tidysdm: leveraging the flexibility of *tidymodels* for Species Distribution Modelling in R

Michela Leonardi[1], Margherita Colucci[2], Andrea Manica[1]*

[1] Evolutionary Ecology Group, Department of Zoology, University of Cambridge, UK

[2] Pan African Evolution Research Group, Max Planck Institute of Geoanthropology Jena, Germany

*Corresponding author: Andrea Manica, am315@cam.ac.uk

Abstract

In species distribution modelling (SDM), it is common practice to explore multiple machine-learning algorithms and combine their results into ensembles. This is challenging in R: different algorithms were developed independently, with inconsistent syntax and data structures. Specialised SDM packages solve this problem by wrapping them into complex functions that tackle their specific requirements. But creating and maintaining such interfaces is time-consuming, and there is no way to easily integrate other methods that may become available.

Here we present *tidysdm*, an R package that solves this problem by taking advantage of the *tidymodels* universe. Being part of the *tidyverse*, it (i) has standardised grammar, data structures and interface for modelling, (ii) includes packages designed for fitting, tuning, and validating various models, and (iii) allows easy integration of new algorithms and methods.

tidysdm grants easy and flexible SDM by supporting standard algorithms, including additional SDM-oriented functions, and allowing the use of any procedure to fit, tune and validate different models. Additionally, it provides functions to easily fit models based on paleo/time-scattered data.

tidysdm includes two vignettes detailing standard procedures for present-day and time-scattered data. Users can utilise any standard-format climatic data as input, but we also showcase the integration with the package *pastclim,* allowing easier access to present, past and future climate. An

additional vignette illustrates how to leverage other *tidyverse* packages to enhance the workflow of *tidysdm*. Finally, a section on the website helps troubleshoot common problems with *tidymodels*.

Introduction

Species Distribution Modelling (SDM, also known as Ecological Niche Modelling, ENM; Habitat Suitability Models, HSM; and other acronyms, Guisan *et al.* 2017) is a class of biological methods using the occurrences of a species and the associated environmental variables to predict how suitable a geographic area is for the species itself (Elith and Leathwick 2009). SDMs have been widely used, *e.g.*, to predict distributions of invasive species in new ranges (Elith 2017), assess habitat suitability under future climatic scenarios (Franklin 2023), reconstruct past ranges based on archaeological/palaeontological remains (*e.g.*, Leonardi *et al.* 2018) or present-day surveys (*e.g.*, Miller, Leonardi *et al.* 2021) and detect niche changes through time (*e.g.*, Leonardi *et al.* 2022).

In computational biology, SDM has been among the first fields to use machine learning (ML), and it is standard practice to explore multiple algorithms and combine their results into ensembles (Araújo and New, 2007). R (R core Team, 2023) is the most adopted framework to fit SDMs, but it has limitations. Individual ML algorithms have been implemented independently: each package provides its own data structure and syntax, which makes it difficult to repeat the same analysis with multiple algorithms and to create ensembles of their predictions. Many SDM packages exist (Sillero *et al.*, 2023) providing a simple and unified interface to multiple algorithms, but the maintainers must keep up with changes in other packages, which is resource-intensive. Furthermore, each SDM package limits the algorithms available and constrains other approaches used in modelling (*e.g.*, the strategy to create ensembles). Finally, any innovation in ML takes longer to be adopted, as it needs to be implemented into a generic package and then to be ported into any specialised SDM packages.

These issues are not peculiar to SDMs but apply to any field using ML algorithms. The *tidymodels* universe of packages (Kuhn and Wickham 2023) has emerged as a field-agnostic solution, providing a standardised interface for modelling in R (as *sklearn* does in Python). *tidymodels* includes packages designed to fit, tune and validate different models. Its strength is given by its being embedded into the *tidyverse*, whose philosophy is to have standardised grammar and data structures. In *tidymodels* every algorithm (no matter its library) must use the same syntax and input and output formats. In brief, it provides a coherent interface to modelling while simultaneously allowing the integration of any new algorithm, method or procedure that may be made available by the community. Packages such as *parsnip* (Kuhn and Vaughan 2023) and *yardstick* (Kuhn *et al.* 2023) provide the infrastructure to add respectively new models and metrics that seamlessly integrate within the *tidymodels* universe.

