

A Julia toolkit for species distribution data

Timothée Poisot¹, Ariane Bussi eres-Fournel¹, Gabriel Dansereau¹ and Michael D. Catchen¹

¹ Universit   de Montr  al, D  partement de Sciences Biologiques, Montr  al QC, Canada

Correspondence to Timoth  e Poisot — timothee.poisot@umontreal.ca

Abstract: (1) Species distribution modeling requires to handle varied types of data, and benefits from an integrated approach to programming. (2) We introduce **SpeciesDistributionToolkit**, a **Julia** package aiming to facilitate the production of species distribution models. It covers various steps of the data collection and analysis process, extending to the development of interfaces for integration of additional functionalities. (3) By relying on semantic versioning and strong design choices on modularity, we expect that this package will lead to improved reproducibility and long-term maintainability. (4) We illustrate the functionalities of the package through several case studies, accompanied by reproducible code.

Keywords: species distribution models, biogeography, occurrence data, land use, climatic data, pseudo-absences

1 Introduction

2 Species Distribution Models [SDMs; Elith and Leathwick (2009)], in addition to being key
3 tools to further our knowledge of biodiversity, are key components of effective conservation
4 decisions (Guisan et al. 2013), planning (McShea 2014), and ecological impact assesment
5 (Baker et al. 2021). The training and evaluation of a SDM is a complex process, with key
6 decisions to make on design and reporting (Zurell et al. 2020). The ability to use the correct
7 data format of representation at of these steps is central to support the correct interpretation
8 of these models (Araújo et al. 2019). This is particularly true since the choice of data source
9 can affect the prediction significantly (Booth 2022, Arenas-Castro et al. 2022), suggesting that
10 there is a need for flexible pipelines in which data sources can be conveniently swapped. In
11 recent years, there has been an increase in the number of software packages and tools to assist
12 ecologists with various steps of the development of species distribution models.

13 As Kass et al. (2024) point out, this increase in the diversity of software tools (most of them
14 in the **R** language) is a good thing. Because the SDMs are a general-purpose methodology, a
15 varied software offers increases the chances that specific decisions can be chained together in
16 the way that best support a specific use case. By making code available for all users, package
17 developers reduce the need for custom implementation of analytical steps, and contribute to
18 the adoption of good practices in the field. However, because building, validating, and apply-
19 ing SDMs requires a diversity of data types, from different sources, many existing packages
20 have been designed independently. Therefore, they may suffer from low interoperability,
21 which can create friction when using multiple tools together. As an illustration, Kellner et al.
22 (2025) highlight that, out of publications on abundance or distribution models that share code
23 and data, about 20% are not reproducible because of issues in package dependencies.

24 To promote interoperability and improve reproducibility, tools that provide an integrated
25 environment are important. In this manuscript, we present **SpeciesDistributionToolkit**
26 (abbreviated as **SDT**), a meta-package for the **Julia** programming language, offering an
27 integrated environment for the retrieval, formatting, and interpretation of data relevant to
28 the modeling of species distributions. **SDT** was in part designed to work within the BON-
29 in-a-Box project (Gonzalez et al. 2023, Griffith et al. 2024), a GEO BON initiative to facilitate
30 the calculation and reporting of biodiversity indicators supporting the Kunming-Montréal
31 Global Biodiversity Framework. A leading design consideration for **SDT** was therefore to
32 maximize interoperability between components and functionalities from the ground up. This

is achieved through three mechanisms. First, by relying on strict semantic versioning: package releases provide information about the compatibility of existing code. Second, through the use of interfaces: separate software components (including ones external to the package) can interact without prior knowledge of either implementation, and without *dependencies* between the components of **SDT**. Finally, through the use of **Julia**'s extension mechanism. These are detailed in Box 1.

In this manuscript, we describe provide a high-level overview of the functionalities of the package(s) forming **SDT**. We then discuss design principles that facilitate long-term maintenance, development, and integration. We finish by presenting four illustrative case studies: extraction of data at known species occurrences, manipulation of multiple geospatial layers, training and explanation of a SDM, and creation of virtual communities to simulate the spatial distribution of ecological uniqueness. This later case study is intended to provide an impression of what using **SDT** as a support for the development of novel analyses feels like. All of the case studies are available as supplementary material, in the form of fully reproducible, self-contained Jupyter notebooks.

Application description

SpeciesDistributionToolkit is released as a package for the **Julia** programming language (Bezanson et al. 2017). It is licensed under the open-source initiative approved MIT license. It has evolved from a previous collection of packages to handle GBIF and raster data (Dansereau and Poisot 2021), and now provides extended functionalities as well as improved performance. The package is registered in the **Julia** package repository and can be downloaded and installed anonymously. It is compatible with the current long-term support (LTS) release of **Julia**. The full source code, complete commit history, plans for future development, and a forum, are 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 tutorials.

An overview of the **SDT** package is given in Figure 1. The project is organized as a “monorepo”, in which separate but interoperable packages (meaning that they can be installed independently, but are designed to work cohesively) reside. This allows expanding the scope of the package by moving functionalities into new component packages, without requiring interventions from users. 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. The monorepo structure has an important advantage for users: the code of all component packages can be found in the same location, and it makes inspecting the internal implementation of any package easier. In addition, users can open an issue describing a problem or desired feature within the monorepo, without needing to understand which component package is the right target for this issue. This both decreases barriers to interact with the software, while also facilitating the work of contributors who can look at all the issues to address in a centralized way. Similarly, monorepo lend themselves to integrated documentation, which is the approach we have chosen with the online **SDT** manual.

