Project Report
BERN01 - Modelling in Computational Science

Random forest models for the prediction of biome types and climate variables

Author
Pim Nelissen
In collaboration with
Benas Juska, Claire Brumat

LUND UNIVERSITY
FACULTY OF SCIENCE

Department of Mathematics
Faculty of Science
Lund University

November 10, 2024

1 Introduction

The LPJ GUESS model is a model developed by several institutes, one of which Lund University. It predicts biomes based on climate variables. In this project, an attempt is made to predict biomes and vegetation parameters using climate variables such as precipitation, temperature and soil data. The modelling approach is random forest machine learning. Several types of models are produced; binary and multi-class classification of biomes and regression models for predicting the vegetation carbon pool (VegC) , which indicates how much carbon is stored in biomass, and Net primary productivity (NPP), which is how much carbon the biomass gains in a period of one year. The approach is explained, results are organised in sections according to the model type. Feature analysis and performance measures are done to see how the models generalise to unseen data.

2 Machine learning and random forest models

The goal of machine learning, in essence, is to build models using existing data (known as training data) which have predictive capabilities on new data. The data is typically comprised of a collection of observations of several values. In the context of climate modelling, such data could for example be the mean precipitation or temperature at a specific location. More formally, let $\boldsymbol{X_{trn}}$ be an $m \times n$ matrix. We then speak of n *features* and m observations. Besides the training data, there must also be a desired result which is somehow correlated to the features. This is typically known as the *target* and is a $m \times l$ matrix $\boldsymbol{y}$, where l is the number of outputs. The machine learning model then, is mathematical function which takes in an observation $\boldsymbol{m_i}$ and produces an output $\boldsymbol{\hat{y}}$ which is an estimate of $\boldsymbol{y}$. In the most general sense, 'training' means optimising the function such that $||\boldsymbol{y} - \boldsymbol{\hat{y}}||$ is minimised. The choice of algorithm to find this ideal function depends largely on the type of problem and the nature of the target $\boldsymbol{y}$. An algorithm can be one of two types: regressive models are suited to continuous target variables; for distinct prediction categories (e.g. cats and dogs), a classification algorithm is more appropriate.

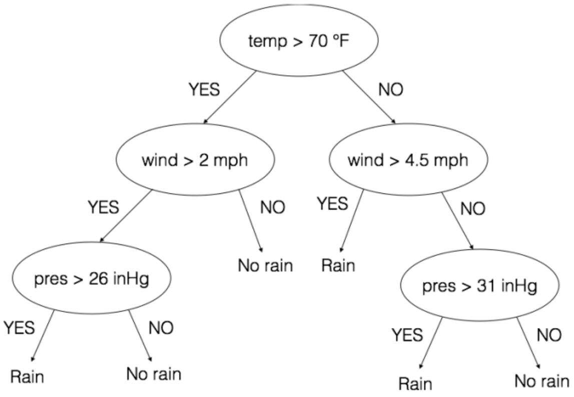

Figure 1: An example of a binary decision tree classifier, to determine whether it will rain or not based on three different weather variables.

How *exactly* the algorithm optimises for the solution is a matter of design, and depends to some extent on the available data and the nature of the patterns within that data. *Random Forest* models offer a good balance between predictive power and interpretability, and are the choice of model in this project. A random forest consists of *decision trees.* One decision tree is a series of comparisons of the input values and a target variable. An example is shown in Figure 1. The idea behind creating a *forest* of trees is that many different sequences of decisions will be generated in order to classify or predict the output. Then, by a system of majority voting of the trees, the class or value that is the outcome of the majority of trees is chosen as the predicted value of a random forest. The advantage of this is that overfitting could be reduced, as 'deep' trees (that is, trees with many decisions), can potentially be biased and have poor generalisation to unseen data by themselves.

Model	$\boldsymbol{X}_{trn}$	$\boldsymbol{X}_{tst}$
Binary classifier	Germany	Benelux[1]
Multi-class classifier	Canada	Nordics[2]
Regression models	Americas	Australia

Table 1: [1] Belgium, Netherlands, Luxembourg. [2] Denmark, Sweden, Norway, Finland.