Here, we introduce *tidysdm*, a package that facilitates fitting SDMs with *tidymodels*. *tidysdm* integrates additional methods to several *tidymodels* functions to handle spatial data. It also implements SDM-relevant algorithms and metrics unavailable from other *tidymodels* packages. By building on the infrastructure of *tidymodels*, *tidysdm* leverages on a large community of developers rather than creating a complete solution from scratch. Objects created within *tidysdm* can be fed to functions from other packages (see Example 3), and any expansion within *tidymodels* will be immediately usable for SDM without waiting for a field-specific implementation.

Finally, *tidysdm* contains functions to easily work with time-scattered data, a task requiring extensive tweaking with mostSDM packages. Access and manipulation of (palaeo)climate are facilitated by the integration with *pastclim* (Leonardi *et al.* 2023): from version 2.0 it includes present-day climate and future reconstructions, in addition to palaeoclimate. *pastclim* is not required, though: users can import environmental data through their preferred pipeline, as *tidysdm* handles rasters with the *terra* package (Hijmans 2023).

We illustrate the functionality of *tidysdm* through two examples, detailed in the respective vignettes available on the website. The first is based on present-day occurrences of an endemic Iberian lizard, *Lacerta schreiberi,* and projects its range in response to future climate change (https://evolecolgroup.github.io/tidysdm/articles/a0_tidysdm_overview.html). The second reconstructs the range of wild horses during the Last Glacial Maximum (21,000 years BP) based on radiocarbon dates (https://evolecolgroup.github.io/tidysdm/articles/a1_palaeodata_application.html). Finally, we illustrate how to integrate additional packages to enhance the functionality in *tidymodels*, such as *DALEX* (Biecek 2018) to interpret models and *stacks* (Couch and Kuhn 2023) to create stacked ensembles (https://evolecolgroup.github.io/tidysdm/articles/a2_tidymodels_additions.html).

