Spatial prediction of soil parameters using machine learning methods

Krzysztof Dyba

krzdyb@amu.edu.pl

Adam Mickiewicz University in Poznań

The National Centre for Research and Development (POIR 04.01.02-00-0110/17-00)

- „Sustainable management of the productivity of agricultural crops using satellite images, based on personalized GIS systems available at a dedicated portal”.
- Contractors:
  - Department of Soil Science and Remote Sensing of Soils, Adam Mickiewicz University,
  - Department of Agricultural Chemistry and Environmental Biogeochemistry, Poznan University of Life Sciences,
  - Asseco Poland S.A.

Issues

- Prediction of the selected soil parameter in places not covered by measurements (data change from discrete to continuous form).
- Delimitation of homogeneous zones on the field (zonal mapping).
- Integration of various data sources.
- Basic statistical methods are insufficient in the face of complex and non-linear environmental processes.

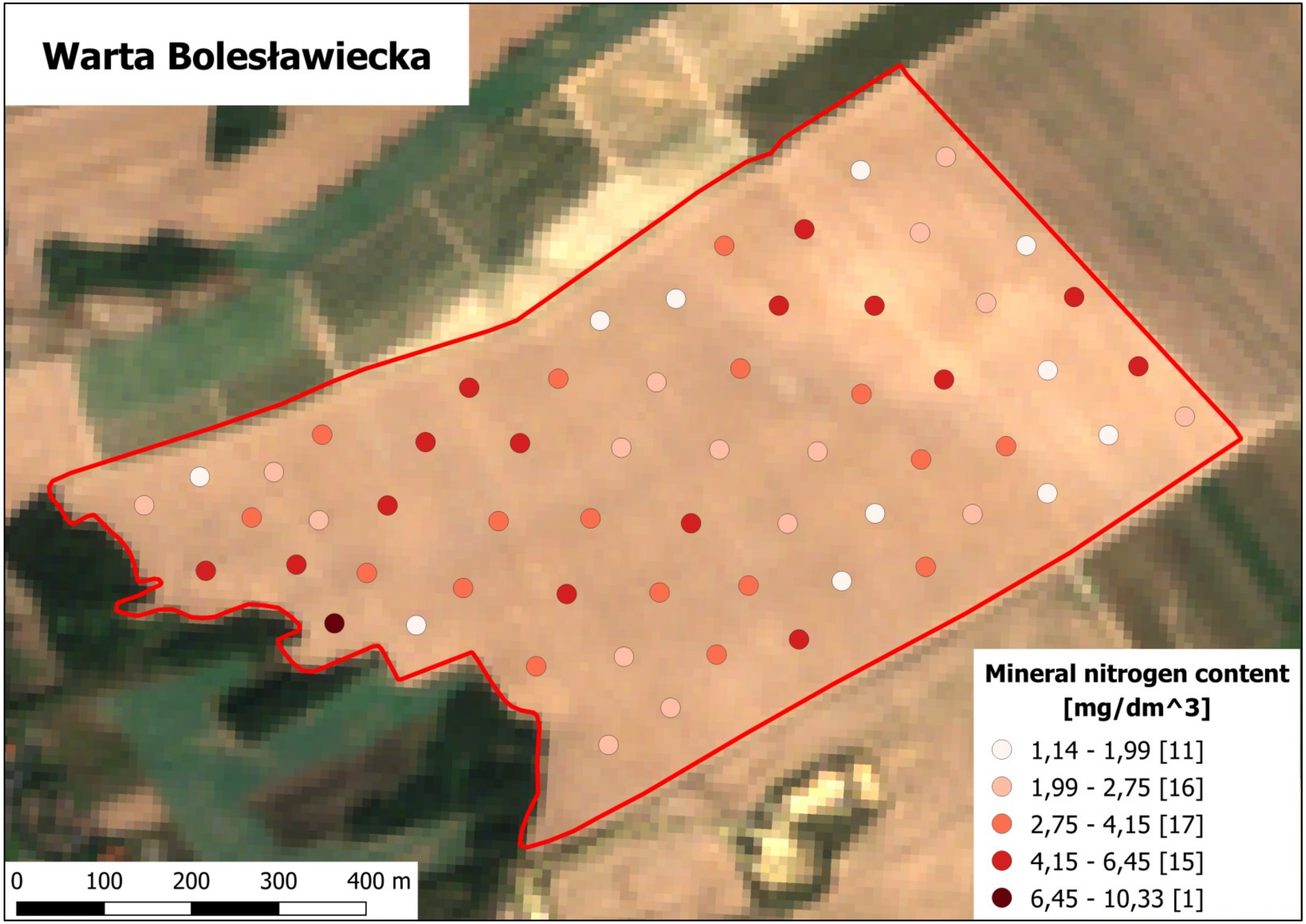
Examples of basic predictive models

- **Inverse distance weighting** - a linearly weighted combination of values, where the weight is the inverse distance function. The influence of the variable decreases as the distance increases.
- **Natural neighbor** – Voronoi diagram is created, and then the values are interpolated with weights proportional to the size of defined areas.

Examples of basic predictive models

- **Polynomian of degree n** – polynomial adjustment to the input data causes trend specification and smoothing of values.
  A higher degree of polynomial means greater complexity, and not necessarily better results (Runge's phenomenon).
- **Spline** – low order polynomials are used, which reduces oscillation errors that occur with higher degree polynomials.

