Land Cover Suitability Modelling Workflow Manual

Author: Simon Thomsen

Date: 03 July 2025

1. Introduction

This manual explains the land cover suitability modelling workflow implemented in the 'suitability' function. The function generates a raster stack with a suitability layer per land cover as an output, which serves as an input for LULCC modelling in CLUMondoPy. The ‘suitability’ function contains:

- feature selection with variable inflation factor (VIF)
- stratified random sampling (if no shapefile is provided
- model parametrization with 5 different algorithms (Logistic Regression, Random Forest (RF), XGBoost (XGB), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP))
- model evaluation (AUC)
- prediction and export of predicted rasters to stack
- optionally: ensemble model

2. Parameters

The 'suitability' function takes the following parameters:

- classification (str): Path to land cover raster. Raster values should start from 0 in ascending order.
- env_vars (list): List of environmental variable raster paths.
- Mode (str or list): Model(s) to use. Can either be a string (single model) or a list of strings (multiple models). Users can choose from the options ‘logistic’, ‘random_forest’, ‘XGBoost’, ‘SVM’ and ‘MLP’
- out_path (str): Output directory.
- n_samples_corr (int): Number of samples for correlation/VIF.
- sample_size_list (list): List of number of samples per class for model training and validation (only necessary if the parameter ‘sample_points_shapefile’ is not provided)
- min_distance (int): Minimum pixel distance between samples (default to 1, only necessary if the parameter ‘sample_points_shapefile’ is not provided)
- sample_points_shapefile (str): Optional path to a point shapefile for sampling.
- vif_threshold (int): VIF threshold to filter collinear variables (default to 5)
- test_fraction (float): Fraction of data used for testing (default to 0.3)
- random_state (int): Seed for reproducibility (default to None)
- dynamic (bool): Whether to use dynamic predictor variables (default to False)
- dyn_years (list): List of years (int) for dynamic predictions (default to None). See further explication under section 4 for more details.

- dyn_vars (list): List of dynamic variable raster paths (str) (default to None). See further explication under section 4 for more details.
- ensemble (bool): Whether to compute ensemble predictions (default to False)
- no_data_value (int): No-data value used in rasters (default to -9999)
- predict_outputs (bool): Whether to generate suitability maps. If false, only model statistics and metadata will be computed (default to False)

3. Output Files

- - Suitability stack per model (same number of layers as classes in classification raster) (.tif).
- - Model logs: AUC scores, coefficients, variable importance (.txt)
- - Ensemble suitability map (if enabled).

4. Dynamic years and variables

The modelling workflow can account for dynamics in predictor variables. This is useful if users want to account for changes in climate or demographics over time. For example, bioclimatic variables are often used for suitability prediction and are also available for future climate change scenarios.

To use the dynamic functionality, users have to provide two parameters:

- The years in which variables should be updated as a list ('dyn_years'), for example [2030,2040,2050].
- The file paths to dynamic variables as a list ('dyn_vars'). These variables should have the same name as the ones provided in the env_vars argument with an additional extension indicating the year. For example, if one of my environmental variables is 'Bioclim01.tif' and I I have an updated version for the year 2030, I should name it 'Bioclim01_2030.tif'. Make sure that years in 'dyn_years' match the ones in 'dyn_vars'. The order of the file paths is irrelevant.