Example 1: SDM pipeline with present-day occurrences

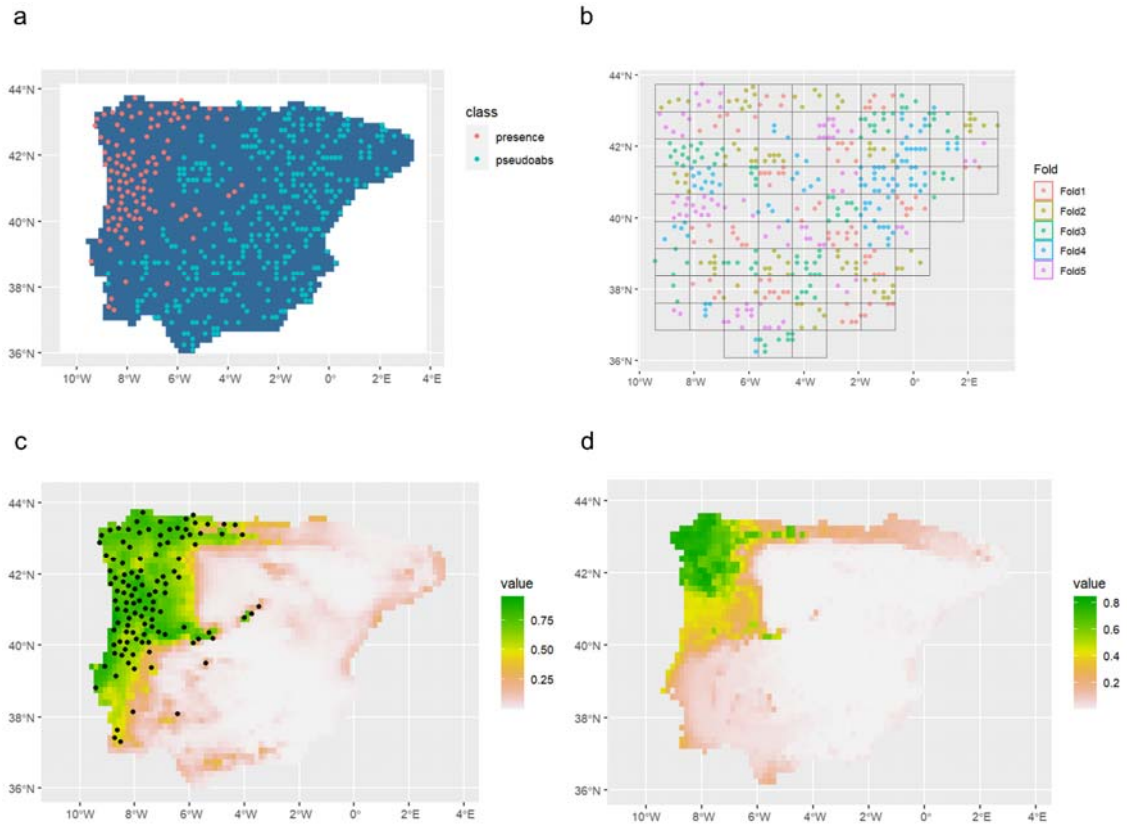

Figure 1: outputs of the SDM pipeline using present-day occurrences (Example 1). a) thinned presences and sampled pseudo-absences; b) spatial blocks for cross-validation; c) projection of the ensemble over the present; d) projection of the ensemble over the future.

One of the most common uses of SDMs is predicting future distribution under global change from present-day presences. The first vignette (https://evolecolgroup.github.io/tidysdm/articles/a0_tidysdm_overview.html) provides an example, using occurrences for the Iberian emerald lizard, *Lacerta schreiberi*, an endemic species of the Iberian Peninsula (Rödder and Schulte 2010). Data is provided as a dataset in *tidysdm* and available as a GBIF Occurrence Download (6 July 2023) https://doi.org/10.15468/dl.srq3b3.

In *tidysdm*, presences are represented as spatial point objects using the library *sf* (Pebesma 2018; Pebesma and Bivand 2023). All functions in *tidysdm* can work with *sf* objects, and the package extends several methods from the *tidymodels* universe to handle them. Environmental raster data, instead, are manipulated with *terra*. These are the most adopted spatial packages to represent point and raster data respectively.

A challenge in most SDM analyses is that only presence data are available, and the choice of suitable pseudo-absences and the definition of the area of interest are key to informative modelling. These steps are outside the remit of *tidymodels*, but *tidysdm* provides functions to facilitate them (see below). There are many available SDM packages (Sillero *et al.* 2023) that can complement the functions provided here; not all of them are compatible with *sf* objects, but it is easy to convert the latter into a plain data frame.

Download the climate: in the vignette, we use *pastclim* to download environmental data from WorldClim (Fick and Hijmans 2017), and *terra* to crop the resulting rasters to the Iberian Peninsula (but users can choose any pipeline to handle climatic data).

Thinning: a common challenge in SDMs is that presences are the collation of haphazard efforts, leading to spatial sampling biases. A common solution is to thin the data (Steen *et al.* 2021), as done by *spThin* (Aiello-Lammens *et al.* 2015). In *tidysdm*, *thin_by_dist()* thins the dataset by distance (removing close-by points), *thin_by_cell()* retains one presence per cell of the raster. Analogous functions work with time-scattered presences (Example 2).

Sampling pseudo-absences/background: as true absences are rarely available, a wide literature arose on how to sample pseudo-absences/background points (*e.g.,* Barbet-Massin *et al.* 2012). The *tidysdm* function *sample_pseudoabs()* provides commonly adopted techniques including random, minimum distance, maximum distance, and a disc identified by minimum and maximum distance.

