`modleR`: a modular **workflow** for ecological niche modeling in R

Ecological Niche Modeling 2020 online course

Andrea Sánchez-Tapia & Sara Mortara

Rio de Janeiro Botanical Garden

May 2020

about

Marinez F. de Siqueira

Sara Mortara

Andrea Sánchez-Tapia

- Scientific Computation Lab @ Rio de Janeiro Botanical Garden

about

Marinez F. de Siqueira

Sara Mortara

Andrea Sánchez-Tapia

- Scientific Computation Lab @ Rio de Janeiro Botanical Garden
- biodiversity informatics, ENM/SDM, **open science**, reproducibility

about

Marinez F. de Siqueira

Sara Mortara

Andrea Sánchez-Tapia

- Scientific Computation Lab @ Rio de Janeiro Botanical Garden
- biodiversity informatics, ENM/SDM, **open science**, reproducibility
- **scientific workflows** based in R for data downloading and cleaning, taxonomic checking

about

Marinez F. de Siqueira

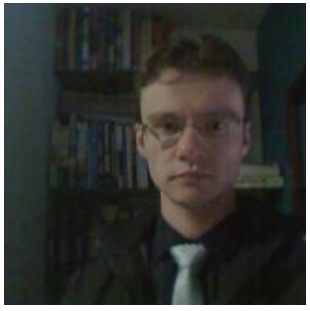
Sara Mortara

Andrea Sánchez-Tapia

- Scientific Computation Lab @ Rio de Janeiro Botanical Garden
- biodiversity informatics, ENM/SDM, **open science**, reproducibility
- **scientific workflows** based in R for data downloading and cleaning, taxonomic checking
- support for IUCN authority (CNCFlora)

other modleR developers





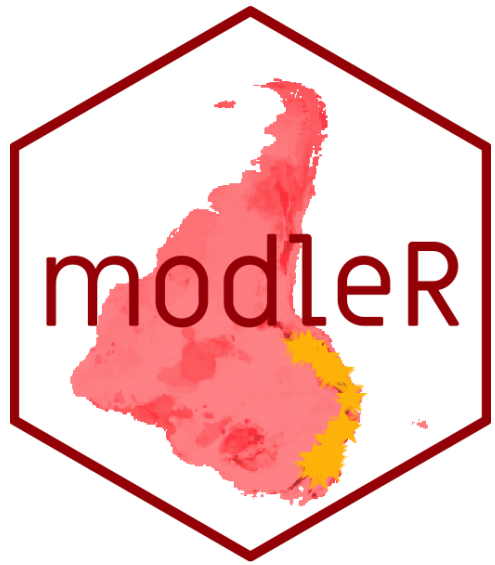

Diogo S.B. Rocha
Maria Luisa Mondelli
Guilherme Gall
Felipe Barros

modleR

- Four-step workflow wrapping around functions in dismo
- Started as a set of scripts to execute niche modeling for species from the Brazilian Atlantic Forest.
- Reformatted into an R package
- Added an initial shiny app
- Added HPC computation and parallelization options
- Added vignettes, documentation, pkgdown page, tests

modleR

A **workflow** based on package `dismo` (Hijmans et al 2017), developed to automatize some of the common steps in ecological niche modeling



https://model-r.github.io/modleR/

SDM/ENM in R

- GIS with raster, sp, maps, rgdal, sf
- Established packages such as dismo (Hijmans et al 2017), BIOMOD2 (Thuiller et al 2007).
- Newer packages for the whole process or parts of it sdm, wallace, ENMeval, spThin etc.

why another ENM package?

- reproducibility: we need workflows rather than just packages
- thorough metadata recording
- flexibility for parametrization
- facility to use in high performance/ high throughput computational frameworks (HPC/HTC)
- comunication with other packages in the R environment (spThin, ENMGadgets, kuenm, ENMTML)

we think about **workflows**

- reproducibility

we think about **workflows**

- reproducibility
- using a single working directory per project

we think about **workflows**

- reproducibility
- using a single working directory per project
- different steps: different subfolders

we think about **workflows**

- reproducibility
- using a single working directory per project
- different steps: different subfolders
- relative rather than absolute paths and no `setwd()`

we think about **workflows**

- reproducibility
- using a single working directory per project
- different steps: different subfolders
- relative rather than absolute paths and no `setwd()`
- enter and exit the workflow at any step

we think about **workflows**

- reproducibility
- using a single working directory per project
- different steps: different subfolders
- relative rather than absolute paths and no `setwd()`
- enter and exit the workflow at any step
- using HD space rather than RAM

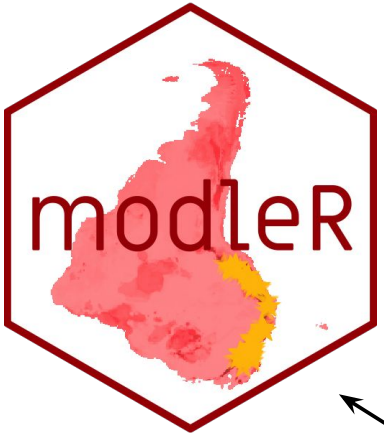
we think about **workflows**

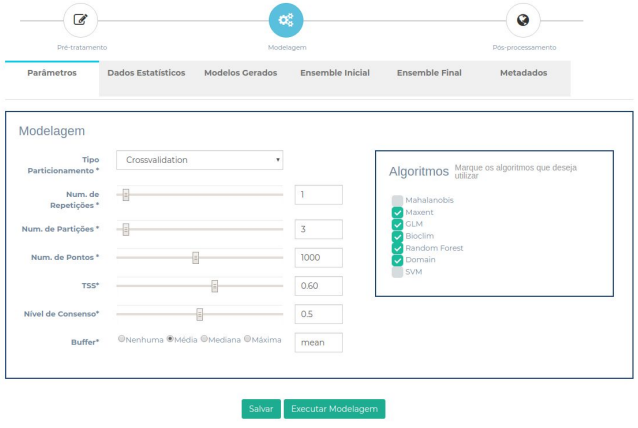
- reproducibility
- using a single working directory per project
- different steps: different subfolders
- relative rather than absolute paths and no `setwd()`
- enter and exit the workflow at any step
- using HD space rather than RAM
- parallelization options

