A Julia toolkit for species distribution data

Timothée Poisot
Département de Sciences Biologiques, Université de Montréal, Montréal, Canada
timothee.poisot@umontreal.ca

Abstract LATER

Introduction

Species Distribution Models (SDMs) are one of the most effective predictive approach to study the global distribution of biodiversity (Elith and 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.

Kellner, Doser, and Belant (2025) versioning problem

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

Application description

SpeciesDistributionToolkit is released as a package for the **Julia** programming language (Bezanson et al. 2017), licensed under the open-source initiative approved MIT license. It has evolved from a previous collection of packages to handle GBIF data (Dansereau and Poisot 2021), and now provides extended functionalities and improved performances. The package is registered in the **Julia** package repository and can be downloaded and installed anonymously. It is compatible with version 1.8 and above. The full source and complete edition history is available at `https://github.com/PoisotLab/SpeciesDistributionToolkit.jl`. This page additionally has a link to the documentation, containing a full reference for the package functions, a series of briefs how-to examples, and longer vignettes showcasing more integrative examples.

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. As **SDT** is registered in the **Julia** package repository, it can be installed by using `add SpeciesDistributionToolkit` when in package mode at the **Julia** prompt.

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.

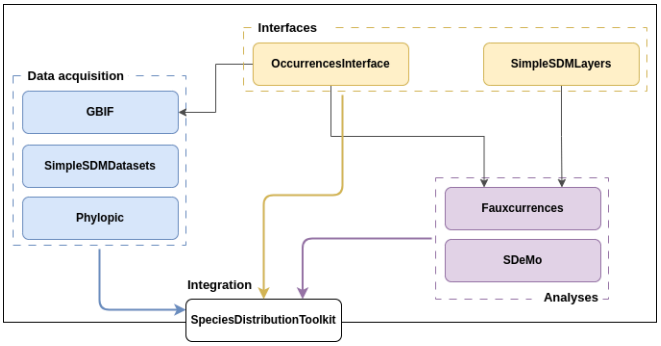

Figure 1: Overview of the packages included in **SpeciesDistributionToolkit**. The packages are color-coded by intended use, and their more specific content is presented in the main text. Note that because the package relies on *interfaces* to facilitate code interoperability, there are only three dependency relationships.

The **SDT** package primarily provides integration between the other packages, through the mechanism of method overloading, allowing to efficiently join packages together (Roesch et al. 2023). Additional functionalities that reside in the top-level package are the generation of pseudo-absences inspired by Barbet-Massin et al. (2012), access to the `gadm.org` database, handling of polygon data, and various quality of life methods. Because of the modular nature of the code, any of these functions can be transparently moved to their own packages in the future.

The **SimpleSDMLayers** package offers a series of types to represent raster data in various projections, and a series of functions to operate on these layers. This package provides the main data representation for most functionalities that **SDT** supports, and handles saving and loading data.

The **OccurrencesInterface** is a light-weight package to provide a common interface for occurrence data. It implements abstract and concrete types to define a single occurrence and a collection thereof, and a series of methods allowing any occurrence data provider or data representation to become fully interoperable with the rest of **SDT**. All **SDT** methods that handle occurrence data do so through the **OccurrencesInterface** interface, allowing future data sources to be integrated without the need for new code.

The **GBIF** package offers access to the `gbif.org` streaming API, including the ability to retrieve, filter, and restart downloads. Although this package returns a rich data representation for occurrence data, all the objects it returns adhere to the **OccurrencesInterface** interface.

SimpleSDMDatasets implements an interface to retrieve and locally store raster data, which can be extended by users to support additional data sources. In addition, it offers access to a series of data sources, including the biodiversity mapping project (Jenkins, Pimm, and Joppa 2013), the EarthEnv collection for land cover (Tuanmu and Jetz 2014) and habitat heterogeneity (Tuanmu and Jetz 2015), Copernicus land cover 100m data (Buchhorn et al. 2020), the PaleoClim (Brown et al. 2018) data, the WorldClim 1 and 2 data (Fick and Hijmans 2017) and their projections under various RCPs and SSP, and part of the CHELSA 1 and 2 data (Karger et al. 2017) and their projections under various RCPs and SSPs.