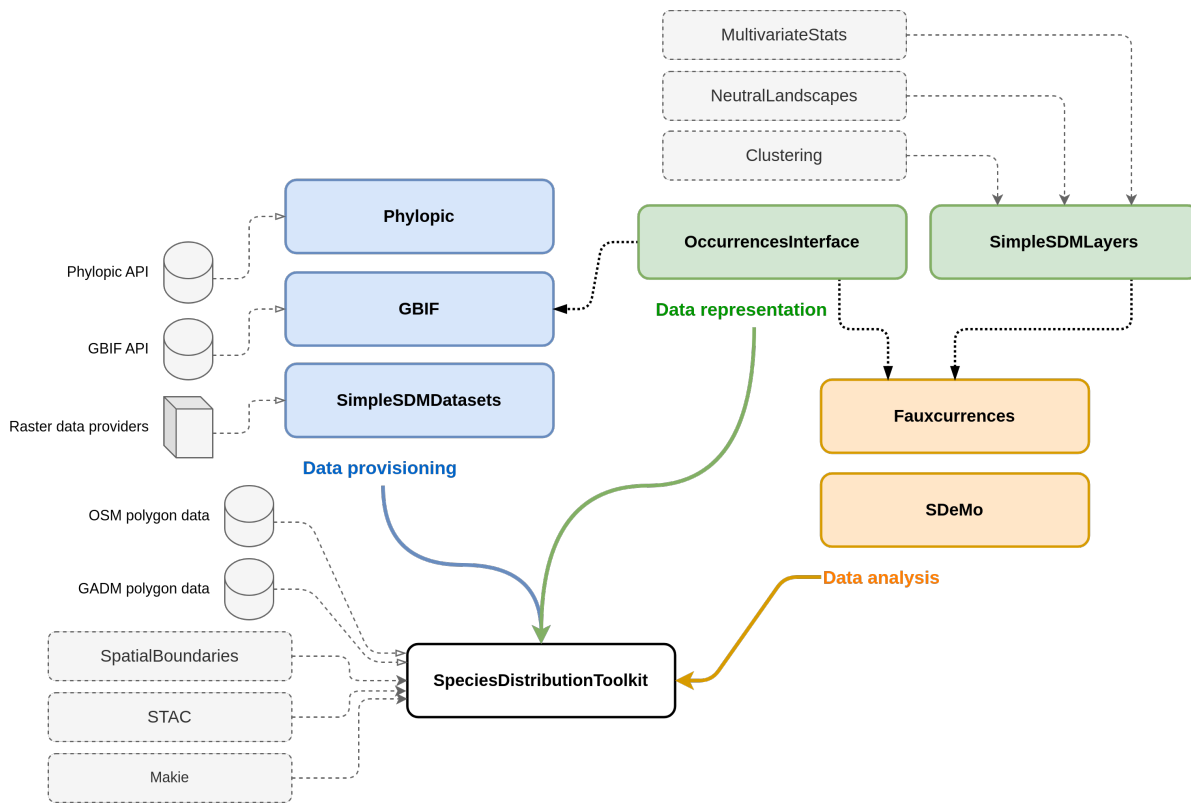


Figure 1: Overview of the packages included in **SpeciesDistributionToolkit**. The packages are color-coded by intended use (acquisition, representation, and analysis of data). The specific content of each package is presented in the main text. Note that because the package relies on *interfaces* to facilitate code interoperability, there are only three dependency relationships (black arrows). Some packages can interact with data sources, represented on the left side of the figure. When loading **SpeciesDistributionToolkit**, all public methods from the package are accessible to the user. Packages that are supported through extensions are in dashed boxes.

SDT uses the built-in **Julia** package manager to keep all dependencies up to date. Furthermore, we use strict semantic versioning: major versions correspond to changes that would break user-developped code; minor versions represent additional functionalities; patch

releases cover minor bug fixes or documentation changes. All component packages are versioned independently, and have their own *CHANGELOG* file documenting each release. This strict reliance on semantic versioning removes 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 new functionalities without needing to adapt existing code. This behavior is extensively tested, both through unit tests and through integration testing generated as part of the online documentation.

Component packages

The **SDT** package primarily provides integration between the other packages via method overloading (reusing method names for intuitive and concise code), 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 (Barbet-Massin et al. 2012), access to the *gadm.org* database, handling of polygon data and zonal statistics, 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 without affecting reproducibility. Note that all packages can still be installed (and would be fully functional) independently.

The **SimpleSDMLayers** package offers a series of types to represent raster data in arbitrary projections defined by a proj string (Evenden et al. 2024). This package provides the main data representation for most spatial functionalities that **SDT** supports, and handles saving and loading data. It also contains utility functions to deal with raster data, including interpolation to different spatial grids and CRS, rescaling and quantization of data, masking, and most mathematical operations that can be applied to rasters.

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 (e.g. GBIF) or data representation to become fully interoperable with the rest of **SDT**. All **SDT** methods that handle occurrence data do so through the interface provided by the **OccurrencesInterface** package, allowing future data sources to be integrated without the need for new code.

The **GBIF** package offers access to the *gbif.org* streaming API (GBIF: The Global Biodiversity Information Facility 2025), including the ability to retrieve, filter, and restart downloads. Although this package provides a rich data representation for occurrence data when access to the full GBIF data schema is required, all the objects it returns adhere to the **OccurrencesIn-**

terface interface. The package also offers the functionality to download datasets from GBIF using their DOI.

SimpleSDMDatasets implements an interface to retrieve and locally store raster data, which can be extended by users to support additional data sources. It offers access to a series of common data sources for spatial biodiversity modeling, including the biodiversity mapping project (Jenkins et al. 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), PaleoClim (Brown et al. 2018) data, WorldClim 1 and 2 (Fick and Hijmans 2017) and CHELSA 1 and 2 (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 (Gearty and Jones 2023).

Fauxcurrences 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 weights for conserving intra and inter-specific occurrence distances.

