BIODIVERSITY BUILDING BLOCKS FOR POLICY

M11 Code development for predictive habitat suitability modelling

18/10/2024

Author(s): Rocìo Beatriz Cortès Lobos, Michele Di Musciano, Matilde Martini, & Duccio Rocchini

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

The B3 project (Biodiversity Building Blocks for Policy) aims at ensuring that monitoring data be easily accessible, reliable, and useful, thereby enhancing the efficiency of future conservation-related decisions. The ongoing global biodiversity crisis needs robust, precise, reliable, and recurrent biodiversity monitoring data for effective policy assessment.

A considerable amount of data has already been collected, such as datasets coming from Habitat Directive reports of the European Union. However, many of these datasets are not easily accessible, with temporal and spatial data often kept separately or not harmonized in terms of taxonomy. Additionally, biodiversity data are often influenced by errors that make their applicability uncertain in modeling species distribution.

This report addresses these challenges by employing the data cube format to organize biodiversity data across spatial, temporal, and taxonomic dimensions while also associating a suitability score. This makes the data easier to use and enhances the efficiency of modeling biodiversity change and status. A key aspect of Task 4.1 is the integration of virtual species and simulations, which were used both for building the **Virtual Suitability Cube (VSC**) and for assessing how sampling bias affects **Species Distribution Models** (**SDMs**). By simulating species occurrences with known ecological characteristics, this approach allows for better control over variables and improves our understanding of the ecological niches and conservation needs of real species.

This report is the outcome of the first Milestone of task 4.1 of B3.

List of abbreviations

EU	European Union
GBIF	Global Biodiversity Information Facility
SDM	Species Distribution Models
AOA	Area of Applicability
VSC	Virtual Suitability Cube

1 Introduction

Species Distribution Models (SDMs) are essential tools for predicting where species are most likely to survive, reproduce, and thrive, both in the present and under future environmental scenarios *(Guisan, Thuiller, 2005*). These models use species occurrence data and environmental variables, such as climate, landscape features, and resource availability, to forecast the distribution of species across different habitats. SDMs are widely used in conservation biology to identify suitable habitats, predict the spread of invasive species, and understand how species distributions may shift due to factors like climate change.

By integrating ecological data with occurrence records, SDMs allow researchers to better understand the environmental conditions that support species survival and inform strategies to protect biodiversity.

In light of these applications, our work focuses on enhancing species distribution modeling through two key objectives: (i) the creation of a Virtual Suitability Cube through the *virtualspecies* R package (*Leroy, 2018*) (VSC) and (ii) the assessment of sampling bias impacts on niche predictions.

The VSC aims to organize species occurrence data and associated suitability scores within a multidimensional framework that reflects key ecological and spatial factors. Meanwhile, the analysis of sampling bias seeks to improve the accuracy of SDM predictions by understanding how biased sampling can distort our view of species distributions and ecological niches.

These objectives align with the broader aim of providing workflows that are repeatable, scalable, and adaptable for diverse end users, while ensuring the results are concrete and reproducible for policy applications.

1.1. Suitability Cube format

To enhance the structuring and accessibility of environmental and species occurrence data, we developed the VSC, a structured, multi-dimensional array that organizes data across key ecological dimensions such as species, geographical coordinates, temporal scales, and suitability scores. This cube format facilitates the integration of environmental data from sources like the Copernicus Program, WorldClim, and other remote sensing datasets, simplifying the process of modeling species distributions under current and future global change scenarios.
By associating a suitability score with each species occurrence, the cube streamlines the process of modeling both present and future species distributions.
Additionally, this cube format not only supports modeling current species distributions but also allows for retrospective analyses of past distributions, creating a more complete picture of species habitat suitability and potential distribution over time. So far, the VSC has been modelled and tested with virtual species; the transition from those species to the real ones is expected to be done by March 2025.

