BioDT species distribution models

– ODMAP Protocol –

2024-07-03

## Overview

#### Authorship

##### Contact email address

[simrol@ceh.ac.uk](mailto:simrol@ceh.ac.uk); [dylcar@ceh.ac.uk](mailto:dylcar@ceh.ac.uk)

##### Link to study (DOI, web address)

<Study link>

#### Model objective

##### Model objective

Mapping and interpolation

##### Main target output: e.g., suitable vs. unsuitable habitat, continuous habitat suitability index, abundance

Maps of species presence

#### Focal Taxon

##### Focal taxon

Linnaea borealis L.

#### Location

##### Location of study area

Cairngorms National Park, United Kingdom

#### Scale of Analysis

| Attribute | Value |
| --- | --- |
| Location name | Cairngorms National Park, United Kingdom |
| Spatial extent minimum longitude | -4.47763888888889 |
| Spatial extent maximum longitude | -2.79986111111111 |
| Spatial extent minimum latitude | 56.7401388888889 |
| Spatial extent maximum longitude | 57.4134722222222 |
| Boundary type | political |
| Spatial resolution (Metres) | 0.1 m |
| Temporal extent start | There is no temporal extent to the analysis |
| Temporal extent end | There is no temporal extent to the analysis |
| Temporal resolution | There is no temporal resolution to the analysis |

#### Biodiversity data

##### Observation type: choose from list or insert new values

citizen science; field survey

##### Response/data type: choose from list or insert new values

point occurrence

#### Predictors

##### Type of predictor variables: choose from list or insert new values

<Predictor types>

#### Hypotheses

##### Hypotheses about species-environment relationships

investigating how environment variables affect the distributions of the species, Linnaea borealis L. in the Cairngorms National Park

#### Assumptions

##### Critical model assumptions

We assume that there is no temporal changes in environmental variables, that the abiotic variables is the sole predictor of species distributions other than biotic variables

#### Algorithms

##### Modelling techniques used

glm; svm; gaussian process model

##### Justification of model complexity

A variety of models were used without making run time too extensive

##### Use of model averaging/ensemble modelling

We calculated a mean weighted average based on model performance

#### Workflow

##### Conceptual description of modelling steps including model fitting, assessment and prediction

Species occurrence data from the Cairngorms, Scotland was obtained by download from GBIF. We filtered environmental variables to only include environment data from within a 5 km buffer of recorded occurrences, and conducted spatial thinning. Using 4-fold partitioning, a series of spatial models were developed and validated. An ensemble model was created from the model series. We corrected for overprediction using posteriori methods

#### Software

##### Modelling platform incl. version and key packages used

Written using R version 4.4.0 (2024-04-24 ucrt) with packages:

rangeModelMetadata version 0.1.5

lubridate version 1.9.3

forcats version 1.0.0

stringr version 1.5.1

purrr version 1.0.2

tidyr version 1.3.1

tibble version 3.2.1

ggplot2 version 3.5.1

tidyverse version 2.0.0

DT version 0.33

readr version 2.1.5

rmarkdown version 2.27

dplyr version 1.1.4

spThin version 0.2.0

knitr version 1.47

fields version 15.2

viridisLite version 0.4.2

spam version 2.10-0

rgbif version 3.7.9

flexsdm version 1.3.4

sf version 1.0-16

terra version 1.7-78

##### Availability of codes, e.g. code links

<https://github.com/BioDT/uc-ces/tree/main/biodiversity_model>

##### Availability of data, e.g. data links

data obtained from GBIF API with DOI: 10.15468/dl.q5uhpc

## Data

#### Biodiversity data

##### Taxon names: e.g., names of subspecies, species, genus, families

Species: Linnaea borealis L., phylum: Tracheophyta, order: Dipsacales, family: Caprifoliaceae

##### Details on taxonomic reference system

GBIF taxonomic backbone

##### Ecological level: choose from list or insert new values

species

##### Details on species data source: e.g., URL/DOI, accession date, database version

data obtained from GBIF API with DOI: your\_doi\_here at datetime: 2024-07-03 12:54:56.952897