Finally, **SDeMo** provides a high-level interface to the training, validation, and interpretation of species distribution modeling. By providing a series of data transformation (PCA, Whitening, z-score) and classifiers (currently BIOCLIM, Naive Bayes, logistic regression, and decision trees), it offers the basic elements to demonstrate training and evaluation of SDMs, as well as techniques related to heterogeneous ensembles and bagging with support for arbitrary consensus (Marmion et al. 2009) and voting (Drake 2014) functions. **SDeMo** promotes the use of interpretable techniques: the package supports regular (Elith et al. 2005) and inflated (Zurell et al. 2012) partial responses, as well as the calculation and mapping of Shapley values (Mesgaran et al. 2014, Wadoux et al. 2023) using the standard Monte-Carlo approach (Mitchell et al. 2021). Counterfactuals (Karimi et al. 2019, Van Looveren and Klaise 2019), representing perturbation of the input data leading to the opposite prediction (*i.e.* “what environmental conditions would lead to the species being absent”) can also be generated. The API of **SDeMo** has been designed to (i) enforce the use of best practices, and (ii) be consistent across analyses, so that the package can be used for educational material. Because **SDeMo** is a generic interface to any predictive model, users can expand it by adding additional packages. This can be done either through a contribution to the **SDT** repository, or as part of the code written by users for a specific analysis.

151 **TODO CITE BOOTH2014 BIOCLIM**

152 **TODO PSEUDOABSENCES**

153 **TODO POLYGONS**

154 **Case studies**

155 In this section, we provide a series of case studies to illustrate the use of the package. The
156 on-line manual offers longer tutorials, as well as a series of how-to vignettes to illustrate the
157 full scope of what the package allows. As the notebooks accompanying this article cover the
158 full code required to run these case studies, we do not present code snippets in the main
159 text (as they are presented with detailed explanations in the Supp. Mat.), but rather focus on
160 explaining how the component packages work together in each example.

161 *Landcover consensus map*

162 In this case study (Supp. Mat. 1), we retrieve the land cover data from Tuanmu and Jetz (2014),
163 clip them to a GeoJSON polygon describing the country of Paraguay (**SDT** can download data
164 directly from `gadm.org`), and apply the `mosaic` operation to figure out which class is the most
165 locally abundant. This case study uses the **SimpleSDMDatasets** package to download (and
166 locally cache) the raster data, as well as the **SimpleSDMLayers** package to provide basic
167 utility functions on raster data. The results are presented in Figure 2.

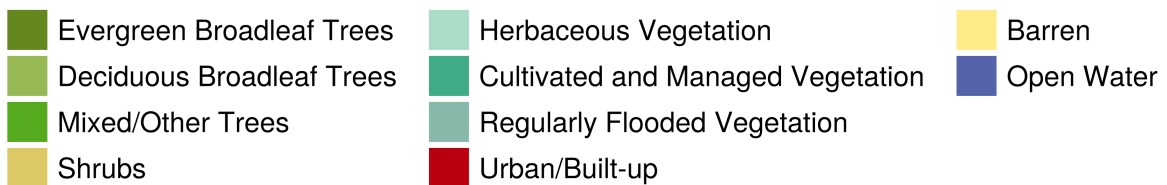
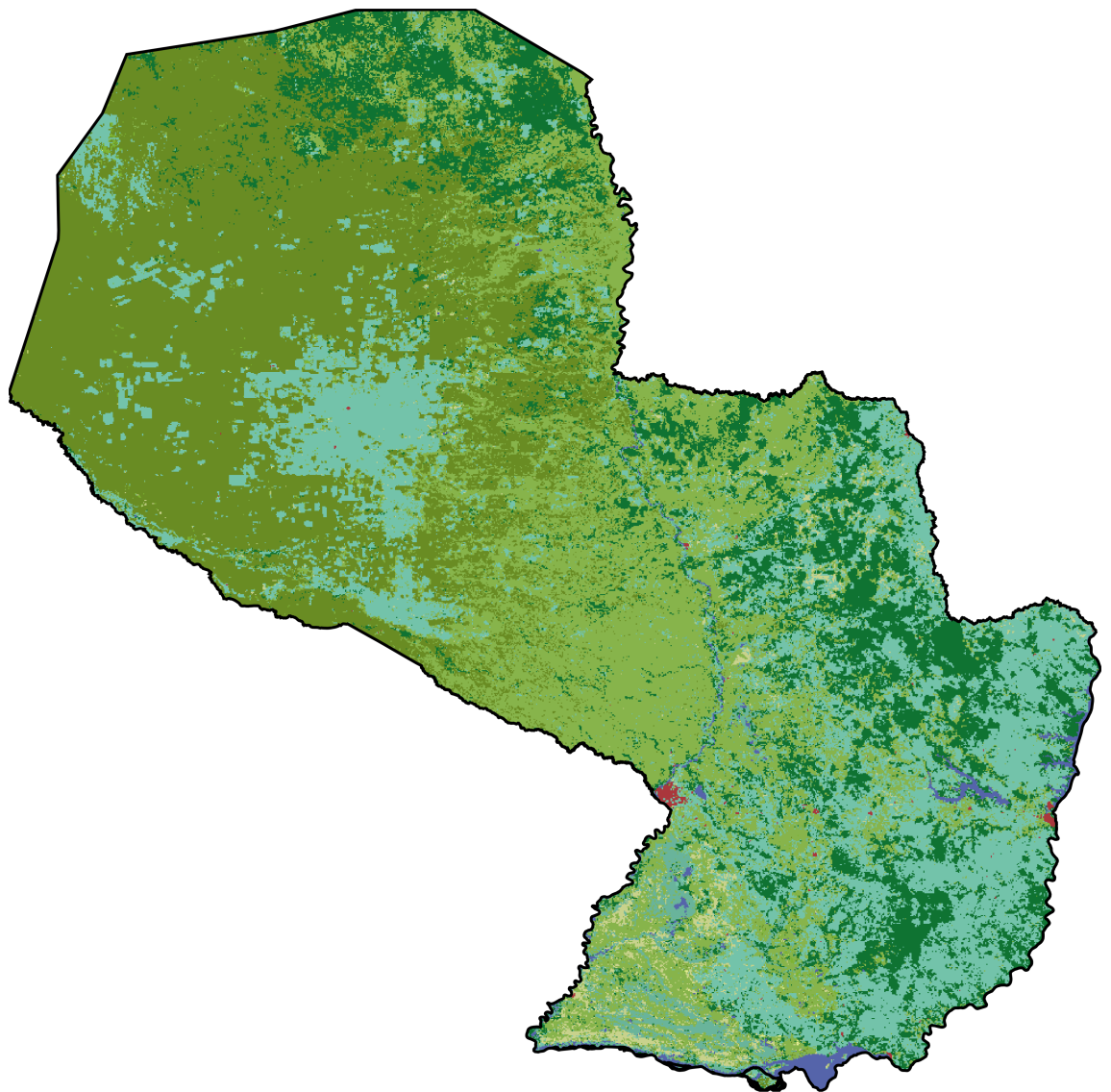