1.2. Sampling Bias

One of the major challenges in biodiversity monitoring is the impact of sampling bias on species distribution models. When occurrence data are collected using non-random or unplanned sampling designs, such as roadside observations, this can skew predictions of species niches and

habitat suitability.
Sampling bias is particularly problematic in large-scale monitoring projects, as it can lead to over- or underestimation of species distributions (*Werkowska et al., 2017)*. The roadside bias, where data points are clustered along roadways, is one of the most common biases and can distort predictions by affecting the completeness of niche modeling (*Kadmon, 2004*). Since roadside bias is one of the most common sources of sampling bias and can be modeled, it will be key to developing a workflow to estimate the incompleteness of the ecological niche. This workflow can be applied to any source of sampling bias.

2 Virtual Species: the key to the workflow

To achieve both objectives of Task 4.1, a **virtual species approach** was adopted, simulating species occurrences based on known ecological characteristics and environmental constraints (*Zurell et al., 2010*).
A virtual species refers to a hypothetical or simulated species created in silico that behaves according to predetermined ecological and environmental constraints. This approach offers significant advantages over using real-world species data, particularly when dealing with the complexities of spatial, temporal, and taxonomic uncertainties that often arise in biodiversity research.
Simulating species occurrences based on known ecological traits and environmental variables allowed for precise, controlled experimentation in ecological modeling and habitat suitability assessments.

- Virtual species to develop the VSC. For demonstrating the structure of the VSC, we generated virtual species with known suitability maps based on climate data. The main steps include combining climate data to calculate the suitability of different species over time and in the same area, then merging these species into a single stars object. For constructing the VSC, virtual species provide a clean, noise-free dataset that enables more accurate and scalable habitat suitability values across different spatial and temporal scales.
- Virtual species to assess sampling bias. In the context of addressing sampling bias in biodiversity data, virtual species help disentangle the effects of sampling bias, improving our understanding of species' ecological niches by mimicking real species' ecological niches without the noise and biases found in real-world data. This approach allows for precise testing of species distribution models (SDMs) while controlling for spatial, temporal, and taxonomic uncertainties.

The workflow developed for the creation of the VSC offers a powerful tool for organizing and analyzing species occurrence data, while the second provides a reproducible framework for assessing the impact of bias in biodiversity monitoring that can be adaptable to real-world biodiversity data (aim of the MS25).

3 Virtual Suitability Cube

To facilitate the observation of suitability for multiple species over time and space by creating a **Virtual Suitability Cube**, we developed a framework that uses virtual data cubes, multidimensional arrays that organize data in a structured way. In this tutorial, we outline the steps to create a stars object, which includes three dimensions: time, space (represented as grid cells), and species, with suitability as the main attribute.

Stars objects can be sliced, aggregated along one of the dimensions, and analyzed, making them ideal for studying species suitability.

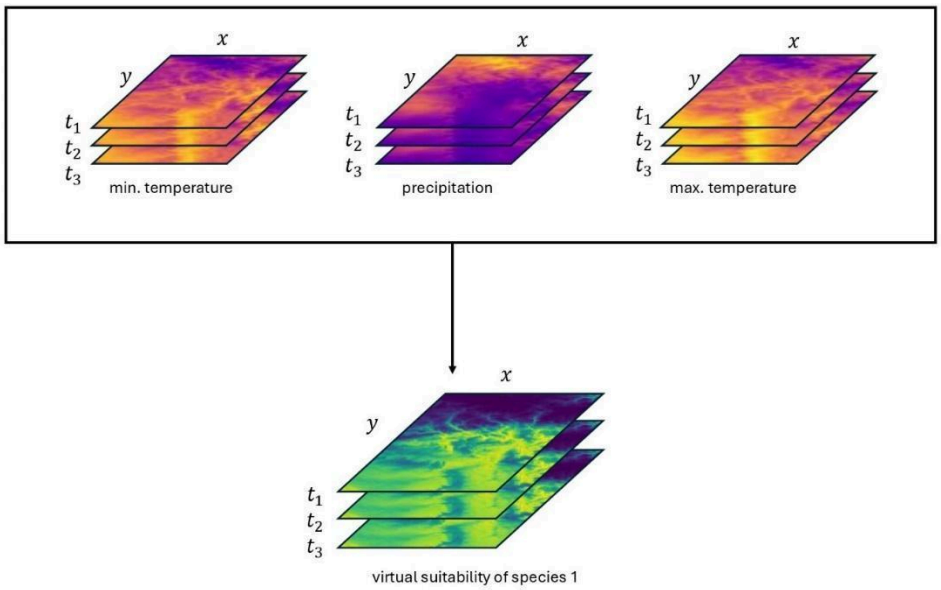

Figure 1: Starting with a time series of climate variables, we combined them to create the suitability for two different virtual species, whose trends over time we wanted to observe.

