Comparison of machine learning algorithms for large-scale land cover fraction estimation

Dainius Masiliūnas, *Nandin-Erdene Tsendbazar, Martin Herold, Jan Verbesselt*

Laboratory of Geo-Information Science and Remote Sensing, Wageningen University

WAGENINGEN
UNIVERSITY & RESEARCH

Fractional land cover mapping

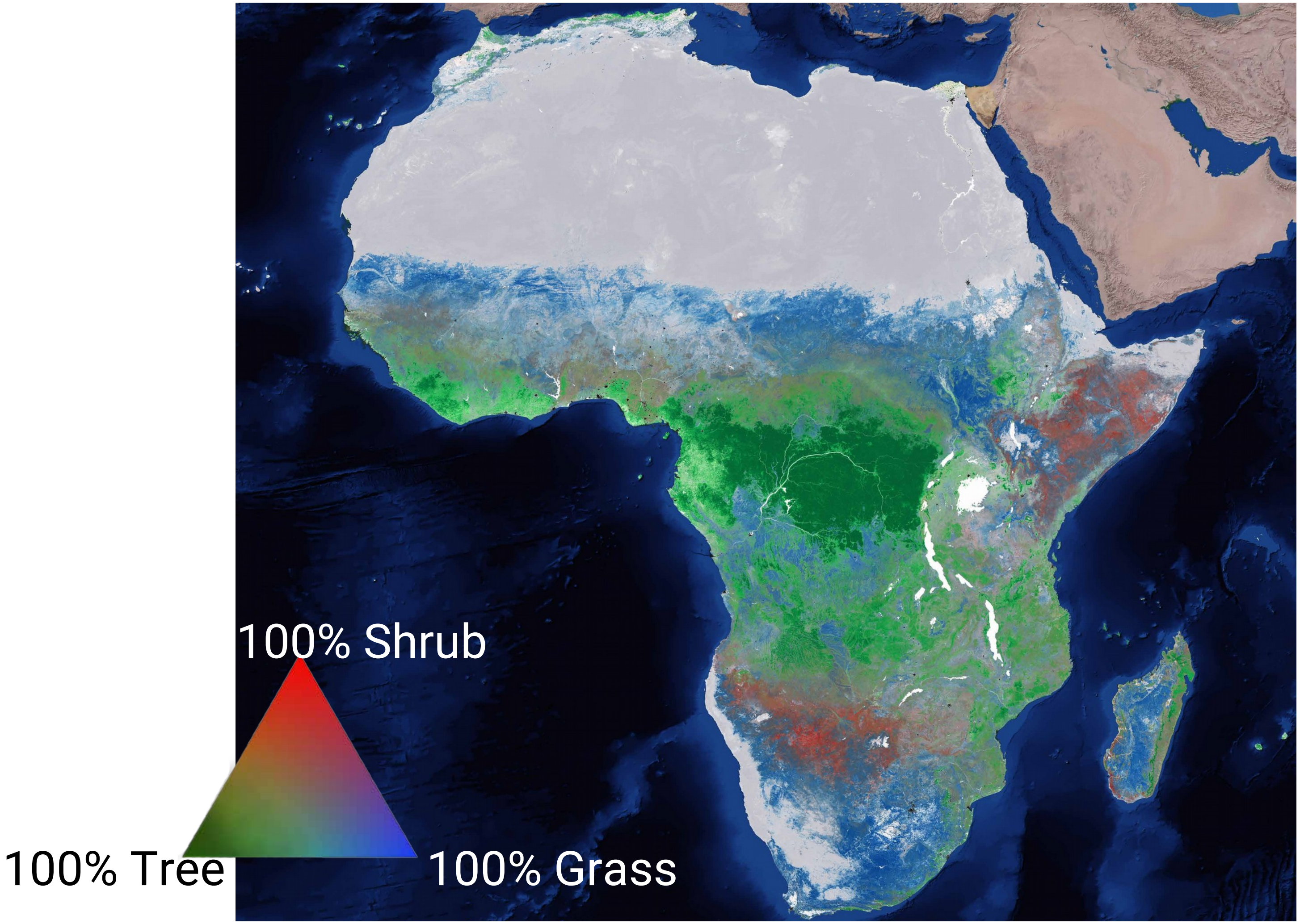
- Traditional land cover (LC) maps assign **one class** to a **pixel**
- **Mixed pixels** cannot be represented!
- Fractional LC mapping: fraction of each class in each pixel