Figure 2: Land cover consensus (defined as the class with the strongest local representation) in the country of Paraguay. Only the classes that were most abundant in at least one pixel are represented. The code to produce this figure is available as Supp. Mat. 2.

SimpleSDMDatasets uses local storage of raster data for future use, to avoid re-downloading data upon repeated use. The location of the data is (i) standardized by the package itself, making the file findable to humans, and (ii) changeable by the user to, *e.g.*, store the data within the project folder rather than in a central location. As much as possible, **SDT** will only read the part of the raster data that is required given the region of interest to the user. This is done by providing additional context in the form of a bounding box (in WGS84, regardless of the underlying raster data projection, in line with the GeoJSON specification). **SDT** has methods to calculate the bounding box for all the objects it supports.

Using data from GBIF

SDT provides strong integration between data on species occurrences and source of geospatial information. To illustrate this, we will collect data on the distribution of *Akodon montensis* (Rodentia, family Cricetidae), a known host of orthohantaviruses (Owen et al. 2010, Burgos et al. 2021), in Paraguay. In Supp. Mat. 2 we (i) request occurrence data using the **GBIF** package, (ii) download the silhouette of the species through **Phylopic**, and (iii) extract temperature and precipitation data at the points of occurrence based on bioclimatic data layers. The results are presented in Figure 3. The full notebook includes information about basic operations on raster data, as well as extraction of data based on occurrence records.

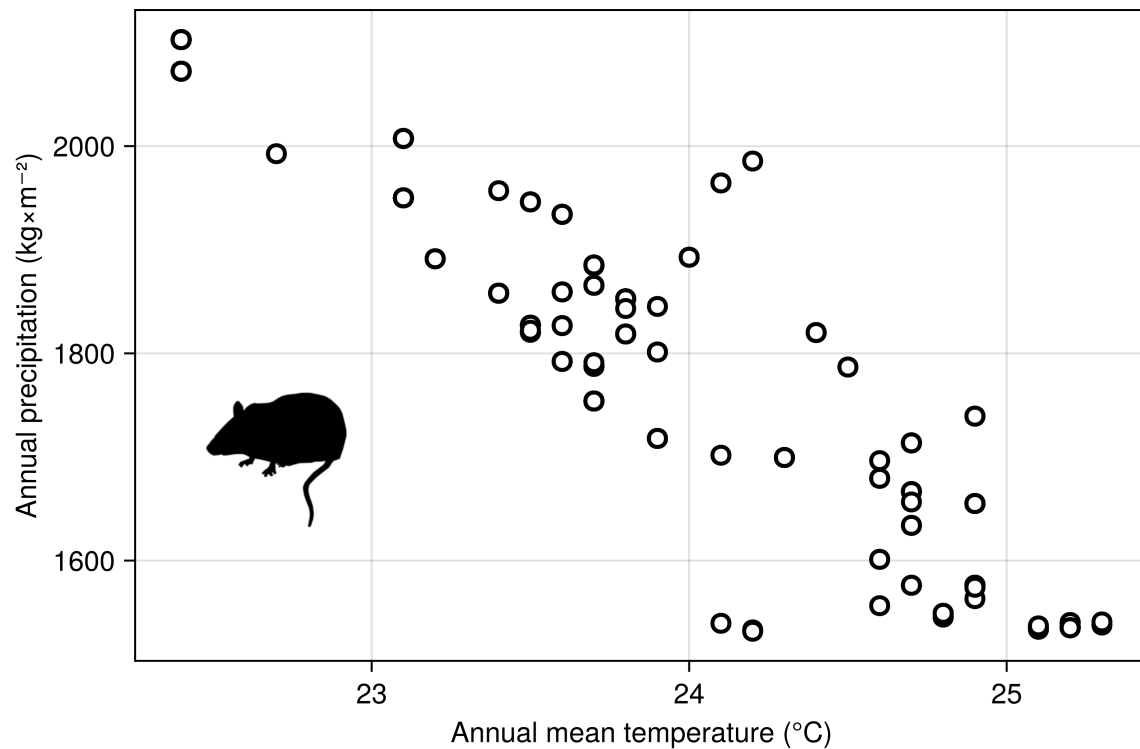


Figure 3: Relationship between temperature and precipitation (BIO1 and BIO12) at each georeferenced occurrence known to GBIF for *Akodon montensis*. The code to produce this figure is available as Supp. Mat. 1.

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 a handful of methods for a custom data type, or by using the convertes built into the package, users can plug-in any occurrence data source or csv file, and enjoy full compatibility with the entire **SDT** functionalities.

