BIODIVERSITY BUILDING BLOCKS FOR POLICY

M12: Code testing for predictive habitat suitability modelling

11/04/2025

Author(s): Rocìo Beatriz Cortès Lobos, Michele di Musciano & 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 B-Cubed 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. Moreover, biodiversity data are often influenced by errors, such as sampling bias, that make their applicability uncertain in modeling species distribution.

Our first contribution to those challenges was developed in Milestone 20 (MS20), in which we showed how to store into the same data cube the suitability information of multiple species, for the same area, and how to study them across time. This was done with virtual species, i.e. species artificially generated and designed for testing sampling schemes or data structures, since factors like disturbance or sampling bias can be easily avoided.

In this second Milestone (MS25), we switched from virtual species to real species, and we started to deal with the generation of Species Distribution Models (SDMs) for each species. The aim of this report is to show the progress that we have made in creating functions that, starting from the occurrences (only presence data) of real species and the environmental predictors, create SDMs out of them and then, using those models, predict the suitability of the same species in a new area. All the information is stored in a single object, called **Suitability Cube**. The advantage of such a structure is that it can be easily sliced across dimensions and allows the user to explore the general suitability of an area for multiple species and to spot relevant patterns.

List of abbreviations

EU	European Union
GBIF	Global Biodiversity Information Facility
SDM	Species Distribution Models
SC	Suitability Cube
MAXENT	Maximum Entropy

Introduction

In a time of fragile balances and great uncertainty about the future of global biodiversity, Species Distribution Models (SDMs) are important tools to predict where species are most likely to survive, reproduce, and thrive, both now and under future environmental scenarios (Guisan & Thuiller, 2005).
By combining ecological data with species occurrence records, SDMs help researchers understand which environmental conditions support species survival and guide actions to protect biodiversity.
To create suitability maps, it is important to use the large amount of data that has been collected over time.

The aim of this step in Task 4.1 is to simplify the process of analyzing the suitability of several real species in a study area where the environmental characteristics are known, but the possible locations of the target species are not.

To improve the organization and accessibility of environmental and species occurrence data, we developed the SC (Suitability Cube), a structured, multi-dimensional array implemented as a **stars** object in R (Pebesma, 2023). The final output, in line with the general goals of the B3 project, is a data cube that contains the suitability of each species, organized by two or three dimensions: space (cells), time, and species.

The cube makes it easier to integrate environmental data from sources like the Copernicus Program, *WorldClim*, and other datasets. This helps simplify the modeling of species distributions under both current conditions and future climate change scenarios.
By linking a suitability score of each species' occurrence to the space and time, the cube provides a practical way to model and compare present and future distributions in a consistent framework.

Additional layers of information, such as uncertainty measures, distances from human infrastructure (e.g., roads), or other relevant variables, can also be included in the same object, making it a flexible tool for ecological analysis.

Moreover, the cube format is not limited to current or future projections: it also supports retrospective analyses, helping to reconstruct past distributions and build a more complete understanding of how species habitats have changed over time.

These goals match the broader aim of developing workflows that are easy to repeat, scale, and adapt for different users.

Input data

The first step involves defining a study area for which we have both species occurrence data and associated environmental information.
The goal is to train species distribution models (SDMs) using this data to make predictions in other regions where the environmental suitability for these species is unknown.

Since we aim to leverage the Dutch Vegetation Database, which hosts information on all plant communities in the Netherlands (Hennekens, 2018), we select the Netherlands as our study area.

For this region, we require climatic variables, the predictors, which can be easily downloaded from WorldClim.