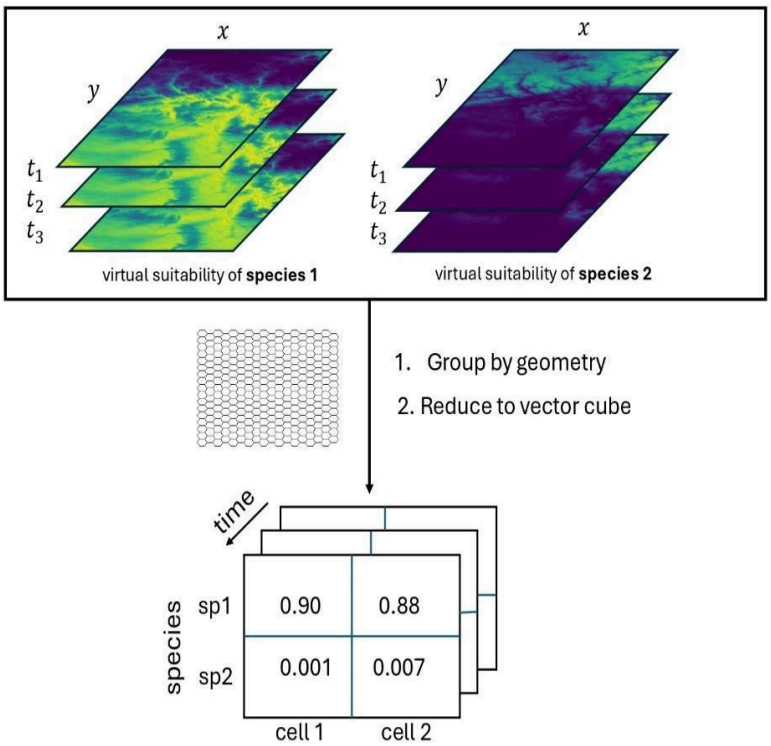

Figure 2: The two data cubes of species suitability were treated as separate entities, which we could then combine by aggregating them over polygons, creating a vector data cube.

This approach makes it easy to visualize and analyze species suitability across time and space for multiple species at once.