Training a species distribution model

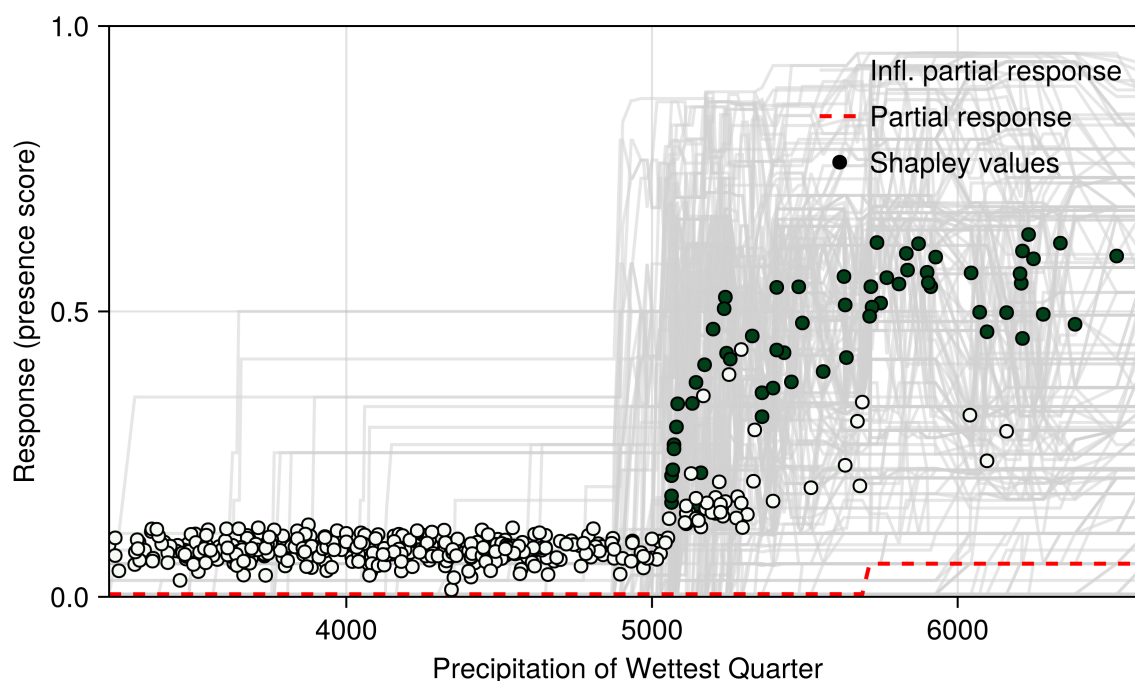
In this case study, we illustrate the integration of **SDeMo** and **SimpleSDMLayers** to train a species distribution model. Specifically, we re-use the data from Figure 3, with additional layers of bioclimatic variables. We train a rotation forest (Bagnall et al. 2018), an homogeneous ensemble of PCA followed by decision trees where each model has a subset of features and training data. The results are presented in Figure 4. The model is built by selecting an optimal suite of BioClim variables, then predicted in space, and the resulting predicted species range is finally clipped by the elevational range observed in the occurrence data. The data transformations in **SDeMo** are always applied in a way that prevents the possibility of data leakage (Stock et al. 2023).

208 **TODO** add splits are balanced by default (but also add PA package)
209 **TODO** CITE BOOTH 2024 CHECKING VARIABLES + Plot partial responses to see if artifacts



210
211 Figure 4: Predicted range of *Akodon montensis* in Paraguay based on a rotation forest trained on GBIF
212 occurrences and the BioClim variables. The predicted range is clipped to the elevational range of the species.
213 The code to produce this figure is available as Supp. Mat. 3.
214 The full notebook (Supp. Mat. 3) has additional information on routines for variable selection,
215 stratified cross-validation, as well as the construction of the ensemble from a single PCA and
216 decision tree. In addition, we report in Figure 5 the partial and inflated partial responses to
217 the most important variable (highlighting an interpretable effect of the variable in the model),
218 as well as the (Monte-Carlo) Shapley values (Mitchell et al. 2021, Wadoux et al. 2023) for
219 each prediction in the training set. Because **SDeMo** works through generic functions, these

220 methods can be applied to any model specified by the user. In practice, generic purpose ML
 221 frameworks **Julia**, notably **MLJ** (Blaom et al. 2020), can also be used and interfaced with **SDT**
 222 by using the classifier and transformer interface.



223
 224 Figure 5: Partial responses (red) and inflated partial responses (grey) to the most important variable. In
 225 addition, the Shapley values for all training data are presented in the same figure; green points are presences,
 226 and pale points are pseudo-absences. Shapley values were added to the average model prediction to be
 227 comparable to partial responses. The code to produce this figure is available as Supp. Mat. 3.

228 *Species and location contribution to beta diversity*

229 In the final case study (Supp. Mat. 4), we simulate the distribution of virtual species (Hirzel
 230 et al. 2001) with a logistic response to two environmental covariate (Leroy et al. 2016). We
 231 then use this simulated sample to perform the decomposition of β -diversity introduced by
 232 Legendre and De Cáceres (2013) and applied by Dansereau et al. (2022) to spatially continuous
 233 data. This simulates the potential distribution of hotspots and coldspots of ecological unique-
 234 ness. The results are presented in Figure 6.

