Fractional land cover mapping of Africa using machine learning techniques on Proba-V image time series

Dainius Masiliūnas[1], *Nandin-Erdene Tsendbazar*[1], *Martin Herold*[1], *Myroslava Lesiv*[2], *Jan Verbesselt*[1]

[1] *Laboratory of Geo-Information Science and Remote Sensing, Wageningen University*

[2] *International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria*

WAGENINGEN UNIVERSITY & RESEARCH

IIASA

Fractional land cover mapping

- Traditional moderate-resolution 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

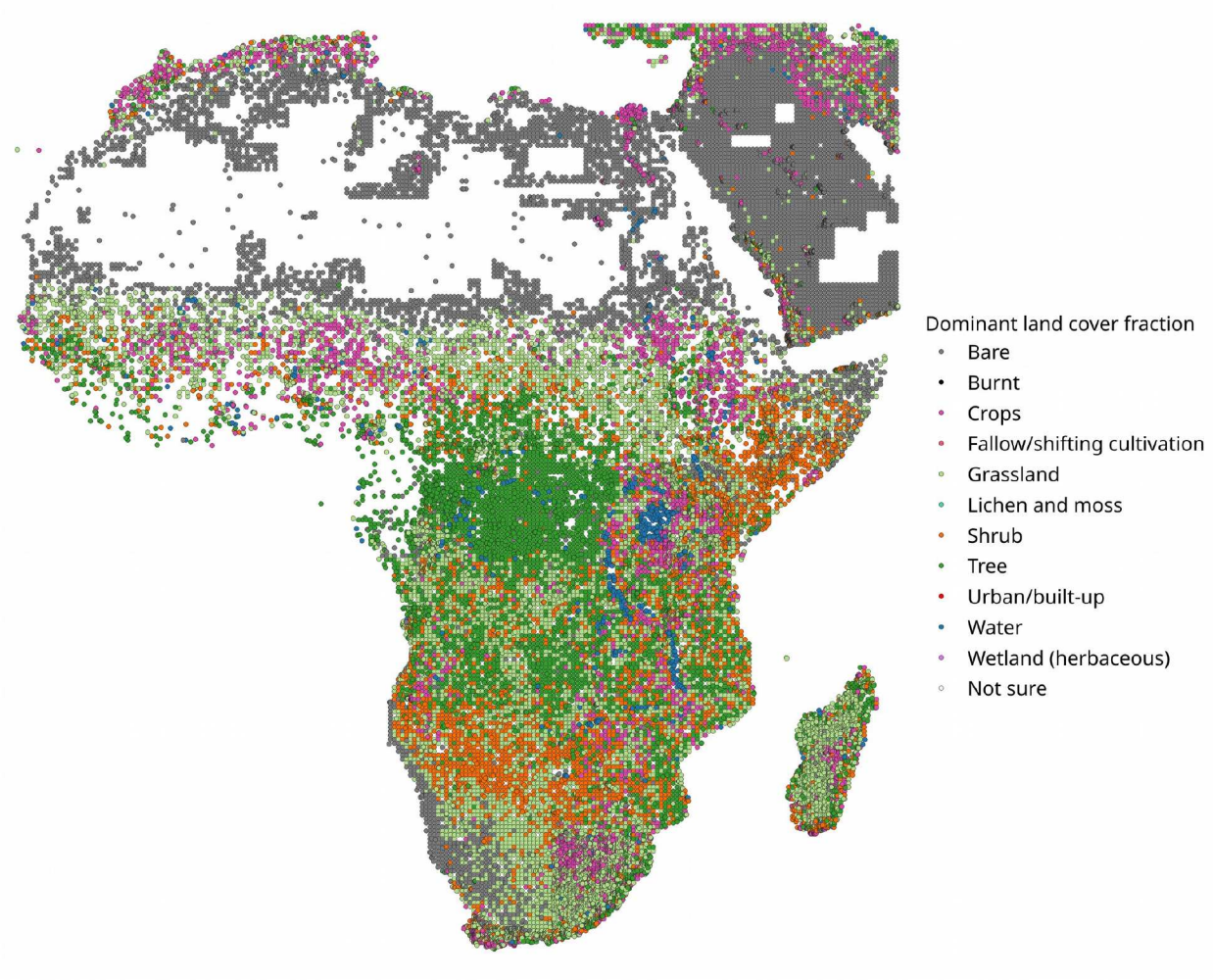
- Develop **methodology** for dealing with **fractional** training data
- **Compare** machine learning regression **algorithm performance** in fractional LC mapping
- Determine which **covariates** are **most important** for fractional LC mapping