The first function in the workflow enables the creation of a data cube that contains all the climatic variables, organized along spatial and temporal dimensions. The climatic variables are attributes that are related to the same spatial and temporal dimensions.

```
> stars_clima
stars object with 3 dimensions and 5 attributes
attribute(s), summary of first 1e+05 cells:
      Min. 1st Qu. Median         Mean 3rd Qu. Max.  NAs
tmin  -0.8    -0.4   -0.1  0.007868939     0.2  2.6 63985
tmax   3.7     4.3    4.6  4.718736640     5.0  6.4 63985
prec  58.0    66.0   68.0 67.980036096    70.0 80.0 63985
tavg   1.6     1.9    2.2  2.363004291     2.6  4.5 63985
wind   4.0     4.6    5.2  5.363057057     6.0  7.8 63985
dimension(s):
     from  to offset     delta refsys point x/y
x       1 540      3  0.008333 WGS 84 FALSE [x]
y       1 420     54 -0.008333 WGS 84 FALSE [y]
time    1  12      1         1     NA    NA
```

Fig. 1: How a stars data cube looks like on R

From the cube that contains all months, we extract May to make predictions.

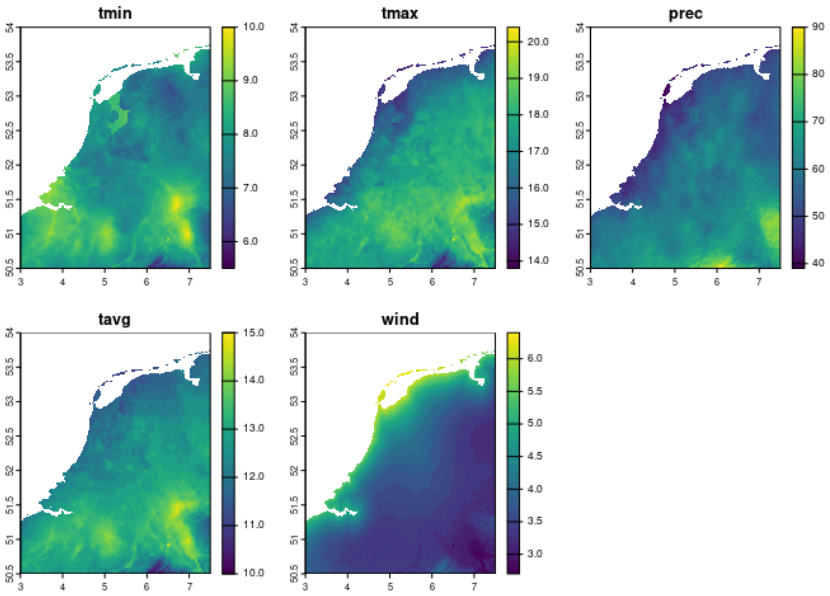

Fig. 2: Climatic variables for The Netherlands on May

Regarding species occurrences, we focus on the following five species:

- *Galium verum* L.
- *Ophrys apifera* Huds.
- *Paris quadrifolia* L.
- *Chrysosplenium alternifolium* L.
- *Anemone nemorosa* L.

