Data Mining in Precision Agriculture

IPMU'2010, Dortmund

Georg Ruß and Alexander Brenning

June 28th, 2010

Precision Agriculture

- GPS technology used in site-specific, sensor-based crop management
- combination of agriculture and information technology
- data-driven approach to agriculture
- lots of data analysis tasks

Data Details – Example Field

Figure: F440 field, depicted on satellite imagery, source: Google Earth

Data Details – Example Sensor

Figure: Yara N-Sensor for vegetation index data collection, source: Agricon GmbH

Data Details – Features

- collect a number of geo-coded, high-resolution features such as:
  - N1, N2, N3: nitrogen fertilizer application rates
  - REIP32, REIP49: vegetation index (red edge inflection point)
  - Yield: winter wheat yield in this year
  - EC25: electrical conductivity of soil, represents information about soil humidity, mineral content, pH value (et al)
- two fields available, 5000/6500 data records in $10 \times 10m$-resolution

Data Details – Temporal Aspects

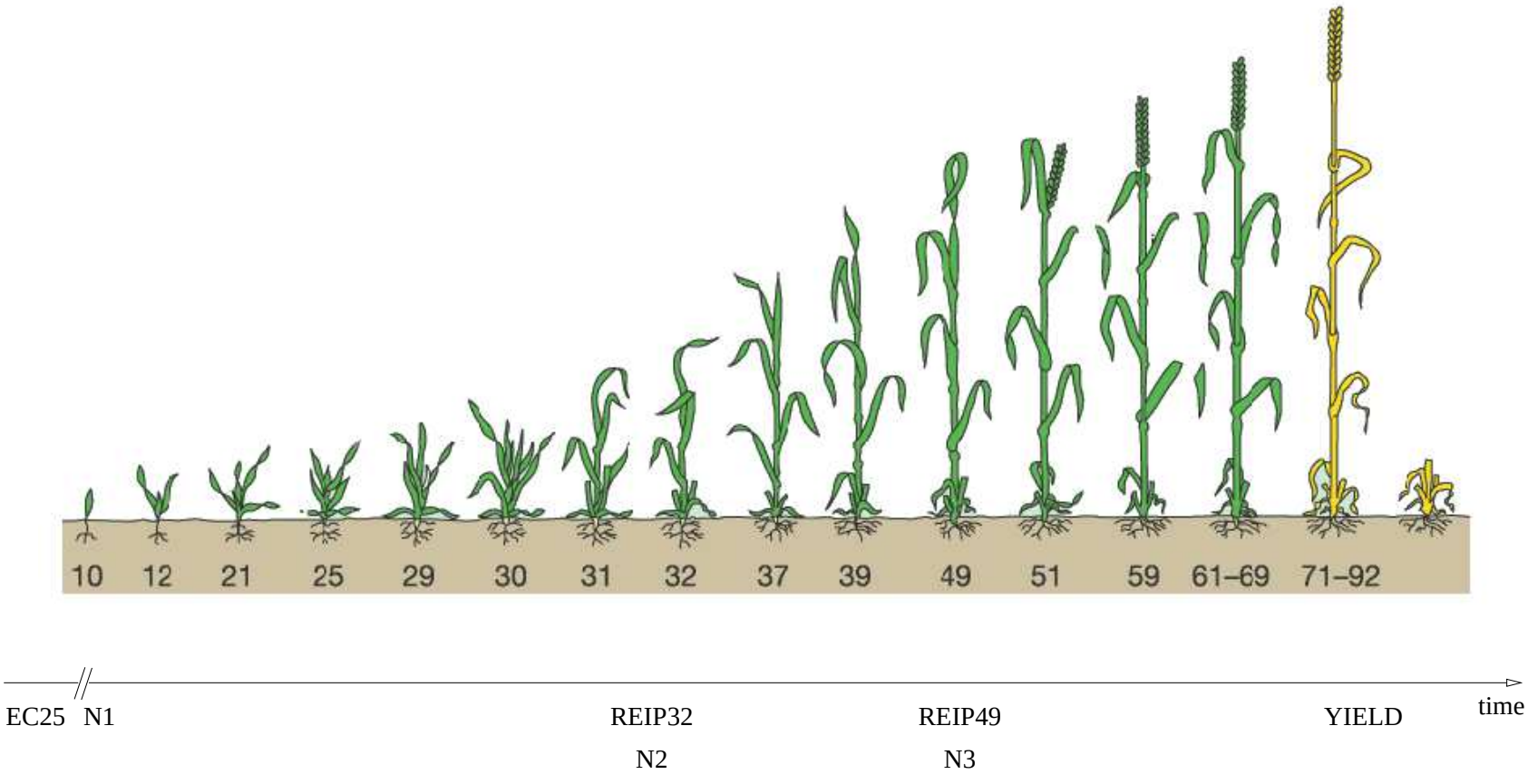


Figure: growing stages of cereals, source: adapted from BBCH

Data Details – Questions

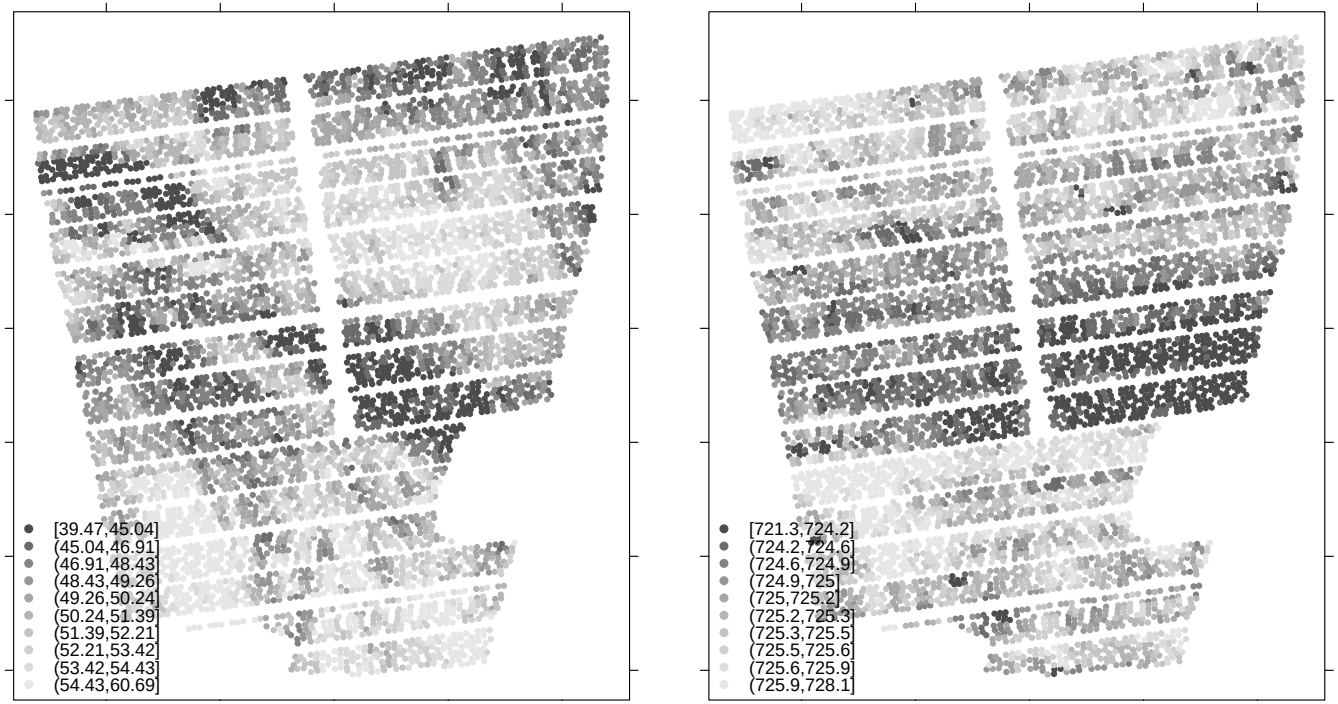
- Can the current year's yield be predicted from the available features?
  - $\rightarrow$ Regression
- We are using spatial, geo-referenced data:
  - $\rightarrow$ *Spatial* Regression

