A Julia toolkit for species distribution data

Timothée Poisot [1]

Abstract: LATER

Introduction

Species Distribution Models (SDMs) are one of the most effective predictive approach to study the global distribution of biodiversity (Elith & Leathwick, 2009). The training and evaluation of a SDM requires many steps, governing both its design and reporting (Zurell et al., 2020) and ultimate use and interpretation (Araújo et al., 2019). In the recent years, there has been an increase in the number of software packages and tools to assist ecologists with the development of species distribution models. As Kass et al. (2024) point out, this increase in the diversity of packages (most of them in the R language) is a good thing, as it can accommodate multiple workflows, and contributes to the adoption of good practices in the field.

Because the practice of species distribution modeling and analysis usually involve many different data types, tools that can provide an integrated environment are important: many existing packages have been designed independently, and therefore may suffer when it comes to interoperability. In this manuscript, we present SpeciesDistributionToolkit (abbreviated as SDT), a meta-package for the Julia programming language, offering an integrated environment for the retrieval, formatting, and interpretation of data relevant to the modeling of species distributions.

Griffith et al. (2024) for large-scale SDM

Application description

...

Component packages

An overview of the SDT package is given in Figure 1. The project is organized as a "monorepo", in which multiple packages live. This allows expanding the scope of the package by moving functionalities into new component packages, without complexifying the installation process. At SDT is registered in the Julia package repository, it can be installed with:

```julia
import Pkg; Pkg.add("SpeciesDistributionToolkit")
```

When loading the SDT package with using SpeciesDistributionToolkit, all component packages are automatically and transparently loaded. Therefore, users do not need to know where a specific method or function resides to use it. In the next section, we discuss how this modular design ensure that we can grow the functionality of the toolkit over time, while maintaining strict backward compatibility and allowing full reproducibility of an analysis.
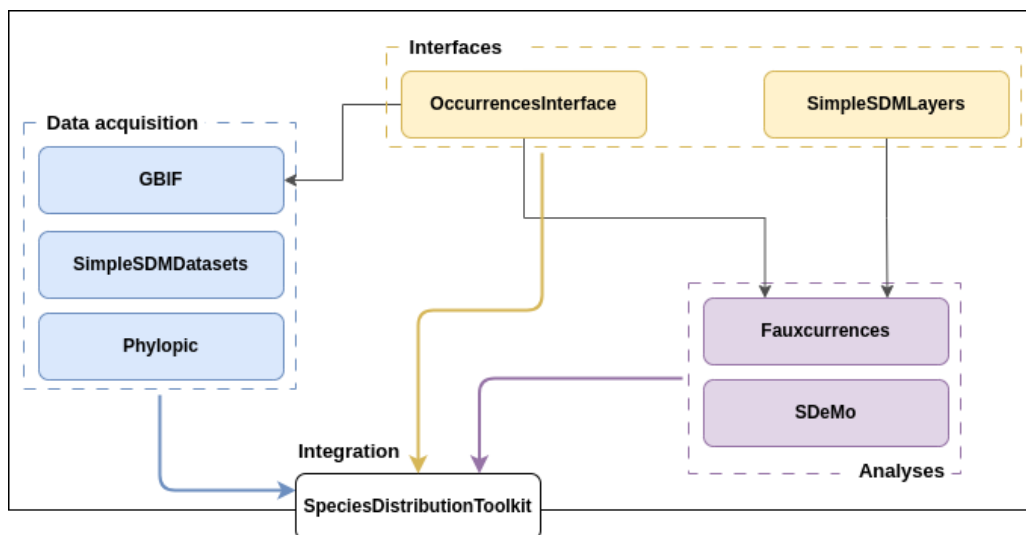

Figure 1: Overview of the packages

GBIF

SimpleSDMDatasets

Phylopic

OccurrencesInterface

SimpleSDMLayers

The Fauxcurrences packages is inspired by the work of Osborne et al. (2022), and

SDeMo

Software information

SDT uses the built-in Julia package manager to ensure that the version of all dependencies are kept up to date. Furthermore, we use strict semantic versioning: major versions correspond to no breaking changes in user-developped code, minor versions increase with additional functionalities, and patch releases cover minor bug fixes or documentation changes. All packages have a CHANGELOG file, which documents what changes are included in each release.

This strict reliance on semantic versioning solves the issues of maintaining compatibility when new functionalities are added: all releases in the v1.x.x branch of SDT depend on component packages in their respective v1.x.x branch, and users can benefit from now functionalities without risking to break existing code. This behavior is extensively tested, both using unit tests, and through integration testing generated as part of the online documentation.