Choice of variables: In SDMs, it is desirable to focus on variables that have biological significance for the focal species. When such knowledge is unavailable, informative variables can be identified by looking for mismatches between the presences and the underlying background (pseudo-absences) (Miller, Green, *et al.* 2021). *plot_pres_vs_bg()* uses violin plots to compare presences and pseudo-absences in their distributions over the variable space, which helps choosing variables for which they differ. *dist_pres_vs_bg()* offers a quantitative approach that ranks variables based on the proportional overlap of the respective density plots. The variables of interest can be then pruned for collinearity by setting a correlation cutoff with the function *filter_high_cor()*. This function provides a sensible set of uncorrelated variables to be used with algorithms such as GLMs and GAMSs, but other such sets will likely exist and might be even better: we recommend that users inspect the variables with care and make decisions based on their biological understanding of the system.

Modelling: once the presences and pseudo-absences (Figure 1a) have been selected and associated with the climatic variables of interest, modelling is performed with *tidymodels.* A *workflow* is created by combining a *recipe* to process the data and a *model specification*. The *model specification* can have a few selected values for the hyperparameters (*i.e.*, model parameters that cannot be learned directly by training the dataset and should be defined by the user) or can prescribe their tuning. A

workflow can then be fitted and tuned to the data. This is usually done using spatial blocking to generate data folds that account for the autocorrelation intrinsic to spatial data (cross-validation, Figure 1b). *tidymodels*, through the additional package *spatialsample* (Mahoney *et al.* 2023), provides simple syntax and infrastructure to perform all these steps (including additional ways of creating data folds, *e.g.,* by clustering points based on similarity in the environmental parameters). For more complex approaches, outputs from the *blockCV* package (Valavi *et al.* 2019) can be converted with *blockcv2rsample()*.

The major advantage of *tidymodels* is that, by changing the model specification, it is possible to fit different models with the same code defining the rest of the *workflow*. *workflowsets* (Kuhn and Couch 2023) standardises this step*:* it easily creates multiple workflows, different in their *recipe* or *model specification,* and fits them to the same data folds. Multiple algorithms are thus easily investigated with shorter and simpler-to-read code that minimises the chances of errors. *tidysdm* provides functions to simplify these steps (*e.g.*, automatically generating model specifications with tuning for hyperparameters usually studied in SDMs) and expands methods from *tidymodels* that are not natively compatible with *sf* objects.

Evaluation: Model tuning requires metrics to assess the goodness-of-fit of different hyperparameter sets. The package *yardstick* is used in *tidymodels* to provide standard metrics, *e.g.,* Area Under the Curve (AUC, Swets 1988); *tidysdm* uses the *yardstick* infrastructure to add metrics commonly used in SDMs, *e.g.*, maximum True Skill Statistics (TSS, Allouche *et al.* 2006), Boyce Continuous Index (BCI, Hirzel *et al.* 2006), and maximum Kappa (K, Cohen 1960). Thanks to the standard grammar provided by *yardstick*, it is possible to explore multiple metrics simultaneously, and *tidysdm* offers a standard set of metrics for SDMs via *sdm_metric_set()*. Note that *yardstick* assumes that, in classification models (as used in SDMs), the first level of the response variable is the event of interest. In SDMs, these are presences, so the response variable needs to be a factor with presences as the first

(reference) level. *check_sdm_presence()* makes sure that the response variable is formatted correctly.

Ensemble: In many SDM analyses, the best-performing models based on a metric of choice are combined into an ensemble (Araújo and New 2007), used to predict ranges. This can be done following different procedures: the easiest is building the ensemble using the best-performing version of each algorithm. In *tidysdm*, we term this as a *simple_ensemble.* There is also the option to use *tidymodels'* stacked ensembles (Couch and Kuhn 2023), as described in Example 3. The simplest way to build an ensemble is to start from a *workflowset* in which several algorithms were tuned; the function *add_member()* will choose and extract the best hyperparameter values from the tuned *workflowset*. It is also possible to assemble an ensemble from individual *workflows* (*e.g.*, for computationally-demanding models where individual workflows are fitted separately on a computer cluster).