3 Methodology

3.1 Training and testing data

A total of four models are implemented; two classifiers and two regressive models. The classifiers are a binary and multi-class classifier for biomes. The regressive models are to predict NPP and VegC values. The first step for each of these models is deciding on a satisfactory training set $\boldsymbol{X}_{trn}$ and testing data set $\boldsymbol{X}_{tst}$. Some general good practices/rules are:

- The more training data the better (many observations m_{trn}),
- ideally one has $m_{trn} \gg m_{tst}$, and
- a model can only predict what it has seen; test data must reflect the training data. For example, if a model is trained to classify tundra and ice, it will never classify a test data set containing tropical rainforest and steppe.

With this in mind, the data sets of choice are presented in Table 1.

3.2 Resampling using SMOTE

One common problem is the imbalance of classes. In short, a machine learning algorithm may show a bias towards over-represented classes, and thus overestimate the presence of these classes in unseen data. Likewise, under-represented classes may be predicted less (or not at all) in test data. One common way of dealing with this is resampling of data. One such method is the Synthetic Minority Oversampling TEchnique (SMOTE). The idea is to generate new synthetic data using k-nearest neighbours. This should in theory improve the performance of our algorithm.

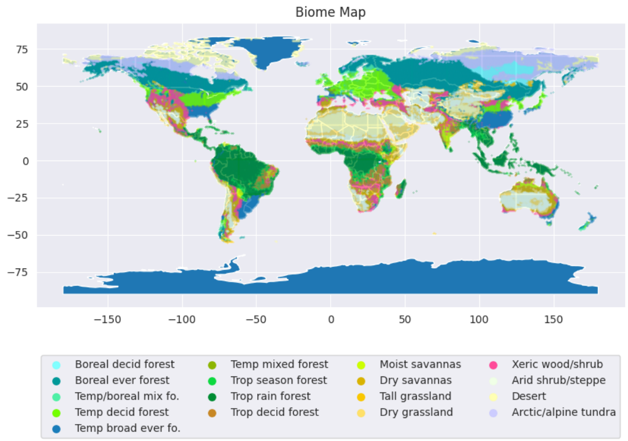

Figure 2: Worldwide distribution of biomes as predicted by the LPJ GUESS model.

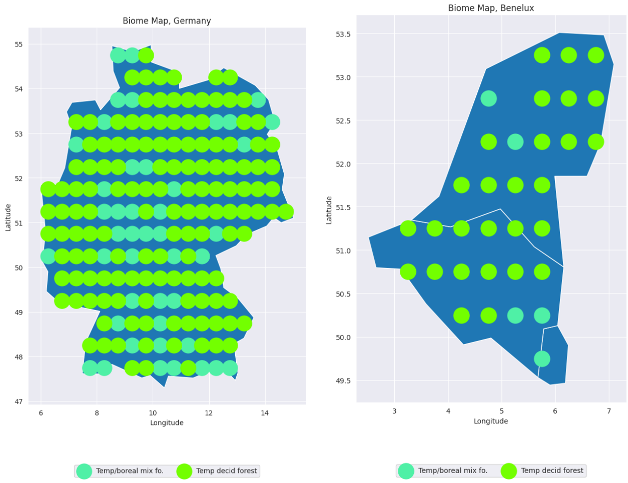

Figure 3: The two targets used for training and testing the binary classifier. The training data shown here is *before* resampling.

4 Results

4.1 Binary classifier

The goal of the binary classifier was to predict the biomes labelled `Biome_obs` in the LPJ GUESS dataset. The two classes in the model are the *temperate/boreal mixed forest* and the *temperate deciduous forest*. the target and test sets $\boldsymbol{y}$ are plotted in Figure 3. Before resampling, an accuracy of 82.8% was achieved, which increased to 85.7% after upsampling the tem-

Class	Precision	Recall
Temp./Boreal Mix	0.3750	0.6000
Temp. Deciduous	0.9259	0.8333

Table 2: Performance measures for the two different classes.