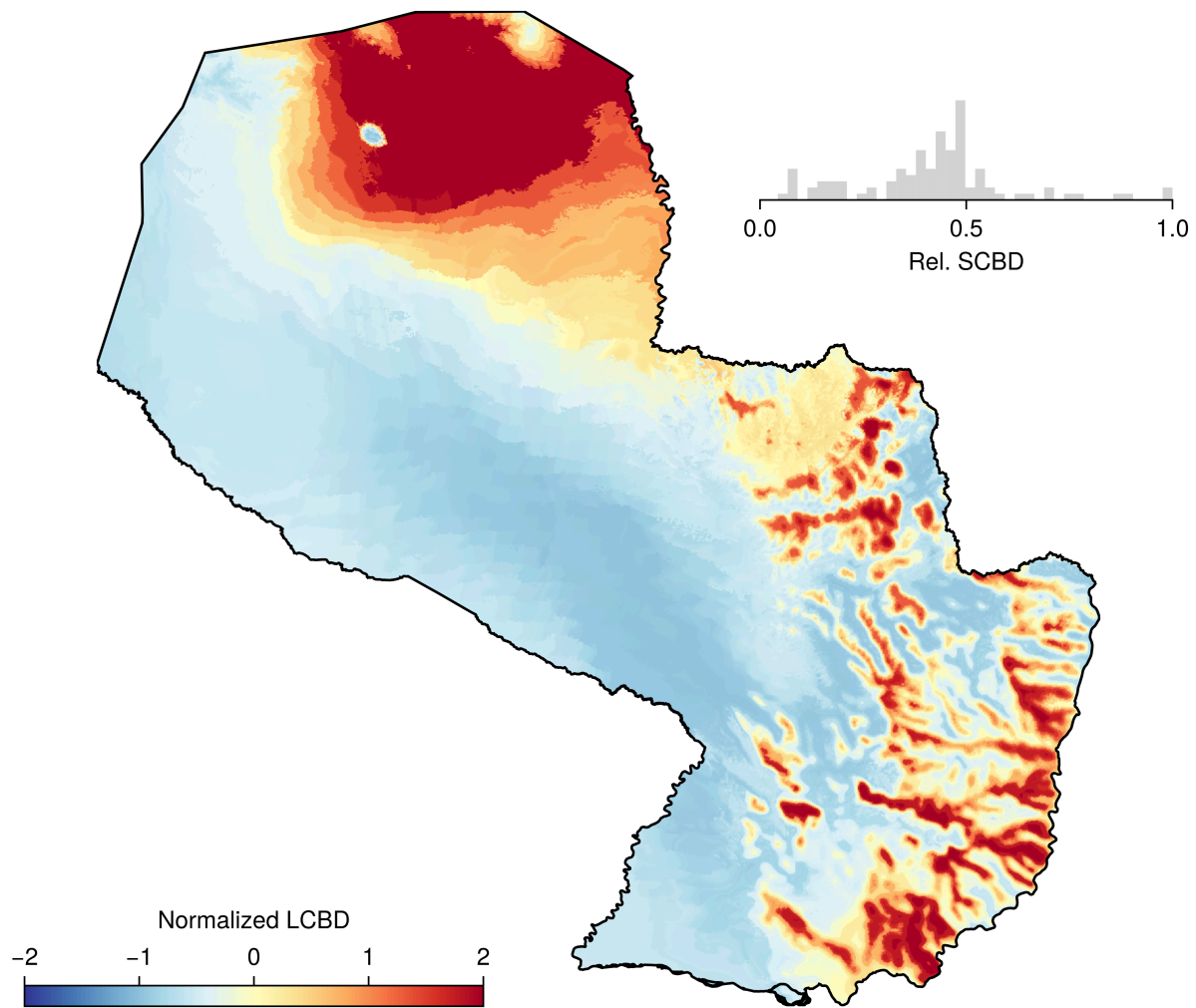


Figure 6: Virtual distribution of normalized (mean of 0 and unit variance) locality contribution to beta-diversity (Legendre and De Cáceres 2013), based on a pool of 100 virtual species. The inset histogram represents the standardized species contribution to beta-diversity. Red areas represent comparatively more unique areas in terms of simulated species composition. The code to produce this figure is available as Supp. Mat. 4.

Because the layers used by **SDT** are broadcastable, we can rapidly apply a function (here, the logistic response to the environmental covariate) to each layer, and then multiply the suitability together. The last step is facilitated by the fact that most basic arithmetic operations are defined for layers, allowing for example to add, multiply, subtract, and divide them by one another.

Conclusion

We have presented **SpeciesDistributionToolkit**, a package for the **Julia** programming language aiming to facilitate the collection, curation, analysis, and visualisation of data commonly used in species distribution modeling. Through the use of interfaces and a modular design, we have made this package robust to changes, easy to add functionalities to, and well integrated to the rest of the **Julia** ecosystem. All code for the case studies can be found in Supp. Mat. 1-4.

Plans for active development of the package are focused on (i) additional techniques for pseudo-absence generations, likely leading to their separate component package, (ii) full compatibility with the **MultivariateStatistics** for transformation, and (iii) additional **SDeMo** functionalities to allow cross-validation techniques with biologically relevant structure (Roberts et al. 2017).

Acknowledgements: TP is funded by an NSERC Discovery grant, a Discovery Acceleration Supplement grant, and a Wellcome Trust grant (223764/Z/21/Z). MDC is funded by an IVADO Postdoctoral Fellowship.

i Box 1 - integration with other **Julia** packages

The **SDT** package benefits from close integration with other packages in the **Julia** universe. Notably, this includes **Makie** [including **GeoMakie**; Danisch and Krumbiegel (2021)] for plotting and interactive data visualisation: all relevant plot types are overloaded for layer and occurrence data. Most data handled by **SDT** can be exported using the **Tables** interface, which allows data to be consumed by other packages like **DataFrames** (Bouchet-Valat and Kamiński 2023) and **MLJ** (Blaom et al. 2020), or directly saved as csv files. Interfaces to internal **Julia** methods are implemented whenever they are pertinent. **SimpleSDMLayers** and **OccurrencesInterface** objects behave like arrays, are iterable, and broadcastable. The **SDeMo** package relies in part on the **StatsAPI** interface, allowing to easily define new data transformation and classifier types to support additional features. Achieving integration with other packages through method overloading and the adherence to well-established interfaces is important, as it increases the chances that additional functionalities external to **SDT** can be used directly or fully supported with minimal addition of code. For situations where interfaces are not sufficient to link with other packages, we rely on **Julia**'s extension mechanism. For instance, **SimpleSDMLayers** objects can be used with **Clustering**, **MultivariateStats**, as well as **SpatialBoundaries** (Strydom and Poisot 2023), with strict version bounds, ensuring that this integration will remain usable regardless of possible changes in external packages.

