**Data preparation & species distribution modelling**

Identifying fine-scale habitat preferences of threatened butterflies using airborne laser scanning

**Jan Peter Reinier de Vries1, Zsófia Koma1, Michiel F. WallisDeVries2,3 & W. Daniel Kissling1,\***

1Institute of Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, the Netherlands.

2De Vlinderstichting / Dutch Butterfly Conservation, Wageningen, the Netherlands

3Plant Ecology and Nature Conservation Group, Wageningen University, Wageningen, the Netherlands

\*Corresponding author: E-mail address: wdkissling@gmail.com (W.D. Kissling).

Analyses are performed in the following order:

Data preparation:

* A: Script1\_ButterflyData\_1ststep.R: Preparation of the butterfly distribution data for analyses in QGIS.
* B: Spatial analyses in QGIS.
* C: Script2\_ButterflyData\_2ndstep.R: Preparation of the butterfly distribution data for SDMs, including spatial thinning.
* D: Script3\_ButterflyAnalysis.R: Preparation of the LiDAR metrics data per focal species.

Modelling:

* E: Script4\_ SDM\_Model.R: Collinearity assessment & SDM models
* F: Script5\_SDM\_Visualisation.R: Visualisation of the SDM RF models.
* G: Script6\_SDM\_Visualisation\_Ensemble.R: Visualisation of the SDM Ensemble models.

***Data preparation***

**A:** Script1\_ButterflyData\_1ststep.R

* Input:
	+ Raw butterfly count data in csv files ‘<speciesname>\_1418.csv’, presenting transect section counts over the years 2014-2018 per species separately. These have 8 attributes: Transect, Tr\_sec, Transect and Section specify the the id of the transects and sections, N\_butterflies specifies the butterfly counts per Transect section, N\_years specifies the years of full counting, x and y are the spatial coordinates of the observations in the RD\_New coordinate system (EPSG:28992).
	+ A csv file ‘AllSp-data\_1418.csv’ that presents all fully counted transect sections of the butterfly monitoring scheme and is used for the selection of absence points. Attributes are the same as described above.
* From the presence data per species, incidental records (i.e., one individual over four years of monitoring) are rejected to avoid records that refer to erroneous identifications or wandering individuals.
* From the all-species data, absence records per species are selected by filtering all transect sections that are not in the presence dataset of that species.
* Output:
	+ Csv-files of the presence data per species, ‘<speciesname>\_presence.csv’.
	+ Csv-files of the absence data per species, ‘<speciesname>\_absence.csv’.

**B:** Spatial analyses in QGIS

* Input:
	+ Set the RD\_New coordinate system (EPSG:28992).
	+ Csv files of the presence and absence data per species as derived in step A, ‘<speciesname>\_presence.csv’ and ‘<speciesname>\_absence.csv’.
	+ National soil classification data. This dataset was produced by Wösten et al. (2013) and downloaded from [www.pdok.nl](http://www.pdok.nl), files BOFEK2012\_versie2\_1.
	+ National classification data of physical-geographical regions (FGR). This data is derived from [www.pdok.nl](http://www.pdok.nl), files fgrshape\_CBS\_20170410.
	+ Provincial borders of the Netherlands, extracted from the GADM database ([www.gadm.org](http://www.gadm.org)), files gadm36\_NLD\_1.
* Add soil data: We use the PAWN classification of 21 soil types plus water, urban and zero (i.e., NA), originally developed by Wösten et al. (1988). The soil data are added to all presence & absence records of each species, using the tool ‘add polygon attributes to points’ (extension SAGA) with settings; select polygon (BOFEK), select attribute (PAWN), save as .shp-file.
* Separate coastal & inland locations: Dune polygons are extracted from the FGR data and added to the presence-absence datasets for *Boloria* & *Hipparchia*, using the tool ‘add polygon attributes to points’ (extension SAGA) with settings; select polygon (Dunes), select attribute (TOEWIJZING), save as .shp-file.
* Derive province borders for mapping: Province & water polygons from the GADM data are split using the tool ‘split vector’ (QGIS tool), split by ‘type 1’ separating water bodies that are rejected while the province polygons are saved as .shp-file.
* Buffer presence records: A 10 km buffer around the presence records is created for each species, using the tool ‘buffer’ (QGIS tool) with settings; distance = 10 km, segments = 500, style = round, results are merged and saved as .shp-file.
* Filter absence records: Absence records within 10 km distance from a presence record are selected, using the tool ‘Extract by location’ (QGIS tool) with settings; falls within the corresponding buffer .shp-file, then saved as new .shp-file.
* Output:
	+ Csv-files of the presence data per species with new attributes PAWN (soil type) and TOEWIJZING (coastal records), ‘<speciesname>\_PAWN.csv’.
	+ Csv-files of the absence data per species filtered by distance (10 km) and with new attributes PAWN and TOEWIJZING, ‘<speciesname>\_absent\_filter10km.csv’.

