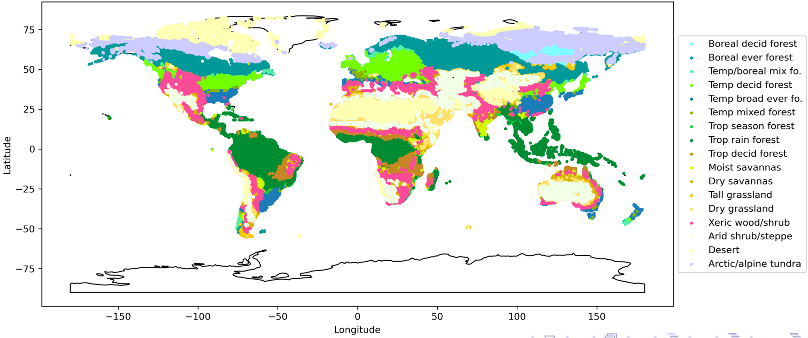
Lund University
BERN01

Using random forest machine learning to predict plant geography and carbon fluxes

Theo Koppenhöfer, Anna Rockstroh, Carmen Lopez Jurado

26.10.2023



Content

1. Binary classification

2. Multiclass classification

3. Regression

Content

1. Binary classification

2. Multiclass classification

3. Regression

Motivation for choosing China and Egypt

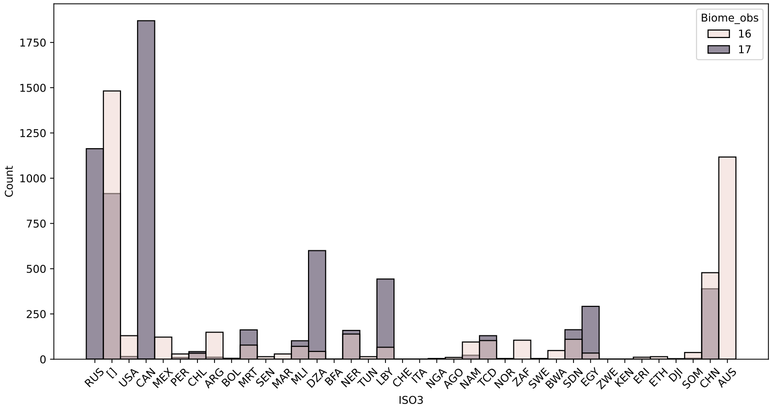


Fig. 1: Arid shrub/stepp (16) and desert (17) locations for some countries.

Binary classification for China and Egypt

We used

- Egypt to train (sample size 326)
- China to test (sample size 867)

Base experiment: we drop the categories `'MaxBiomeLAI','Biome_obs','Biome_LAI','Biome_Cmax','Lon','Lat','Pan_2007','ISO3','UN','MaxBiomeCmax'`

Basic results

Truth Predicted	16	17
16	477	12
17	1	377

Tabelle 1: Confusion matrix.

	precision	recall	f1-score	support
16	0.975460	0.997908	0.986556	478.000000
17	0.997354	0.969152	0.983051	389.000000
accuracy	0.985006	0.985006	0.985006	0.985006
macro avg	0.986407	0.983530	0.984804	867.000000
weighted avg	0.985284	0.985006	0.984984	867.000000

Tabelle 2: Output of the classification report function.

An series of experiments

Experiment	accuracy on the test data
base	0.9850
drop medians	0.9919
drop climate	0.9931
drop spring	0.9781
drop summer	0.9308
drop fall	0.9334
drop winter	0.9769
drop precipitation	0.9850
drop temperature	0.9839
drop tswrf (radiation)	0.9850
only climate	0.5502

Tabelle 3: Accuracy on the test data for various experiments

What happens if we only use climate data?

Truth Predicted	16	17
16	477	389
17	1	0

Tabelle 4: Confusion matrix.

- Table 8: that all the desert (17) in China was classified as arid shrub (16)
- Explanation: China's deserts have a different climate to Egypt's deserts

Content

1. Binary classification

2. Multiclass classification

3. Regression

Motivation for choosing Russia and Canada

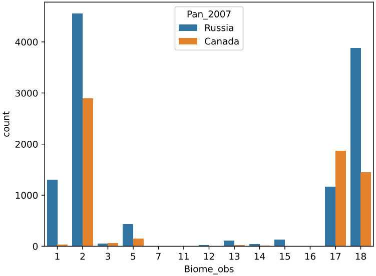


Fig. 2: Biome_obs distribution for Russia and Canada.

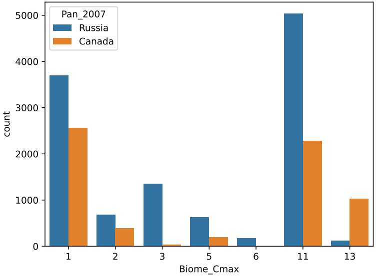


Fig. 3: Biome_Cmax distribution for Russia and Canada.