Bibliography

- Araújo MB, Anderson RP, Márcia Barbosa A, et al (2019) Standards for distribution models in biodiversity assessments. *Science advances* 5:eaat4858. <https://doi.org/10.1126/sciadv.aat4858>
- Arenas-Castro S, Regos A, Martins I, et al (2022) Effects of input data sources on species distribution model predictions across species with different distributional ranges. *Journal of biogeography* 49:1299–1312. <https://doi.org/10.1111/jbi.14382>
- Bagnall A, Flynn M, Large J, et al (2018) Is rotation forest the best classifier for problems with continuous features?. *arXiv [csLG]*
- Baker DJ, Maclean IMD, Goodall M, Gaston KJ (2021) Species distribution modelling is needed to support ecological impact assessments. *The journal of applied ecology* 58:21–26. <https://doi.org/10.1111/1365-2664.13782>

Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (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* 3:327–338. <https://doi.org/10.1111/j.2041-210x.2011.00172.x>

Bezanson J, Edelman A, Karpinski S, Shah VB (2017) Julia: A fresh approach to numerical computing. *SIAM review Society for Industrial and Applied Mathematics* 59:65–98. <https://doi.org/10.1137/141000671>

Blaom A, Kiraly F, Lienart T, et al (2020) MLJ: A Julia package for composable machine learning. *Journal of open source software* 5:2704. <https://doi.org/10.21105/joss.02704>

Booth TH (2022) Checking bioclimatic variables that combine temperature and precipitation data before their use in species distribution models. *Austral ecology* 47:1506–1514. <https://doi.org/10.1111/aec.13234>

Bouchet-Valat M, Kamiński B (2023) DataFrames.Jl: Flexible and fast tabular data in Julia. *Journal of statistical software* 107:. <https://doi.org/10.18637/jss.v107.i04>

Brown JL, Hill DJ, Dolan AM, et al (2018) PaleoClim, high spatial resolution paleoclimate surfaces for global land areas. *Scientific data* 5:180254. <https://doi.org/10.1038/sdata.2018.254>

Buchhorn M, Smets B, Bertels L, et al (2020) Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019: Globe

Burgos EF, Vadell MV, Bellomo CM, et al (2021) First evidence of Akodon-borne orthohantavirus in northeastern Argentina. *EcoHealth* 18:429–439. <https://doi.org/10.1007/s10393-021-01564-6>

Danisch S, Krumbiegel J (2021) Makie.jl: Flexible high-performance data visualization for Julia. *Journal of open source software* 6:3349. <https://doi.org/10.21105/joss.03349>

Dansereau G, Legendre P, Poisot T (2022) Evaluating ecological uniqueness over broad spatial extents using species distribution modelling. *Oikos (Copenhagen, Denmark)* 2022:e9063. <https://doi.org/10.1111/oik.09063>

Dansereau G, Poisot T (2021) SimpleSDMLayers.Jl and GBIF.Jl: A framework for species distribution modeling in Julia. *Journal of open source software* 6:2872. <https://doi.org/10.21105/joss.02872>

Drake JM (2014) Ensemble algorithms for ecological niche modeling from presence-background and presence-only data. *Ecosphere (Washington, DC)* 5:1–16. <https://doi.org/10.1890/es13-00202.1>

325 Elith J, Ferrier S, Huettmann F, Leathwick J (2005) The evaluation strip: A new and robust
 326 method for plotting predicted responses from species distribution models. *Ecological*
 327 *modelling* 186:280–289. <https://doi.org/10.1016/j.ecolmodel.2004.12.007>

328 Elith J, Leathwick JR (2009) Species distribution models: Ecological explanation and prediction
 329 across space and time. *Annual review of ecology, evolution, and systematics* 40:677–697.
 330 <https://doi.org/10.1146/annurev.ecolsys.110308.120159>

331 Evenden GI, Rouault E, Warmerdam F, et al (2024) PROJ

332 Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for
 333 global land areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS. *International*
 334 *journal of climatology: a journal of the Royal Meteorological Society* 37:4302–4315.
 335 <https://doi.org/10.1002/joc.5086>

336 GBIF: The Global Biodiversity Information Facility (2025) \textit{What is GBIF?}

337 Gearty W, Jones LA (2023) rphylopic: An R package for fetching, transforming, and visualising
 338 PhyloPic silhouettes. *Methods in ecology and evolution* 14:2700–2708. [https://doi.org/10.](https://doi.org/10.1111/2041-210x.14221)
 339 [1111/2041-210x.14221](https://doi.org/10.1111/2041-210x.14221)

340 Gonzalez A, Vihervaara P, Balvanera P, et al (2023) A global biodiversity observing system
 341 to unite monitoring and guide action. *Nature ecology & evolution* 1–5. [https://doi.org/10.](https://doi.org/10.1038/s41559-023-02171-0)
 342 [1038/s41559-023-02171-0](https://doi.org/10.1038/s41559-023-02171-0)

343 Griffith J, Lord J-M, Catchen MD, et al (2024) BON in a Box: An Open and Collaborative
 344 Platform for Biodiversity Monitoring, Indicator Calculation, and Reporting. [https://doi.](https://doi.org/10.32942/X2M320)
 345 [org/10.32942/X2M320](https://doi.org/10.32942/X2M320)

346 Guisan A, Tingley R, Baumgartner JB, et al (2013) Predicting species distributions for conser-
 347 vation decisions. *Ecology letters* 16:1424–1435. <https://doi.org/10.1111/ele.12189>

348 Hirzel AH, Helfer V, Metral F (2001) Assessing habitat-suitability models with a virtual species.
 349 *Ecological modelling* 145:111–121. [https://doi.org/10.1016/s0304-3800\(01\)00396-9](https://doi.org/10.1016/s0304-3800(01)00396-9)