Methodology

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

Reference data: 31256 training + 3600 validation points (collected for CGLS-LC100)

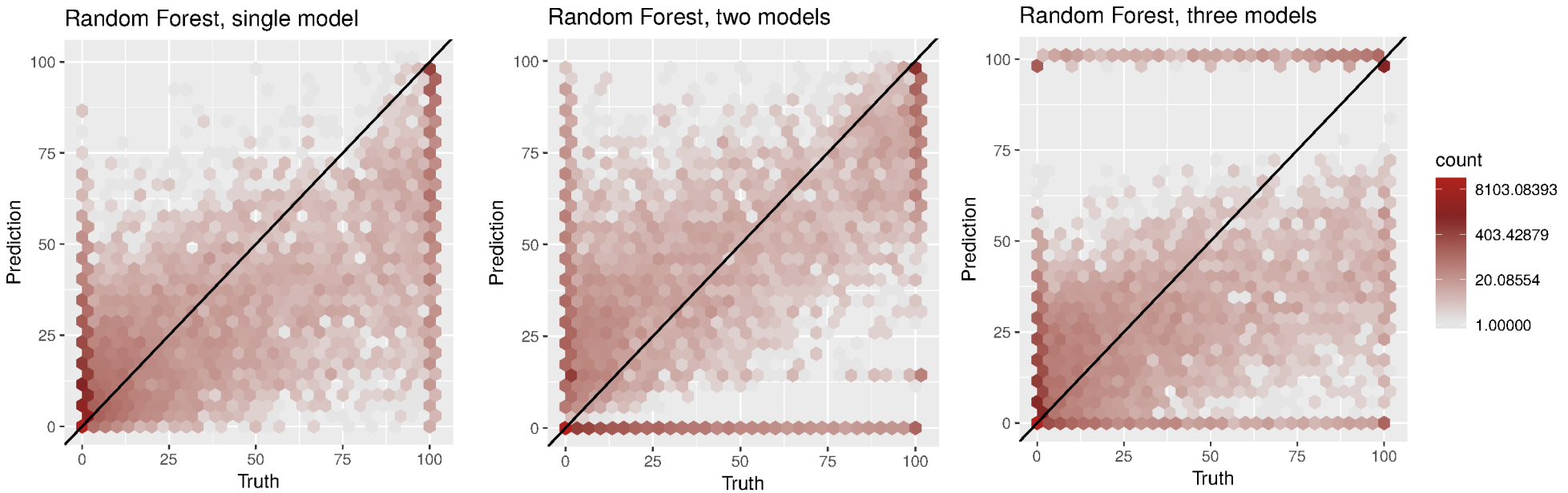
Global Land Operations



WAGENINGEN UNIVERSITY & RESEARCH

Multimodel method

- Fractional, so training data **imbalanced** towards 0. Tested potential solutions:
  - Two models: one to classify **zeroes**, one for **non-zeroes**
  - Three models: one to classify whether it is **pure/non-pure**, one **classification** and one **regression**
  - Histogram matching