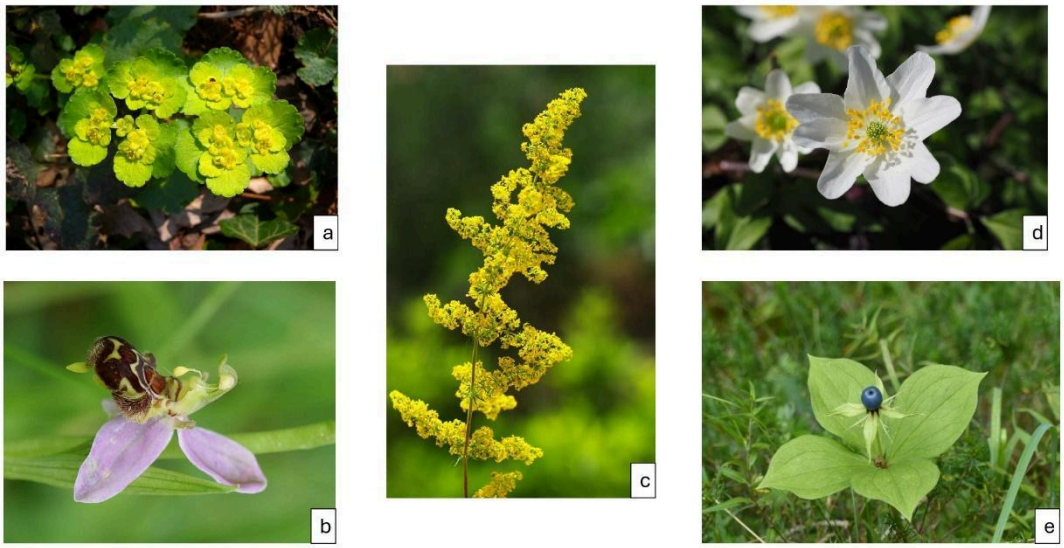

Fig. 3: Chosen species: **a)** *Chrysosplenium alternifolium* L., **b)** *Ophrys apifera* Huds., **c)** *Galium verum* L., **d)** *Anemone nemorosa* L., **e)** *Paris quadrifolia* L.

For each species, we filter occurrence records from the year 2000 to 2017.
The number of occurrence points is highly variable over time for each species.

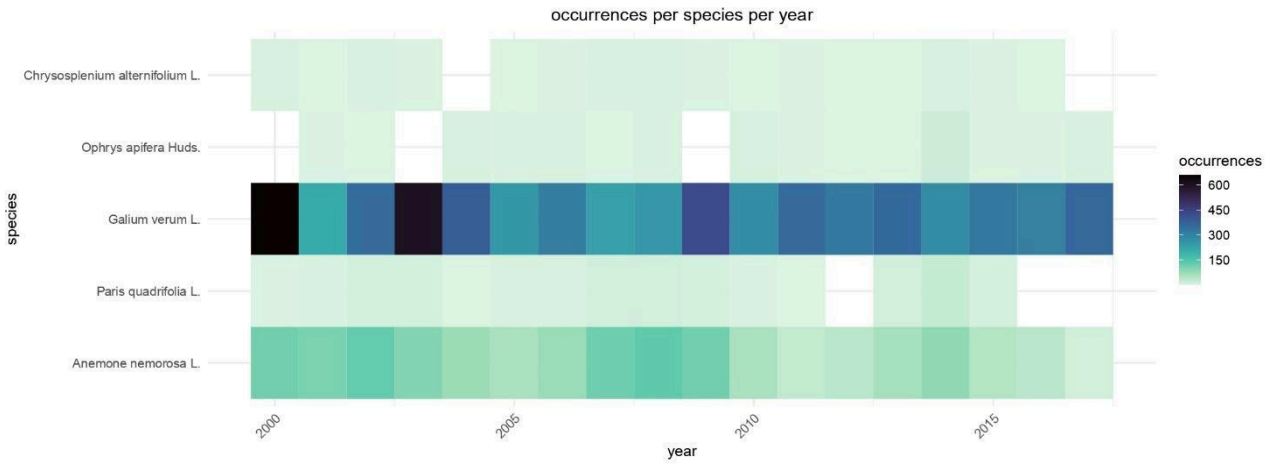

Fig. 4: Species occurrences per year from 2000 to 2017 show that *Galium verum* L. has significantly more records than the other species.

Starting from a dataframe containing the scientific name, latitude, longitude, and year, the *split_species_data* function allows us to create a list in which each element corresponds to a species and contains its specific occurrence dataframe.

```
> species <- split_species_data(occ)
Done
> typeof(species)
[1] "list"
> names(species)
[1] "Anemone nemorosa L."             "Chrysosplenium alternifolium L." "Galium verum L."
  "Ophrys apifera Huds."
[5] "Paris quadrifolia L."
```

Fig. 5: The function *split_species_data* will give you a list containing all the occurrences divided in species

SDMs: MaxEnt

MaxEnt offers several advantages, including the fact that it requires only presence data, along with environmental information covering the entire study area.
As a test case for our workflow, we chose to use MaxEnt, implemented through the enmSdmX package (Smith et al., 2023).
The final version of the function will incorporate the DeepMaxent implementation.