350 Jenkins CN, Pimm SL, Joppa LN (2013) Global patterns of terrestrial vertebrate diversity and
 351 conservation. *Proceedings of the National Academy of Sciences of the United States of*
 352 *America* 110:E2602–10. <https://doi.org/10.1073/pnas.1302251110>

353 Karger DN, Conrad O, Böhner J, et al (2017) Climatologies at high resolution for the earth's
 354 land surface areas. *Scientific data* 4:170122. <https://doi.org/10.1038/sdata.2017.122>

355 Karimi A-H, Barthe G, Balle B, Valera I (2019) Model-agnostic counterfactual explanations for
 356 consequential decisions. *arXiv [csLG]*

357 Kass JM, Smith AB, Warren DL, et al (2024) Achieving higher standards in species distribution
 358 modeling by leveraging the diversity of available software. *Ecography*. [https://doi.org/10.](https://doi.org/10.1111/ecog.07346)
 359 1111/ecog.07346

360 Kellner KF, Doser JW, Belant JL (2025) Functional R code is rare in species distribution and
 361 abundance papers. *Ecology* 106:e4475. <https://doi.org/10.1002/ecy.4475>

362 Legendre P, De Cáceres M (2013) Beta diversity as the variance of community data: dissim-
 363 ilarity coefficients and partitioning. *Ecology letters* 16:951–963. [https://doi.org/10.1111/](https://doi.org/10.1111/ele.12141)
 364 ele.12141

365 Leroy B, Meynard CN, Bellard C, Courchamp F (2016) virtualspecies, an R package to generate
 366 virtual species distributions. *Ecography* 39:599–607. <https://doi.org/10.1111/ecog.01388>

367 Marmion M, Parviainen M, Luoto M, et al (2009) Evaluation of consensus methods in predic-
 368 tive species distribution modelling. *Diversity & distributions* 15:59–69. [https://doi.org/10.](https://doi.org/10.1111/j.1472-4642.2008.00491.x)
 369 1111/j.1472-4642.2008.00491.x

370 McShea WJ (2014) What are the roles of species distribution models in conservation planning?.
 371 *Environmental conservation* 41:93–96. <https://doi.org/10.1017/s0376892913000581>

372 Mesgaran MB, Cousens RD, Webber BL (2014) Here be dragons: a tool for quantifying novelty
 373 due to covariate range and correlation change when projecting species distribution
 374 models. *Diversity & distributions* 20:1147–1159. <https://doi.org/10.1111/ddi.12209>

375 Mitchell R, Cooper J, Frank E, Holmes G (2021) Sampling Permutations for Shapley Value
 376 Estimation. *arXiv [statML]*

377 Osborne OG, Fell HG, Atkins H, et al (2022) Fauxcurrence: simulating multi-species occur-
 378 rences for null models in species distribution modelling and biogeography. *Ecography*
 379 2022:e5880. <https://doi.org/10.1111/ecog.05880>

380 Owen RD, Goodin DG, Koch DE, et al (2010) Spatiotemporal variation in *Akodon montensis*
 381 (Cricetidae: Sigmodontinae) and hantaviral seroprevalence in a subtropical forest ecosys-
 382 tem. *Journal of Mammalogy* 91:467–481. <https://doi.org/10.1644/09-MAMM-A-152.1>

383 Roberts DR, Bahn V, Ciuti S, et al (2017) Cross-validation strategies for data with temporal,
 384 spatial, hierarchical, or phylogenetic structure. *Ecography* 40:913–929. [https://doi.org/10.](https://doi.org/10.1111/ecog.02881)
 385 1111/ecog.02881

386 Roesch E, Greener JG, MacLean AL, et al (2023) Julia for biologists. *Nature methods* 20:655–
 387 664. <https://doi.org/10.1038/s41592-023-01832-z>

388 Stock A, Gregr EJ, Chan KMA (2023) Data leakage jeopardizes ecological applications of
 389 machine learning. *Nature ecology & evolution* 7:1743–1745. [https://doi.org/10.1038/s](https://doi.org/10.1038/s41559-023-02162-1)
 390 [41559-023-02162-1](https://doi.org/10.1038/s41559-023-02162-1)

391 Strydom T, Poisot T (2023) SpatialBoundaries.jl: edge detection using spatial wombling.
 392 *Ecography* 2023:. <https://doi.org/10.1111/ecog.06609>

393 Tuanmu M-N, Jetz W (2014) A global 1-km consensus land-cover product for biodiversity and
 394 ecosystem modelling: Consensus land cover. *Global ecology and biogeography: a journal*
 395 *of macroecology* 23:1031–1045. <https://doi.org/10.1111/geb.12182>

396 Tuanmu M-N, Jetz W (2015) A global, remote sensing-based characterization of terrestrial
 397 habitat heterogeneity for biodiversity and ecosystem modelling: Global habitat hetero-
 398 geneity. *Global ecology and biogeography: a journal of macroecology* 24:1329–1339.
 399 <https://doi.org/10.1111/geb.12365>

400 Van Looveren A, Klaise J (2019) Interpretable counterfactual explanations guided by proto-
 401 types. *arXiv [csLG]*

402 Wadoux AMJ-C, Saby NPA, Martin MP (2023) Shapley values reveal the drivers of soil organic
 403 carbon stock prediction. *SOIL* 9:21–38. <https://doi.org/10.5194/soil-9-21-2023>

404 Zurell D, Elith J, Schröder B (2012) Predicting to new environments: tools for visualizing
 405 model behaviour and impacts on mapped distributions. *Diversity & distributions* 18:628–
 406 634. <https://doi.org/10.1111/j.1472-4642.2012.00887.x>

407 Zurell D, Franklin J, König C, et al (2020) A standard protocol for reporting species distribution
 408 models. *Ecography* 43:1261–1277. <https://doi.org/10.1111/ecog.04960>