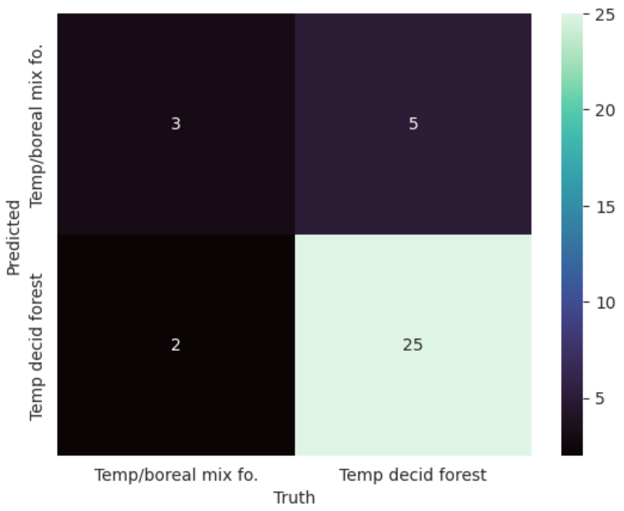

Figure 4: Confusion matrix for the binary classifier predicting on the test data. The model was trained on resampled training data.

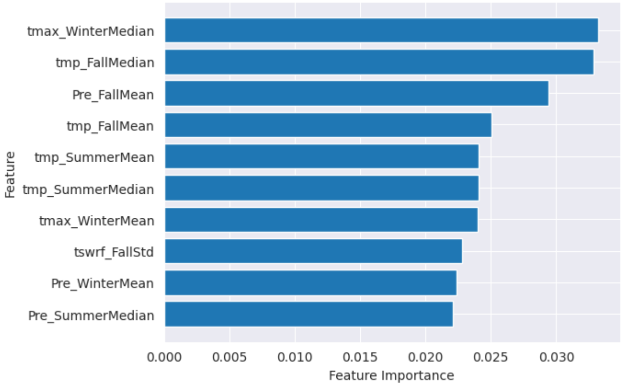

Figure 5: The ten most important features for the binary classification. Temperature and precipitation seem to dominate in this model.

perate/boreal mixed forest using SMOTE. Other performance measures for the model using resampled data are displayed in Table 2 and Figure 4. Finally, feature importance analysis was performed on the model, with the result being displayed in Figure 5.

4.2 Multi-class classifier

Climate data for Canada and the Nordics were used for the training and testing respectively. A display of the biome distribution can be found in Appendix A. For this model, two different targets were used for training. Initially, `Biome_obs` (the observed biomes) were used. As per the project tasks, the model was then retrained using grid search cross-validation. The resulting confusion matrices are displayed in Figure 6 and 7. The weighted average f1 score for both models was about 0.68 and 0.65 respectively, which is unexpected, as a grid search should in theory increase the performance of the model. The full classification reports are appended in Appendix B.

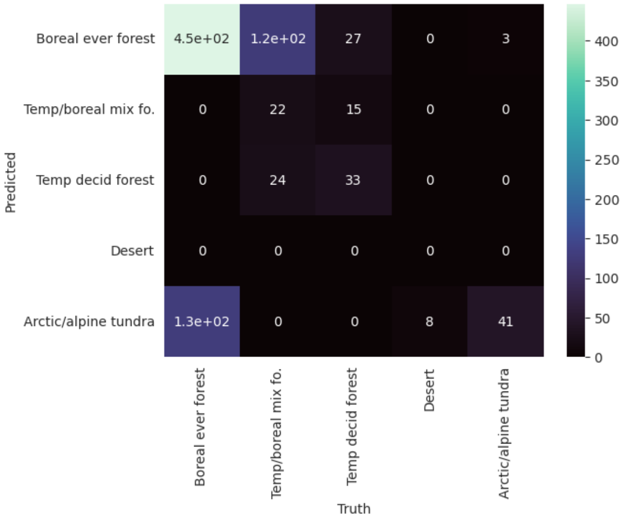

Figure 6: Multiclass confusion matrix without tuning.

Then, the model was retrained using `Biome_Cmax` instead, which is the biome prediction of LPJ GUESS based on the maximum biomass. Here, the weighted average of f1-scores is 0.52, slightly worse than the models trained on `Biome_obs`. However, because of the reduced complexity, we see more correct predictions for temperate deciduous forests, while the new biome class of ‘moist savanna’ has rather poor prediction.

