Spatial Variable Importance Assessment for Yield Prediction in Precision Agriculture

IDA'2010, Biosphere II, Arizona

Georg Ruß and Alexander Brenning

May 19th-21st, 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

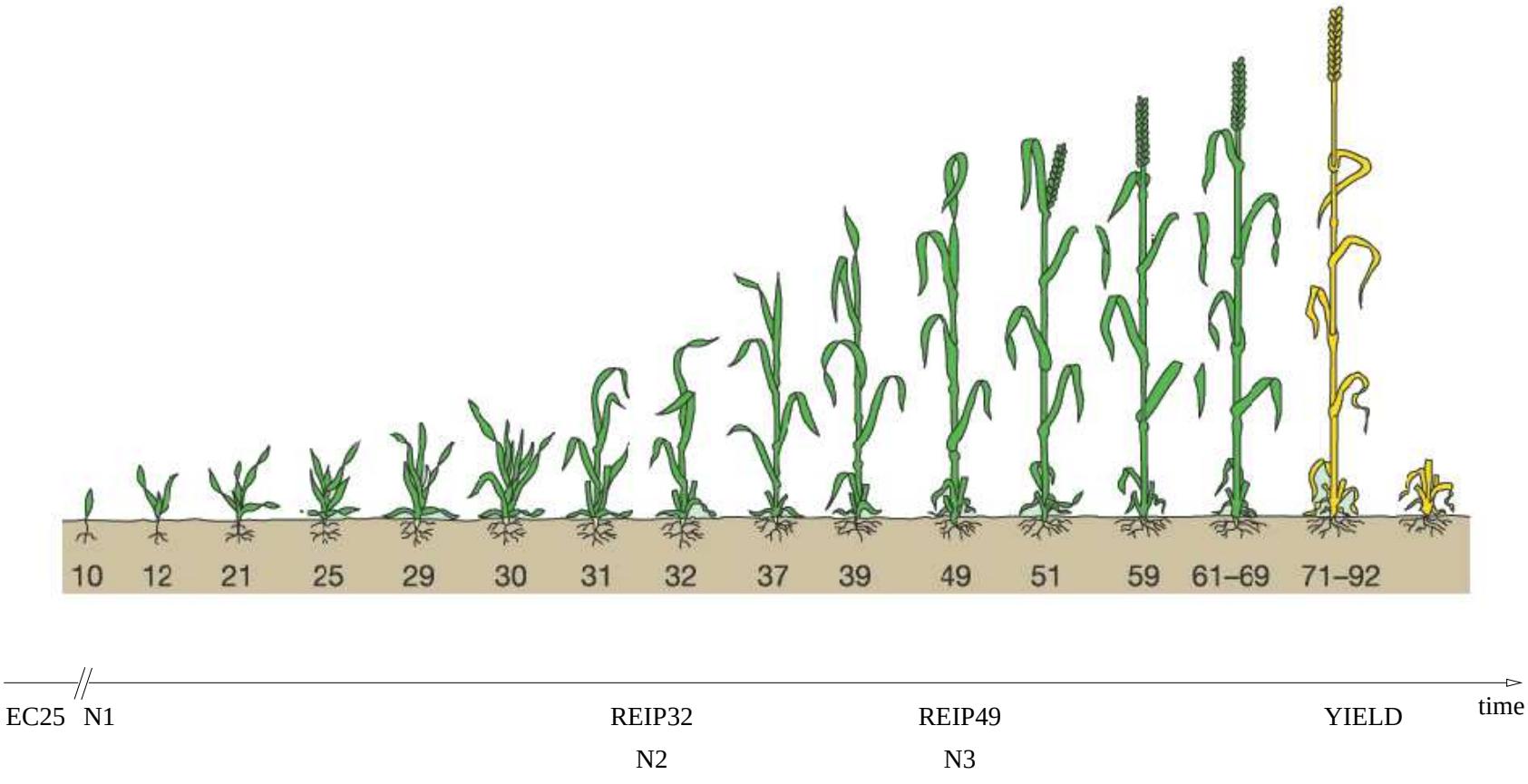


Figure: growing stages of cereals, source: adapted from BBCH

Data Details – Questions

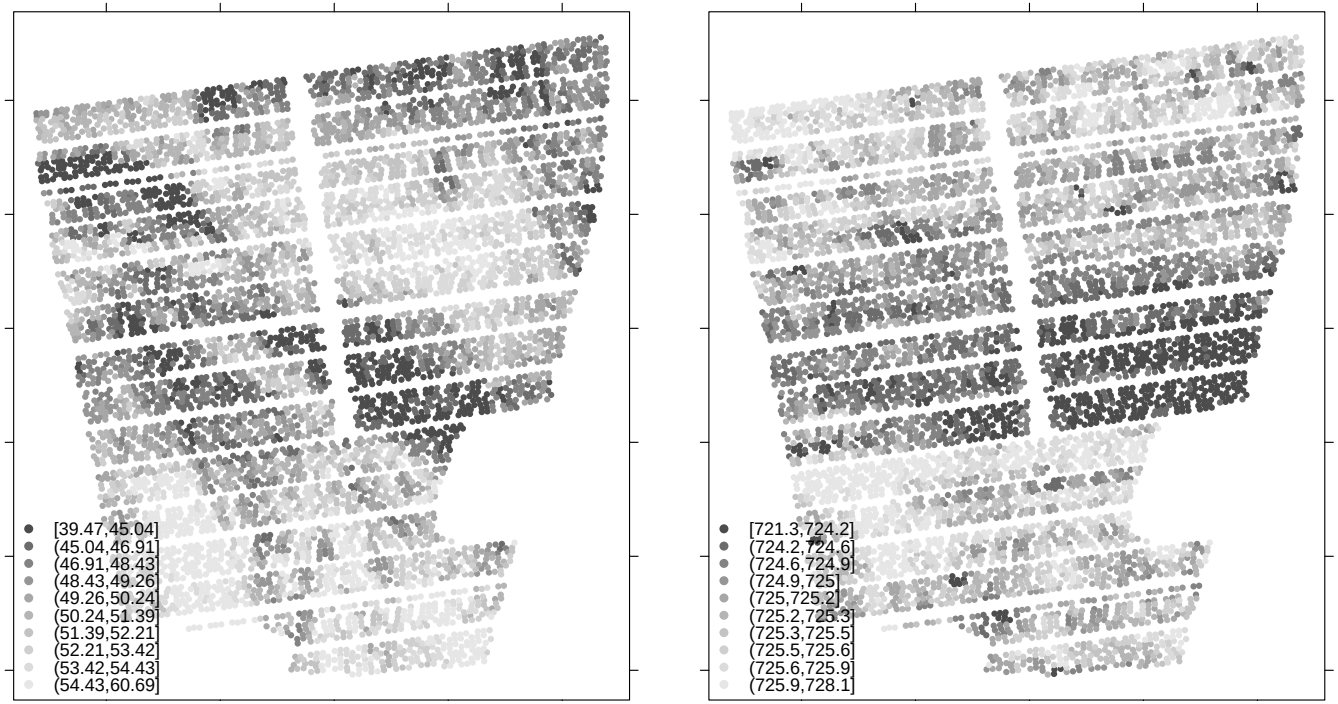
- Can the current year's yield be predicted from the available features?
  - $\rightarrow$ Spatial Regression
- Which of the features are important for the above yield prediction?
  - $\rightarrow$ Spatial Variable Importance

(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
- linear model (usually as a baseline and with linear dependencies in data)
- support vector regression (support vector machine)
- random forest, bagging, regression tree (tree-based models)

(Spatial) Regression – Issue

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



(a) EC25 (b) REIP32

Figure: F440, EC25/REIP32 predictor

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 – Figure

Figure: Tessellation of F440 using k-means, $k = 20$

Spatial Variable Importance – Principle

- new data are collected: decide whether they're useful for yield prediction
- traditionally: feature selection (wrapper/filter approach)
- but: interdependencies among the variables
- novel variable importance approach:
  - choose one variable and permute its values in the test set
  - measure the increase in prediction error on the test set
  - low/high increase: low/high importance (depending on data and model)