Study goals

- Develop **methodology** for dealing with **fractional** training data
- **Compare** machine learning regression **algorithm performance** in fractional LC mapping
- Determine which **covariates** are **most important** for fractional LC mapping

WAGENINGEN UNIVERSITY & RESEARCH

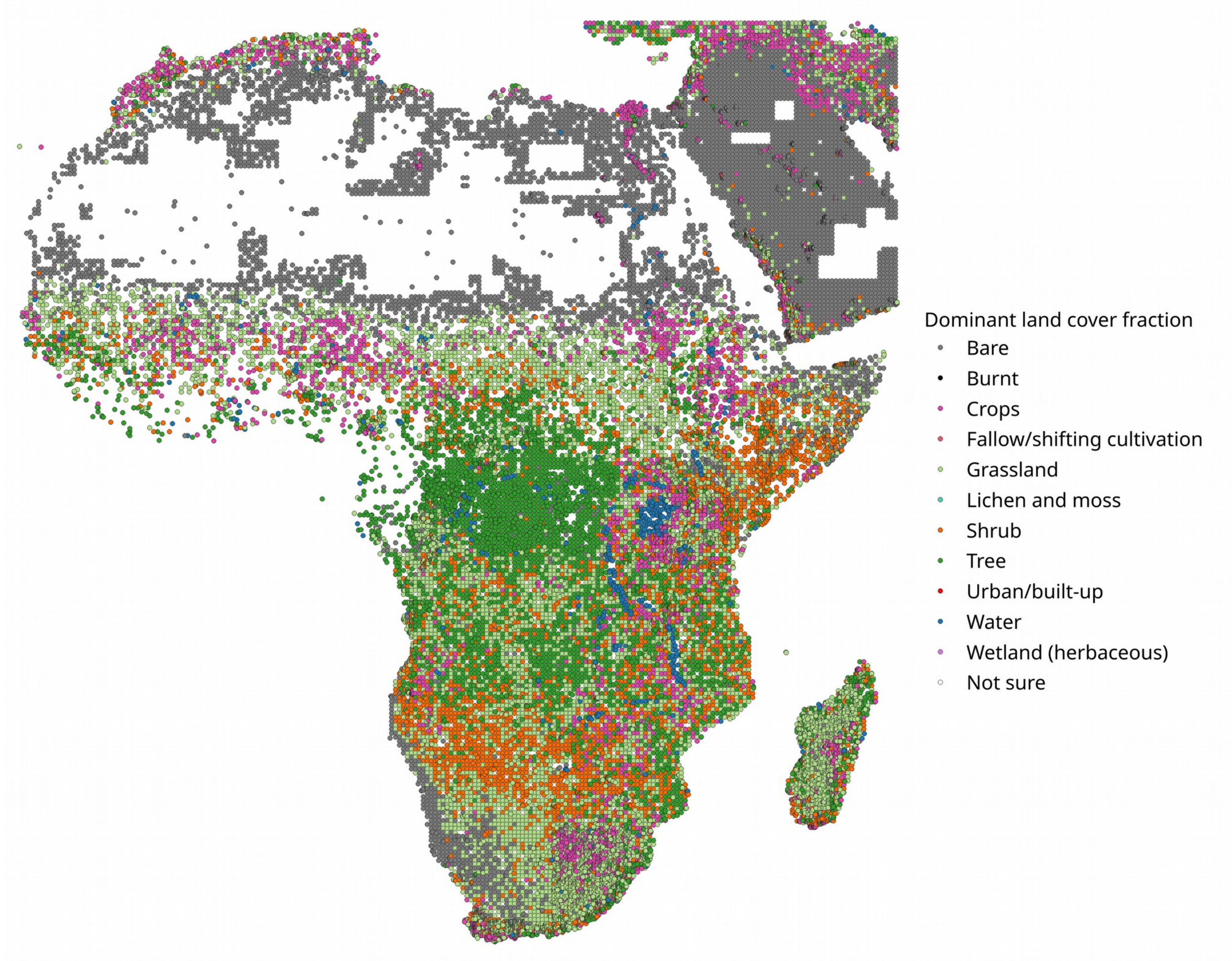
Land cover fractions (CGLS-LC100)