Phylopic offers a wrapper around the `phylopic.org` API to download silhouettes for taxonomic entities. It also provides utilities for citation of the downloaded images. Its functionalities are similar to the **rphylopic** package (**Gearty2023?**).

The **Fauxcurrences** packages is inspired by the work of Osborne et al. (2022), and allows generating a series of simulated occurrence data that have the same statistical structure as observed ones. The package supports multi-species data, with user-specified relative weight of intra and inter-specific distances conservation.

The **SDeMo** package is aimed at providing tools to use as part of training and education material on species distribution modeling. By providing a series of data transformation (PCA, Whitening, z-score) and classifiers (BIOCLIM, Naive Bayes, and decision trees), it offers the basic elements to demonstrate training and evaluation of SDMs, as well as techniques related to ensembles and bagging. In addition, to promote the use of interpretable techniques, the package supports regular (Elith et al. 2005) and inflated (Zurell, Elith, and Schröder 2012) partial responses, as well as the calculation and mapping of Shapley values (Wadoux, Saby, and Martin 2023; Mesgaran, Cousens, and Webber 2014), and the generation of counterfactuals Karimi et al. (2019).

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. Following a constructive cost model analysis (Kemerer 1987) of the version described in this publication, the package represents approx. 11k lines of active code (no blank lines, no comments), for an estimated development cost of approx. 325k USD.

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, Doser, and Belant (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 occur-

rence 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.

Beyond supporting external interfaces, **SDT** defines its own internally. Access to raster data is supported by a trait-based interface for **SimpleSDMDatasets**.

Internal use of other interfaces like **StatsAPI** in **SDeMo**

one of the component packages (**OccurrencesInterface**) implements a minimalist interface to facilite the consumption of occurrence data.

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 and 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.

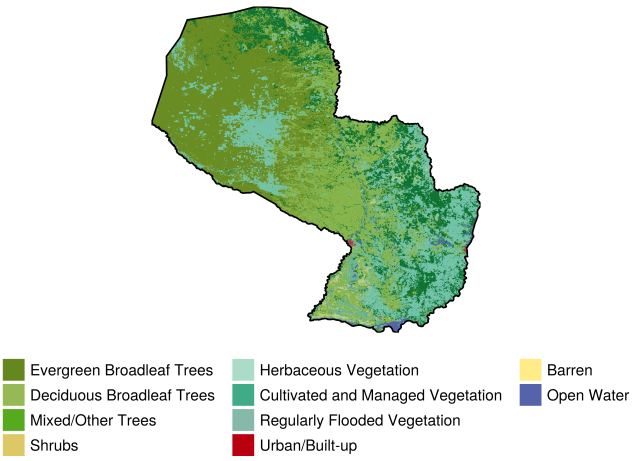

Figure 2: yeah

Using data from GBIF

(GBIF: The Global Biodiversity Information Facility 2025)

(Karger et al. 2017)