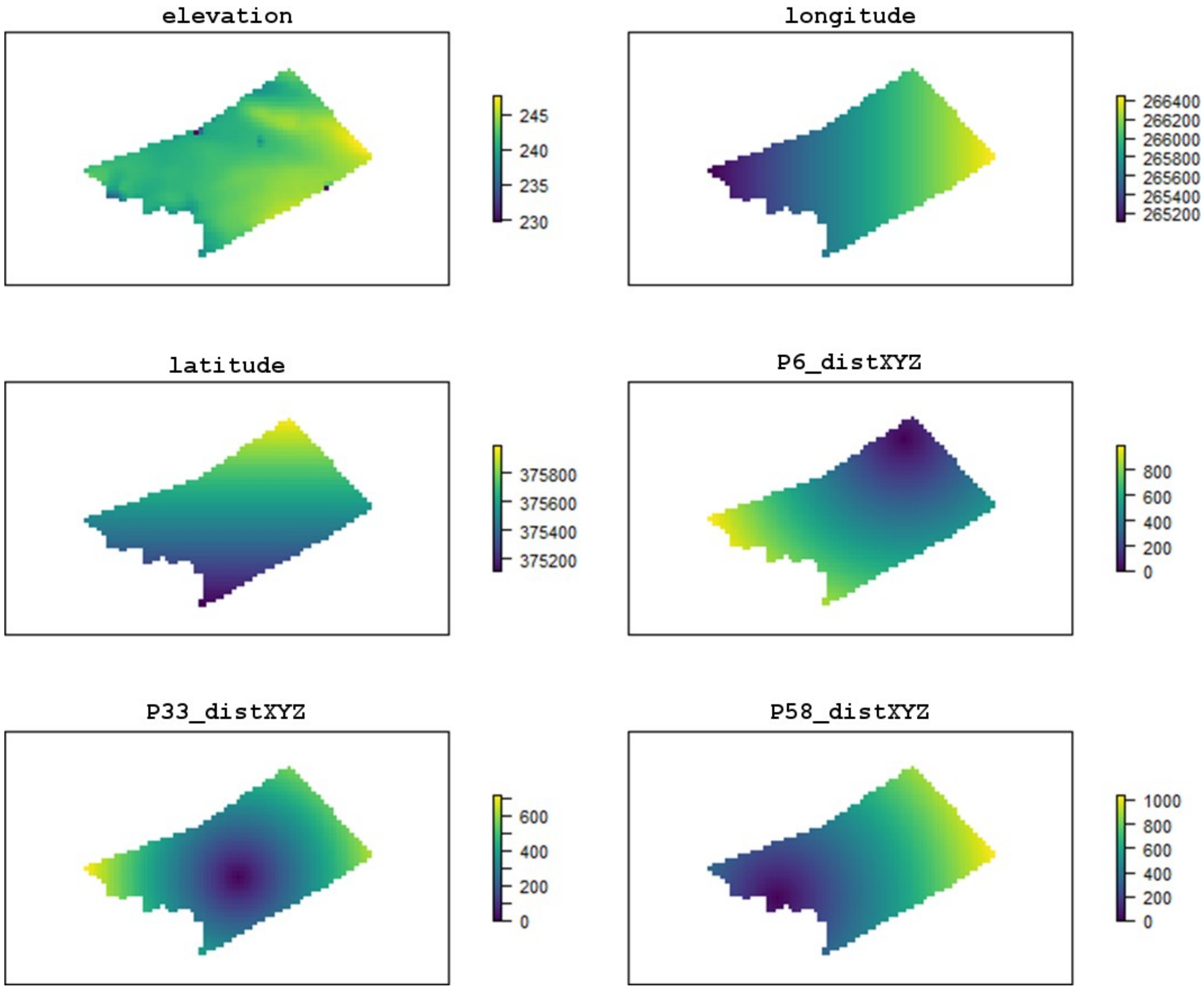
Field survey



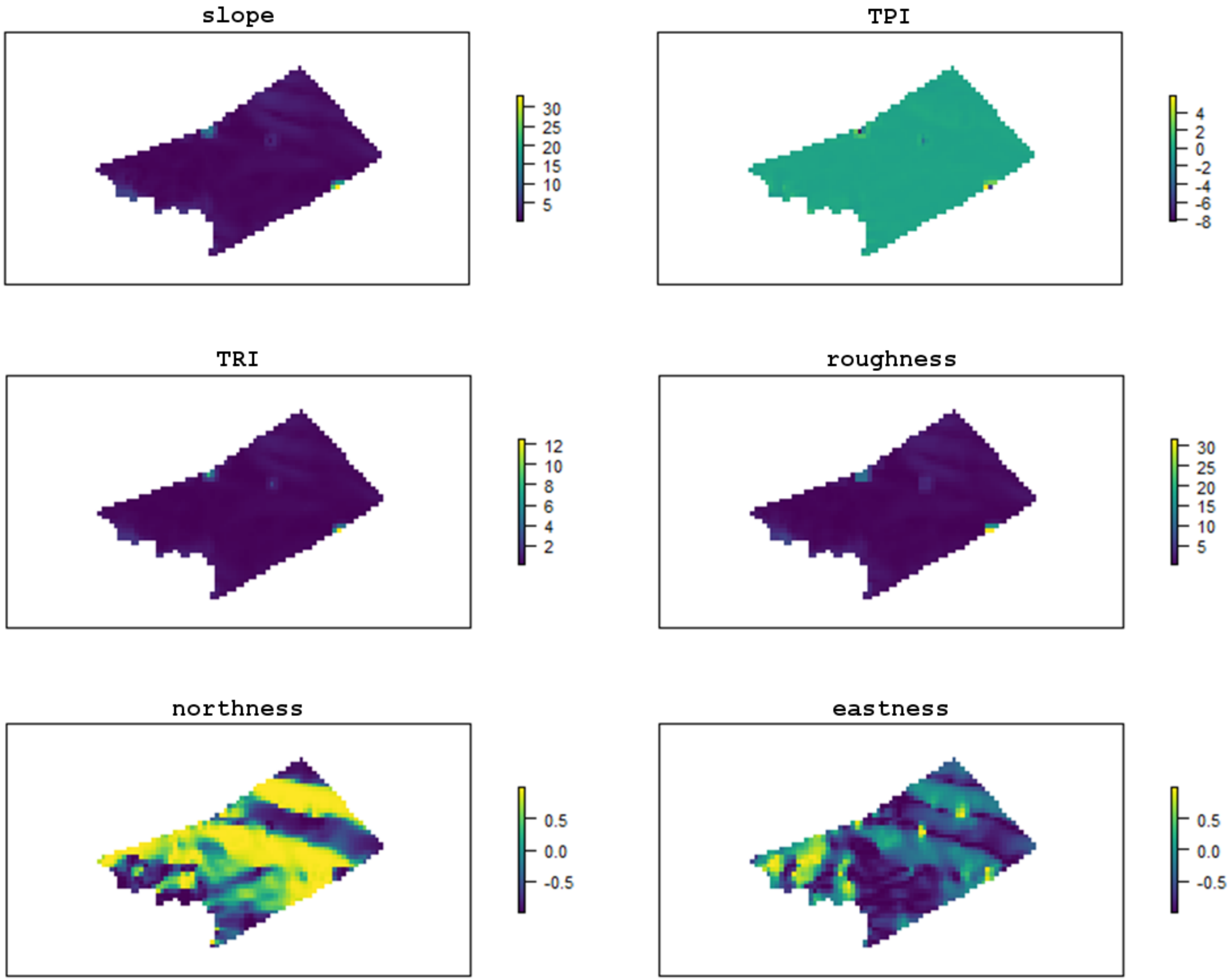
Input data

- **The sum of mineral nitrogen in a 90 cm deep profile [mg/dm^3],**
- Spatial variables:
  - terrain elevation [m ASL],
  - latitude and longitude in the metric reference system [m],
  - three-dimensional Euclidean distance between points [m],
- Geomorphometric variables:
  - slope of the surface [°],
  - Topographic Position Index [-],
  - Terrain Ruggedness Index [-],
  - Terrain Roughness [-],
  - east-west and north-south surface exposition [-]),
- Multispectral satellite image Senitnel 2 (European Space Agency).