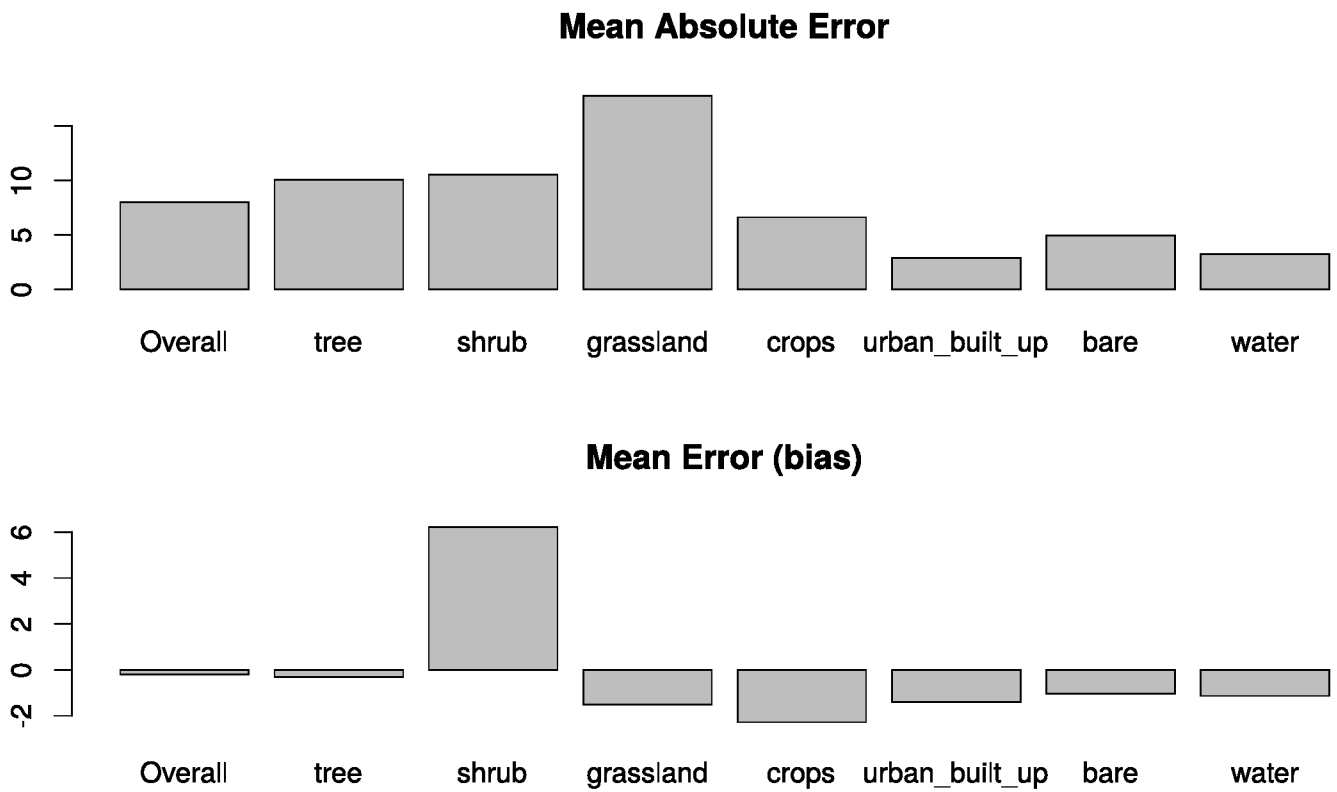


Results

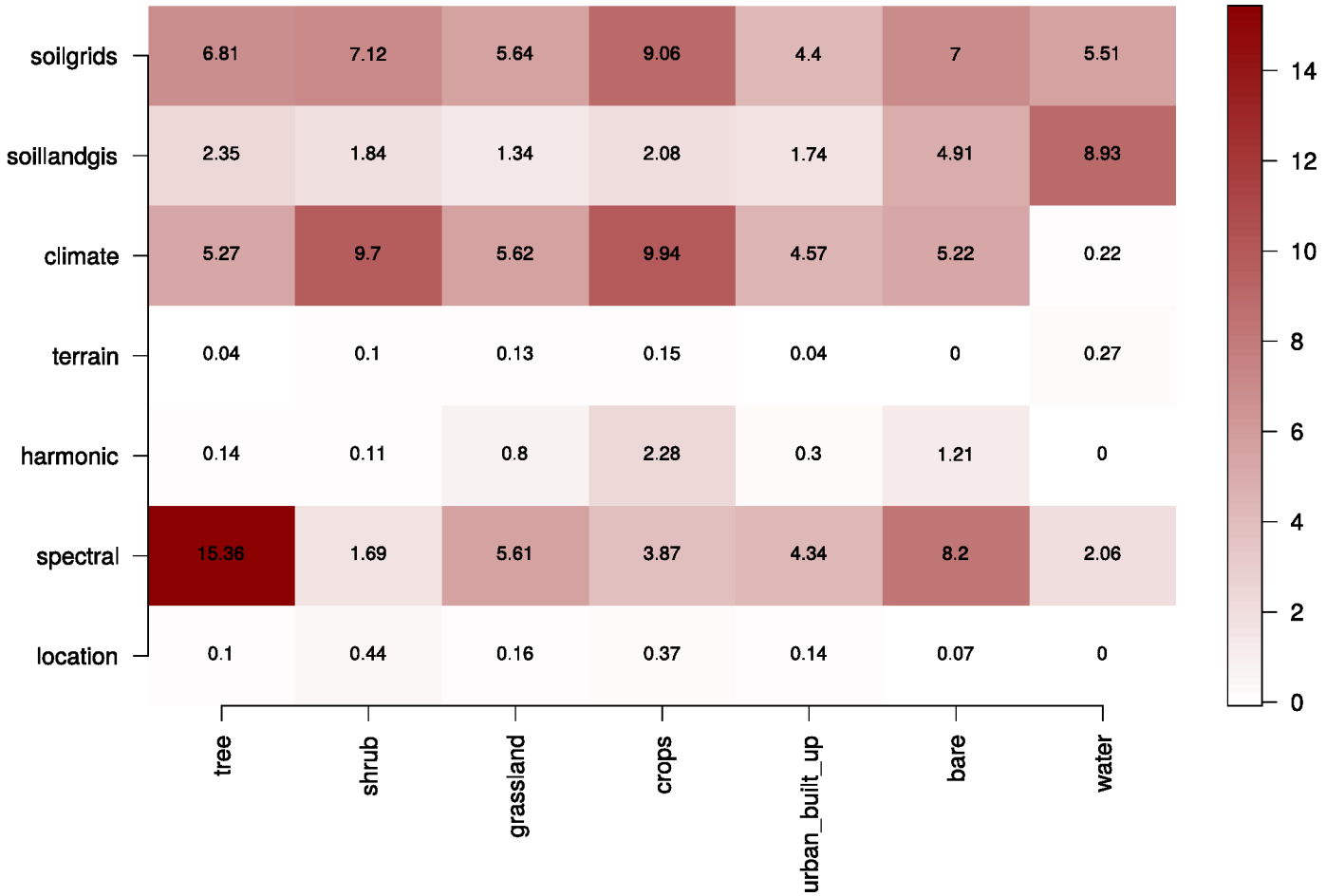
Model	Root Mean Squared Error (%)	Mean Absolute Error (%)	Overall Accuracy (%)	Nash-Sutcliffe Efficiency
RF single model	**16.6**	9.2	68±4	**0.67**
RF two models	18.3	**8.0**	**72±2**	0.59
RF three models	18.6	8.2	71±4	0.59
Logistic regression	19.7	10.7	63±5	0.54
Intercept only	30.7	21.8	24±4	0

- Multi-model methods improve MAE and OA, but hurt RMSE and R^2
- Histogram matching does not help
- Logistic regression trained on hard data (dominant LC) and 15 covariates performs surprisingly well

Errors per class, Random Forest, two models



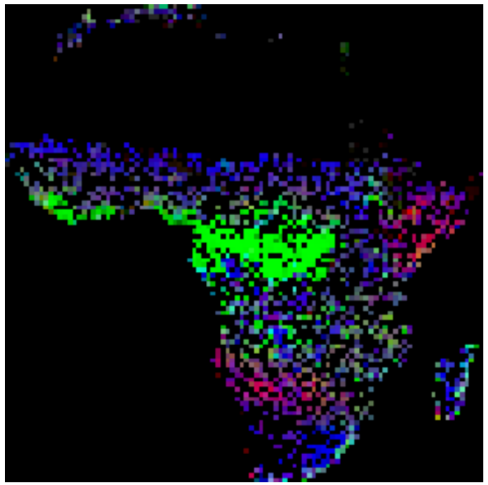
Random Forest covariate permutation importance (decrease in RMSE when permuting all covariates in the group)