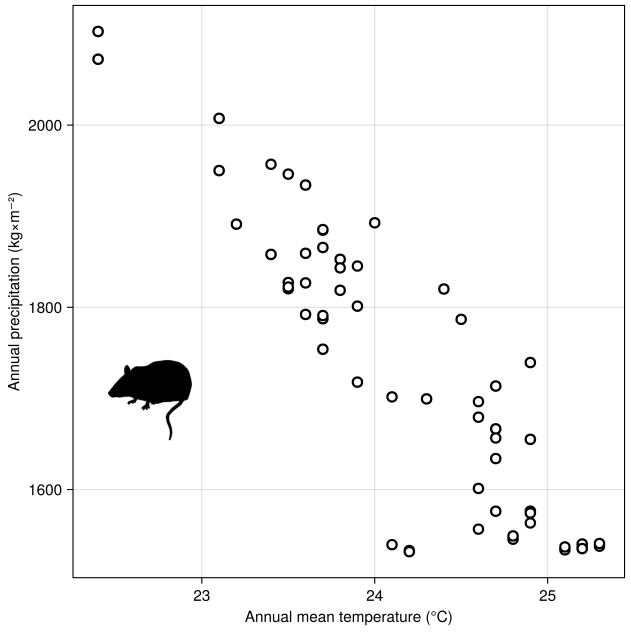

Figure 3: yeah

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)

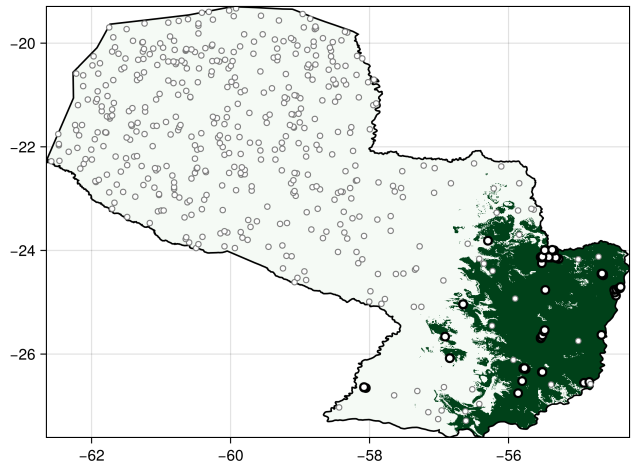

Figure 4: also yeah

Generating the distribution of a virtual species

(Leroy et al. 2016)

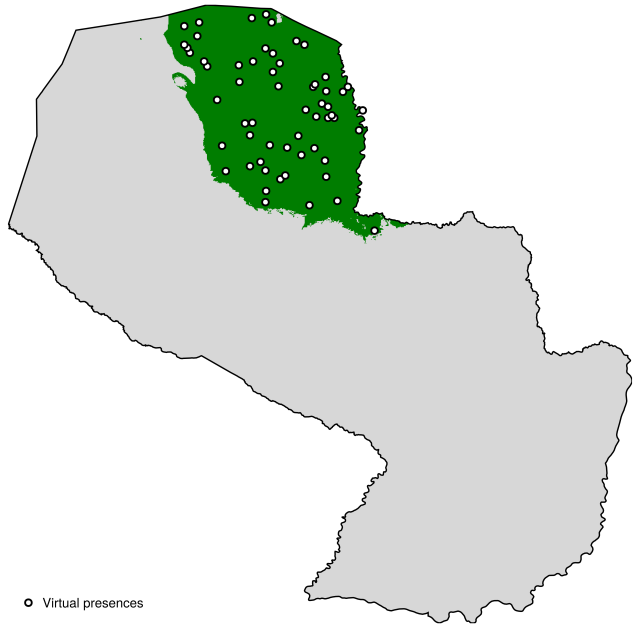

Figure 5: yeah

References

Araújo, Miguel B, Robert P Anderson, A Márcia Barbosa, Colin M Beale, Carsten F Dormann, Regan Early, Raquel A Garcia, et al. 2019. "Standards for Distribution Models in Biodiversity Assessments." *Science Advances* 5 (January): eaat4858. https://doi.org/10.1126/sciadv.aat4858.

Bagnall, A, M Flynn, J Large, J Line, A Bostrom, and G Cawley. 2018. "Is Rotation Forest the Best Classifier for Problems with Continuous Features?" *arXiv [Cs.LG]*, September.

