Species distribution models to predict the habitat suitability of alien species - workflow

Aim: This should be an automated workflow that is **easy** to apply by an external user (with little/medium R-knowledge) and that runs robust/ yields **robust results**. It will be linked to a model that simulates the spread of invasive species along traffic networks (terrestrial: train tracks and roads, aquatic: water ways; mainly in Germany, potentially beyond). Specifically, the output of the SDM-workflow should be the climate and habitat suitability (probability of occurrence) for a given focal species along these networks (*i.e.* along the georeferenced train tracks, roads and waterways).

Step 1: Data input and preparation

- **occurrence data**
  - input (user/automated GBIF download)
  - quality check (sufficient occurrence points)
  - format
- **absence data**
  - selection of pseudoabsences (terrestrial; five random sets of PAs)
  - format
- **environmental data**
  - input (automated download of climate data of choice, automated loading of land cover data)
  - correlation test of environmental predictors
  - format

Step 2: Model fitting and validation

- **input of objects from step 1** (occurrences, pseudoabsences, environmental variables)
- **model fitting and validation**
  - algorithm: GAM
  - five random 70-30 data splits
  - five runs for each set of PA (25 runs in total)
  - model fit on 70% of the data, validation on 30%
  - computation of AUC for each model
- **save models and their AUCs as object**

Step 3: Prediction of suitability

- **input of objects from step 3** (final models)
- **input of environmental info** (environment around traffic network, linked to CASPIAN)
- **selection of models with a good AUC value**
- **computation of environmental suitability**
- **average suitability** over the different model runs (algorithm and PA; to be linked with CASPIAN)
- **plot suitability**
- **compute average suitability for traffic network**

This workflow is implemented as a series of R-functions that are called and run in one R-script that is structured according to the three main steps of the workflow as indicated above.

The main steps of the workflow:

Step 1: Data input and preparation
Occurrence data: either the user can insert a dataset with occurrences of the species, or the user can insert the scientific name of a species and occurrence data for that species can be downloaded from GBIF and processed automatically. User should get notifications if the number of occurrence points is too low to yield reliable models. Not yet implemented: Option to combine user-data and GBIF data; option to clip GBIF data to a certain extent, *e.g.* Europe only
Absence data: five random sets of pseudoabsences will be selected automatically in a way that is suitable for the model fitting algorithm that will be used in step 2.
Environmental data: the user can insert the names of the environmental variables of interest and those will be downloaded automatically. Currently, the workflow uses a resolution of 2.5 minutes for the environmental data. WorldClim climate data is used. In addition, the user can select Corine land cover variables that will be provided by us, aggregated to a 2.5 min spatial resolution and giving the percentage land cover per grid cell.

Step 2: Model fitting and validation
Model fitting: Based on the environmental, occurrence and pseudoabsence data prepared in the previous step, models will be fitted automatically using preselected model algorithms (generalized additive models, GAM). Not yet implemented: a second model algorithm, *e.g.* GBM
Model validation: Model validation is done based on five random 70-30 data splits, *i.e.* for each pseudoabsence dataset, the model is fitted based on a random subset of 70% of the data and validated based on the other 30%. For validation, the area under the curve (AUC) is computed.

Step 3: Prediction of suitability
Environmental data: Thus far, the suitability is predicted for the same study area as the model is fitted (*i.e.* global with the option to clip to European extent). The respective environmental layers will be loaded automatically.
Environmental suitability: Based on these environmental data and the fitted models from the previous step an environmental suitability is computed for each grid cell of the study area. This is only done for models with an AUC > 0.7.
Model averaging: The average suitability over the model runs is computed automatically.
Plotting: Two plots are created and stored as pdf on the user's computer. One is a map showing the probability of occurrence of the species in the focal area as predicted by the SDM workflow. The other one plots the original occurrence data on top of the modelled occurrence probability.
Average suitability per link in the traffic network: In a last step the mean and standard deviation of environmental suitability for each link (line segment) in the traffic network can be computed and returned as a shapefile that can be stored on the user's computer.