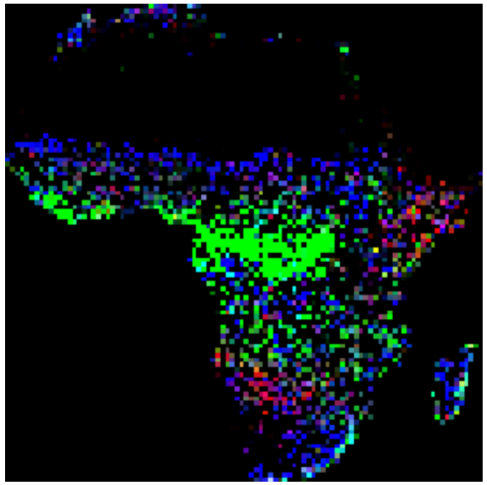
Discussion, next steps

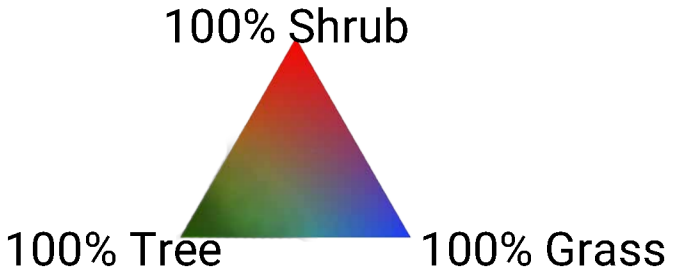
- Covariate imbalance: 3 location, 6 terrain, 10 harmonic, 14 spectral, **112** climate and **168** soil covariates: 313 total
  - Manually removed covariates correlated > 0.9 to get 86 total
- Model optimisation: what else could help model closer to 1:1 (increase NSE)
- Upscaling to the whole world
- Producing wall-to-wall raster predictions

Random Forest, two models



Validation data





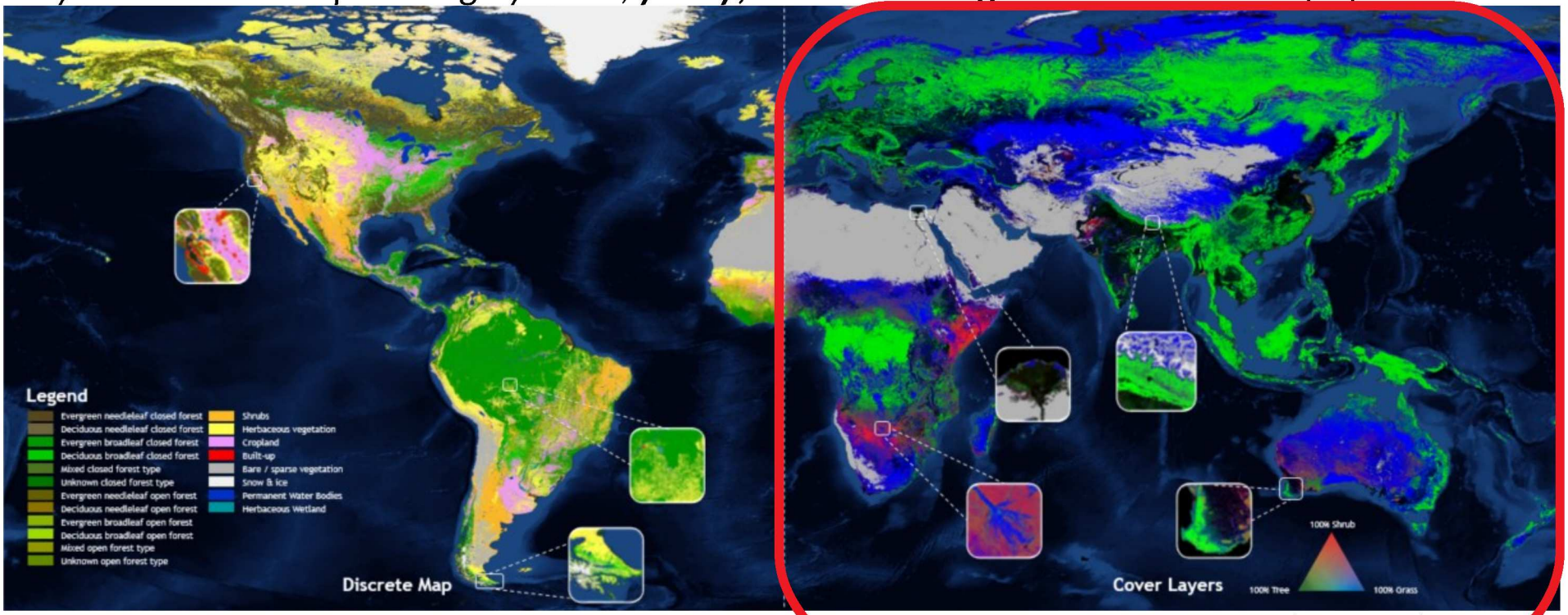
Conclusion

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

Land cover fractions (CGLS-LC100)