modleR belongs to a larger framework

backend

frontend





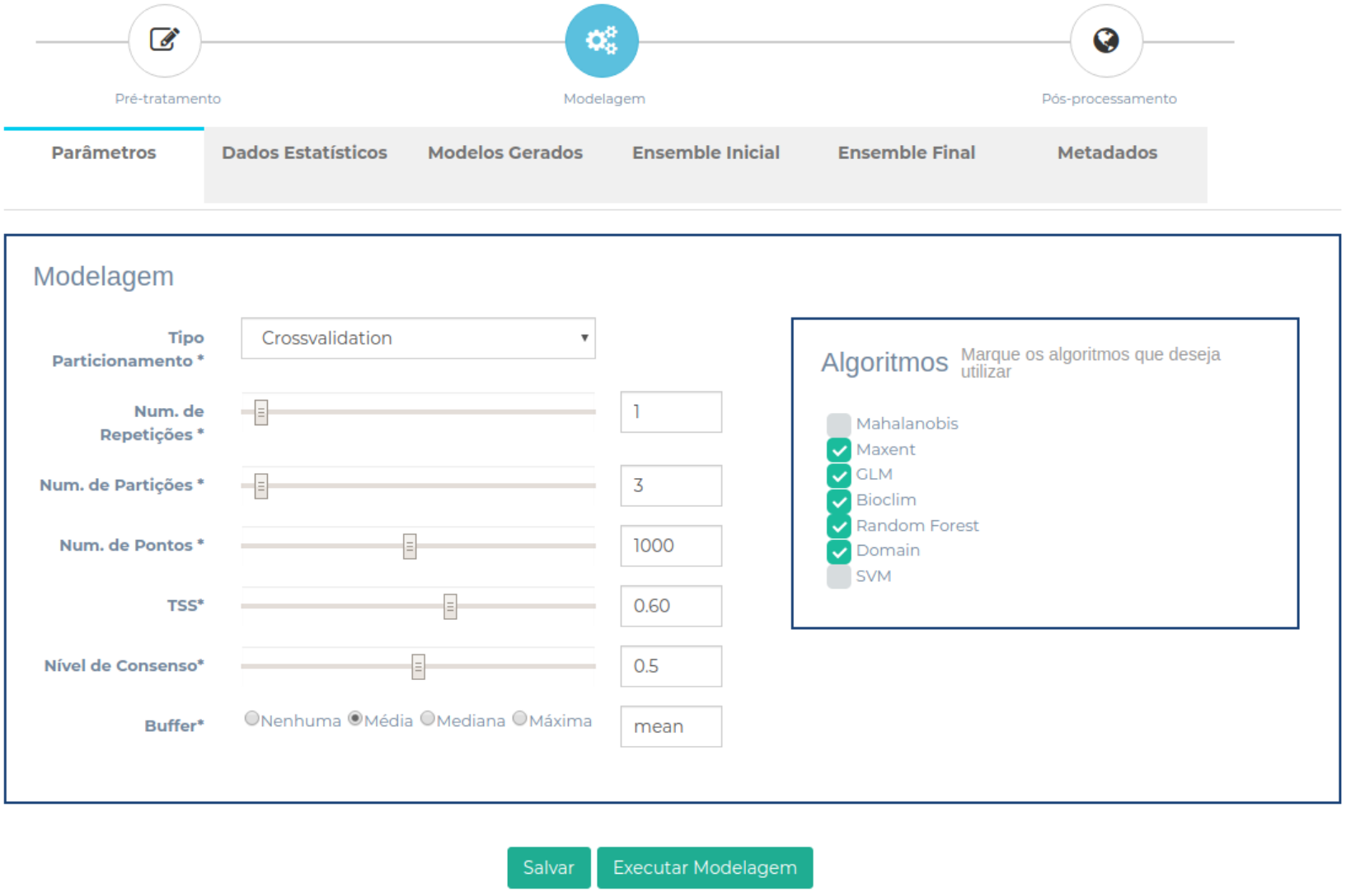
modleR web

HPC processing

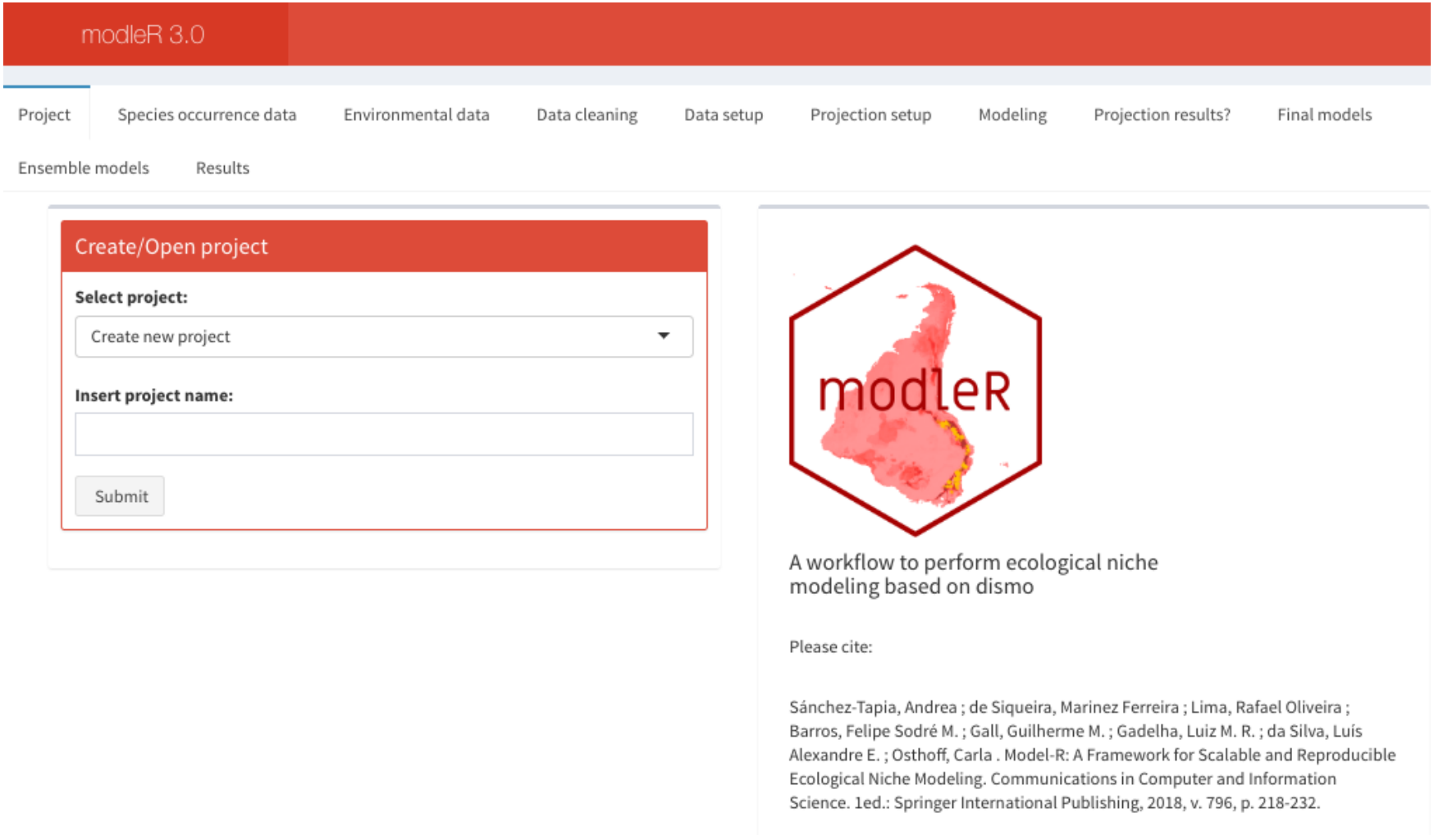
Santos Dumont supercomputer

Laboratório Nacional de Computação Científica

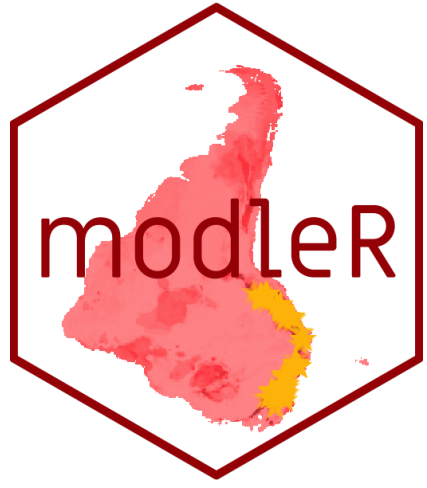
web interface:https://model-r.jbrj.gov.br/