The Functions:

Step 1: Data input and preparation

'userdatacheck'

- ✓ *what:* checks the quality of the user data and notifies the user
- ✓ *input:* table with occurrence records in predefined format (by user)

'GBIFdownclean'

- ✓ *what:* automatically loads, cleans and checks quality of GBIF occurrence data for the species of interest, plot occurrences, saves table on users computer
- ✓ *input:* scientific name of focal species as character (by user)

'addenvir2'; 'addenvirLC2'

- ✓ *what:* automatically loads environmental predictors of interest, extract environmental data for occurrence records, test for correlation between environmental predictors, notifies user
- ✓ *input:* names of environmental predictors of interest as character (by user), occurrence table from previous step

Not yet implemented:

- ✓ sub-setting occurrences and environment to a smaller extent, *e.g.* Europe (when only climate data is used, this is not yet implemented, could be implemented as an option for the user)
- ✓ distance weighing for PA sampling (not yet implemented, optional)
- ✓ option to subset occurrences to the invasive range (not yet implemented optional)

'PAsampleParallel', 'PAsampleParallelLC'

- ✓ *what:* samples 5 times random pseudoabsences, extracts environmental info for these locations, creates a list of tables with occurrences, pseudoabsences and environmental information
- ✓ *input:* name of environmental predictors of interest as character (automatically, use the one from before), occurrence and environment table from previous step (automatically)

Not yet implemented:

- ✓ spatial blocking for cross validation (not yet implemented, optional)

Step 2: Model fitting and validation

'GAMfitting', 'GAMfittingLC2'

- ✓ *what:* fits 5 GAMs per pseudoabsence sample (based on 5 different random 70-30 data splits), cross validates, computes the AUC and saves models and their AUC in a list
- ✓ *input:* the list of PA samples from the previous script (loaded automatically)

Not yet implemented:

- ✓ a second model algorithm (optional)

Step 3: Prediction of suitability

'loadBaseEnv', 'loadBaseEnvLC2'

- ✓ *what:* loads the environmental info for the area of interest, stacks them, optionally crops them and creates a table that can be used to predict
- ✓ *input:* name of environmental predictors of interest as character (automatically from previous), optionally: cropping mask (user)

'predictGAMParallel'

- ✓ *what:* predicts suitability based on the different model runs and returns a table with the results
- ✓ *input:* list of model runs (automatically, from previous), environmental info table (automatically from previous)

'modelaverageParallel', 'modelaverageParallelLC'

- ✓ *what:* computes mean and standard deviation over the predictions and adds them to the table from the previous step
- ✓ *input:* table with model predictions from previous step

'plotSuitabilities'

- ✓ *what*: creates two plots (modelled suitability; modelled suitability plus occurrence points) and stores them as a pdf on the user's computer.
- ✓ *input*: average suitability from previous step

'SuitabilityNet'

- ✓ *what*: computes for each link (i.e. line feature) in the traffic network the mean and standard deviation of the environmental suitability along that link. The output of this can optionally be stored as a shapefile on the user's computer.
- ✓ *input*: average suitability from previous, shapefile (SpatialLinesDataFrame) with the traffic network as uploaded by the user

Connection with CASPIAN

- ✓ **the input** for the SDM workflow is the name of a species of interest and the name of a selection of environmental variables of interest
- ✓ **the output** is a raster file with probabilities of occurrence/ habitat suitability for Europe (or, if chosen, as smaller extent) at a 2.5 min resolution
- ✓ **one way** to **combine this with CASPIAN** would be to **keep both work flows separately** and combine them via the output file, *i.e.*, if the user wishes to he/she can run the SDM workflow to compute the habitat suitability, these then need to be combined with the shapefile that contains information for the traffic network; otherwise the simple version of habitat suitability that is currently implemented in CASPIAN will be utilized.
- ✓ **combination with the shapefile for the traffic network**: In the CASPIAN it is checked whether the column ***'Env_suit'*** exists in the shapefile of the traffic network, if it does not exist, the 'simple' version of habitat suitability is computed. Therefore, one possibility would be to write a function that fills the column ***'Env_suit'*** with the occurrence probability/

habitat suitability that has been computed in the SDM workflow. To this end, the latitude – longitude information for each node needs to be provided in the shapefile, so that a habitat suitability value for each can be extracted from the raster-file that is the output of the SDM workflow. This might be something Hanno and Mauricio need to prepare.

- ✓ **another option** (more complex and NOT yet implemented) would be to compute the spread within CASPIAN not per node, but per grid cell (from grid cell to grid cell), this might make most sense when the grid cells are smaller than the road/ railway segments? It would offer the option to also include spread away from the network in the future. For this, there might be some changes to the CASPIAN workflow necessary, I need to check the CASPIAN workflow in greater detail.

Github

https://github.com/LarissaNow/SDM-workflow-CASPIAN