al., 2020).

This dataset has been used to evaluate and compare various methods (Elith* et al., 2006; Phillips et al., 2009; Valavi et al., 2022), allowing for comparisons with existing SDM methods. All the species in each biological group in each region are used to form models based on presence data only. The models are then evaluated with presence-absence data using the Area Under Curve (AUC) criterion. Finally, AUC values are averaged by region or for all regions.

Table 1: Details of the Elith dataset where each line corresponds to the data used to create a model.

Code	Location	Biological Group	Species number	Occurrences number(PO)	Occurrences number(PA)
AWT	Australian wet tropics	bird	40	3105	340
AWT	Australian wet tropics	plant	40	701	102
CAN	Ontario, Canada	bird	20	5063	14571

NSW	New South Wales	bate	54	187	570
NSW	New South Wales	bird	54	1781	1839
NSW	New South Wales	plant	54	680	5329
NSW	New South Wales	reptile	54	675	1008
NZ	New Zealand	plant	52	3088	19120
SA	South America	plant	30	2220	152
SWI	Switzerland	tree	30	35105	10013

4. References

Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species Distribution Modeling for Machine Learning Practitioners: A Review. *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, 329–348. https://doi.org/10.1145/3460112.3471966

Bonnet, P., Joly, A., Faton, J., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J., Mary, L., Vignau, C., & Munoz, F. (2020). How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence*, *1*(2), Article 2. https://doi.org/10.1002/2688-8319.12023

Callaghan, C. T., Mesaglio, T., Ascher, J. S., Brooks, T. M., Cabras, A. A., Chandler, M., Cornwell, W. K., Cristóbal Ríos-Málaver, I., Dankowicz, E., & Urfi Dhiya'ulhaq, N. (2022). The benefits of contributing to the citizen science platform iNaturalist as an identifier. *PLoS Biology*, *20*(11), Article 11.

Deneu, B., Joly, A., Bonnet, P., Servajean, M., & Munoz, F. (2022). Very high resolution species distribution modeling based on remote sensing imagery: How to capture fine-grained and large-scale vegetation ecology with convolutional neural networks? *Frontiers in Plant Science*, *13*, 839279.

Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ford, A., Guisan, A., Hijmans, R. J., Huettmann, F., Lohmann, L., Loiselle, B., Moritz, C., Overton, J., Peterson, A. T., Phillips,

S., Richardson, K., Williams, S., Wiser, S. K., Wohlgemuth, T., & Zimmermann, N. E. (2020). Presence-only and Presence-absence Data for Comparing Species Distribution Modeling Methods. *Biodiversity Informatics*, *15*(2), Article 2. https://doi.org/10.17161/bi.v15i2.13384

Elith*, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), Article 2. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Estopinan, J., Bonnet, P., Servajean, M., Munoz, F., & Joly, A. (2024). *Modelling Species Distributions with Deep Learning to Predict Plant Extinction Risk and Assess Climate Change Impacts* (arXiv:2401.05470; Issue arXiv:2401.05470). arXiv. http://arxiv.org/abs/2401.05470

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), Article 4. https://doi.org/10.1111/2041-210X.12242

Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., & Kubota, Y. (2020). Sampling bias correction in species distribution models by quasi-linear Poisson point process. *Ecological Informatics*, *55*, 101015. https://doi.org/10.1016/j.ecoinf.2019.101015

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), Article 1. https://doi.org/10.1890/07-2153.1

Seo, E., Hutchinson, R. A., Fu, X., Li, C., Hallman, T. A., Kilbride, J., & Robinson, W. D. (2021).

StatEcoNet: Statistical Ecology Neural Networks for Species Distribution Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), Article 1. https://doi.org/10.1609/aaai.v35i1.16129

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1), Article 1. https://doi.org/10.1002/ecm.1486