(Spatial) Regression – Basics

- multivariate regression: usually a cross-validation setup
  - divide data into training and test sets
  - train regression model on training set
  - report error on independent (!) test set
- support vector regression (support vector machine)
- random forest, bagging, regression tree (tree-based models)

(Spatial) Regression – Spatial Autocorrelation

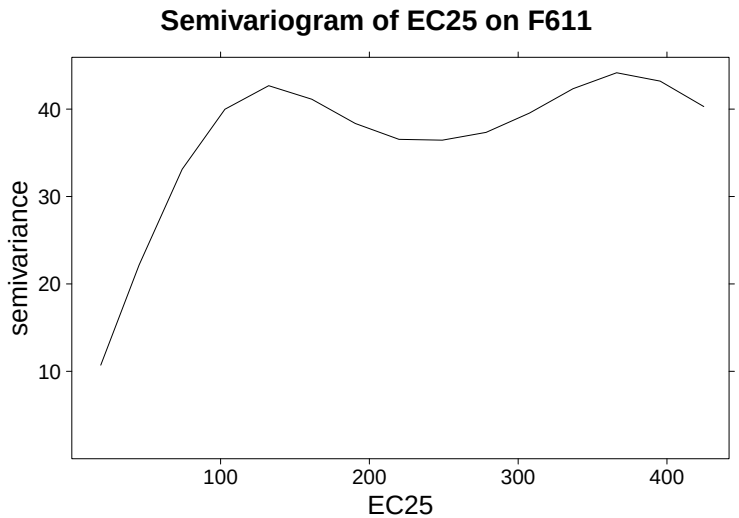
Are (spatial) data records independent of each other?
(Do we have spatial autocorrelation?)



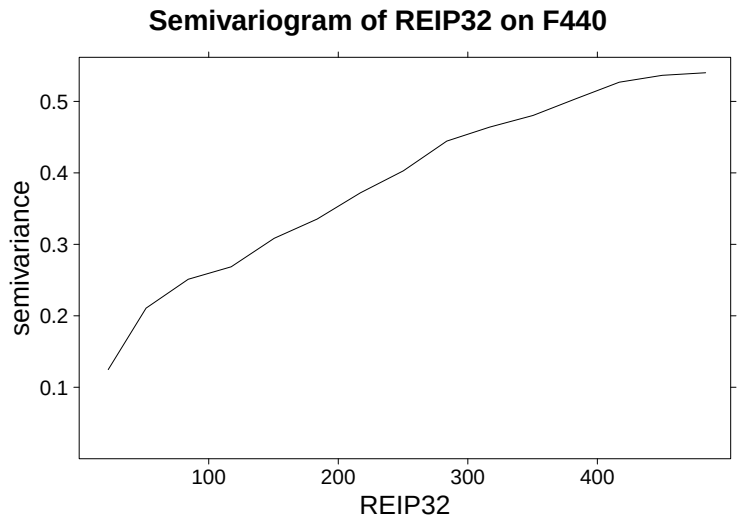
(a) EC25 (b) REIP32

Figure: F440, EC25/REIP32 predictor

(Spatial) Regression – Spatial Autocorrelation



(a) EC25



(b) REIP32

Figure: F440, EC25/REIP32 semivariograms, variance as a function of distance (omnidirectional)

Spatial Regression – Idea

- for spatial data: develop spatial cross-validation approach:
  - *don't* sample test and training sets randomly
  - instead: sample using spatial relationships between records
- idea: subdivide the field into contiguous zones
  - use k-means on the data records' coordinates
  - select training and test sets from this set of zones
  - continue with the (now spatial) standard cross-validation approach

Spatial Regression – Tessellation Figure

Figure: Tessellation of F440 using k-means, $k = 20$ (colors are for illustration only, no further meaning here)