##### Sampling design: spatial design: e.g. random, uniform, stratified), temporal design, nestedness

opportunistic data

##### Sample size per taxon (incl. prevalence): e.g., number of observations/counts, prevalence

species: Linnaea borealis L., sample size = 622

##### Country/region mask, if applicable

No mask was used

##### Details on scaling, if applicable: e.g., rasterisation of polygon maps, spatial and temporal thinning, measures to address spatial uncertainties

Spatial thinning: TRUE

Thinned occurrences based on environmental space

temporal thinning: FALSE

##### Details on data cleaning/filtering steps, if applicable: e.g., taxonomically, outlier presence/treatment

no cleaning/filtering steps

##### Details on absence data collection, if applicable

not applicable

##### Details on background data derivation, if applicable: e.g., spatial and temporal extent, spatial and temporal buffer, bias correction (e.g. target group sampling)

Species occurrences plotted for only species: Linnaea borealis L.

Spatial buffer: TRUE

Established spatial buffers from occurrences with 5 km radius

##### Details on potential errors and biases in data, if applicable: e.g., detection probability, misidentification potential, geo-referencing errors, sampling bias

#### Data partitioning

##### Selection of training data (for model fitting)

random partitioning: TRUE

Conducted in flexsdm using 4 fold random partitioning

##### Selection of validation data (withheld from model fitting, used for estimating prediction error for model selection, model averaging or ensemble): e.g., cross-validation method

we calculate TSS, the threshold at which sensitivity and specificity are equal, as the performance metric used for selecting the best combination of hyper-parameter values in the tuned Maximum Entropy model

##### Selection of test (truly independent) data , sensu Hastie, et al. (2009)

Random partitioning: TRUE

Conducted in flexsdm using 4 fold random partitioning

#### Predictor variables

##### State predictor variables used

bio01, bio02, bio05, bio06, bio12, bio13, bio14, bio15, elevation, slope, aspect, hillshade, Percent\_Tree\_Cover, NDVI, ph, soil\_carbon, grass, flooded\_vegetation, crops, shrub\_and\_scrub, built, bare, snow\_and\_ice, prox\_water, prox\_grass, prox\_flooded\_vegetation, prox\_crops, prox\_shrub\_and\_scrub, prox\_built, prox\_bare, prox\_snow\_and\_ice

##### Details on data sources: e.g., URL/DOI, accession date, database version

Google earth engine

##### Spatial extent of raw data

Spatial extent: -4.47763888888889, -2.79986111111111, 56.7401388888889, 57.4134722222222 (xmin, xmax, ymin, ymax)

##### Spatial resolution of raw data

0.1 km

##### Coordinate reference system (CRS), e.g. proj4 string, EPSG code, ESRI PE string

WGS 84

##### Temporal extent of raw data

No temporal extent

##### Temporal resolution of raw data, if applicable

No temporal resolution

##### Details on data processing and on spatial, temporal and thematic scaling: e.g. upscaling/downscaling, transformations, normalisations, thematic aggregations (e.g. of land cover classes), measures to address spatial uncertainties

no upscaling/downscaling

##### Details on measurements errors and bias, when known

##### Details on dimension reduction of variable set, if applicable - if model-based, this should be contained in Model section (element: Details on pre-selection of variables)

#### Transfer data

##### Details on data sources: e.g., URL/DOI, accession date, database version

data obtained from GBIF API with DOI: your\_doi\_here at datetime: 2024-07-03 12:54:56.952897

##### Spatial extent of transfer data

<Spatial extent>

##### Spatial resolution of transfer data

Not applicable

##### Temporal extent of transfer data

Not applicable

##### Temporal resolution of transfer data, if applicable

Not applicable

##### Models and scenarios used

Not applicable

##### Details on data processing and scaling (see section Predictor variables)

Not applicable

##### Quantification of novel environmental conditions and novel environmental combinations: e.g., distance to training data

Not applicable

## Model

#### Variable pre-selection

##### Details on pre-selection of variables, if applicable

Not applicable

#### Multicollinearity

##### Methods for identifying and dealing with multicollinearity (Dormann, et al. 2013) or justification if multicollinearity is not explicitly dealt with