shiny application:
https://github.com/Model-R/modleR_shiny_app



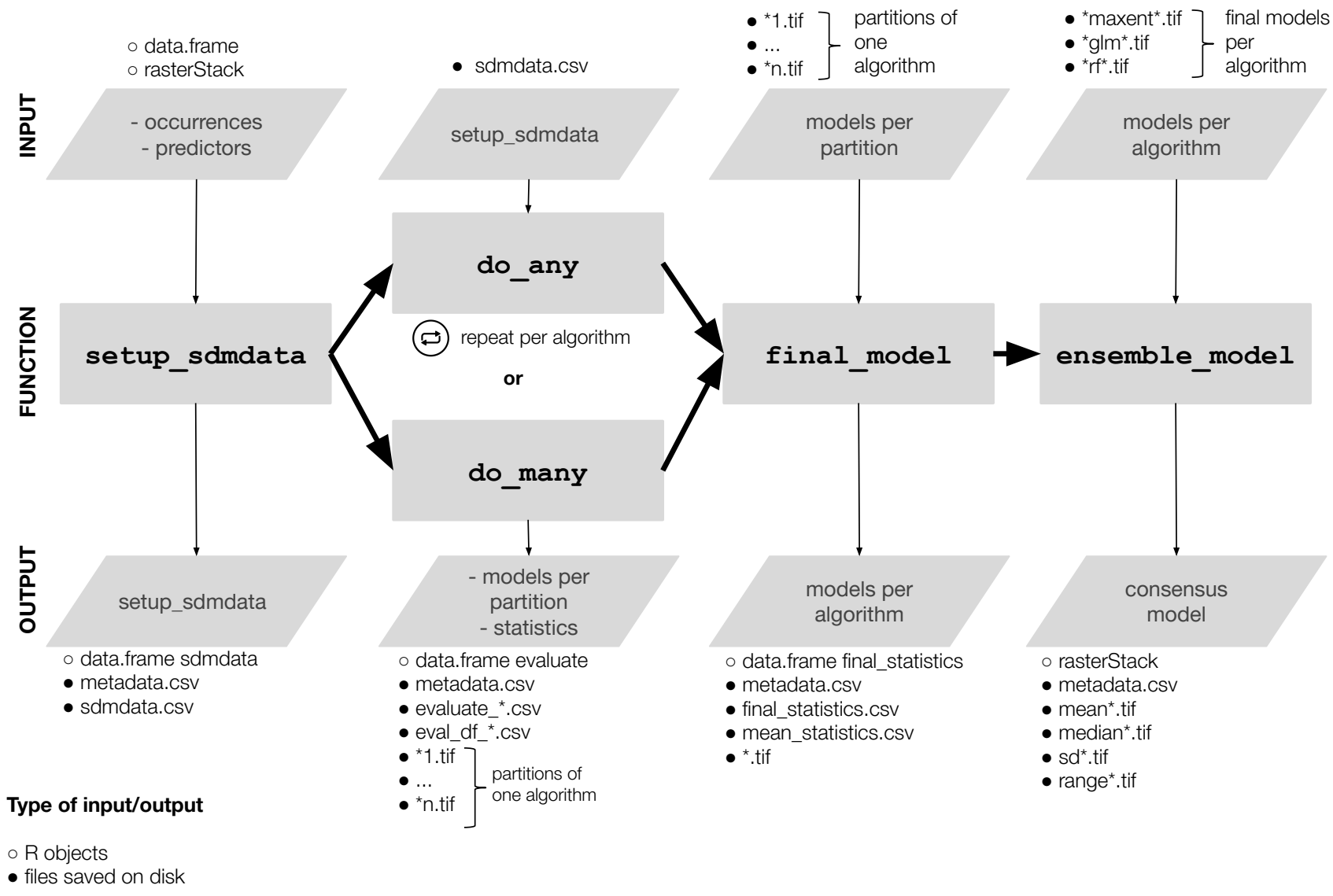
the workflow



a **four-step** workflow

1. `setup_sdmdata()`: data setup
2. `do_(m)any()`: model fitting, projecting and evaluating
3. `final_model()`: joining partitions
4. `ensemble_model()`: algorithm consensus