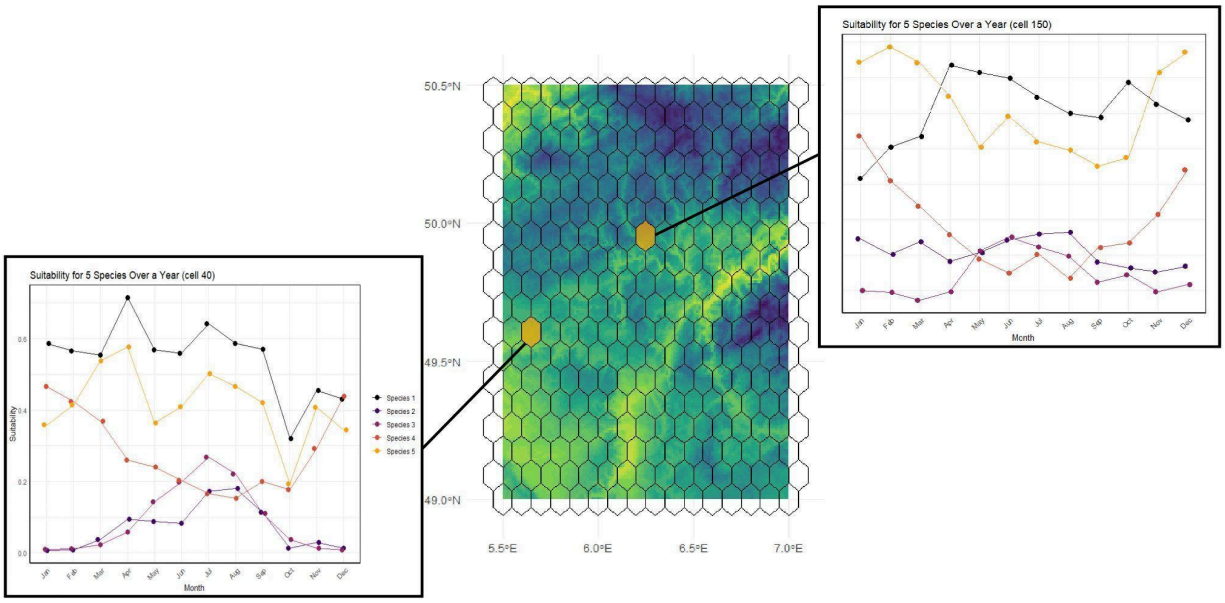

Figure 3: In the example, the suitability of 5 species across 12 months in 2 areas is shown.

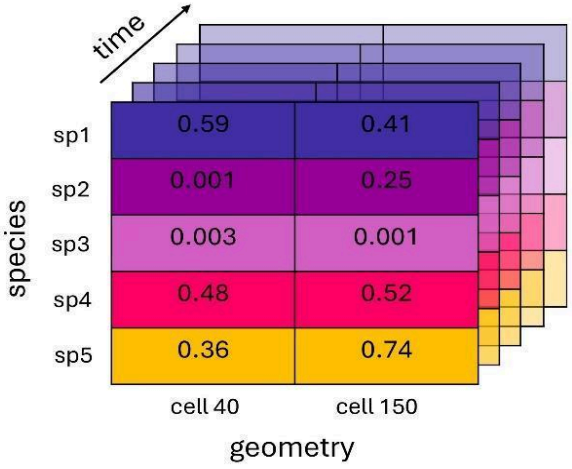

Figure 4: This is how we can imagine the data cube related to the example

4 Hypervolume and Area of Applicability

To understand how subsampling due to roadside bias impacts niche completeness and whether it leads to underestimating species niches or inaccurately projecting species suitability to new areas or under future climate scenarios, we created, for the same simulated species, two datasets: one randomly sampled, the other one clustered along the roads. For both, we measured:

- The ecological niche as an **hypervolume** in an n-dimensional space, which can therefore be quantified; we do this through the hypervolume R package (*The n-dimensional hypervolume, Benjamin Blonder*).
- The **Area of Applicability** (AOA), defined as the region where a predictive model can reliably be applied based on the relationships it learned from the training data (Meyer, 2021),

We aim to understand not only the incompleteness of the ecological niche but also if a biased dataset results in a reduced spatial area where species distribution estimates are unreliable.

The developed framework starts from the simulation, for the same species, of an **ideal random sampling** and a **biased sampling** along roads, by generating the species environmental suitability.

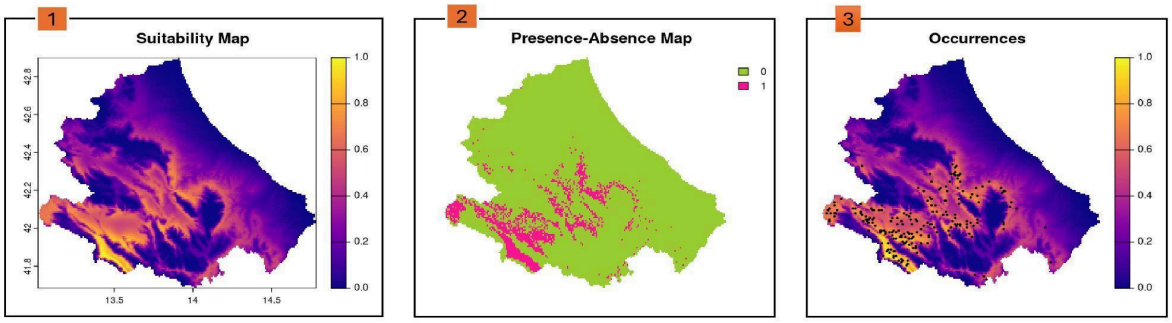

Figure 5: Starting from suitability map (randomly generated by the virtualspecies R package), we simulate a set of unbiased occurrences.