The current function allows the creation of SDMs with MaxEnt for multiple species within the same region, using a common set of predictors.
It takes as input:

- The bioclimatic variable stack previously created
- A list containing the occurrence points for each species
- The number of background points required for MaxEnt modeling
- The predictors within the stack that we want to use. If not specified, all of them will be use

As output, in addition to suitability maps for each species across the study area, we obtain a trained model for each species that can be applied to new environmental predictors in a different area.
Disclaimer: it's highly probable that the ecological meaning of this brief tutorial is not relevant. The aim is just to show the data structure

```
# function
sdms <- create_sdm_for_species_list(species, clima_train, background_points = 10000, predictors =
names(clima_train))

# maxent model
sdms$models$`Anemone nemorosa L.`

# suitability map
plot(rast(sdms$predictions))
```

Fig. 6: Outputs of *create_sdm_for_species* function

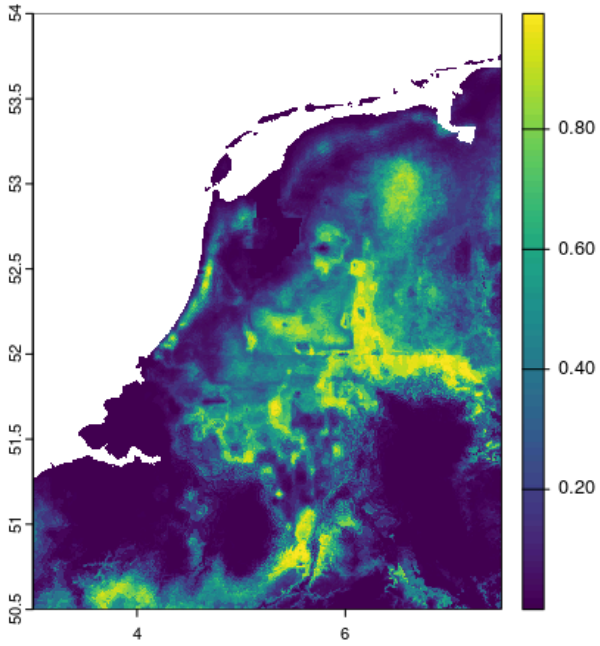

Fig. 7: Suitability map for *Anemone nemorosa* L. in The Netherlands

Prediction and aggregation

For each species, we obtain a model capable of predicting its suitability in a new area, different from the training region.
We selected Belgium as our test area, for which we have no prior information regarding the presence of the target species.
By using the same bioclimatic variables as those employed for the Netherlands, our goal is to predict the environmental suitability for the five species.