a **four-step** workflow



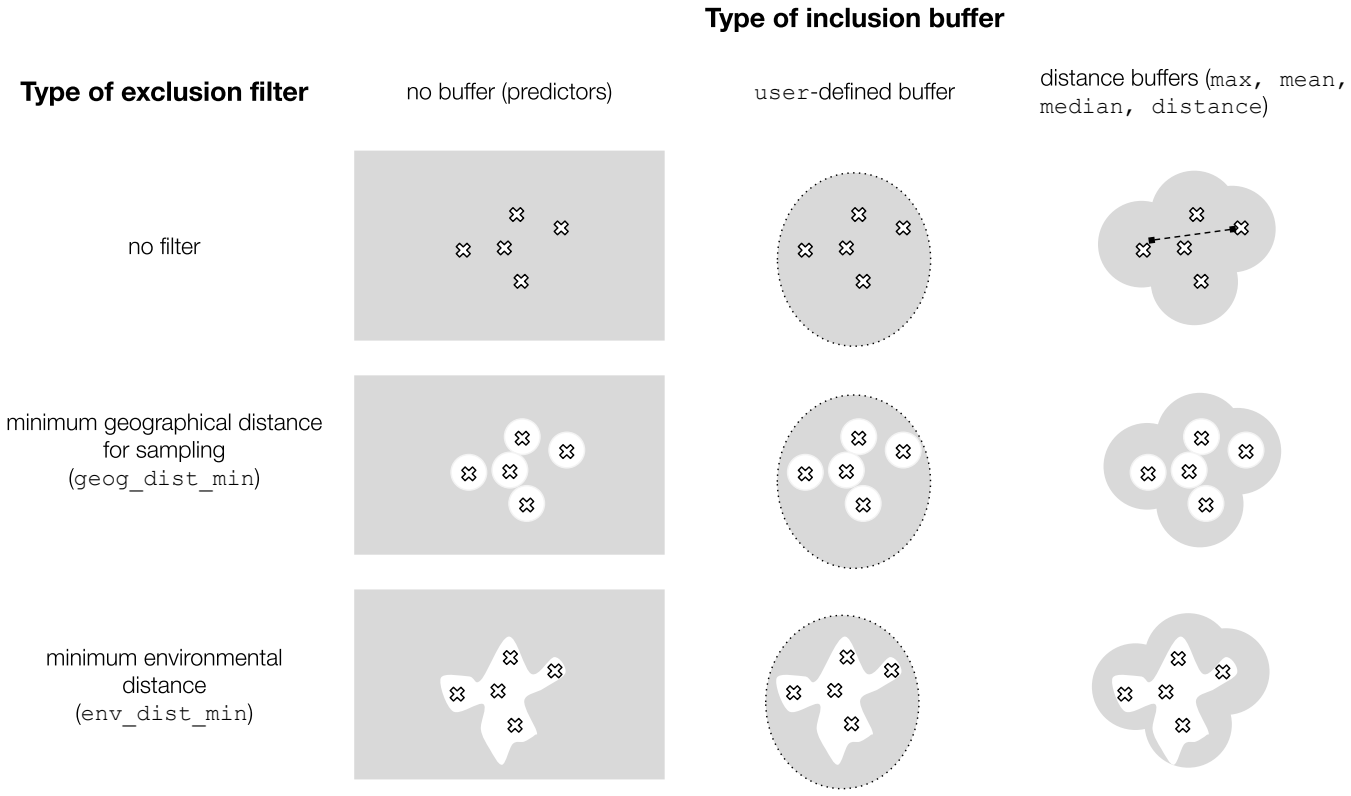
`setup_sdmdata`: data preparation

Data preparation and cleaning should be performed previously

- data cleaning checks: exact duplicates, NAs and one-per-pixel.
- pseudo-absence sampling
- experimental design: *bootstrap*, cross-validation