Projection: besides a conventional *predict()* function, *predict_raster()* allows to generate a map by taking a raster of environmental predictors and returning a raster of the ensemble prediction (by default, the mean of the models within the ensemble, but the median and weighted versions of the two are also available). The ensemble can also be subsetted to include only the best algorithms specifying a minimum threshold in the metric (*e.g.*, Boyce continuous index, Figure 1c). Finally, *tidysdm* allows the creation of a *repeated_ensemble* to explore the role of stochasticity in thinning and sampling. The simplest route is to loop over the code to create a *simple_ensemble*, adding each iteration to a list, which is then converted into a *repeated_ensemble* object using *add_repeat()*.

tidysdm also offers an easy way to convert the predicted probability of occurrence to binary format (presences/absences): an optimal threshold based on a given metric is calibrated with *calib_class_thresh()*, then the prediction is performed again with *predict_raster()* specifying *type="class"* and the desired threshold. In the vignette, *predict_raster()* is used both to predict the present range, and make future projections for a given Shared Socio-economic Pathways (SSP)

(Figure 1d). For convenience, we rely on *pastclim* to download and manipulate the WorldClim data, but the user can bring in their own environmental data, which can be easily read with *terra*.

Example 2: SDM pipeline with time-scattered occurrences

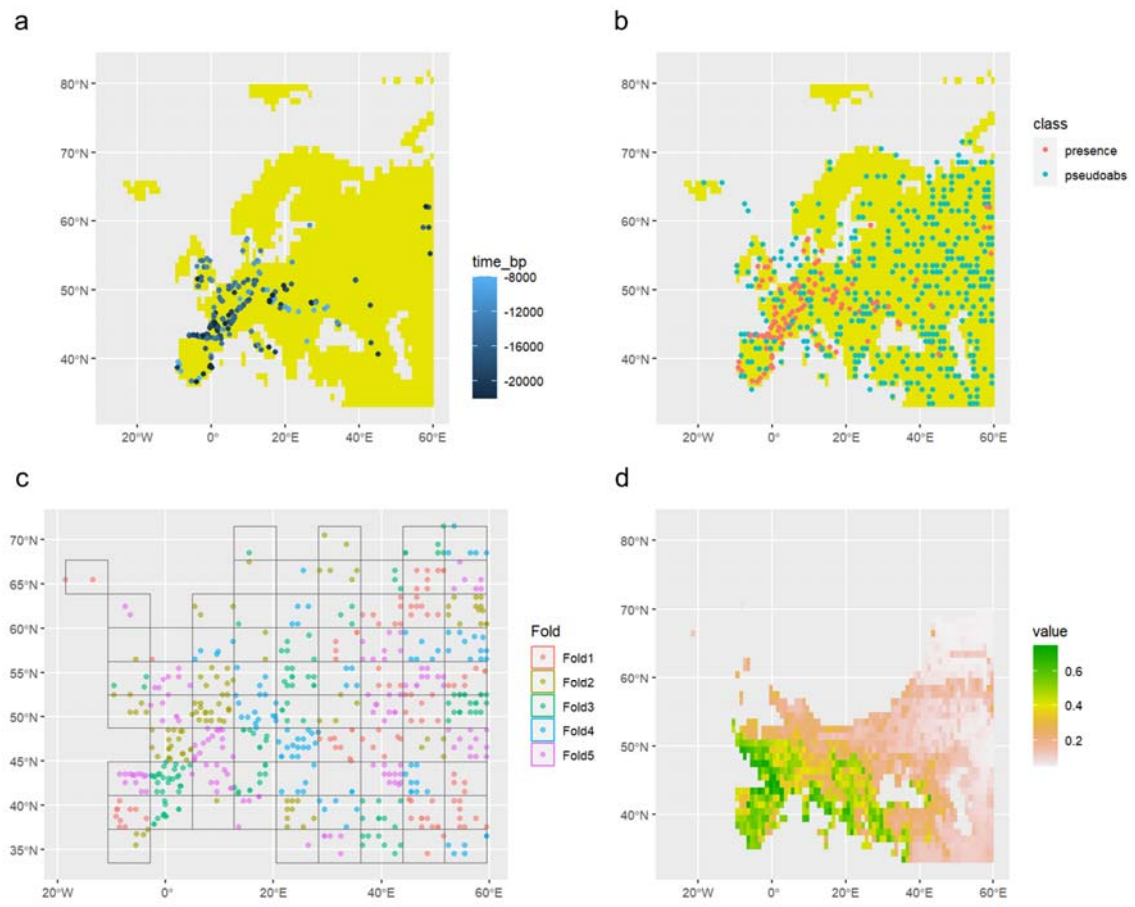

Figure 2: outputs of the SDM pipeline using time-scattered occurrences (Example 2). a) Presences included in the analyses; b) thinned presences and sampled pseudo-absences; c) Spatial blocks for cross-validation; d) Projection of the ensemble over the Last Glacial Maximum (21,000 years BP)