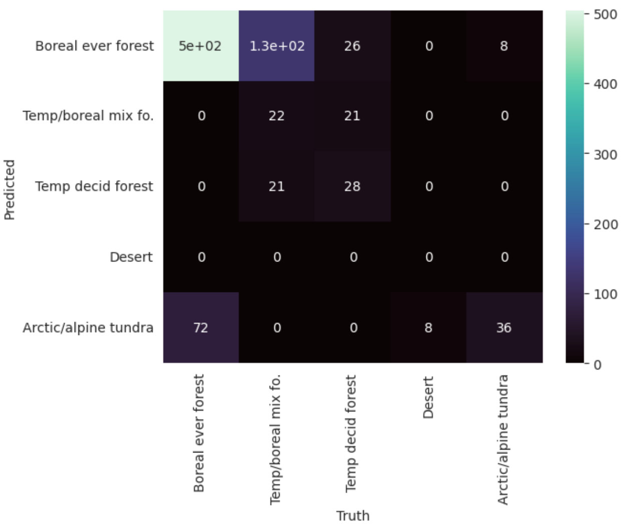

Figure 7: Multi-class confusion matrix with grid search cross-validation.

4.3 VegC & NPP Regression models

Finally, regression models for two continuous variables were implemented. To measure performance, the absolute difference between the model prediction and the LPJ GUESS prediction is plotted in Figure 11 and 12. The root-mean square error (RMSE) was 0.1213 kg C m^{-2} year^{-1} for NPP and 2.8879 kg C m^{-2} for VegC respectively. To make them more comparable, one can normalise the RMSE by taking RMSE/(max(x)−min(x)) where x is the target data for NPP and VegC. This gives NRMSE$_{NPP}$ $\approx$ 0.108 and NRMSE$_{VegC}$ $\approx$ 0.131.

As before, feature importance graphs were generated and are in Figure 13 and 14.

5 Discussion

Binary classifier

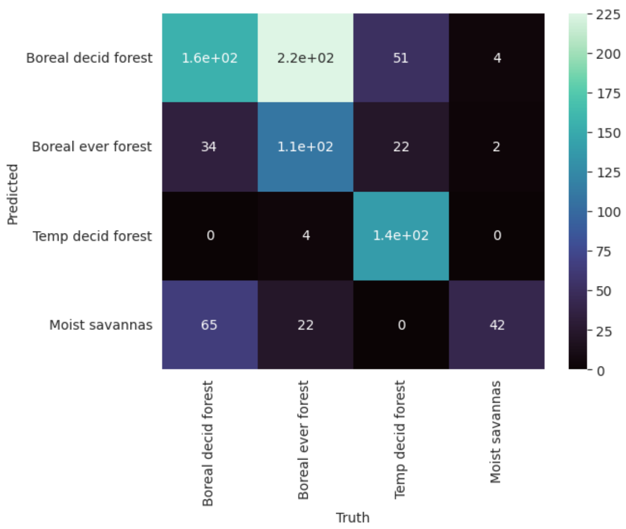

Figure 8: Multi-class confusion matrix for the model trained on `Biome_Cmax`.

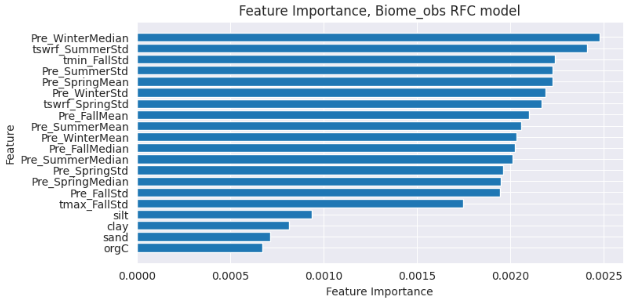

Figure 9: Feature importance plot for the multi-class classifier trained with `Biome_obs` as a target.

The binary classifier was a good first attempt at making a random forest classifier, and with SMOTE resampling, the predictions are somewhat good at 86%. The most common misclassification was temperate deciduous forest as temperate/boreal mixed forest. In hindsight, these two biomes are perhaps too similar, and after discussion we conclude that the climate indicators may not be enough to separate these two biomes well. It could also be that resampling the data did not have as much of an effect as is desired, however, it was effective in the sense that the recall for the mixed forest class went from 0 (not a single prediction)to 0.6, meaning the class was finally predicted sometimes by the model. When it comes to feature importance, we find that the maximum temperature in the winter, as well as the temperature and precipitation in the fall are the most important features. However, there is not any features which are decisively more important than any other, which is a supportive argument for the statement earlier; that these two biomes may not be easily separable based