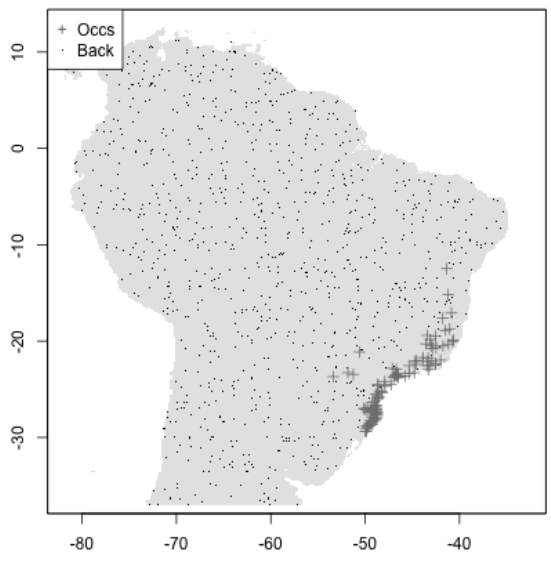
pseudoabsence sampling

pseudoabsence sampling options

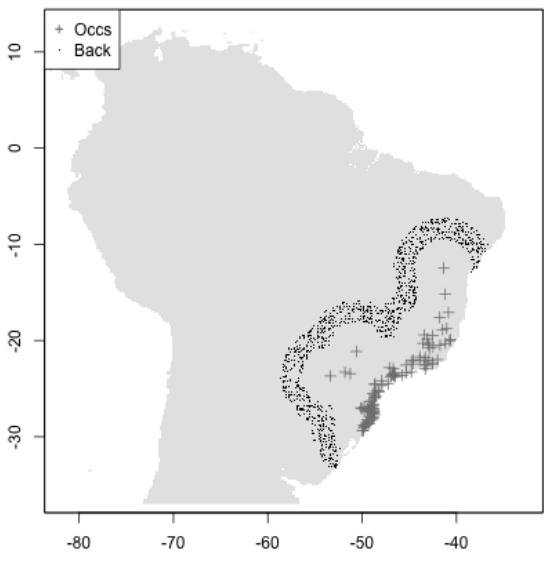


pseudoabsence sampling

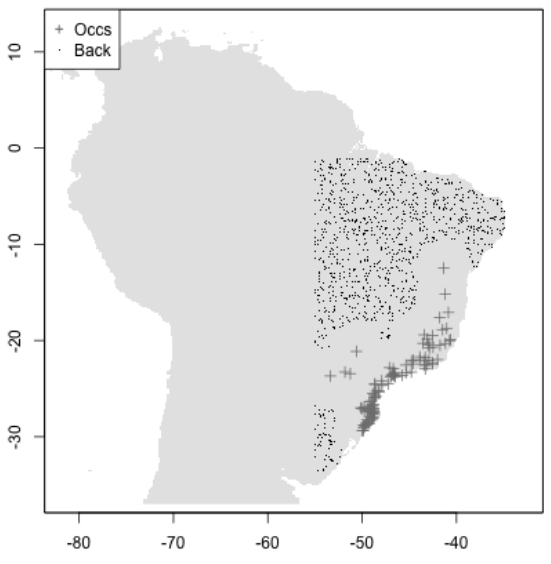
no buffer



mean distance buffer and euclidean distance filter

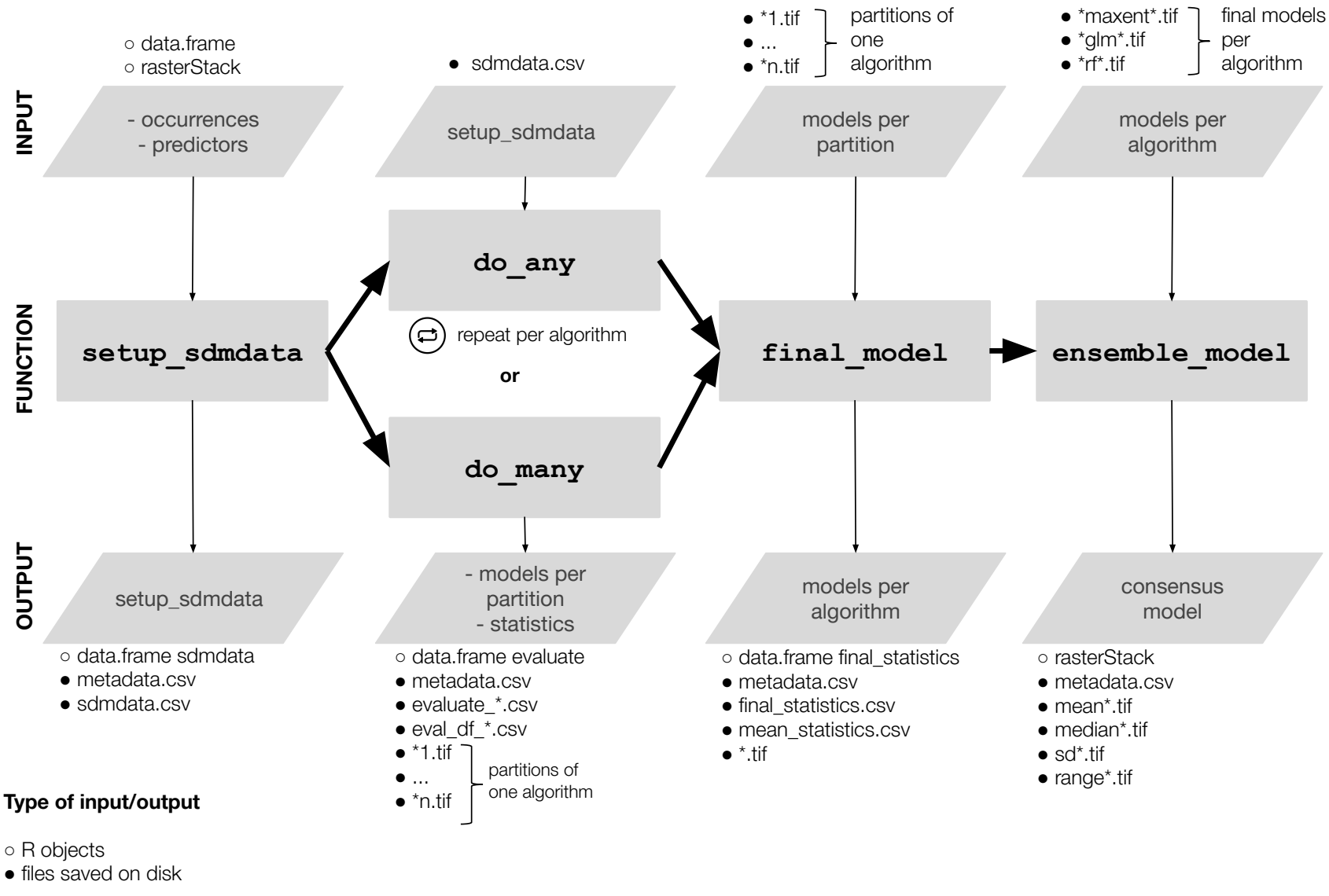


user-defined buffer (M) and euclidean distance filter



sdmdata structure

2. do_(m)any: model fitting and projection



2. do_(m)any: model fitting and projection

do_any() fits the ENM for one algorithm and partition; optionally, do_many() calls do_any() to fit multiple algorithms

- do_any for one algorithm; do_many for many
- parametrization

2. do_(m)any: model fitting and projection

do_any() fits the ENM for one algorithm and partition; optionally, do_many() calls do_any() to fit multiple algorithms

- do_any for one algorithm; do_many for many
- parametrization
- the user can apply a mask

2. do_(m)any: model fitting and projection

do_any() fits the ENM for one algorithm and partition; optionally, do_many() calls do_any() to fit multiple algorithms

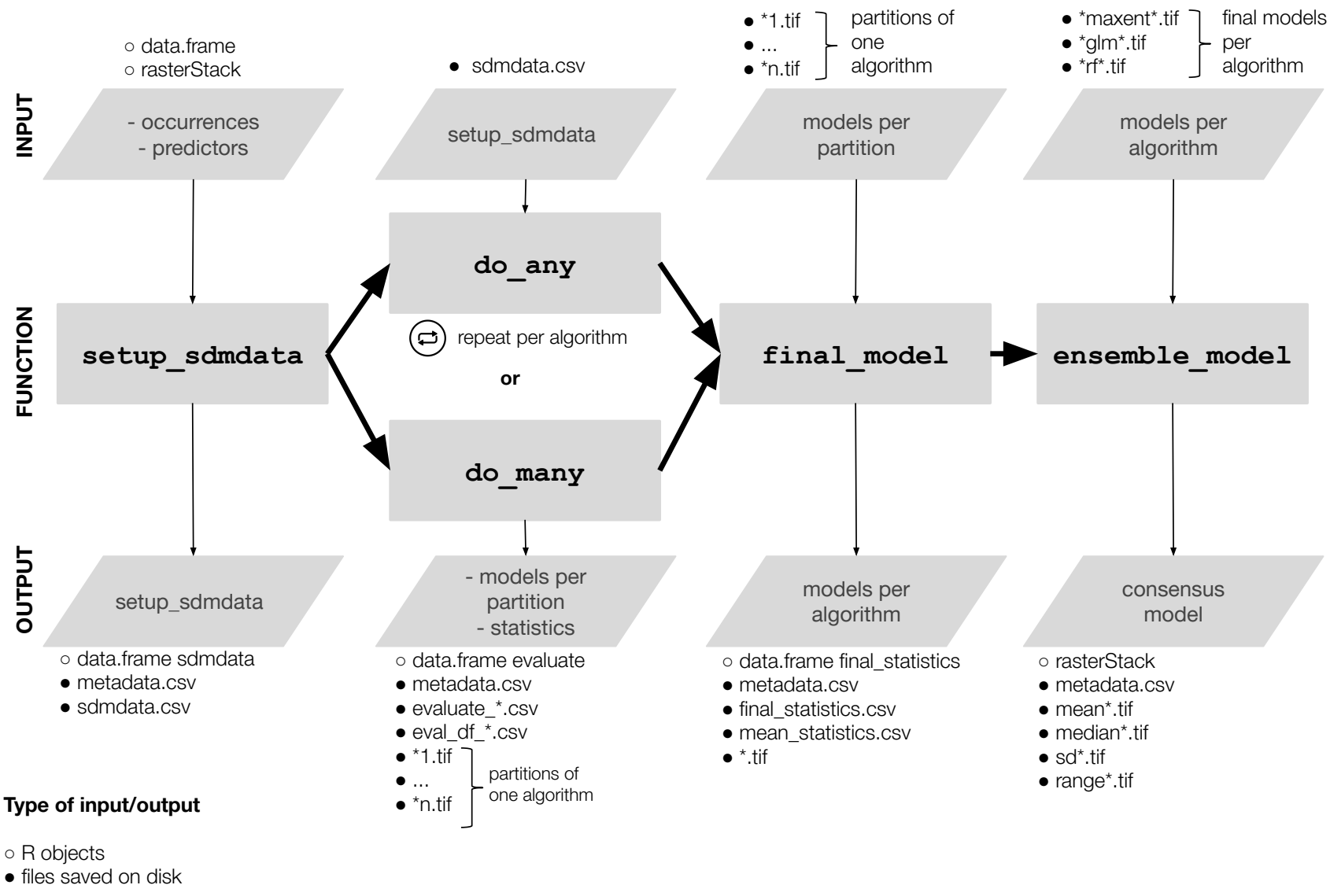
- do_any for one algorithm; do_many for many
- parametrization
- the user can apply a mask
- projection to different datasets (in time or space)