Results for Biome_obs

Notes:

- training data: Russia (11696 samples), test data: Canada (6497 samples)
- drop of the same features as in the binary case
- Application of the basic RandomForestClassifier
- Decision against tuning hyperparameters

(Balanced) accuracy of the RandomForestClassifier:

	Train	Test
Accuracy	1.0	0.96
Balanced accuracy	1.0	0.66

Averaged precision, recall and f1-score:

	precision	recall	f1-score	support
Macro avg	0.78	0.66	0.72	6497
Weighted avg	0.96	0.96	0.96	6497

Feature importance

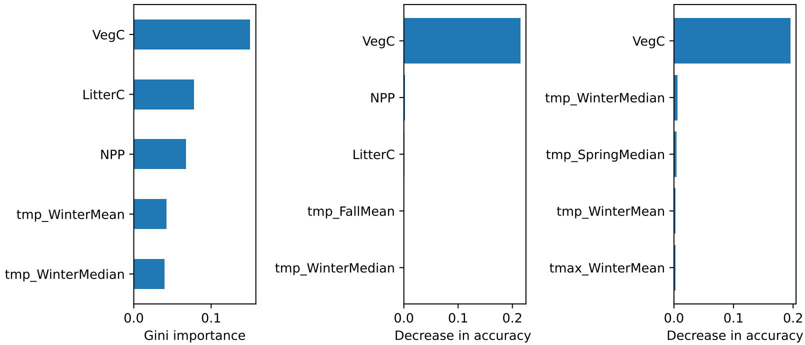


Fig. 4: Impurity importance and permutation importance for train data (first, second plot) and permutation importance for test data (third plot).

Conclusion:

- VegC most important variable
- Importance of features differs for test and train data
- Multicollinearity between features

Clustering and various experiments

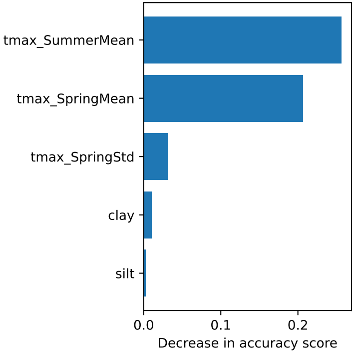


Fig. 5: Permutation importances on through clustering selected subset of features.

Experiment	acc.	bal. acc.
Basic	0.96	0.66
VegC, LitterC, NPP	0.93	0.41
Drop climate	0.93	0.38
Only climate	0.85	0.47

Tabelle 5: Accuracy on the test data for various experiments

Comparison with LPJ_guess output

1. Training on Biome_obs, Testing on Biome_Cmax gives 0.09 accuracy
2. Training on Biome_Cmax, Test on Biome_obs gives 0.14 accuracy

Conclusion:

- Accuracy is bad
- No prediction of Biome_Cmax with Biome_obs model possible, and v. v.
- Reason: Significant differences Biome_Cmax and Biome_obs data
- Conclusion:
  - LPJ-Guess output is not accurate
  - Random Forest method trains the model for the desired label and than can only predict this type of label.

Content

1. Binary classification

2. Multiclass classification

3. Regression

Regression for Russia and Canada

We use

- Canada (sample size 6499) to train
- Russia (sample size 11696) to test

to predict the continuous parameters

- NPP (net primary productivity)
- VegC (vegetation carbon pool)

Basic results for VegC

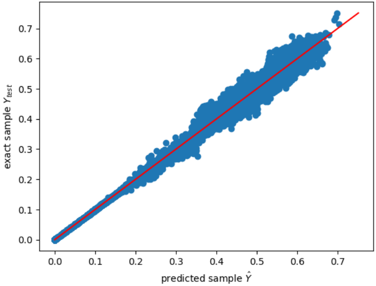


Fig. 6: NPP predicted vs. real.

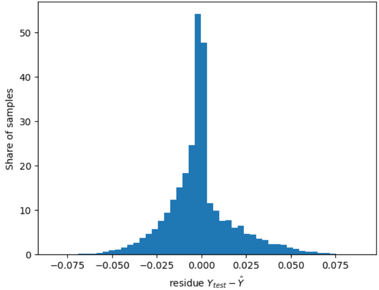


Fig. 7: Distribution of the residual for NPP.

Basic results for NPP

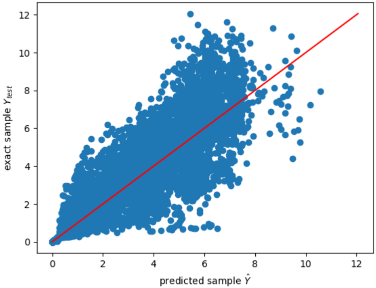


Fig. 8: VegC predicted vs. real.

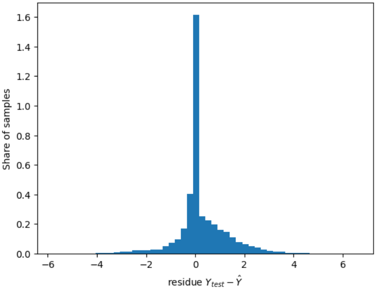


Fig. 9: Distribution of the residual for VegC.

The results for a series of experiments

experiment name	sqrt(MSE)
base	0.990894
drop meadians	0.991130
drop weather	0.992043
drop Fall	0.990528
drop Summer	0.990358
drop Winter	0.991410
drop Spring	0.991412
drop pre	0.991112
drop tmp|tmin|tmax	0.990946
drop tswrf	0.991115
only weather	0.832905

Tabelle 6: Accuracy on the test data for various experiments

Thank you for listening!

Any Questions?