Kellner et al. (2025) reported that about 20% of failures to reproduce species distribution or abundance modeling code was related to package issues. The strict reliance on semantic versioning, alongside technical choices in the Julia package manager and repository, means that it is possible to specify the full version of all dependencies used in a project, which addresses this important obstacle to reproducibility.

Integration with other packages

The SDT package benefits from close integration with other packages in the Julia universe. Notably, this includes Makie (and all related backends) for plotting and data visualisation, where usual plot types are overloaded for layer and occurrence data. Most data can be exported using the Tables interface, which allows data to be consumed by other packages like DataFrames and MLJ. Interfaces internal to Julia are also implemented whenever they make sense. Layers behave like arrays, are iterable, and broadcastable; occurrences collections are arrays and iterables. Beyong supporting external interfaces, SDT defines its own internally. Access to raster data is supported by a trait-based interface for SimpleSDMDatasets, and one of the component packages (OccurrencesInterface implements a minimalist interface to facilite the consumption of occurrence data.

Internal use of other interfaces like StatsAPI

Illustrative case studies

In this section, we provide a series of case studies, meant to illustrate the use of the package. The on-line documentation offers longer tutorials, as well as a series of how-to vignettes to illustrate the full scope of what the package allows. The code for each of these case studies is available as fully independent Jupyter notebooks, forming the supplementary material of this article. The example we use throughout is the distribution of Akodon montensis (Rodentia, family Cricetidae), and a host or orthohantaviruses (Burgos et al., 2021; Owen et al., 2010), in Paraguay. As the notebooks accompanying this article cover the full code required to run these examples, we do not present code snippets in the main text, and instead focus on explaining which component packages are used in each example.

Landcover consensus map

In this case study, we retrieve the land cover data from Tuanmu & Jetz (2014), clip them to a GeoJSON polygon describing the country of Paraguay (SDT can download data directly from gadm.org), and apply the mosaic operation to figure out which class is the most locally abundant. This case study uses the SimpleSDMDatasets package to download (and locally cache) the raster data, as well as the SimpleSDMLayers package to provide basic utility functions on raster data.
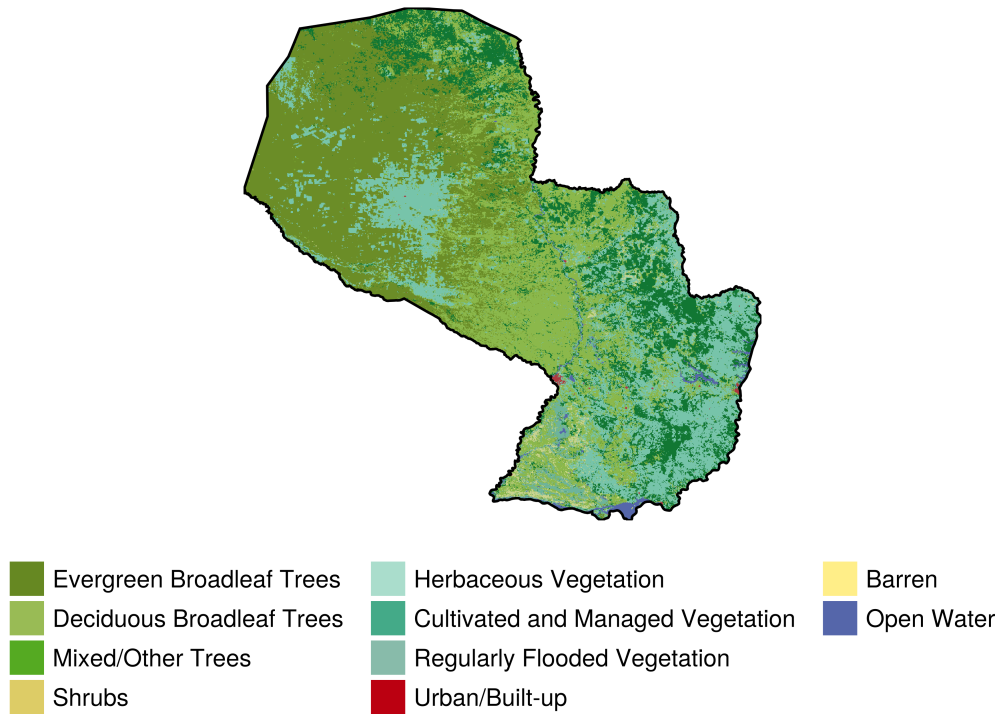