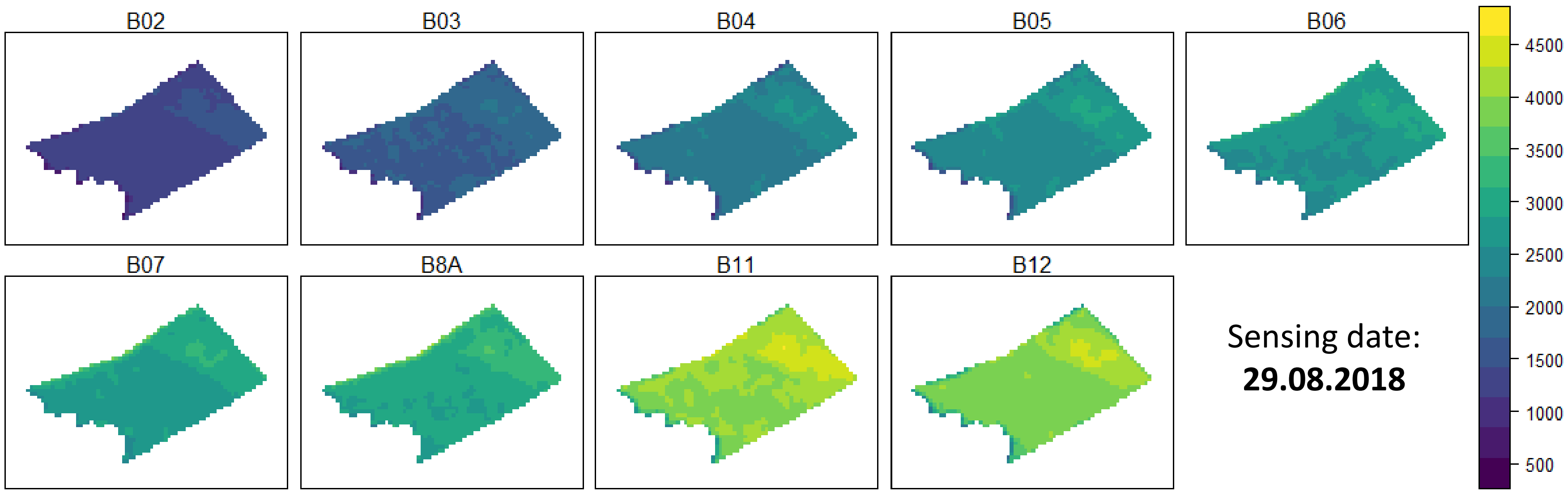
Spatial variables



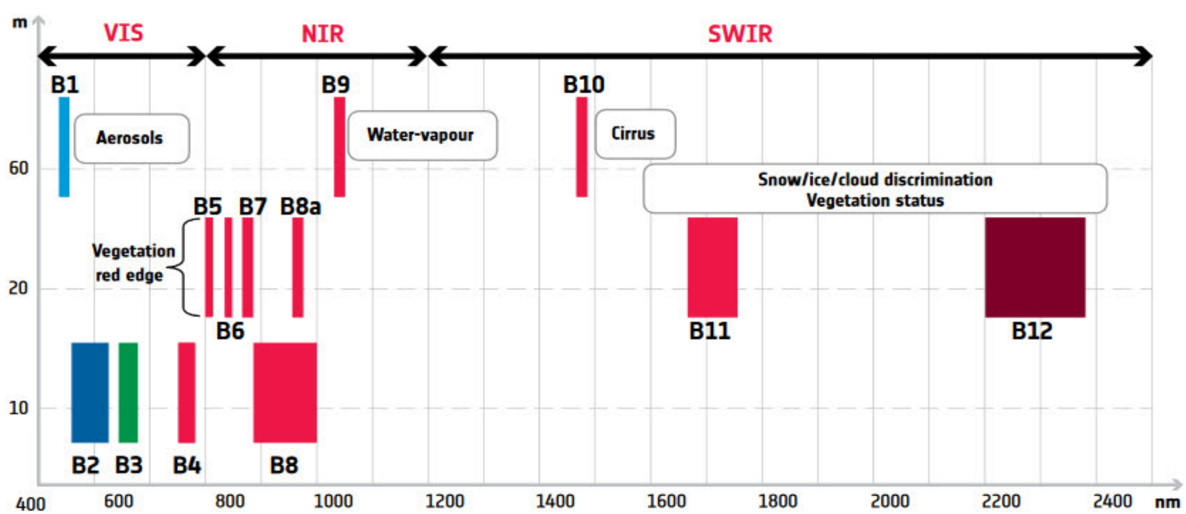
Geomorphometric variables



Multispectral satellite image



Sensing date:
29.08.2018



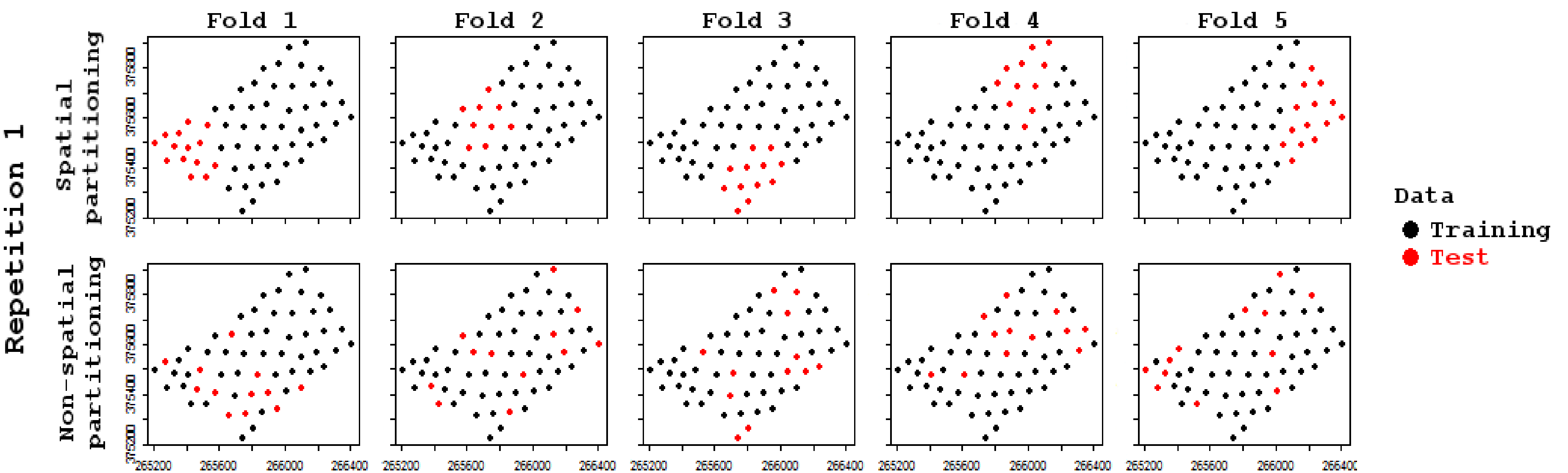
https://directory.eoportal.org/documents/163813/4592192/S2-2019_Auto16.jpeg

Model training

- Model:
  - Random Forest, implementation ranger in R,
  - 78 explanatory variables,
- Optimizing of model parameters in the grid:
  - *„number of trees"*: [100 - 500],
  - *„mtry"* (number of randomly selected candidate variables for node division): [4 - 8],