From the ideal random sampled dataset, we created a biased dataset, using the road network as a source of bias: as the distance from the road network increases, the probability of sampling a species will decrease.
We select the points that have a 100% probability of being sampled. In this way, we obtained a subsample of points with sampling bias.

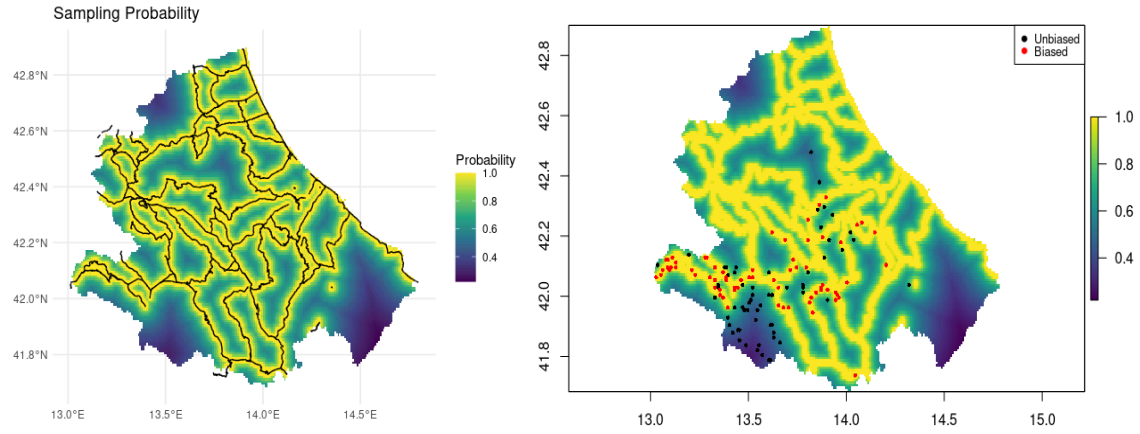

Figure 6: The road network allows us to simulate the biased dataset by giving to a point the probability to be sampled by a 'lazy sampler' who made the in situ sampling close to the roads.

4.1 Hypervolume

According to Hutchinson (1957), the ecological niche is a **hypervolume** in an n-dimensional space, which can therefore be quantified. For both datasets (unbiased and biased), we observe the trend of the hypervolume as occurrences increase and accumulate (*Arlè et al., 2024*). The future steps will focus on the comparison of the two accumulation curves for many simulated species in order to see if there is a relationship between the number of occurrences and the unbiased-biased gap.