Barbet-Massin, Morgane, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. 2012. "Selecting Pseudo-absences for Species Distribution Models: How, Where and How Many?: How to Use Pseudo-Absences in Niche Modelling?" *Methods in Ecology and Evolution* 3 (April): 327–38. https://doi.org/10.1111/j.2041-210x.2011.00172.x.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review. Society for Industrial and Applied Mathematics* 59 (January): 65–98. https://doi.org/10.1137/141000671.

Brown, Jason L, Daniel J Hill, Aisling M Dolan, Ana C Carnaval, and Alan M Haywood. 2018. "PaleoClim, High Spatial Resolution Paleoclimate Surfaces for Global Land Areas." *Scientific Data* 5 (November): 180254. https://doi.org/10.1038/sdata.2018.254.

Buchhorn, Marcel, Bruno Smets, Luc Bertels, Bert De Roo, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, and Steffen Fritz. 2020. "Copernicus Global Land Service: Land Cover 100m: Collection 3: Epoch 2019: Globe." Zenodo. https://doi.org/10.5281/ZENODO.3939050.

Burgos, E F, M V Vadell, C M Bellomo, V P Martinez, O D Salomon, and I E Gómez Villafañe. 2021. "First Evidence of Akodon-Borne Orthohantavirus in Northeastern Argentina." *EcoHealth* 18 (December): 429–39. https://doi.org/10.1007/s10393-021-01564-6.

Dansereau, Gabriel, and Timothée Poisot. 2021. "SimpleSDMLayers.jl and GBIF.jl: A Framework for Species Distribution Modeling in Julia." *Journal of Open Source Software* 6 (January): 2872. https://doi.org/10.21105/joss.02872.

Elith, Jane, Simon Ferrier, Falk Huettmann, and John Leathwick. 2005. "The Evaluation Strip: A New and Robust Method for Plotting Predicted Responses from Species Distribution Models." *Ecological Modelling* 186 (August): 280–89. https://doi.org/10.1016/j.ecolmodel.2004.12.007.

Elith, Jane, and John R Leathwick. 2009. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution, and Systematics* 40 (December): 677–97. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Fick, Stephen E, and Robert J Hijmans. 2017. "WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS." *International Journal of Climatology: A Journal of the Royal Meteorological Society* 37 (October): 4302–15. https://doi.org/10.1002/joc.5086.

GBIF: The Global Biodiversity Information Facility. 2025. "*What Is GBIF?*" 2025.

Griffith, Jory, Jean-Michel Lord, Michael D Catchen, Maria Isabel Arce-Plata, Manuel Fernandez Galvez Bohorquez, Matusan Chandramohan, María Camilla Diaz-Corzo, et al. 2024. "BON in a Box: An Open and Collaborative Platform for Biodiversity Monitoring, Indicator Calculation, and Reporting," October. https://doi.org/10.32942/X2M320.

Jenkins, Clinton N, Stuart L Pimm, and Lucas N Joppa. 2013. "Global Patterns of Terrestrial Vertebrate Diversity and Conservation." *Proceedings of the National Academy of Sciences of the United States of America* 110 (July): E2602–10. https://doi.org/10.1073/pnas.1302251110.

Karger, Dirk Nikolaus, Olaf Conrad, Jürgen Böhner, Tobias Kawohl, Holger Kreft, Rodrigo Wilber Soria-Auza, Niklaus E Zimmermann, H Peter Linder, and Michael Kessler. 2017. "Climatologies at High Resolution for the Earth's Land Surface Areas." *Scientific Data* 4 (September): 170122. https://doi.org/10.1038/sdata.2017.122.

Karimi, Amir-Hossein, Gilles Barthe, Borja Balle, and Isabel Valera. 2019. "Model-Agnostic Counterfactual Explanations for Consequential Decisions." *arXiv [Cs.LG]*, May.

Kass, Jamie M, Adam B Smith, Dan L Warren, Sergio Vignali, Sylvain Schmitt, Matthew E Aiello-Lammens, Eduardo Arlé, et al. 2024. "Achieving Higher Standards in Species Distribution Modeling by Leveraging the Diversity of Available Software." *Ecography*, November. https://doi.org/10.1111/ecog.07346.