Existing SDM packages are often designed to automatically associate points with a single raster. This may be problematic when using palaeontological/archaeological data that need time series of palaeoclimatic reconstructions. *tidysdm* is designed to handle this task with ease, as shown in this

example (vignette available at

https://evolecolgroup.github.io/tidysdm/articles/a1_palaeodata_application.html).

We perform SDM on wild horses (*Equus ferus*) dated between 22,000 and 8,000 years BP (extracted from Leonardi *et al.* 2022, Figure 2a). For computational speed, we use a coarse time series (the Example dataset in *pastclim*), with reconstructions every 5k years and a spatial resolution of 1 arc-degree; note that much better reconstructions can be downloaded with *pastclim*.

Many steps are equivalent to the ones described in Example 1; here, we focus on functions specific to time series. For this task, *tidysdm* integrates well with *pastclim*, but users can utilise any *terra* SpatRasterDataset, in which each climatic variable is a *terra* SpatRaster with a time dimension.

Thinning: the thinning functions are equivalent to the ones described earlier but include the time. *thin_by_dist_time()* allows setting the minimum distance in space (*dist_min*) and time (*interval_min*) between presences, and *thin_by_cell_time()* keeps one presence per cell and time slice.

Sampling pseudo-absences/background: *sample_pseudoabs_time()* samples points to match the availability of presences through time (*i.e.*, matching sampling effort on the time axis), with the same strategies available for *sample_pseudoabs()*. We use the *location_slice_from_region_series()* function from *pastclim* to associate presences and pseudo-absences to the correct time steps of the SpatRasterDataset.

Modelling and ensemble: procedures for fitting models and creating ensemble are identical to the one described for contemporary data (Example 1). Note that it is better to remove the time columns from the *sf* object before fitting models: time is not a direct predictor of occurrence but is needed to extract the climate for the associated time step.

Projection: similarly to Example 1, after obtaining a *simple_ensemble()*, predicting for any given time step is straightforward, as the input climate for the time step is represented as a *terra SpatRaster*.

pastclim::region_slice() easily extracts climate for any time step: here, we predict the range 21,000 years BP (Last Glacial Maximum).

Example 3: Additional *tidymodels* features

In this final example (vignette available at https://evolecolgroup.github.io/tidysdm/articles/a2_tidymodels_additions.html) we illustrate additional features from *tidymodels* and related packages that can enhance SDM pipelines.

Explaining the behaviour: exploring and explaining the behaviour of ML algorithms is an important step to understand the predictions made with SDM. The package *DALEX* provides tools to explain models, and it can handle objects created with *tidymodels; tidysdm* provides further functionality to use *DALEX* directly on ensembles.

Initial split: in *tidymodels*, an initial split of the data into a training and testing set is advocated. Whilst this is considered good practice in ML, it is seldom used in SDMs. *tidysdm* provides functions to create an initial split with spatial data, thus applying a more formal approach to ML modelling in SDMs.

Workflowsets: the modular approach used in defining *workflowsets* allows to use different recipes for each algorithm, *e.g.*, to filter correlated variables for algorithms that require uncorrelated ones (*e.g.*, GLM or GAM), and use all possible predictors for the others (*e.g.*, random forests and maxent).

Stacked ensembles: stacking (Opitz and Maclin 1999) is a powerful approach to combine models. It uses a meta-learning algorithm to learn how to best combine multiple models, including multiple versions of the same algorithm with different hyperparameters. The package *stacks* builds stacked ensembles directly from the fitted *workflowsets*, providing an alternative method to the *simple_ensemble* approaches available in *tidysdm*.