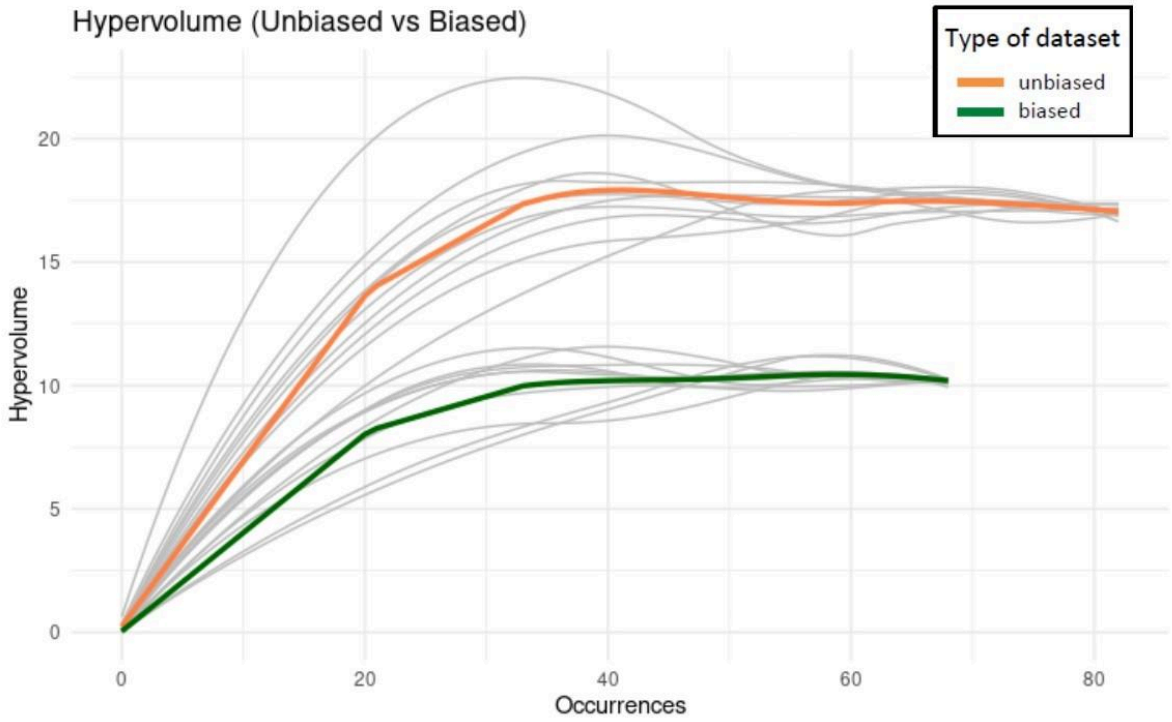

Figure 7: The biased dataset describes an incomplete ecological niche. There is a gap of information between unbiased and biased sampling.

4.2 Area of Applicability (AOA)

The calculation of the hypervolume provides a measure of how much the ecological niche is underrepresented by preferential sampling along the roads, but we lack any spatial information.
The **Area of Applicability** (AOA) is defined as the region where the model can reliably be applied based on the relationships it learned from the training data (Meyer, 2021).
Our focus is on the initial data: we will generate the same model using the two different datasets, obtaining a null model and a biased model.

Specifically, we aim to verify whether the same model can provide more spatial information, in terms of pixels, when constructed on an unbiased dataset.

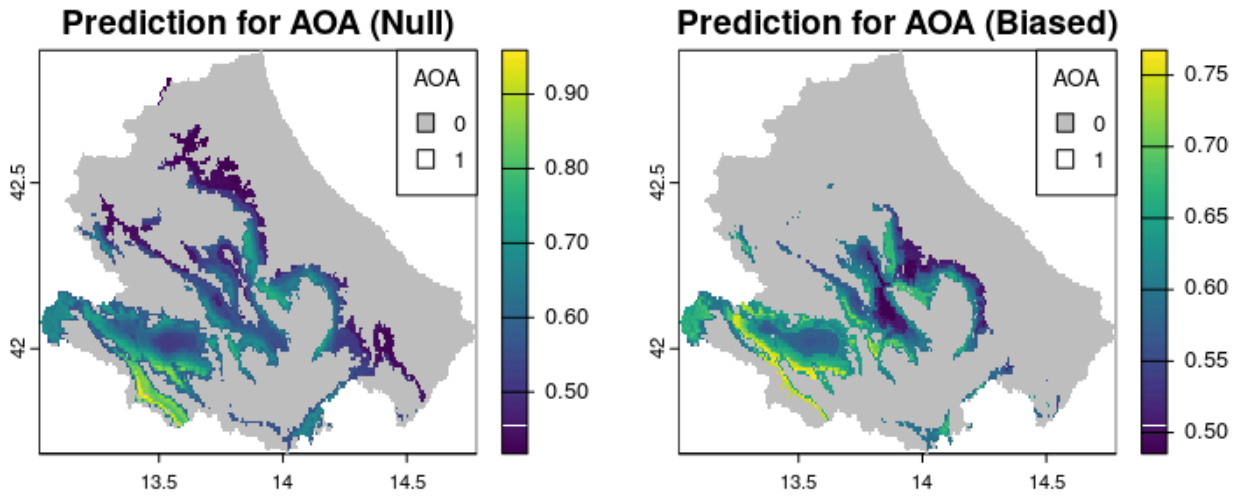

Figure 8: A model (Random Forest) trained on randomly sampled data has a much higher predictive capacity than the same model trained on data collected only along the roads.