Global Land Operations

*A systematic **service** providing dynamic, **yearly, user-oriented global** land cover maps from 2015*

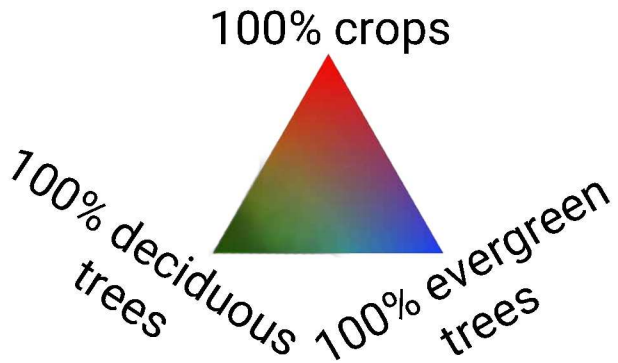


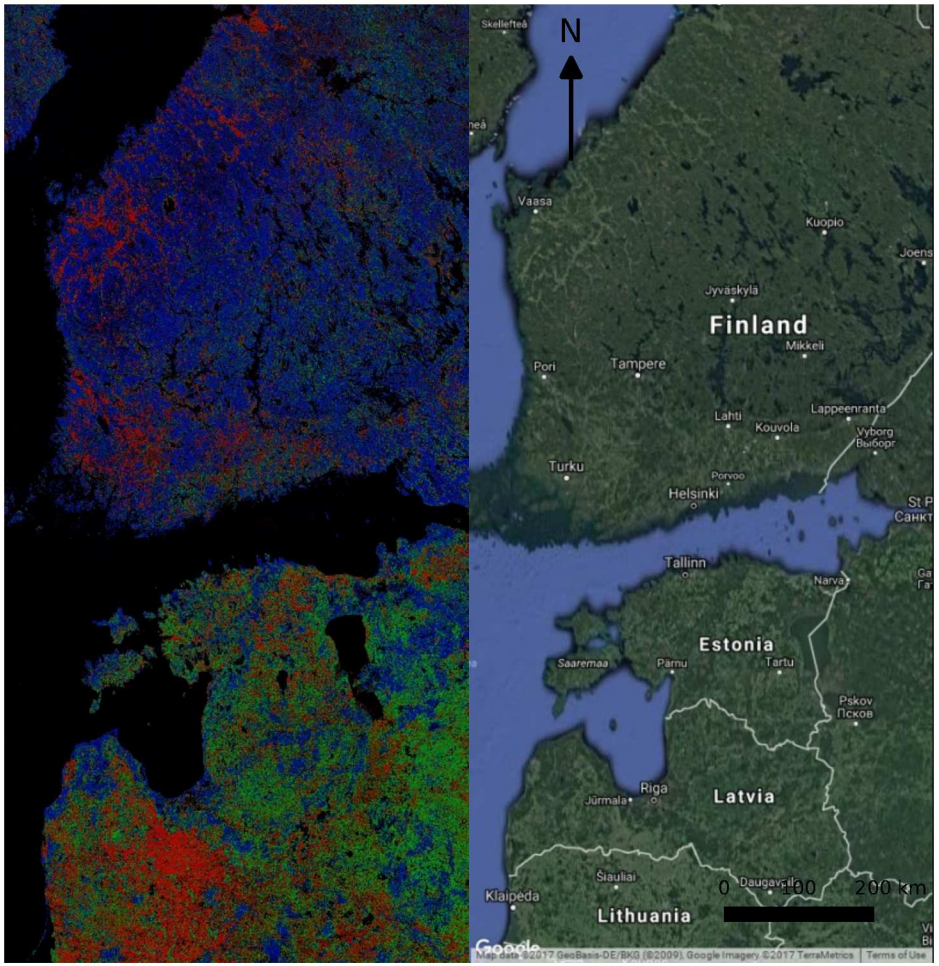
Discrete Map (21 classes)

10 Continuous Covers (0-100%)

Permanent water is derived from GSW (Pekel et al.)
Built-up is derived from WSF (Marconcini et al.)

Thank you for your attention!





We'd like to thank the Copernicus Global Land Services VITO team for ideas and suggestions:

Bruno Smets
Marcel Buchhorn
Ruben van de Kerchhove