These are only a few examples of how *tidysdm* can be enhanced with features from *tidymodels* and companion packages; we envisage many more applications thanks to the widespread use of *tidymodels* in multiple fields.

Conclusions

tidysdm provides a flexible framework that leverages the power of the *tidymodels* universe, with applications for both modern species distribution modelling (*e.g.*, in ecology, conservation, biogeography, zoology) and paleo-communities (including palaeoecology, archaeology, archaeozoology, palaeontology, evolutionary ecology, macroevolution, and population genetics/genomics). A major advantage is that any addition to the *tidymodels* universe (*e.g.,* stacked ensembles) is immediately usable for SDMs, allowing developments from the broader community of machine-learning practitioners to be quickly adopted.

Package installation and availability

The package is open-source and available from CRAN https://cran.r-project.org/web/packages/tidysdm/index.html.It has a dedicated website https://github.com/EvolEcolGroup/tidysdm, containing a manual, the three vignettes described here, and all details about the licence and the dependencies. When using *tidysdm*, please cite the present work.

Author's contribution

A.M. wrote the R package; M.L. defined the SDM pipelines with inputs from A.M. and M.C.; A.M. and M.C. wrote the vignettes with inputs from M.L.; M.L. designed the figures; all authors tested the package; A.M. wrote the function documentation and built the website; all authors wrote the manuscript.

Acknowledgements

We thank Johanna Paijmans and Sophie Dennis for testing the package during its development.

Funding

M.L. and A.M. were funded by the Leverhulme Research Grant RPG-2020-317. M.C. was funded by the Lise Meitner Pan-African Evolution Research Group.

References

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B. and Anderson, R.P. 2015, spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography, 38: 541-545 https://doi.org/10.1111/ecog.01132

Allouche, O., Tsoar, A. and Kadmon, R. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43: 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22: 42–7. https://doi.org/10.1016/j.tree.2006.09.010.

Barbet-Massin, M., Jiguet, F., Albert, C. H. and Thuiller, W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3: 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Beyer, R. M., Krapp, M. and Manica, A. 2020. High-resolution terrestrial climate, bioclimate and vegetation for the last 120,000 years. Sci. Data 7: 1–9. https://doi.org/10.1038/s41597-020-0552-1

Biecek P. 2018. *DALEX*: Explainers for Complex Predictive Models in R. Journal of Machine Learning Research, 19(84), 1-5. https://doi.org/10.48550/arXiv.1806.08915

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20: 37– 46. https://doi.org/10.1177/001316446002000104

Couch, S., and Kuhn, M. 2023. stacks: Tidy Model Stacking. https://stacks.tidymodels.org/, https://github.com/tidymodels/stacks.

Elith, J. and Leathwick, J. R. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu. Rev. Ecol. Evol. Syst. 40: 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith, J. 2017. Predicting Distributions of Invasive Species. In A. Robinson, T. Walshe, M. Burgman, & M. Nunn (Eds.), Invasive Species: Risk Assessment and Management (pp. 93-129). Cambridge University Press. https://doi.org/10.1017/9781139019606.006

Fick, S.E. and Hijmans, R.J. 2017, WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol, 37: 4302-4315. https://doi.org/10.1002/joc.5086

Franklin, J. 2023. Species distribution modelling supports the study of past, present and future biogeographies. Journal of Biogeography, 00, 1– 13. https://doi.org/10.1111/jbi.14617

Guisan, A., Thuiller, W. and Zimmermann, N. E. 2017. Habitat suitability and distribution models: With applications in R. Cambridge University Press. https://doi.org/10.1111/jbi.14617

Hernangómez, D. 2023. tidyterra: tidyverse Methods and ggplot2 Helpers for terra Objects (Version 0.4.0). https://doi.org/10.5281/zenodo.6572471

Hijmans, R. 2023. terra: Spatial Data Analysis. R package version 1.7-41. https://github.com/rspatial/terra

Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C. and Guisan, A. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Modell. 199: 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Kuhn, M. and Couch, S. 2023. workflowsets: Create a Collection of 'tidymodels' Workflows. https://github.com/tidymodels/workflowsets, https://workflowsets.tidymodels.org.