  - 50 iterations,
- Resampling:
  - repeated spatial cross-validation,
- Measure of performance:
  - Root Mean Squared Error.

ranger: „A Fast Implementation of Random Forests for High Dimensional Data in C++ and R" Wright M. and Ziegler A., 2017. Journal of Statistical Software, 77(1), p. 1-17.

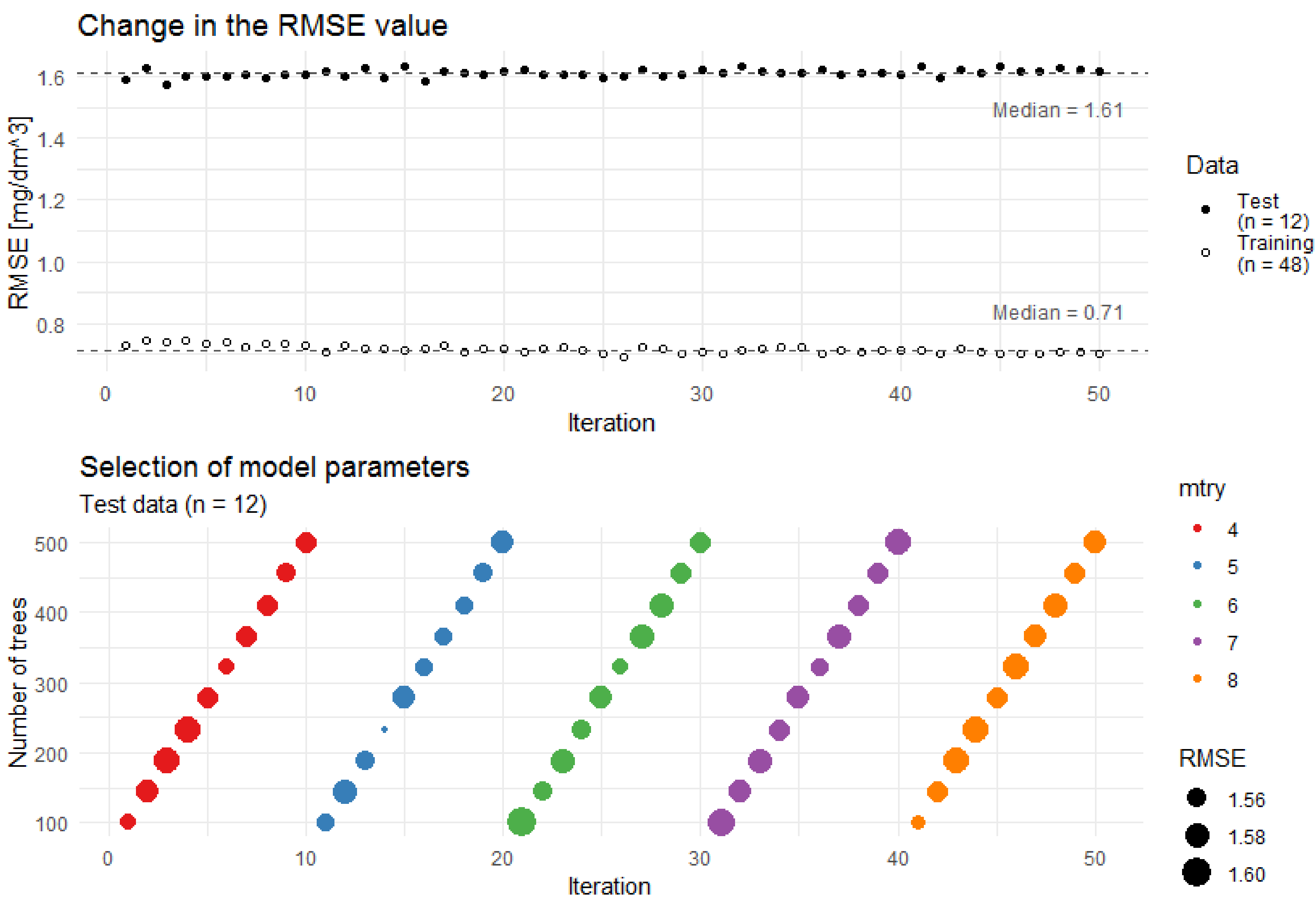
Spatial cross-validation



Due to the spatial aspect of the data, random cross-validation can be ineffective and return too optimistic results. The solution to this problem is spatial cross-validation, which should reduce the bias of the model.

„Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package 'sperrorest'" Brenning A., 2012. IEEE International Symposium on Geoscience and Remote Sensing IGARSS.

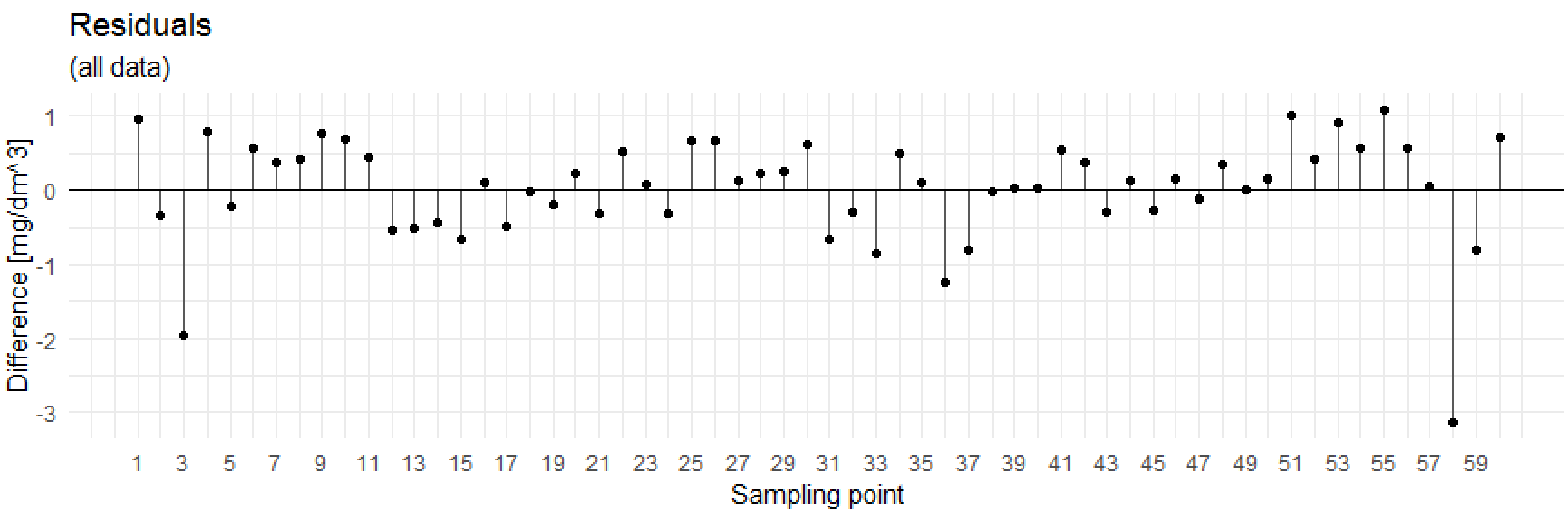
Model training



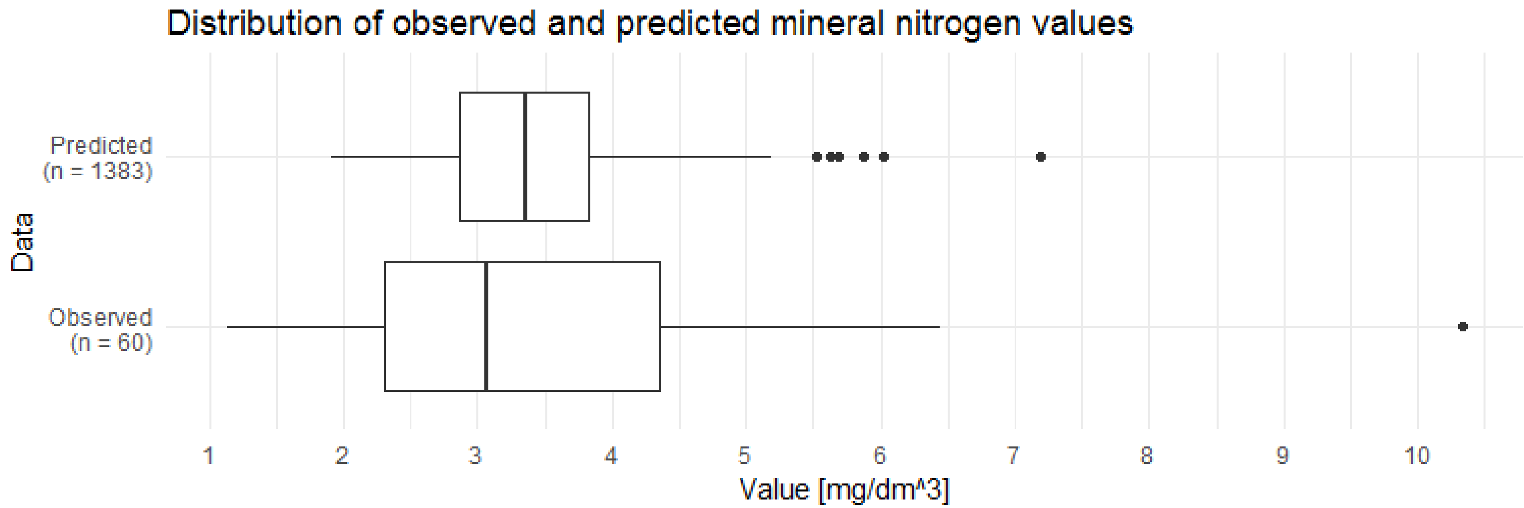
Evaluation

The best model results were obtained for these parameters:
number of trees = 233
mtry = 5