Methodology

- **7 models**: Random Forest regression, Multilayer Perceptron, partial least squares regression, fuzzy nearest centroid, lasso regression, logistic regression, intercept model
- **5 groups of covariates**: Spectral data from **Proba-V**, its temporal metrics, elevation and terrain parameters, climate biophysical parameters, location
- **7 classes**: bare soil, crops, trees, shrubs, grass, urban, water
- **Validation** using RMSE, MAE, ME, R^2, fuzzy confusion matrix

Reference data: 26351 training + 3152 validation points (collected for CGLS-LC100)

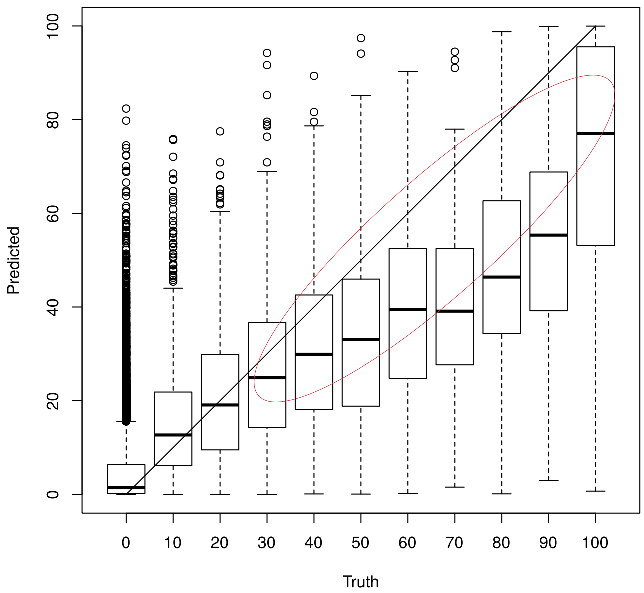


WAGENINGEN UNIVERSITY & RESEARCH

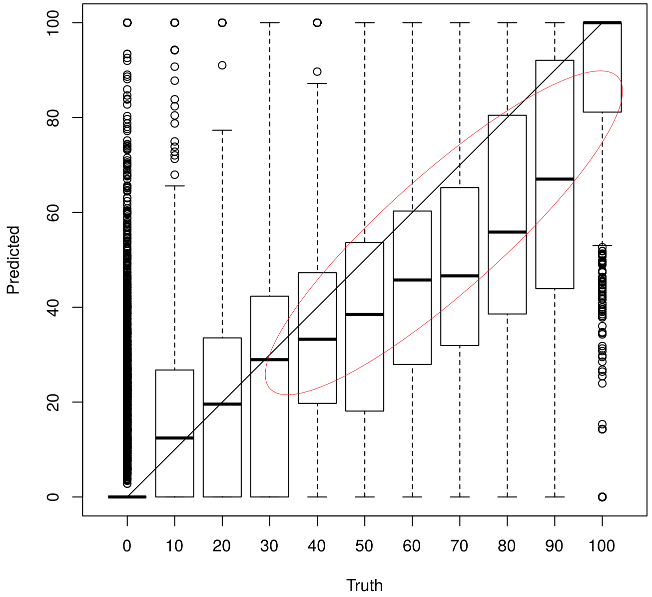
Multimodel method

- Fractional, so training data **imbalanced** towards 0
- Two models: one to classify **zeroes**, one for **non-zeroes**