Kuhn, M. and Wickham, H. 2023. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org/

Kuhn, M. and Vaughan, D. 2023. parsnip: A Common API to Modeling and Analysis Functions. https://github.com/tidymodels/parsnip, https://parsnip.tidymodels.org/.

Kuhn, M., Vaughan, D., and Hvitfeldt, E. 2023. yardstick: Tidy Characterizations of Model Performance. https://github.com/tidymodels/yardstick

Leonardi, M., Boschin, F., Giampoudakis, K., Beyer, R. M., Krapp, M., Bendrey, R., Sommer, R., Boscato, P., Manica, A., Nogues-Bravo, D. and Orlando, L. 2018. Late Quaternary horses in Eurasia in the face of climate and vegetation change. Sci. Adv. 4: eaar5589. https://doi.org/10.1126/sciadv.aar5589

Leonardi, M., Boschin, F., Boscato, P. and Manica, A. 2022. Following the niche: the differential impact of the last glacial maximum on four European ungulates. Commun. Biol. 2022 51 5: 1–10. https://doi.org/10.1038/s42003-022-03993-7

Leonardi, M., Hallett, E. Y., Beyer, R., Krapp, M. and Manica, A. 2023. pastclim 1.2: an R package to easily access and use paleoclimatic reconstructions. Ecography (Cop.). 2023: e06481. https://doi.org/10.1111/ecog.06481

Mahoney, M. J., Johnson, L. K., Silge, J., Frick, H., Kuhn, M. and Beier, C. M. 2023. Assessing the performance of spatial cross-validation approaches for models of spatially structured data. https://doi.org/10.48550/arXiv.2303.07334

Miller, E. F., Leonardi, M., Beyer, R., Krapp, M., Somveille, M., Somma, G. L., Delser, P. M. and Manica, A. 2021. Post-glacial expansion dynamics, not refugial isolation, shaped the genetic

structure of a migratory bird, the Yellow Warbler (Setophaga petechia). bioRxiv: 2021.05.10.443405. https://doi.org/10.1101/2021.05.10.443405

Miller, E. F., Green, R. E., Balmford, A., Maisano Delser, P., Beyer, R., Somveille, M., Leonardi, M., Amos, W. and Manica, A. 2021. Bayesian Skyline Plots disagree with range size changes based on Species Distribution Models for Holarctic birds. Mol. Ecol. 30: 3993–4004. https://doi.org/10.1111/mec.16032

Opitz, D., and maclin, R. 1999. Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research 11: 169-198

Pebesma, E., and Bivand, R. 2023. Spatial Data Science: With Applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

R Core Team 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rödder, D. and Schulte, U. 2010. Potential loss of genetic variability despite well established network of reserves: The case of the Iberian endemic lizard *Lacerta schreiberi*. Biodivers. Conserv. 19: 2651–2666. https://doi.org/10.1007/s10531-010-9865-2

Sillero, N., Campos, J. C., Arenas-Castro, S. and Barbosa, A. M. 2023. A curated list of R packages for ecological niche modelling. Ecol. Modell. 476: 110242. https://doi.org/10.1016/j.ecolmodel.2022.110242

Steen, V. A., Tingley, M. W., Paton, P. W. C. and Elphick, C. S. 2021. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. Methods Ecol. Evol. 12: 216–226. https://doi.org/10.1111/2041-

210X.13525

Swets, J. A. 1988. Measuring the Accuracy of Diagnostic Systems. Science. 240: 1285–1293. https://doi.org/10.1126/science.3287615

Valavi, R., Elith, J., Lahoz-Monfort, J. J., Guillera-Arroita, G. 2019. *blockCV*: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 10: 225–232. https://doi.org/10.1111/2041-210X.13107

Xu, C., Kohler. T. A., Lenton, T. M., Svenning J. C. and Scheffer M. 2020. Future of the human climate niche. Proc. Natl Acad. Sci. USA 117 (21) 11350-11355. https://doi.org/10.1073/pnas.1910114117