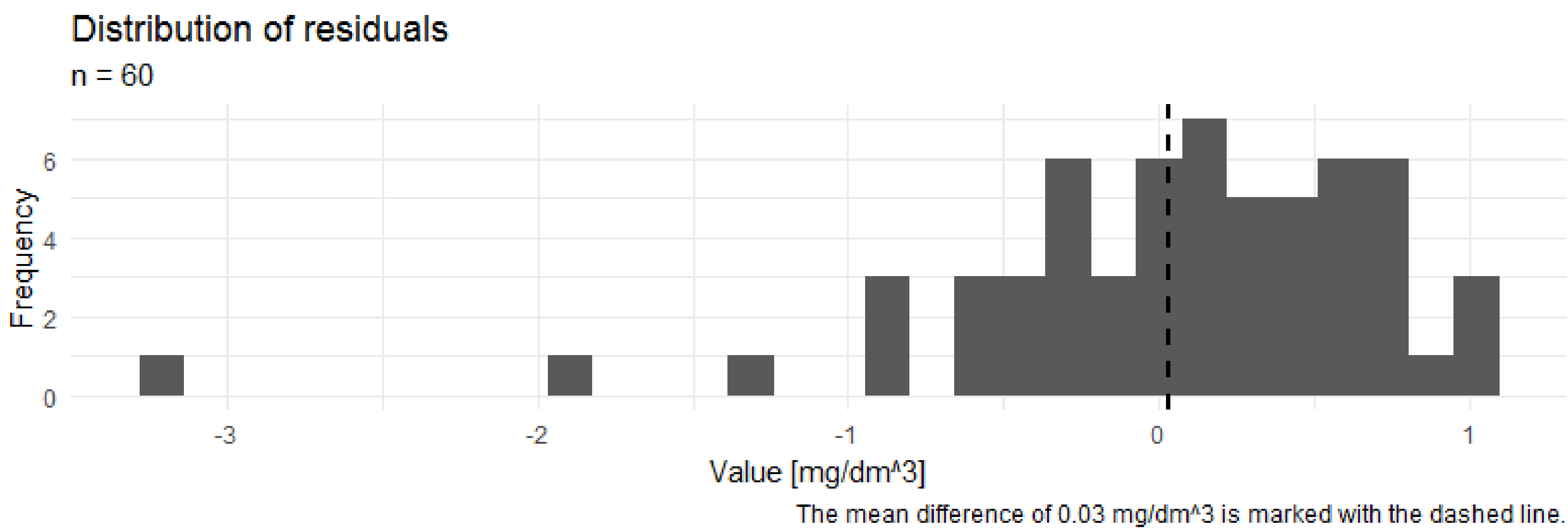
RMSE (test data) $[mg/dm^3]$	RMSE (training data) $[mg/dm^3]$	MAPE (all data) [%]	R^2 (all data) [-]
1,54	0,71	17,87	0,81



Evaluation

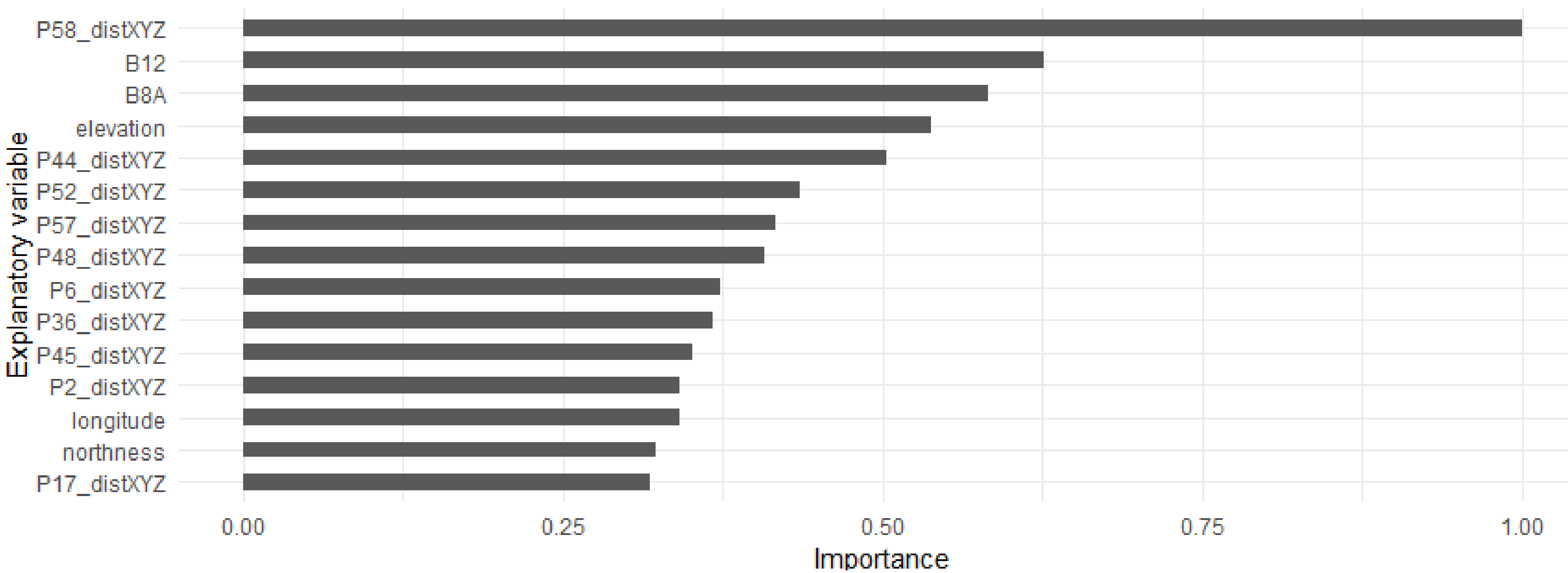


Distribution of observed and predicted mineral nitrogen values
Predicted
(n = 1383)
Observed
(n = 60)
Data
1
2
3
4
5
6
7
8
9
10
Value [mg/dm^3]