2. do_(m)any: model fitting and projection

do_any() fits the ENM for one algorithm and partition; optionally, do_many() calls do_any() to fit multiple algorithms

- do_any for one algorithm; do_many for many
- parametrization
- the user can apply a mask
- projection to different datasets (in time or space)
- returns table with performance statistics → TSS, AUC, pROC, FNR, Jaccard...

3. final_model: a model per algorithm per species



3. `final_model()`: a model per algorithm per species

- the basics: a central tendency measure and uncertainty between partitions
- uncertainty: range between partitions

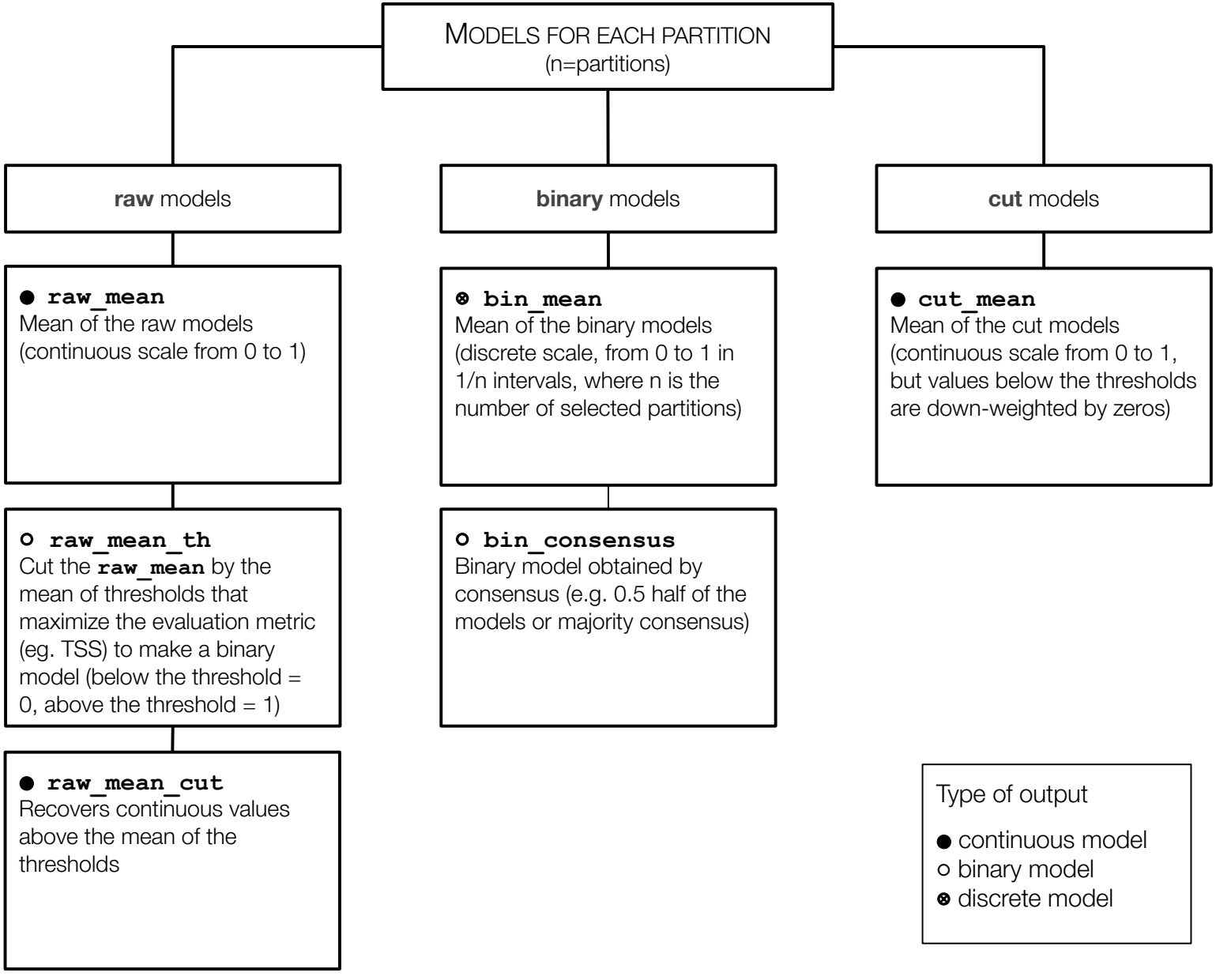
3. `final_model()`: a model per algorithm per species

- the basics: a central tendency measure and uncertainty between partitions
- uncertainty: range between partitions
- which models to join? (the raw continuous model, the binary, o binário)

3. `final_model()`: a model per algorithm per species

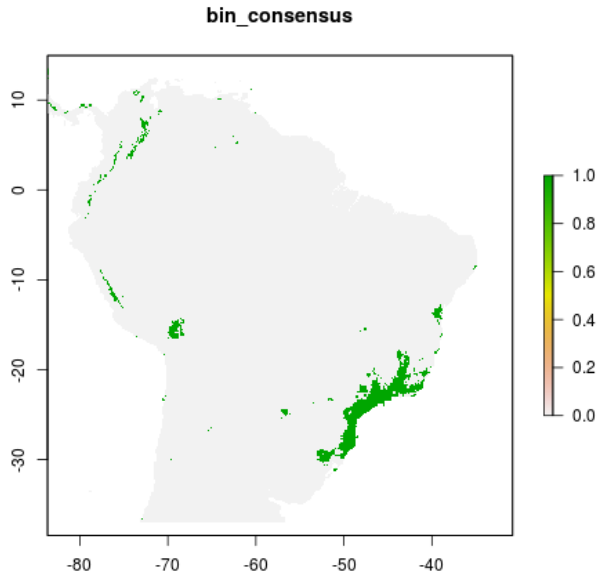
- the basics: a central tendency measure and uncertainty between partitions
- uncertainty: range between partitions
- which models to join? (the raw continuous model, the binary, o binário)
- some additional operations: consensus between binary models