Evergreen Broadleaf Trees   Herbaceous Vegetation   Barren
Deciduous Broadleaf Trees   Cultivated and Managed Vegetation   Open Water
Mixed/Other Trees   Regularly Flooded Vegetation
Shrubs   Urban/Built-up

Figure 2: yeah

Using data from GBIF

(GBIF: The Global Biodiversity Information Facility, 2025)
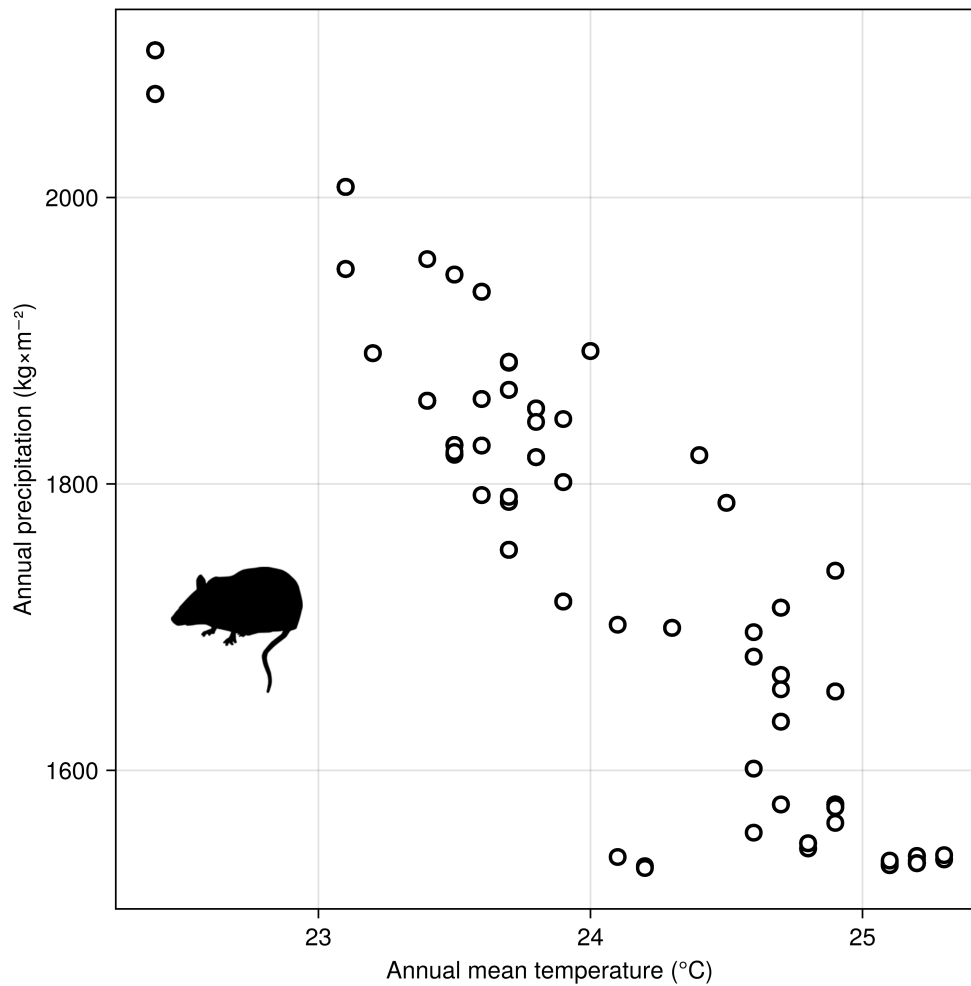
(Karger et al., 2017)

Figure 3: yeah

In practice, although the data are retrieved using the GBIF package, they are used internally by SDT through the OccurrencesInterface package. This package defines a small convention to handle georeferenced occurrence data, and allows to transparently integrate additional occurrence sources. By defining five methods for a custom data type, users can plug-in any occurrence data source and enjoy full compatibility with the entire SDT functionalities.