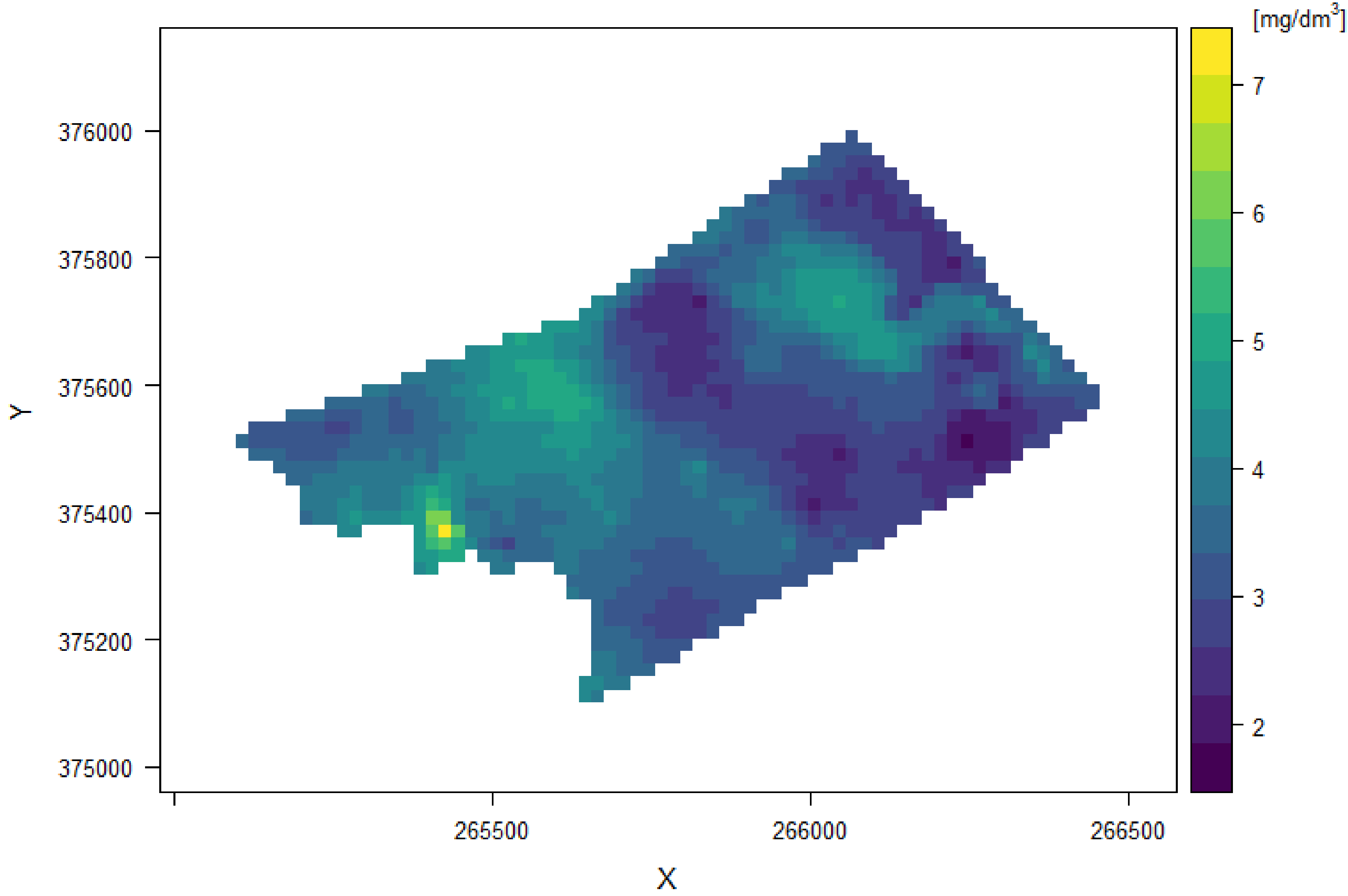


Distribution of residuals
n = 60
Frequency
0
2
4
6
-3
-2
-1
0
1
Value [mg/dm^3]
The mean difference of 0.03 mg/dm^3 is marked with the dashed line.