3. `final_model()`: a model per algorithm per species

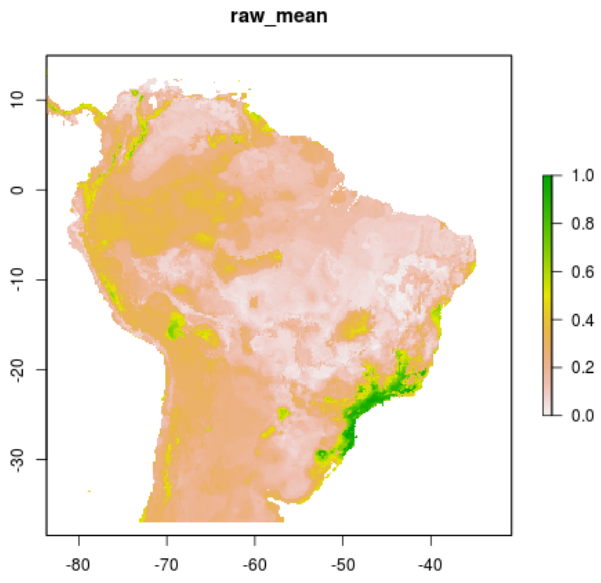


3. `final_model()`: a model per algorithm per species

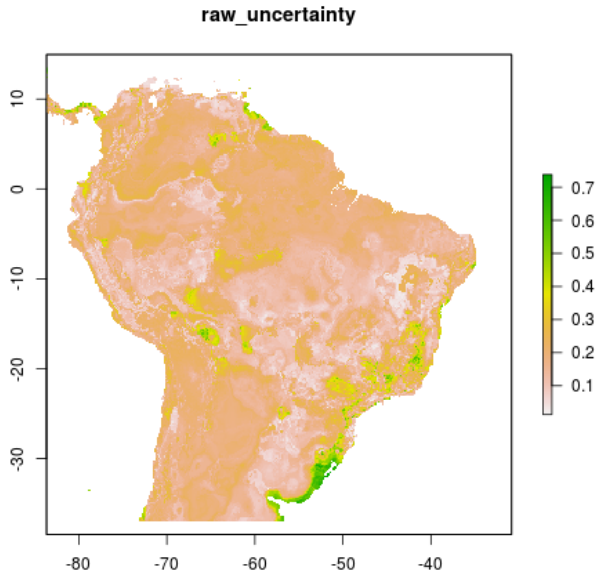
binário



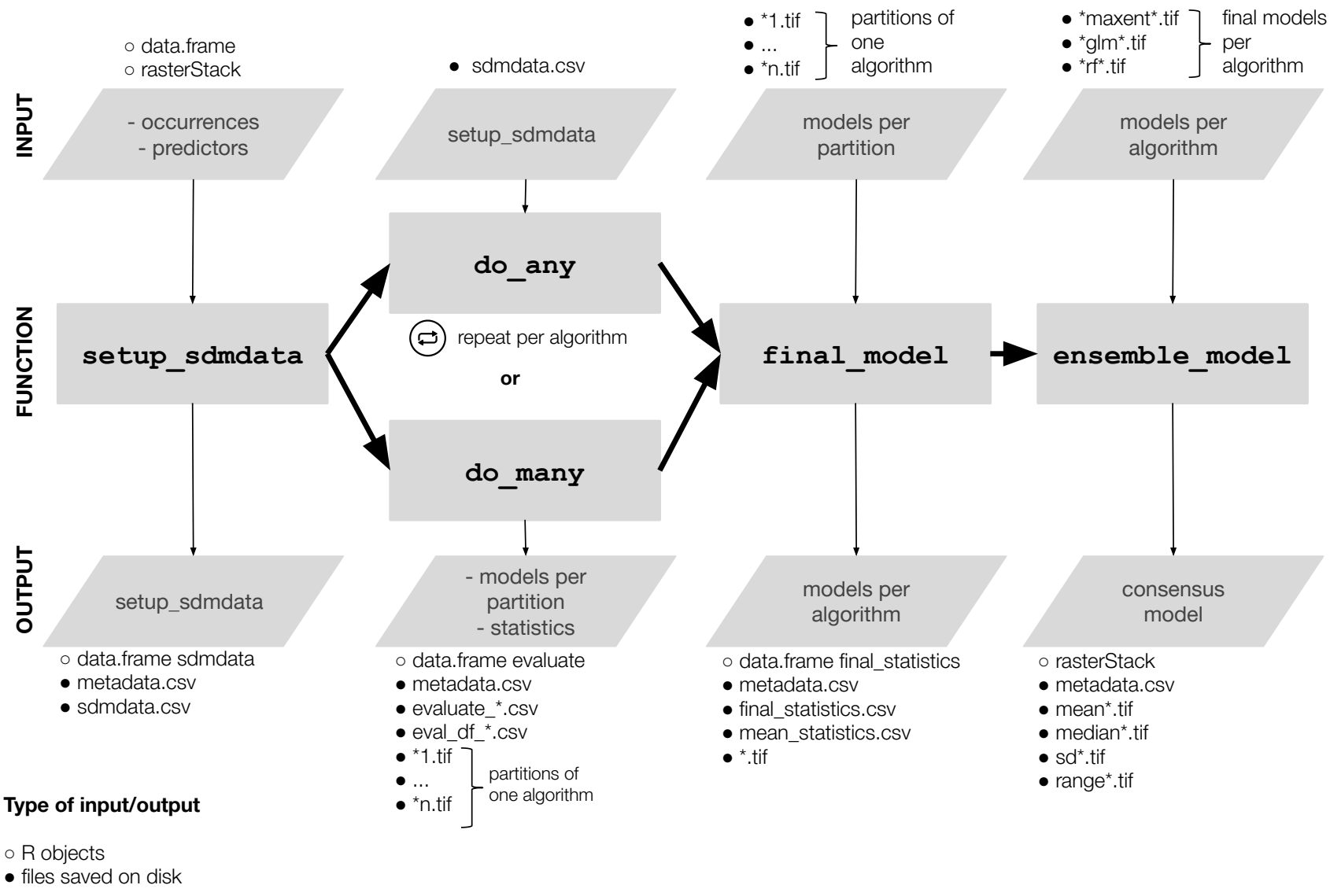
contínuo



incerteza



4. ensemble_model



Type of input/output

○ R objects
● files saved on disk

`ensemble_model`: consenso algorítmico

- média entre os resultados de `final_model`

`ensemble_model`: consenso algorítmico

- média entre os resultados de `final_model`
- desvio padrão, valores mínimos e máximos e incerteza (intervalo - range)