Training a species distribution model
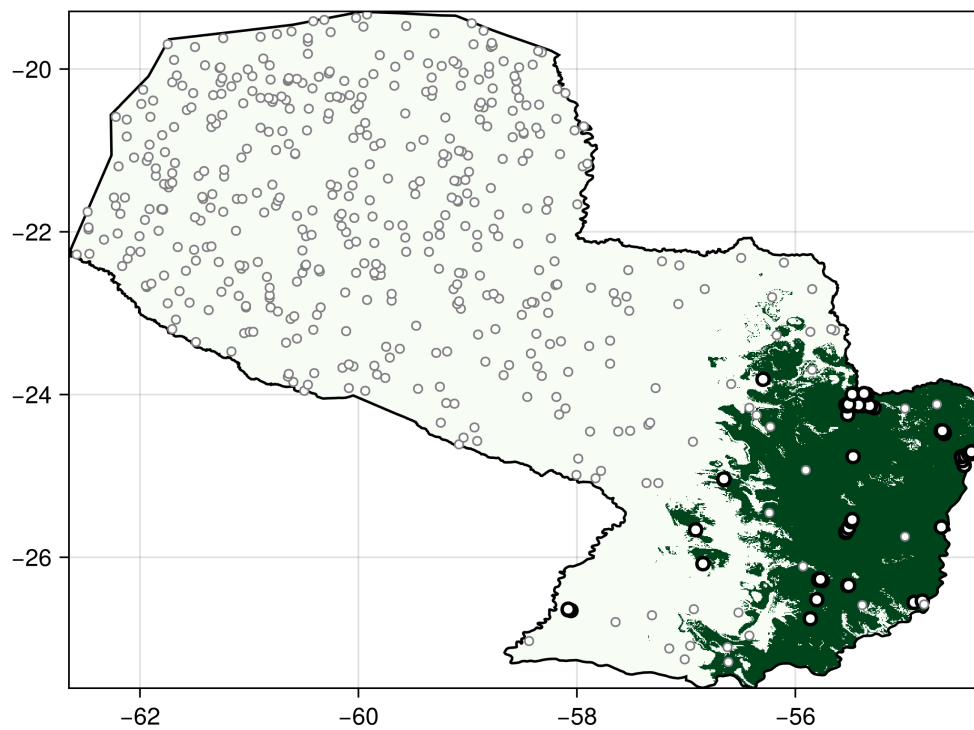
(Bagnall et al., 2018)

Figure 4: also yeah

Generating the distribution of a virtual species
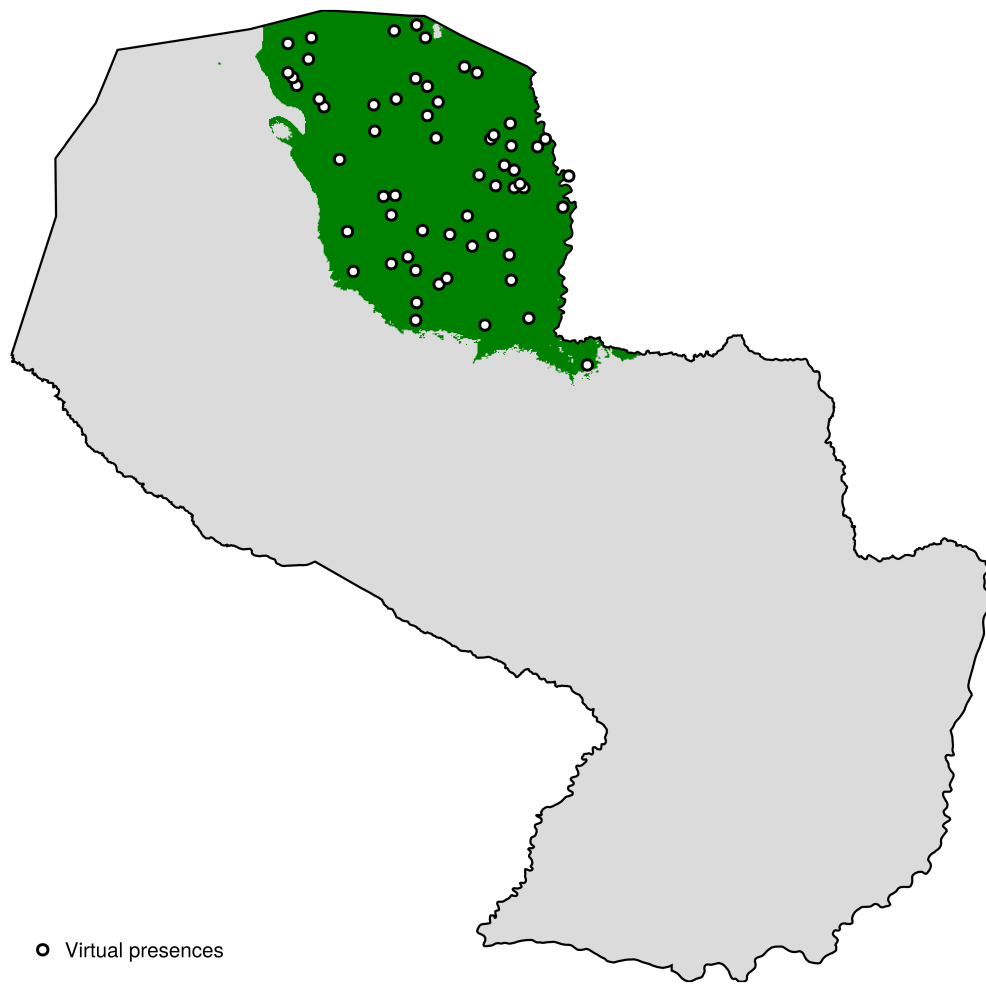
(Leroy et al., 2016)