Random Forest, single model



Random Forest, two models

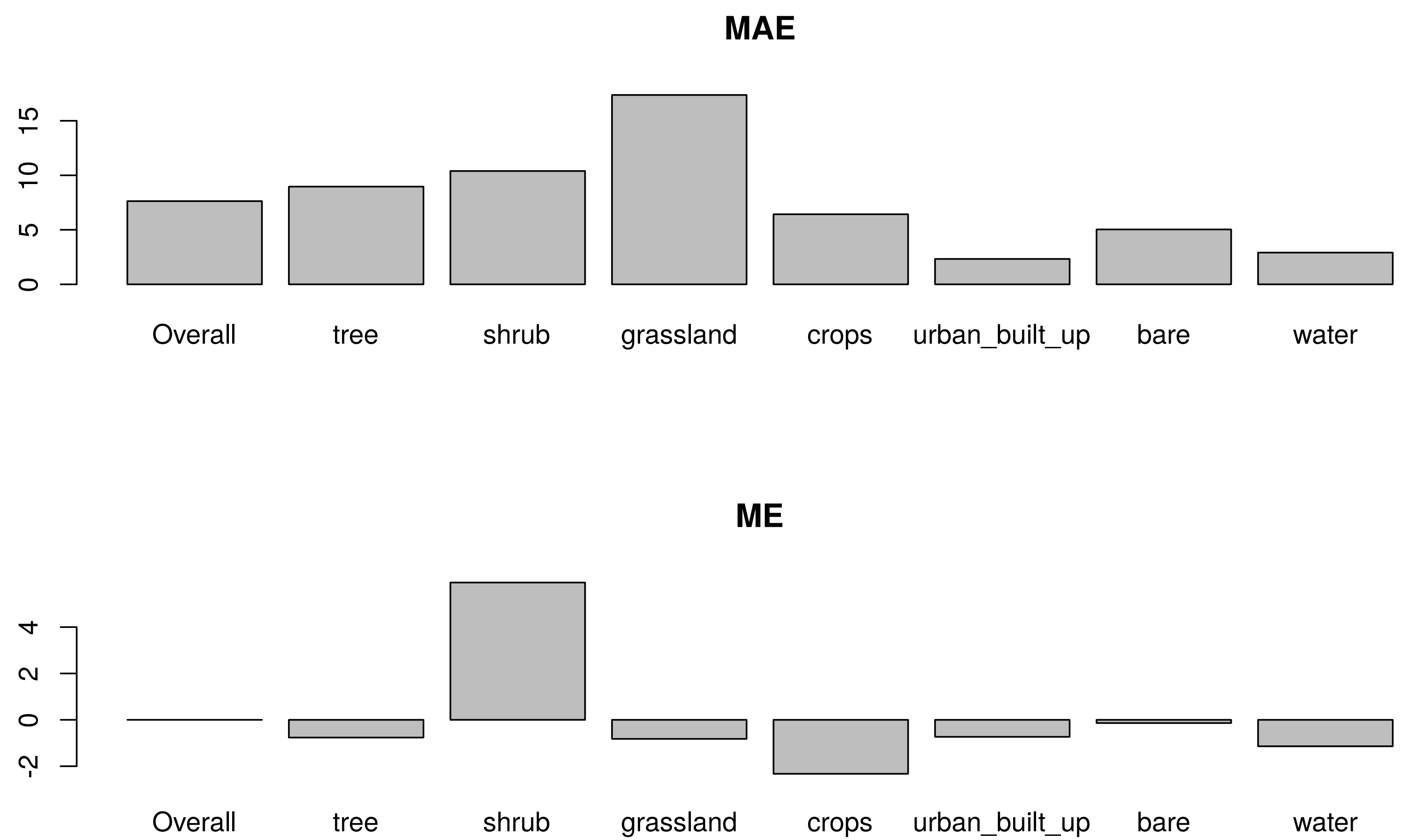


WAGENINGEN
UNIVERSITY & RESEARCH

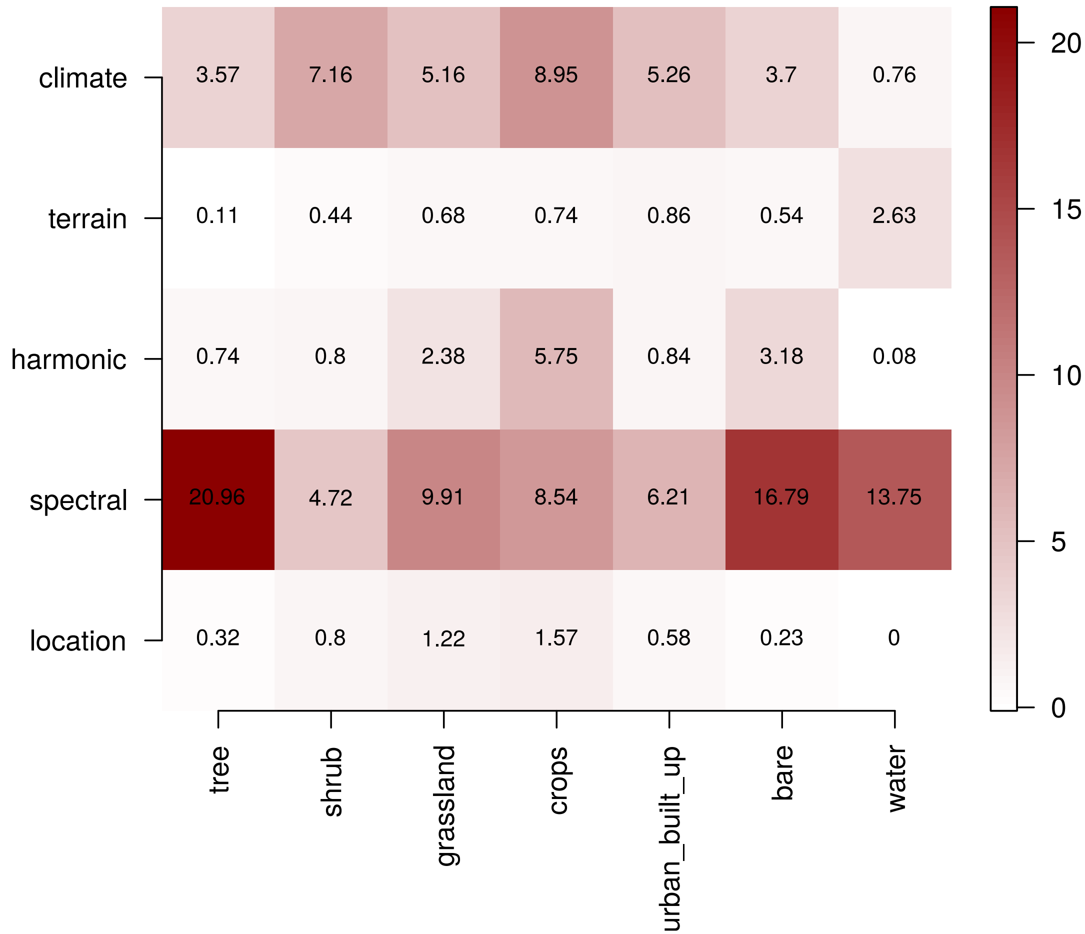
Results

- **Model performance:**
  - **Best**: Random Forest regression with a 2-step model (MAE 7.9%, overall accuracy 72%±2)
  - Logistic regression trained on hardened data: surprisingly good considering (MAE 9.8%, OA 66%±4)
  - Intercept model: MAE 21.7%, OA 25%±4
- Two-model method improves MAE and OA, but hurts RMSE and R^2

Errors per class, Random Forest, two models



Random Forest covariate permutation importance



Discussion, next steps

- **Three-model** approach: one model to determine whether the pixel is **pure**, one for **regression** (if no), one for **classification** (if yes)
- Covariate imbalance: 2 location, 6 terrain, 10 harmonic, 14 spectral, **103** climate and ~**130** soil covariates
- Model optimisation
- Upscaling to the whole world
- Producing wall-to-wall raster predictions

Conclusion

- A two-step model helps improve underestimation of large fractions, at the cost of more erroneous zero predictions
- Random Forest regression with a two-step model performs the best
- Spectral covariates are overall most important, but it varies per class

Thank you for your attention!