No methods used to handle collinearity

#### Model settings

##### Model settings for all selected algorithms (including default settings of specific platforms/packages)

| Model | Family | Formula | Weights | Notes |
| --- | --- | --- | --- | --- |
| Gaussian | gaussian | predictors: bio01; bio02; bio05; bio06; bio12; bio13; bio14; bio15; elevation; slope; aspect; hillshade; Percent\_Tree\_Cover; NDVI; ph; soil\_carbon; grass; flooded\_vegetation; crops; shrub\_and\_scrub; built; bare; snow\_and\_ice; prox\_water; prox\_grass; prox\_flooded\_vegetation; prox\_crops; prox\_shrub\_and\_scrub; prox\_built; prox\_bare; prox\_snow\_and\_ice | none |  |
| GLM | gaussian | predictors: bio01; bio02; bio05; bio06; bio12; bio13; bio14; bio15; elevation; slope; aspect; hillshade; Percent\_Tree\_Cover; NDVI; ph; soil\_carbon; grass; flooded\_vegetation; crops; shrub\_and\_scrub; built; bare; snow\_and\_ice; prox\_water; prox\_grass; prox\_flooded\_vegetation; prox\_crops; prox\_shrub\_and\_scrub; prox\_built; prox\_bare; prox\_snow\_and\_ice | none |  |
| SVM | NA | predictors: bio01; bio02; bio05; bio06; bio12; bio13; bio14; bio15; elevation; slope; aspect; hillshade; Percent\_Tree\_Cover; NDVI; ph; soil\_carbon; grass; flooded\_vegetation; crops; shrub\_and\_scrub; built; bare; snow\_and\_ice; prox\_water; prox\_grass; prox\_flooded\_vegetation; prox\_crops; prox\_shrub\_and\_scrub; prox\_built; prox\_bare; prox\_snow\_and\_ice | none |  |

##### Model settings for extrapolation beyond sample range, if applicable: e.g., clamping

Not applicable

#### Model estimates

##### Assessment of model coefficients, e.g. median or mean posterior

Not applicable

##### Details on quantification of parameter uncertainty, e.g. resampling

No quantification

##### Assessment of variable importance

No assessment

#### Model selection - model averaging - ensembles

##### Model selection strategy: e.g. information-theoretic approach for variable selection, shrinkage and regularization

We included all environment variables recorded in a model input raster spanning the Cairngorms, Scotland

##### Method for model averaging: e.g. derivation of weights

No variable weights were used

##### Ensemble method: e.g. initial conditions (input data), model classes, model parameters, boundary conditions

Occurrences obtained from the Global Biodiversity Information Facility (GBIF), with pseudo replication of absences. See model settings table for model classes and parameters

#### Analysis and Correction of non-independence

##### Method for addressing spatial autocorrelation in residuals

No method

##### Method for addressing temporal autocorrelation in residuals

No method

##### Method to account for nested data: e.g., fixed and random effects

No method

#### Threshold selection

##### Details on threshold selection, if applicable: transforming continuous predictions into binary predictions

Not applicable

## Assessment

#### Performance statistics

##### Performance statistics estimated on training data: choose from list or insert new values

Not applicable

##### Performance statistics estimated on validation data (from data partitioning): choose from list or insert new values

For each model, we selected three threshold values to generate binary suitability predictions: the threshold that maximizes TSS (max\_sens\_spec), the threshold at which sensitivity and specificity are equal (equal\_sens\_spec), and the threshold at which the Sorenson index is highest (max\_sorenson).

##### Performance statistics estimated on test (truly independent) data: choose from list or insert new values

Not applicable.

#### Plausibility check

##### Response plots, e.g. partial response plots, evaluation strips, inflated response plots

No response plots

##### Expert judgements, e.g. map display

No expert judgements

## Prediction

#### Prediction output

##### Prediction unit

Species proportional occurrence

##### Post-processing, e.g. clipping, reprojection

Adjustments for overprediction: TRUE

The overprediction of SDMs was corrected for based on occurrence records and suitability patterns.

#### Uncertainty quantification

##### Algorithmic uncertainity, if applicable

Not applicable

##### Uncertainty in input data, if applicable

The models are trained using GBIF datasets. There may be biases introduced by the method(s) of data collection and source contributor(s)

##### Effect of parameter uncertainty, error propagation, if applicable

Not applicable

##### Uncertainty in scenarios (e.g. climate models, land use models, storylines)

<Scenario uncertainty>

##### Visualization/treatment of novel environments: e.g., masking

No visualization or treatment