Kellner, Kenneth F, Jeffrey W Doser, and Jerrold L Belant. 2025. "Functional R Code Is Rare in Species Distribution and Abundance Papers." *Ecology* 106 (January): e4475. https://doi.org/10.1002/ecy.4475.

Kemerer, Chris F. 1987. "An Empirical Validation of Software Cost Estimation Models." *Communications of the ACM* 30 (May): 416–29. https://doi.org/10.1145/22899.22906.

Leroy, Boris, Christine N Meynard, Céline Bellard, and Franck Courchamp. 2016. "Virtualspecies, an R Package to Generate Virtual Species Distributions." *Ecography* 39 (June): 599–607. https://doi.org/10.1111/ecog.01388.

Mesgaran, Mohsen B, Roger D Cousens, and Bruce L Webber. 2014. "Here Be Dragons: A Tool for Quantifying Novelty Due to Covariate Range and Correlation Change When Projecting Species Distribution Models." *Diversity & Distributions* 20 (October): 1147–59. https://doi.org/10.1111/ddi.12209.

Osborne, Owen G, Henry G Fell, Hannah Atkins, Jan van Tol, Daniel Phillips, Leonel Herrera-Alsina, Poppy Mynard, et al. 2022. "Fauxcurrence: Simulating Multi-species Occurrences for Null Models in Species Distribution Modelling and Biogeography." *Ecography* 2022 (July): e05880. https://doi.org/10.1111/ecog.05880.

Owen, Robert D, Douglas G Goodin, David E Koch, Yong-Kyu Chu, and Colleen B Jonsson. 2010. "Spatiotemporal Variation in Akodon Montensis (Cricetidae: Sigmodontinae) and Hantaviral Seroprevalence in a Subtropical Forest Ecosystem." *Journal of Mammalogy* 91 (April): 467–81. https://doi.org/10.1644/09-MAMM-A-152.1.

Roesch, Elisabeth, Joe G Greener, Adam L MacLean, Huda Nassar, Christopher Rackauckas, Timothy E Holy, and Michael P H Stumpf. 2023. "Julia for Biologists." *Nature Methods* 20 (May): 655–64. https://doi.org/10.1038/s41592-023-01832-z.

Tuanmu, Mao-Ning, and Walter Jetz. 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 (September): 1031–45. https://doi.org/10.1111/geb.12182.

———. 2015. "A Global, Remote Sensing-based Characterization of Terrestrial Habitat Heterogeneity for Biodiversity and Ecosystem Modelling: Global Habitat Heterogeneity." *Global Ecology and Biogeography: A Journal of Macroecology* 24 (November): 1329–39. https://doi.org/10.1111/geb.12365.

Van Looveren, Arnaud, and Janis Klaise. 2019. "Interpretable Counterfactual Explanations Guided by Prototypes." *arXiv [Cs.LG]*, July.

Wadoux, Alexandre M J-C, Nicolas P A Saby, and Manuel P Martin. 2023. "Shapley Values Reveal the Drivers of Soil Organic Carbon Stock Prediction." *SOIL* 9 (January): 21–38. https://doi.org/10.5194/soil-9-21-2023.

Zurell, Damaris, Jane Elith, and Boris Schröder. 2012. "Predicting to New Environments: Tools for Visualizing Model Behaviour and Impacts on Mapped Distributions." *Diversity & Distributions* 18 (June): 628–34. https://doi.org/10.1111/j.1472-4642.2012.00887.x.

Zurell, Damaris, Janet Franklin, Christian König, Phil J Bouchet, Carsten F Dormann, Jane Elith, Guillermo Fandos, et al. 2020. "A Standard Protocol for Reporting Species Distribution Models." *Ecography* 43 (September): 1261–77. https://doi.org/10.1111/ecog.04960.