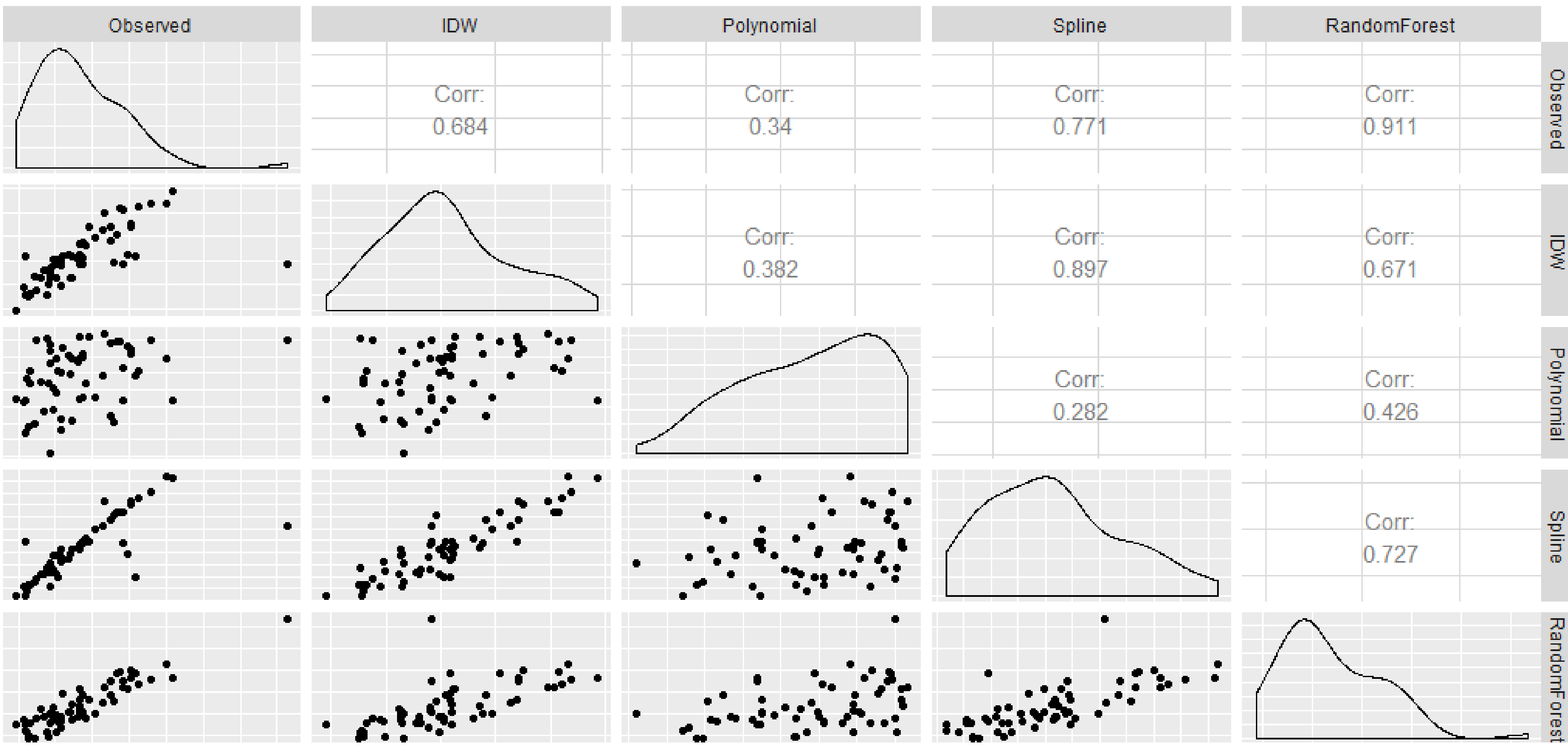
The most important explanatory variables



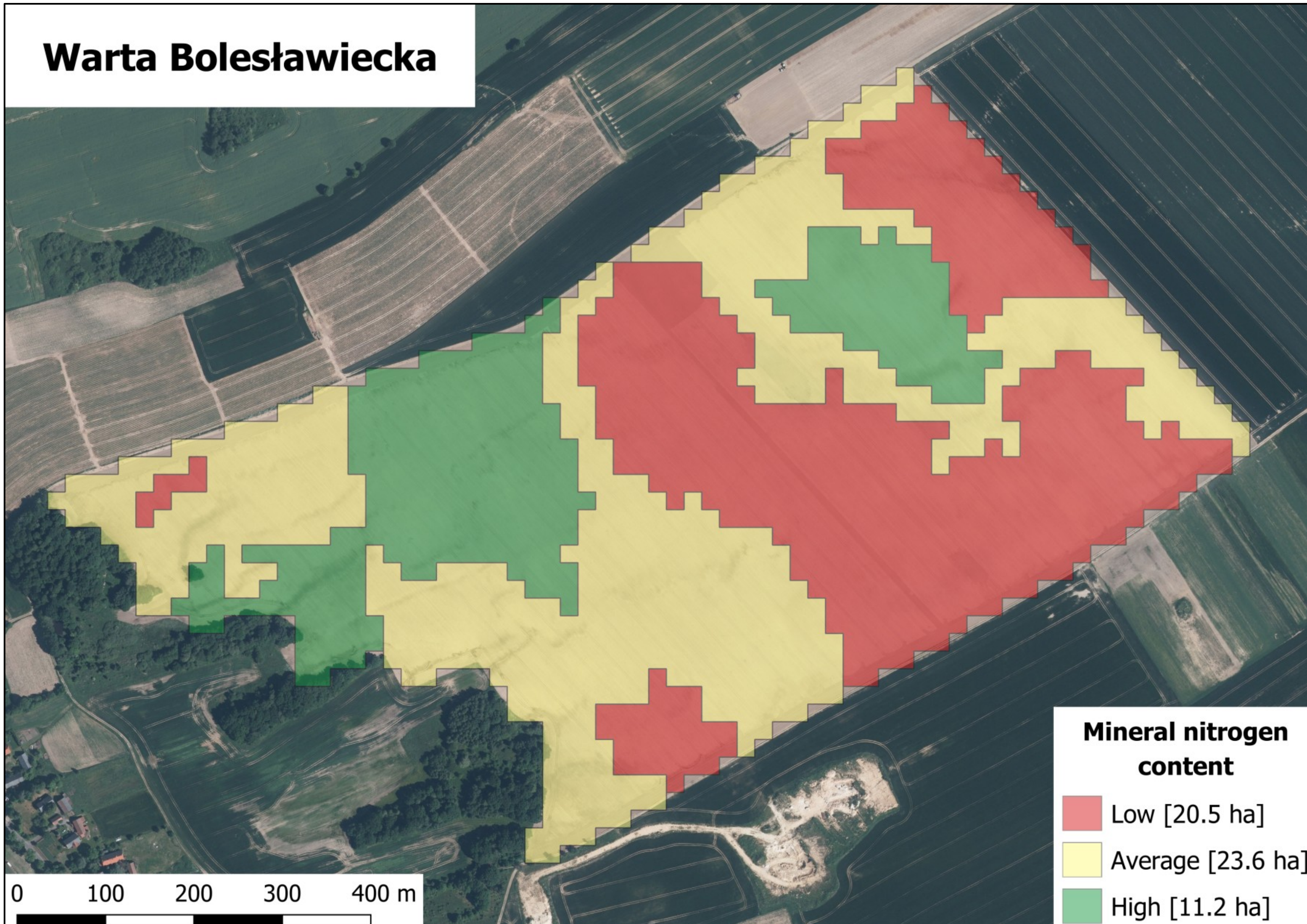
Prediction of mineral nitrogen content



Models comparison



Homogeneous zones on the field



Applications

- Understanding the spatial variability of soil parameters on the field (its most fertile and problematic zones).
- Optimization of agrotechnical operations and rational application of fertilizers (financial benefits and reduction of pressure on the natural environment).
- Source of explanatory variables (information) for crop yield forecasting.

Thank you!