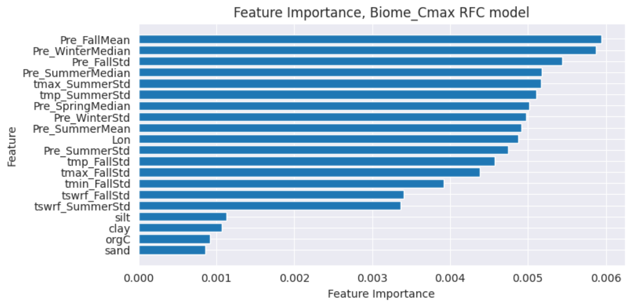

Figure 10: Feature importance plot for the multi-class classifier trained with `Biome_Cmax` as a target.

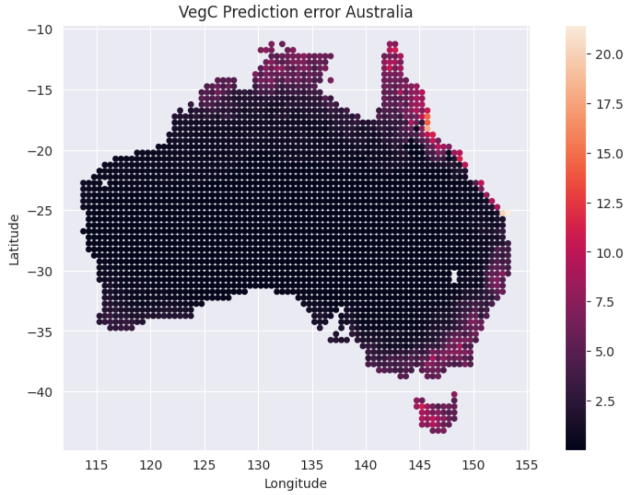

Figure 11: Absolute prediction error of the `VegC` variable for the test data..

on the climate data.

Multi-class classifier

The models both before and after grid search cross validation do not perform particularly well. There are especially many misclassifications for under-represented classes. The worst cases of misclassification before tuning the model were arctic/alpine tundra being classified as boreal ever forest, and temperate/boreal mixed forest being misclassified as boreal ever forest. Overall, there seems to be a strong bias towards this class, which was expected because this is the dominating class in our model. After retuning, the misclassifications of alpine tundra as boreal ever forest reduce significantly. However, we find worse precision in alpine tundra and temperate deciduous forest. Overall, this leads to a lower f1 score of 0.65. This was somewhat unexpected, as finding a model through grid-search cross validation should in theory increase the performance.

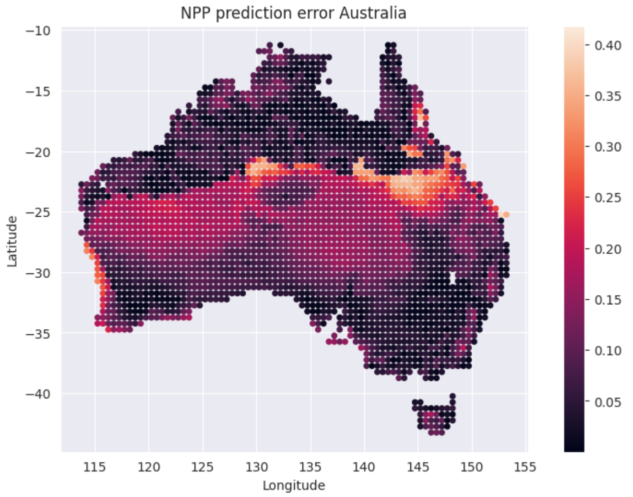

Figure 12: Absolute prediction error of the `NPP` variable for the test data.

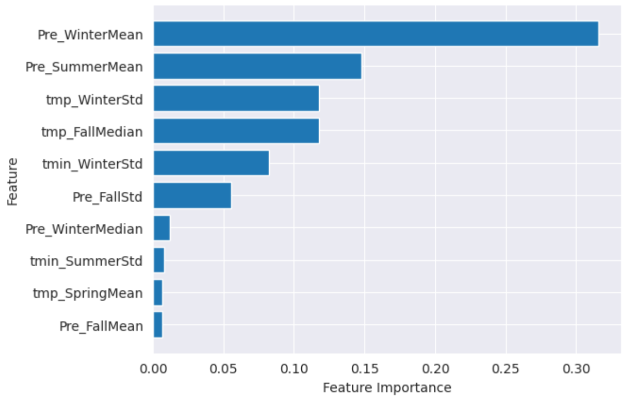

Figure 13: Feature importance for the `NPP` model.