Spatial vs. Non-Spatial Regression – Results: 1st Dataset

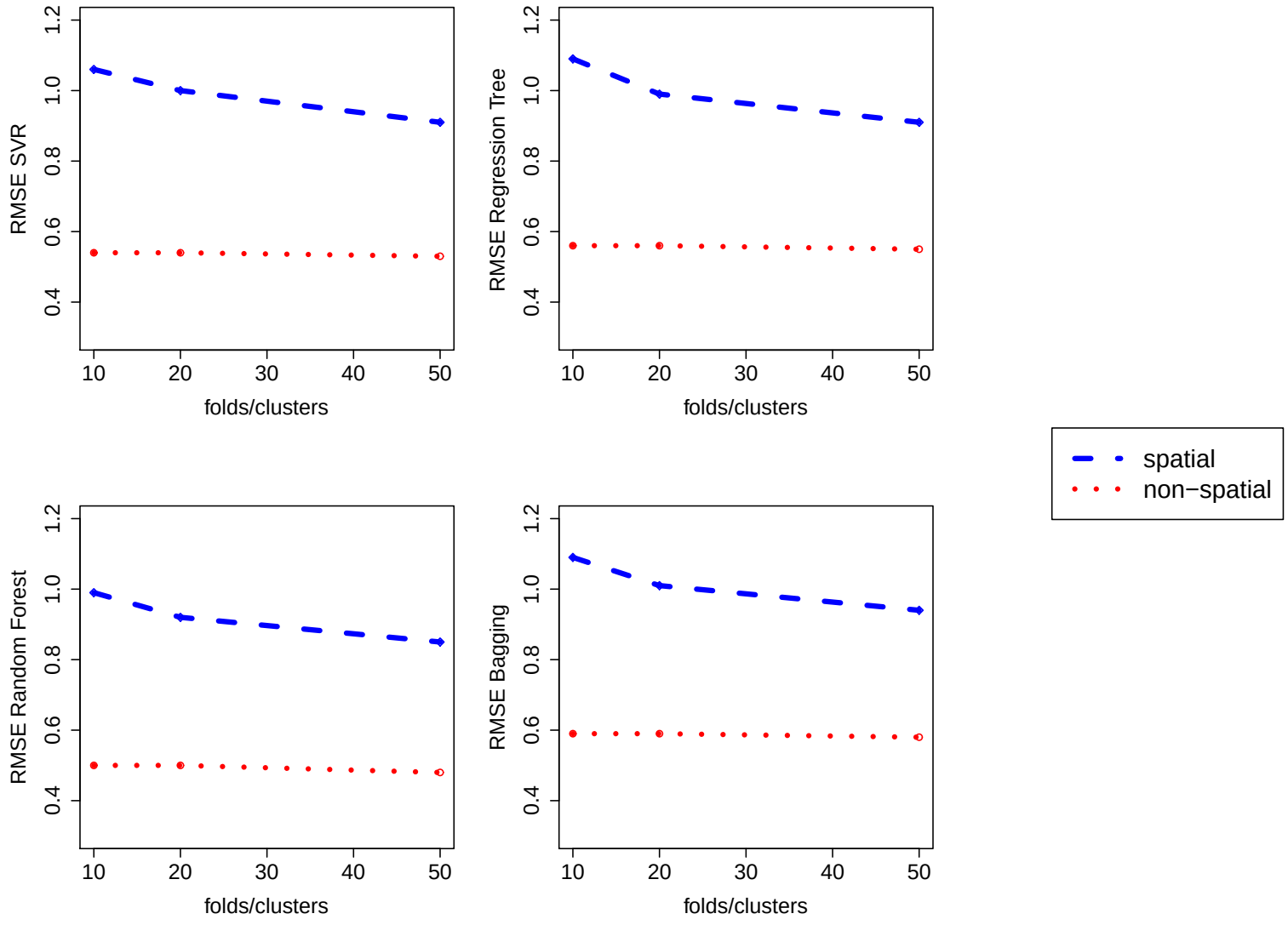


Figure: Results F440, models vs. spatial/non-spatial vs. folds/clusters

Spatial vs. Non-Spatial Regression – Results: 2nd Dataset

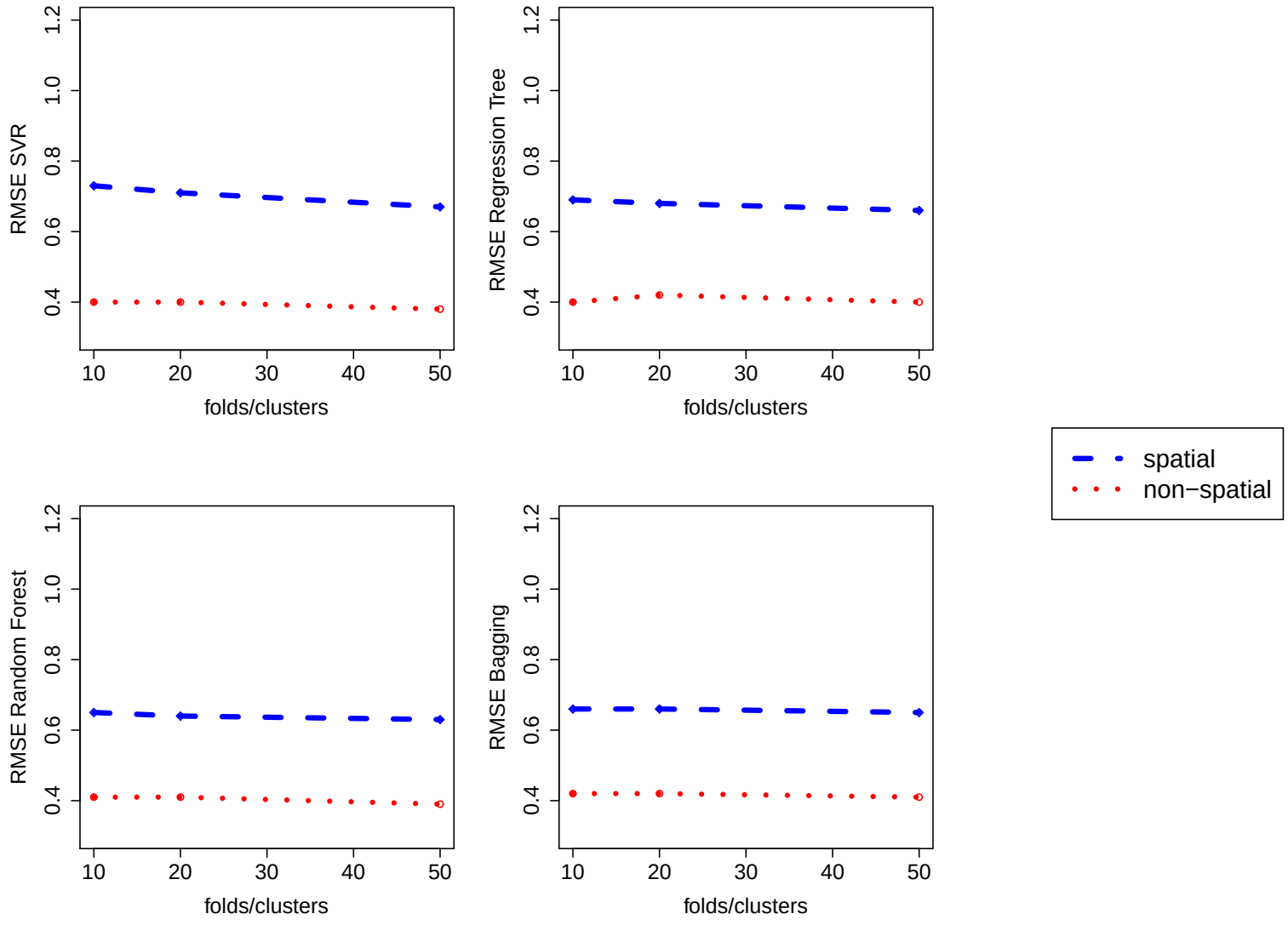


Figure: Results F611, models vs. spatial/non-spatial vs. folds/clusters

Summary

- precision agriculture as a data-driven approach
- spatial, geo-referenced data records in large amounts
- yield prediction solved as spatial cross-validation (regression)
- important difference between spatial and non-spatial data treatment $\Rightarrow$ use models which are fit for spatial tasks

Time for . . .

Questions?

- contact: `georg.russ@ovgu.de`
- slides, R scripts and further info at `http://research.georgruss.de`