**C:** Script2\_ButterflyData\_2ndstep.R

* Input:
	+ Csv files of the presence & absence data per species with geo-attributes as derived in step B, ‘<speciesname>\_PAWN.csv’ and <speciesname>\_absent\_filter10km.csv’.
	+ A csv file presenting the LiDAR metrics from a 25 m radius as derived in LiDAR processing steps, step E, ‘Butterfly\_lidarmetrics\_v6\_25m.csv’.
* Presence & absence information is added to the presence & absence data, respectively.
* Select absence points per species: soil types of water, urban & ‘zero’ (i.e. NA) are rejected, and points with soil types (PAWN) matching the soil types of the presence records are selected. The counts of soil types in the presence data per species are derived.
* Presence & absence data are merged. The variable ‘TOEWIJZING’ (indicating coastal locations) is kept for *Boloria* and *Hipparchia*.
* Incidental mismatches with the LiDAR data presented in ‘Butterfly\_lidarmetrics\_v6\_25m.csv’ and duplicates are avoided. For Hipparchia, incidental mismatches of the coastal locations classification (i.e., coastal locations that fall outside, but are adjacent to, dune areas) are manually corrected.
* Spatial thinning is applied to the data of each species. For this, the coordinates of all transect sections are converted to the WGS84 system (EPSG:4326) and thinned with a distance threshold of 100 m. Each thinning is repeated 10.000 times for optimisation. Outputs are saved as .csv files and imported. The thinned data points are extracted for each species.
* The coastal and inland presence-absence records of *Boloria* & *Hipparchia* are derived in new data frames using the variable ‘TOEWIJZING’. Hereafter, ‘TOEWIJZING’ is rejected from the original thinned data frames of both species.
* Output: Csv-files of the presence-absence data per species, ‘<speciesname>\_thinned \_0420.csv’.

**D:** Script3\_ButterflyAnalysis.R

* Input:
	+ Csv files of the thinned presence-absence data per species as derived in step C, ‘<speciesname>\_thinned \_0420.csv’.
	+ Csv files presenting the LiDAR metrics from a 25 m radius or 100 m radius as derived in LiDAR processing steps, step E-F, ‘Butterfly\_lidarmetrics\_v6\_25m.csv’ and ‘Butterfly\_lidarmetrics\_v6\_100m.csv’.
* The 25 m and 100 m LiDAR metrics are merged into one data frame and renamed. Thereafter, the metrics sets per species are derived.
* The final datasets for species distribution modelling are derived by merging the metrics and species presence data for each species.
* Output: Csv-files of the LiDAR metrics per species, ‘<speciesname>\_metrics\_0420.csv’.

***Modelling***

**E:** Script4\_ SDM\_Model.R

* Input: Csv files of the LiDAR metrics per species as derived in step C, ‘<speciesname>\_metrics \_0420.csv’.
* The datasets are prepared by selecting & ordering the variables needed.
* Collinearity among metrics is assessed per species; collinearity patterns are visualised by Spearman rank correlation plots, and correlated metrics are rejected through a stepwise VIF procedure (Spearman-based). The selections of weakly correlated metrics (r < 0.7) are derived based on VIF outputs, while selecting the metrics that best reflect the species’ ecology from correlated pairs of metrics.
* sdmData-files are derived for each species and sdm models are run, using the algorithms ‘glm’, ‘rf’ and ‘maxent’, bootstrap replication, 100 repetitions per model and a test percentage of 30%. All models are saved as .rds-files.
* Model performance statistics (AUC, COR, TSS and Deviance plus standard deviations) are derived for each model.
* Output:
	+ SDM models of each species saved .rds files, ‘model<species>\_21.rds’.
	+ Spearman rank correlation plots.
	+ Model performance statistics.

**F:** Script5\_SDM\_Visualisation.R

This script produces visualisations based on random forest (rf) models, as shown in the main text.

* Input: Rds files of the SDM models per species as derived in step C, ‘model<species>\_21.rds’.
* Receiver Operating Characteristic (ROC) curves of the models are derived, visualising model performance.
* Variable importance graphs are derived for all models and saved in combined plots for grassland species, woodland species and the inland & coastal populations of *Boloria* & *Hipparchia*.
* Response curves are derived for all variables, visualised in response plots, and saved in combined plots per species. This is repeated for all models.
* The final plots of the grassland models, woodland models and inland/coastal models are built from the variable importance and response plots of each model.
* Output: ROC plots and variable importance & response plots per species, saved as high-resolution tiff files.

**G:** Script6\_SDM\_Visualisation\_Ensemble.R

This script follows the same steps as described in step F, but instead of the random forest (rf) model, ensemble models of the glm, rf and maxent algorithms are visualised here.

* Input: Rds files of the SDM models per species as derived in step C, ‘model<species>\_21.rds’.
* Receiver Operating Characteristic (ROC) curves are visualised for the glm and maxent models separately.
* Variable importance & response plots are based on ensemble models and used to create final plots of the ensemble models.
* Output: ROC plots and variable importance & response plots per species, saved as high-resolution tiff files.

***References***

Wösten, J. H. M., de Vries, F., Denneboom, J. & van Holst, A. F. (1988). Generalisatie en bodemfysische vertaling van de Bodemkaart van Nederland 1 : 250 000 ten behoeve van de PAWN-studie. Rapport 2055, STIBOKA, Wageningen, the Netherlands.

Wösten, J. H. M., de Vries, F., Hoogland, T., Massop, H. T. L., Veldhuizen, A. A., Vroon, H. R. J., ... & Bolman, A. (2013). BOFEK2012, de nieuwe bodemfysische schematisatie van Nederland. Rapport 2387, Alterra, Wageningen, The Netherlands.