Finally, training the model with `Biome_Cmax` as a target, there is much worse performance, although there is some better precision for some classes. This may just be due to the simplified model, having less biomes. This however does not explain why the performance is bad.

For both models, one of the most common misclassification is temperate/boreal mixed forest as boreal evergreen forest. Like with the binary classifier model, this may be due to the climate variables not being enough to separate

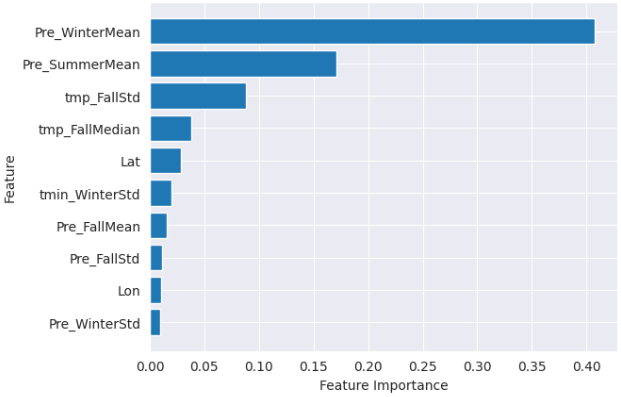

Figure 14: Feature importance for the `VegC` model.

these classes. When it comes to features, there is once again no clear distinguishing features in both models, but a large spread of features all having a similar impact.

Regression models

The regression models, unlike the classifier models, seem to have clear features that are considered important. Firstly, the mean precipitation in winter is an important prediction for both the NPP and VegC models. This is expected; the available biomass and hence NPP will strongly correlate to the amount of water available. In both training and testing data sets, there are regions with more and less precipitation, so this variable forms a clear separator. Precipitation in the summer is the second most important feature. Both the NPP and VegC model have larger errors in predictions in coastal areas. In the north-east of Queensland, near Cairns, we see a particular hotspot of error in both models. Upon further research, we determined that ocean currents may play a role in the wrong prediction here, as this may not be fully reflected in the climate data measured on land. After normalising the RMSE, we find that the NPP model is performing slightly better than the VegC model on average, despite some hotspots with very poor prediction in both the coastal areas as well as desert areas. Perhaps the combination of training and testing data was not appropriate here; the Americas (training data) have relatively little desert and a lot of rainforest, whereas Australia (test data) has a lot of desert. In future work, perhaps the entire world (sans Australia) could be used for training in order to improve the regression models.

6 Conclusion

This project has shown that machine learning as a method of modelling has several aspects. While the technological aspect are well taken care of by extensive libraries, the real challenge is in having enough, and relevant, training data. On top of that, testing data needs to reflect what was trained on. This is the primary factor to a model's performance, and this is likely where our models could be improved in the future; by more careful selection of training data. The feature importance ambiguity for most models seems to also imply that data selection was not optimal. Hyperparameter tuning and techniques like resampling and cross-validation are in theory ways to improve the performance of models, but these techniques had mixed results in our case. Overall, machine learning can be an effective method, especially when it is explainable and interpretable. With more time, a re-evaluation of the training/test data selection and an assessment of the specific input features could be done to get better results.