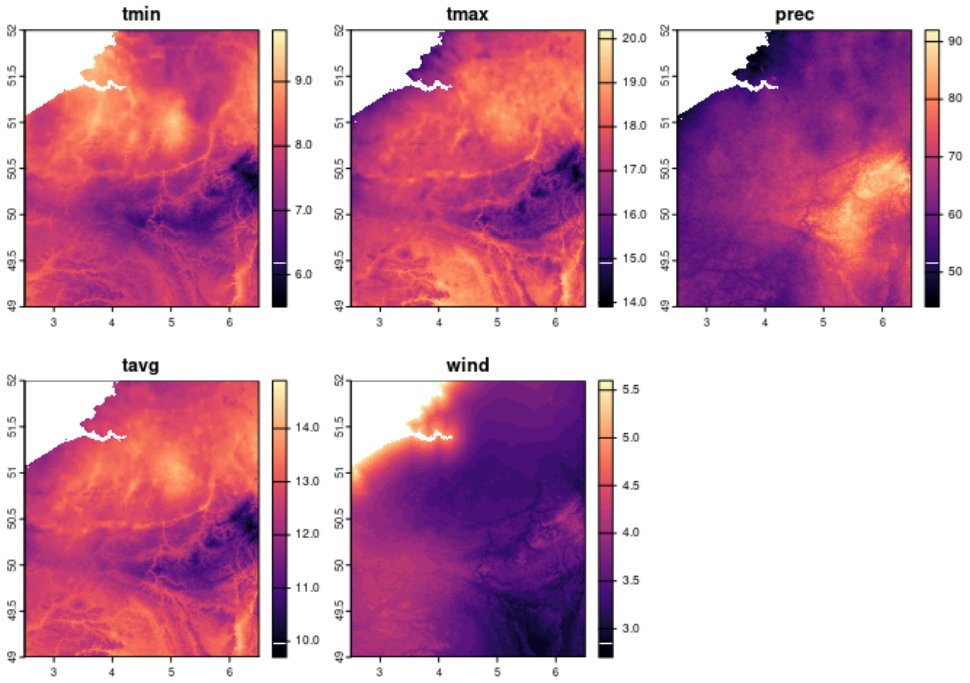

Fig. 8: Climatic predictors for Belgium

The function for this task is *predict_sdm_for_new_area*, which takes as input the trained models and the new set of environmental predictors.
Its output is a data cube where suitability is stored as an attribute, and the dimensions are *x*, *y*, and *species*.

```
 # Prediction over new area
new_predictions_may <- predict_sdm_for_new_area(sdms$models, clima_train_bel_may)

SDM for: Anemone nemorosa L.
SDM for: Chrysosplenium alternifolium L.
SDM for: Galium verum L.
SDM for: Ophrys apifera Huds.
SDM for: Paris quadrifolia L.

# output
new_predictions_may

stars object with 3 dimensions and 1 attribute
attribute(s):
              Min.    1st Qu.    Median      Mean   3rd Qu.      Max.  NA's
suit  5.867529e-13 0.03843691 0.1433858 0.2354462 0.3725311 0.9999935 67380
dimension(s):
        from  to offset     delta refsys                                        values x/y
x          1 480    2.5  0.008333 WGS 84                                          NULL [x]
y          1 360     52 -0.008333 WGS 84                                          NULL [y]
species    1   5     NA        NA     NA Anemone nemorosa L.,...,Paris quadrifolia L.
```

Fig. 9: Data Cube with the suitability prediction for all species

The temporal dimension has not yet been implemented but will be included in the final product alongside the integration of DeepMaxent.

The final step involves polygon-based aggregation. In line with the structure proposed by the B-Cubed framework, we aim to summarize species suitability information within a grid covering the study area.