○ Virtual presences

Figure 5: yeah

References

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. Science Advances, 5, eaat4858. https://doi.org/10.1126/sciadv.aat4858

Bagnall, A., Flynn, M., Large, J., Line, J., Bostrom, A., & Cawley, G. (2018). Is rotation forest the best classifier for problems with continuous features? arXiv [Cs.LG].

Burgos, E. F., Vadell, M. V., Bellomo, C. M., Martinez, V. P., Salomon, O. D., & Gómez Villafañe, I. E. (2021). First evidence of akodon-borne orthohantavirus in

northeastern argentina. EcoHealth, 18, 429–439.
https://doi.org/10.1007/s10393-021-01564-6

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

GBIF: The Global Biodiversity Information Facility. (2025). What is GBIF?

Griffith, J., Lord, J.-M., Catchen, M. D., Arce-Plata, M. I., Bohorquez, M. F. G., Chandramohan, M., Diaz-Corzo, M. C., Gravel, D., Gonzalez, L. F. U., Gutiérrez, C., Helfenstein, I., Hoban, S., Kass, J. M., Laroque, G., Laikre, L., Leigh, D., Leung, B., Mastretta-Yanes, A., Millette, K., … Gonzalez, A. (2024). BON in a box: An open and collaborative platform for biodiversity monitoring, indicator calculation, and reporting. https://doi.org/10.32942/X2M320

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. Scientific Data, 4, 170122. https://doi.org/10.1038/sdata.2017.122

Kass, J. M., Smith, A. B., Warren, D. L., Vignali, S., Schmitt, S., Aiello-Lammens, M. E., Arlé, E., Márcia Barbosa, A., Broennimann, O., Cobos, M. E., Guéguen, M., Guisan, A., Merow, C., Naimi, B., Nobis, M. P., Ondo, I., Osorio-Olvera, L., Owens, H. L., Pinilla-Buitrago, G. E., … Zurell, D. (2024). Achieving higher standards in species distribution modeling by leveraging the diversity of available software. Ecography. https://doi.org/10.1111/ecog.07346

Kellner, K. F., Doser, J. W., & Belant, J. L. (2025). Functional R code is rare in species distribution and abundance papers. Ecology, 106, e4475. https://doi.org/10.1002/ecy.4475

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). Virtualspecies, an R package to generate virtual species distributions. Ecography, 39, 599–607. https://doi.org/10.1111/ecog.01388

Osborne, O. G., Fell, H. G., Atkins, H., Tol, J. van, Phillips, D., Herrera-Alsina, L., Mynard, P., Bocedi, G., Gubry-Rangin, C., Lancaster, L. T., Creer, S., Nangoy, M., Fahri, F., Lupiyaningdyah, P., Sudiana, I. M., Juliandi, B., Travis, J. M. J., Papadopulos, A. S. T., & Algar, A. C. (2022). Fauxcurrence: Simulating multi-species occurrences for null models in species distribution modelling and biogeography. Ecography, 2022, e05880. https://doi.org/10.1111/ecog.05880

Owen, R. D., Goodin, D. G., Koch, D. E., Chu, Y.-K., & Jonsson, C. B. (2010). Spatiotemporal variation in akodon montensis (cricetidae: Sigmodontinae) and hantaviral seroprevalence in a subtropical forest ecosystem. Journal of Mammalogy, 91, 467–481. https://doi.org/10.1644/09-MAMM-A-152.1

Tuanmu, M.-N., & Jetz, W. (2014). A global 1-km consensus land-cover product for biodiversity and ecosystem modelling: Consensus land cover. Global Ecology and Biogeography: A Journal of Macroecology, 23, 1031–1045. https://doi.org/10.1111/geb.12182

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Townsend Peterson, A., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species

distribution models. Ecography, 43, 1261–1277. https://doi.org/10.1111/ecog.04960