BioDT species distribution models

– ODMAP Protocol –

2024-02-23

## Overview

#### Authorship

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

<Study link>

#### Model objective

Model objective: Mapping and interpolation

Target output: Maps of species presence

#### Focal Taxon

Focal Taxon: Linnaea borealis L.

#### Location

Location: Cairngorms National Park, United Kingdom

#### Scale of Analysis

| Attribute | Value |
| --- | --- |
| Location name | [Cairngorms National Park, United Kingdom](https://cairngorms.co.uk/) |
| 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: citizen science; field survey

Response data type: point occurrence

#### Predictors

<Predictor types>

#### Hypotheses

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

#### Assumptions

Model assumptions: glm; svm; gaussian process model

#### Algorithms

<Modelling techniques>

<Model complexity>

<Model averaging>

#### Workflow

Model workflow: Pulled data from GBIF, conducts data thinning, fits model, and builds ensemble from models and predicts in space.

#### Software

Software: Written using R version 4.3.1 (2023-06-16 ucrt) with packages, rangeModelMetadata version 0.1.4

lubridate version 1.9.2

forcats version 1.0.0

stringr version 1.5.1

purrr version 1.0.2

tidyr version 1.3.0

tibble version 3.2.1

ggplot2 version 3.4.4

tidyverse version 2.0.0

DT version 0.32

readr version 2.1.4

rmarkdown version 2.25

dplyr version 1.1.4

spThin version 0.2.0

knitr version 1.45

fields version 15.2

viridisLite version 0.4.2

spam version 2.10-0

rgbif version 3.7.7

flexsdm version 1.3.3

sf version 1.0-14

terra version 1.7-39

devtools version 2.4.5

usethis version 2.2.3

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

Data availability: data obtained from GBIF API with DOI: 50c9509d-22c7-4a22-a47d-8c48425ef4a7 data obtained from GBIF API with DOI: 7a3679ef-5582-4aaa-81f0-8c2545cafc81 data obtained from GBIF API with DOI: 0a013f89-5381-4578-9d82-5f28fd5f1ef6 data obtained from GBIF API with DOI: 14d5676a-2c54-4f94-9023-1e8dcd822aa0 data obtained from GBIF API with DOI: 8a863029-f435-446a-821e-275f4f641165 data obtained from GBIF API with DOI: efb3f1a0-43d5-4ae6-a58f-4fd61043756a data obtained from GBIF API with DOI: 53f13c8f-413e-4537-bda6-98666cd7975f data obtained from GBIF API with DOI: 86d0b5c1-804c-4b98-8104-adeb1bff43c4 data obtained from GBIF API with DOI: 8ef95309-811f-4c3b-a329-a85368e774c9 data obtained from GBIF API with DOI: 0a7c5a4e-db9b-4349-a793-ef4b3ef9e025 data obtained from GBIF API with DOI: 887e6209-ea70-46a7-9652-fc94ca97ae0a data obtained from GBIF API with DOI: 6f86b84a-4a96-4c56-905c-3bfdd6ca9023

## Data

#### Biodiversity data

Taxon names: species: Linnaea borealis L., phylum: Tracheophyta, order: Dipsacales, family: Caprifoliaceae

Taxonomic reference system: GBIF taxonomic backbone

Ecological level: species

Data sources: data obtained from GBIF API with DOI: 50c9509d-22c7-4a22-a47d-8c48425ef4a7 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 7a3679ef-5582-4aaa-81f0-8c2545cafc81 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 0a013f89-5381-4578-9d82-5f28fd5f1ef6 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 14d5676a-2c54-4f94-9023-1e8dcd822aa0 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 8a863029-f435-446a-821e-275f4f641165 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: efb3f1a0-43d5-4ae6-a58f-4fd61043756a accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 53f13c8f-413e-4537-bda6-98666cd7975f accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 86d0b5c1-804c-4b98-8104-adeb1bff43c4 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 8ef95309-811f-4c3b-a329-a85368e774c9 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 0a7c5a4e-db9b-4349-a793-ef4b3ef9e025 accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 887e6209-ea70-46a7-9652-fc94ca97ae0a accessed at: 2024-02-23 14:16:34

data obtained from GBIF API with DOI: 6f86b84a-4a96-4c56-905c-3bfdd6ca9023 accessed at: 2024-02-23 14:16:34

Sampling design: opportunistic data

Sample size: species: Linnaea borealis L., sample size = 970

Clipping: No mask was used

Scaling: spatial thinning: TRUE Thins occurrences based on environmental space, with function occfilt\_env()

temporal thinning: FALSE

Cleaning: no cleaning/filtering steps

Absence data: not applicable

Background data: Species occurences plotted for only species: Linnaea borealis L.

Spatial buffer: TRUE spatial buffer established with function calib\_area()

#### Data partitioning

Training data: random\_partitioning: TRUE

conducted automatically in flexsdm using 4 fold random partitioning

Validation data: 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.

Test data: random\_partitioning: TRUE

conducted automatically in flexsdm using 4 fold random partitioning

#### Predictor variables

Predictor variables: 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

Data sources: Google earth engine

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

Spatial resolution: 0.1 km

Coordinate reference system: WGS 84

Temporal extent: No temporal extent

Temporal resolution: No temporal resolution

Data processing: no upscaling/downscaling

#### Transfer data

Data sources: random\_partitioning: TRUE

Model was validated using only 4 fold random partitioning

<Spatial extent>

Spatial resolution: Not applicable

Temporal extent: Not applicable

Temporal resolution: Not applicable

Models and scenarios: Not applicable

Data processing: Not applicable

Quantification of Novelty: Not applicable

## Model

#### Variable pre-selection

Variable pre-selection: Not applicable

#### Multicollinearity

Multicollinearity: No methods used to handle collinearity

#### Model settings

| 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 (extrapolation): Not applicable

#### Model estimates

Coefficients: Not applicable

Parameter uncertainty: No quantification

Variable importance: No assessment

#### Model selection - model averaging - ensembles

Model selection: We included all environment variables recorded in a raster spanning the Cairngorms, Scotland.

Model averaging: No variable weights were used.

Model ensembles: Occurences 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

Spatial autocorrelation: No method.

Temporal autocorrelation: No method.

Nested data: No method.

#### Threshold selection

Threshold selection: 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).

## Assessment

#### Performance statistics

Performance on training data: No response plots

Performance on validation data:

Performance on test data: Not applicable.

#### Plausibility check

<Response shapes>

<Expert judgement>

## Prediction

#### Prediction output

Prediction unit: species proportional occurence.

Post-processing: Overprediction of SDMs corrected for based on occurrence records and suitability patterns.

#### Uncertainty quantification

Algorithmic uncertainty: Not applicable.

Input data uncertainty: Not applicable.

Parameter uncertainty: Not applicable.

<Scenario uncertainty>

Novel environments: No visualisation.