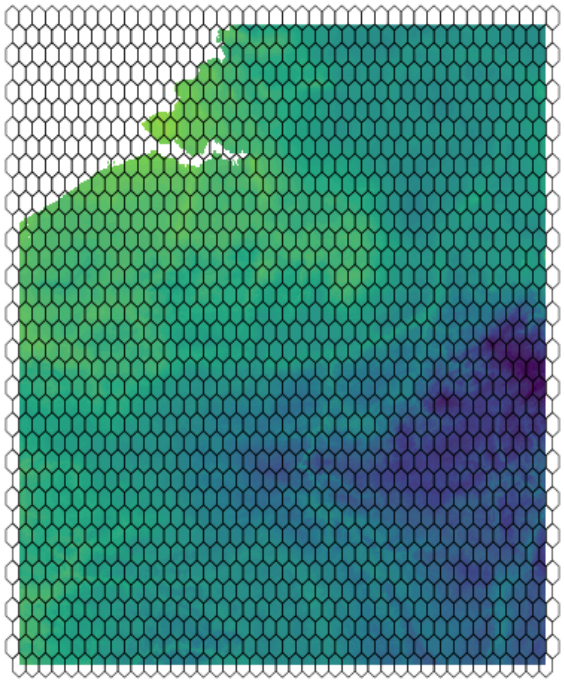

Fig. 10: Grid for Belgium

After creating a grid of polygons using *st_make_grid*, the *aggregate_suitability* function takes this grid and the previously generated data cube as input, computes the mean suitability within each polygon by an average, and returns a stars object.

```
# function
stars_predictions_aggregated <- aggregate_suitability(new_predictions_may, bel_grid)

Aggregating suitability for: Anemone nemorosa L.
Aggregating suitability for: Chrysosplenium alternifolium L.
Aggregating suitability for: Galium verum L.
Aggregating suitability for: Ophrys apifera Huds.
Aggregating suitability for: Paris quadrifolia L.

# print
stars_predictions_aggregated

stars object with 2 dimensions and 1 attribute
attribute(s):
                   Min.    1st Qu.    Median     Mean  3rd Qu.      Max. NA's
suitability  1.540887e-07 0.05111451 0.1581302 0.234773 0.370952 0.9726111  930
dimension(s):
        from   to refsys point                                                values
x          1 1494 WGS 84 FALSE POLYGON ((2.45 49.02887, ...,...,POLYGON ((6.55 51.97335, ...
species    1    5     NA    NA                 Anemone nemorosa L.,...,Paris quadrifolia L.
```

Fig. 11: Final data cube. Now, the 'x' dimension refers to the cells that summarize the geographical information

This object retains all relevant information and allows easy extraction of species-specific suitability values for any given point.

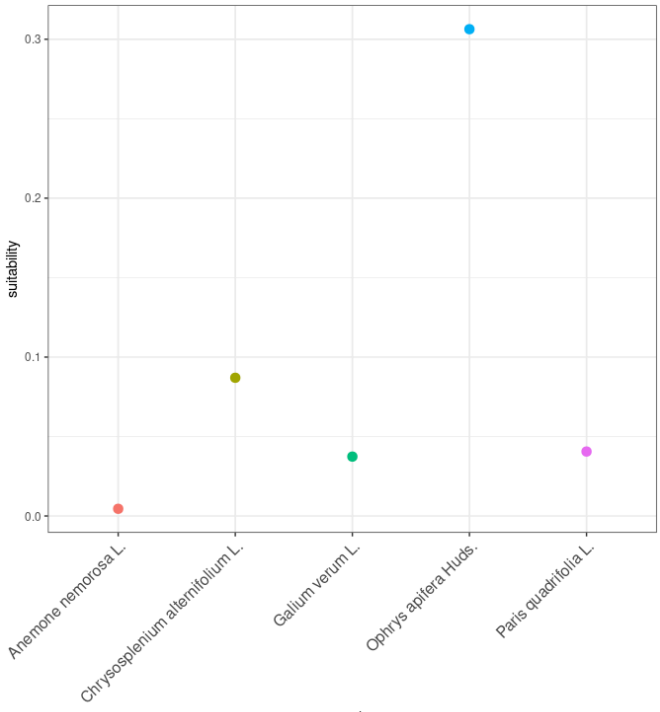

Fig. 12: Suitability of each species in a given cell (685, that contains the city of Bruxelles)

Next steps

The code is available here: https://github.com/b-cubed-eu/virtual-suitability-cube
Next steps will be the following:

- Add the temporal dimensione, as done in the Virtual SC. This is essential for enabling future projections.
- Add an uncertainty measure for each suitability prediction, using the R package dubicube
- Add the Dissimilarity Index as an attribute to measure the distance between training points and predictors (Meyer, 2021)
- Incorporate DeepMaxent: https://github.com/RYCKEWAERT/deepmaxent

References

- Smith, A.B., Murphy, S.J., Henderson, D., and Erickson, K.D. 2023. Including imprecisely georeferenced specimens improves accuracy of species distribution models and estimates of niche breadth. *Global Ecology and Biogeography* In press
- 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, Hanna, and Edzer Pebesma. "Predicting into unknown space? Estimating the area of applicability of spatial prediction models." *Methods in Ecology and Evolution* 12.9 (2021): 1620-1633.

- Steven J. Phillips, Robert P. Anderson, Robert E. Schapire, Maximum entropy modeling of species geographic distributions, Ecological Modelling, Volume 190, Issues 3–4, 2006, Pages 231-259, ISSN 0304-3800, https://doi.org/10.1016/j.ecolmodel.2005.03.026.
- Guisan, Antoine, and Wilfried Thuiller. "Predicting species distribution: offering more than simple habitat models." *Ecology letters* 8.9 (2005): 993-1009.