A Other figures

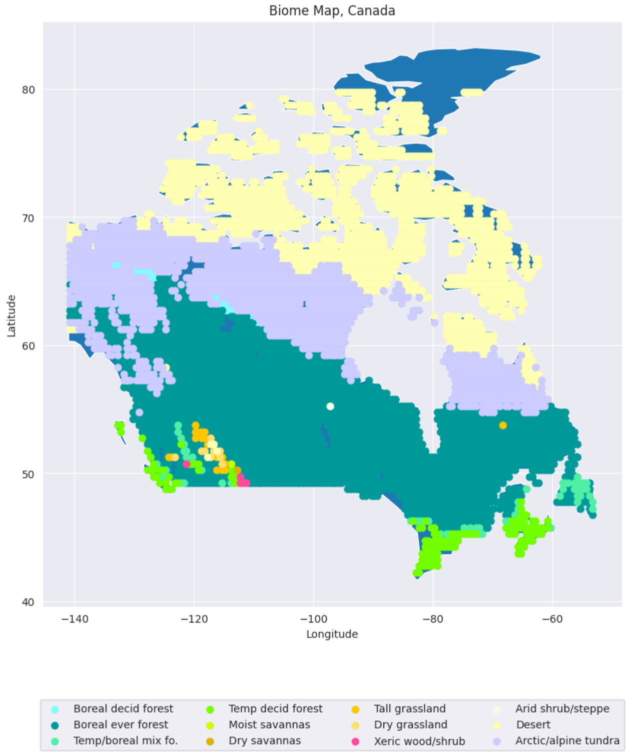

Figure 15: Biome distribution according to observations for Canada, used as the training target for the multi-class algorithm trained on `Biome_obs`.

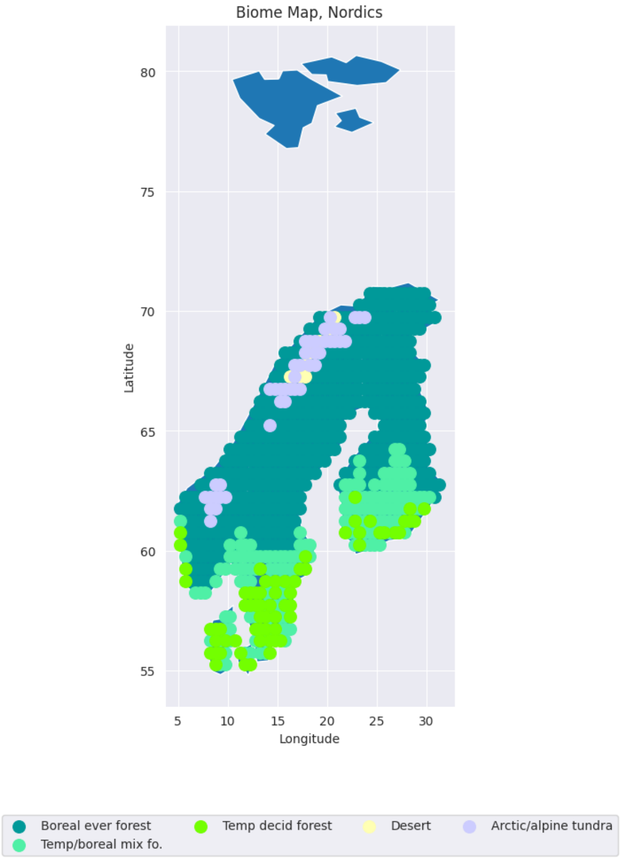

Figure 16: Biome distribution according to observations for the Nordics, used as the test target for the multi-class algorithm trained on `Biome_obs`.

B Multiclass classification reports

	precision	recall	f1-score	support
Boreal ever forest	0.781575	0.913194	0.842274	576
Temp/boreal mix fo.	0.757576	0.146199	0.245098	171
Temp decid forest	0.584416	0.600000	0.592105	75
Desert	1.000000	0.000000	0.000000	8
Arctic/alpine tundra	0.373626	0.772727	0.503704	44
accuracy	0.720824			
macro avg	0.699439	0.486424	0.436636	874
weighted avg	0.741423	0.720824	0.679213	874

Table 3: Classification report for the multiclass model trained on `Biome_obs`, before tuning of the model using cross-validation.

	precision	recall	f1-score	support
Boreal ever forest	0.766412	0.871528	0.815597	576
Temp/boreal mix fo.	0.591837	0.169591	0.263636	171
Temp decid forest	0.596154	0.413333	0.488189	75
Desert	1.000000	0.000000	0.000000	8
Arctic/alpine tundra	0.305085	0.818182	0.444444	44
accuracy	0.684211			
macro avg	0.651898	0.454527	0.402373	874
weighted avg	0.696559	0.684211	0.653359	874

Table 4: Classification report for the multiclass model trained on `Biome_Obs`, AFTER tuning of the model using cross-validation.