`ensemble_model`: consenso algorítmico

- média entre os resultados de `final_model`
- desvio padrão, valores mínimos e máximos e incerteza (intervalo - range)
- dependendo do output que foi escolhido no passo anterior (`which_final`)

`ensemble_model`: consenso algorítmico

- média entre os resultados de `final_model`
- desvio padrão, valores mínimos e máximos e incerteza (intervalo - range)
- dependendo do output que foi escolhido no passo anterior (`which_final`)
- performance necessariamente melhor do que algoritmos individuais? (Zhu & Peterson 2017)

`ensemble_model`: consenso algorítmico

- média entre os resultados de `final_model`
- desvio padrão, valores mínimos e máximos e incerteza (intervalo - range)
- dependendo do output que foi escolhido no passo anterior (`which_final`)
- performance necessariamente melhor do que algoritmos individuais? (Zhu & Peterson 2017)
- testar outras metodologias de consenso e comparar com o melhor algoritmo (em progresso)

some related studies

Model-R: A Framework for Scalable and Reproducible Ecological Niche Modeling

Andrea Sánchez-Tapia[1], Marinez Ferreira de Siqueira[1], Rafael Oliveira Lima[1], Felipe Sodré M. Barros[2], Guilherme M. Gall[3], Luiz M. R. Gadelha Jr.[3], Luís Alexandre E. da Silva[1], and Carla Osthoff[3]

[1] Botanic Garden of Rio de Janeiro, Rio de Janeiro, Brazil
{andreasancheztapia, marinez, rafael, estevao}@jbrj.gov.br
[2] International Institute for Sustainability, Rio de Janeiro, Brazil
f.barros@iis-rio.org
[3] National Laboratory for Scientific Computing, Petrópolis, Brazil
{gmgall, lgadelha, osthoff}@lncc.br

Botanical Journal of the Linnean Society, 2017, **183**, 348–359. With 3 figures.

Palaeodistribution of epiphytic bromeliads points to past connections between the Atlantic and Amazon forests

JEFFERSON RODRIGUES MACIEL[1,3*], ANDREA SÁNCHEZ-TAPIA[2], MARINEZ FERREIRA DE SIQUEIRA[2] and MARCCUS ALVES[3]

[1]*Jardim Botânico do Recife, Km 7,5 da BR 232, s/n, Curado 50000–230, Recife, PE, Brazil*
[2]*Instituto de Pesquisas do Jardim Botânico do Rio de Janeiro, Rua Pacheco Leão 915, Jardim Botânico, Rio de Janeiro 22460-030, RJ, Brazil*
[3]*Universidade Federal de Pernambuco, Laboratório de Morfo-Taxonomia Vegetal, Av. Moraes Rego, s.n., CDU, 50670–930 Recife, PE, Brazil*

Received 21 June 2016; revised 10 November 2016; accepted for publication 28 November 2016

Botanical Journal of the Linnean Society, 2018, **XX**, 1–20. With 9 figures.

Environmental and geographical space partitioning between core and peripheral *Myrsine* species (Primulaceae) of the Brazilian Atlantic Forest

ANDREA SÁNCHEZ-TAPIA[1], MÁRIO L. GARBIN[2], MARINEZ F. SIQUEIRA[1], KARLO G. GUIDONI-MARTINS[3,4], FABIO R. SCARANO FLS[5,6] and TATIANA T. CARRIJO[7]*

[1]*Jardim Botânico do Rio de Janeiro—JBRJ, Rua Pacheco Leão 915, Jardim Botânico, Rio de Janeiro, RJ, Brazil*
[2]*Universidade Vila Velha, Programa de Pós-Graduação em Ecologia de Ecossistemas, Laboratório de Ecologia Vegetal, Vila Velha, Espírito Santo, Brazil*
[3]*Universidade Federal do Espírito Santo, Centro Universitário Norte do Espírito Santo, 10 Rodovia BR 101 Norte, Km 60, Bairro Litorâneo, 29932-540, São Mateus, Espírito Santo, Brazil*
[4]*Programa de Pós-Graduação em Ecologia e Evolução, Universidade Federal de Goiás, CP 131, Goiânia, GO 74001-970, Brazil*
[5]*Fundação Brasileira para o Desenvolvimento Sustentável, Rua Engenheiro Álvaro Niemeyer 76, São Conrado, 22610-180, Rio de Janeiro, RJ, Brazil*
[6]*Universidade Federal do Rio de Janeiro, Departamento de Ecologia, IB, CCS, Ilha do Fundão, 21941-970, Rio de Janeiro, RJ, Brazil*
[7]*Universidade Federal do Espírito Santo, Centro de Ciências Naturais e da Saúde, Rua Alto Universitário s.n., Laboratório de Botânica, 29500-000, Alegre, ES, Brazil*

Received 28 May 2017; revised 11 December 2017; accepted for publication 26 April 2018

ARTICLES
https://doi.org/10.1038/s41559-018-0743-8
nature ecology & evolution

Strategic approaches to restoring ecosystems can triple conservation gains and halve costs

Bernardo B. N. Strassburg[1,2,3*], Hawthorne L. Beyer[4], Renato Crouzeilles[1,2,3], Alvaro Iribarrem[1,2], Felipe Barros[2], Marinez Ferreira de Siqueira[5], Andrea Sánchez-Tapia[5], Andrew Balmford[6], Jerônimo Boelsums Barreto Sansevero[7], Pedro Henrique Santin Brancalion[8], Eben North Broadbent[9], Robin L. Chazdon[2,10,11], Ary Oliveira Filho[12], Toby A. Gardner[2,13], Ascelin Gordon[14], Agnieszka Latawiec[1,2,15,16], Rafael Loyola[17], Jean Paul Metzger[18], Morena Mills[19], Hugh P. Possingham[20,21], Ricardo Ribeiro Rodrigues[22], Carlos Alberto de Mattos Scaramuzza[23], Fabio Rubio Scarano[3,24], Leandro Tambosi[25] and Maria Uriarte[26]

currently

- Cerrado species for priorization in conservation
- Cerrado species for priorization in conservation