By calculating the AOA of the same model trained first on the unbiased dataset and then on the biased dataset, we obtain the difference in the number of pixels between the two outputs. We expect that the area of applicability of a model calibrated on a randomly sampled dataset will be larger compared to that of a biased dataset.

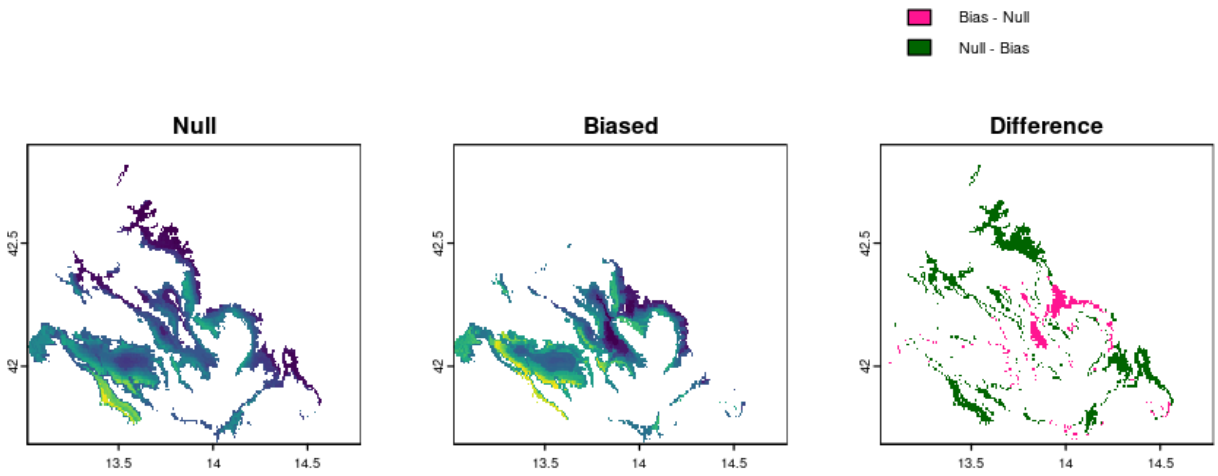

Figure 9: A good dataset, in this example, can provide predictions on 2127 km^2 more than a biased dataset.

5 Materials

A repository can be found at the following address on GitHub: https://github.com/b-cubed-eu/virtual-suitability-cube.
The codes in the repository have been developed exclusively using the R programming language and are still under development.

6 References

- virtualspecies, an R package to generate virtual species distributions (2018), Leroy B. et al., DOI
- Pebesma E, Bivand R (2023). Spatial Data Science: With applications in R. Chapman and Hall/CRC, London. DOI:10.1201/9780429459016, https://r-spatial.org/book/
- Meyer H, Milà C, Ludwig M, Linnenbrink J, Schumacher F (2024). CAST: 'caret' Applications for Spatial-Temporal Models. R package version 1.0.2, https://hannameyer.github.io/CAST/
- Predicting species distribution: offering more than simple habitat models, Guisan A., Thuiller W., 2005, DOI
- A practical overview of transferability in species distribution modeling, Werkowska W., Marquez A., Real R., Acevedo P., 2017, DOI
- Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models, Kadmon R., Farber O., Danin A., 2004, DOI
- The virtual ecologist approach: simulating data and observers, Zurell et al., 2010, DOI
- The n-dimensional hypervolume, Blonder et al., 2014, DOI
- The cumulative niche approach: a framework to assess the performance of ecological niche model projections, Arlè et al., 2024, DOI