Spatial Variable Importance – Results

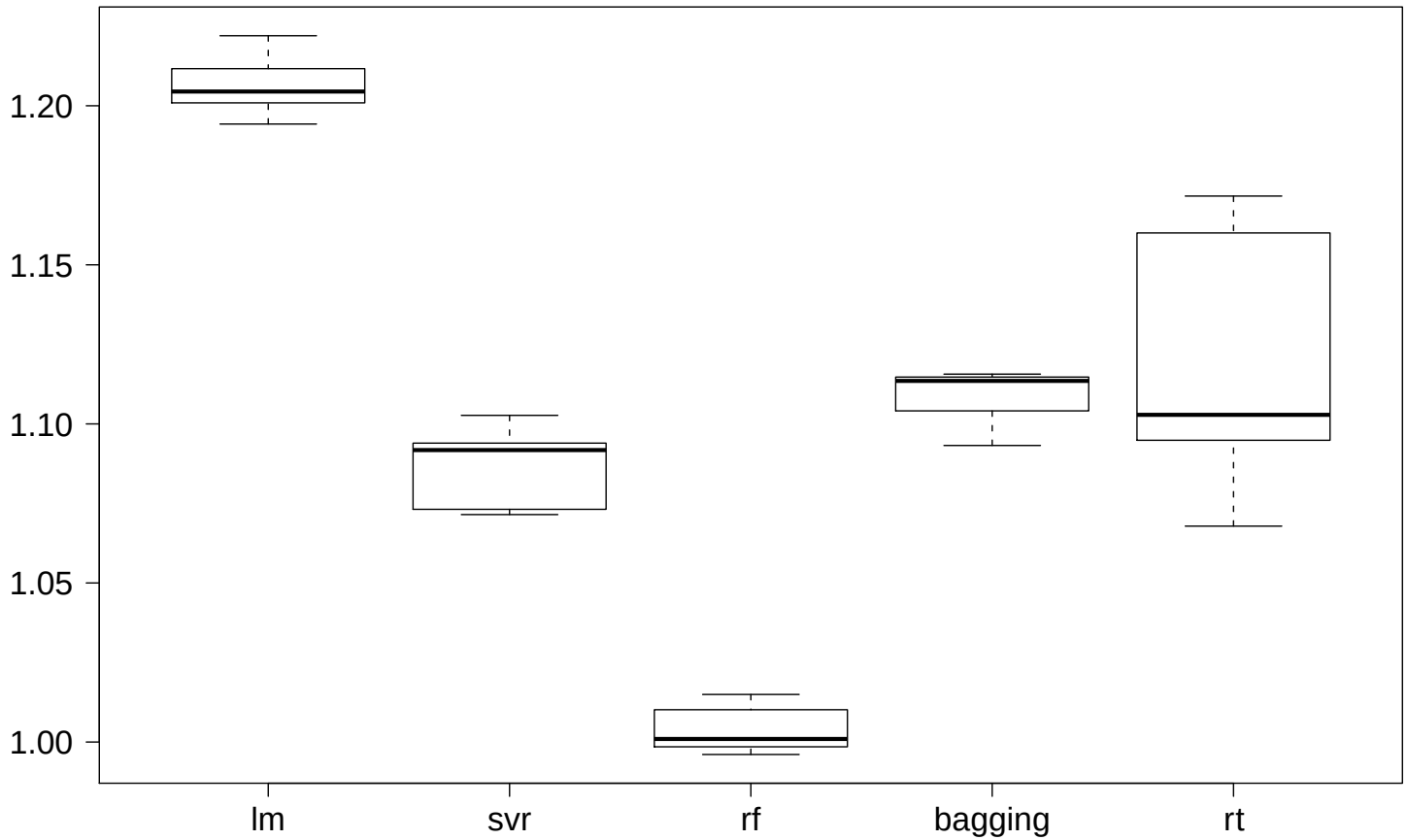


Figure: F440, RMSE of models

Spatial Variable Importance – Results

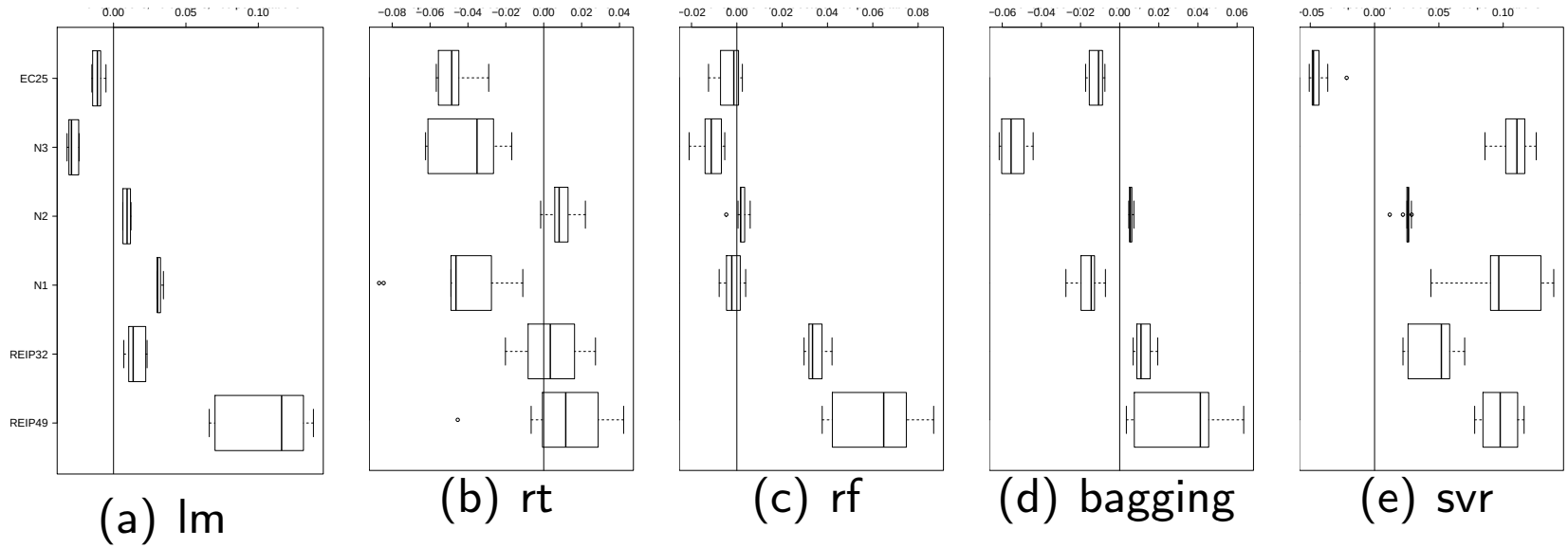


Figure: F440, RMSE increase of models after permuting one variable

Spatial Variable Importance – Results

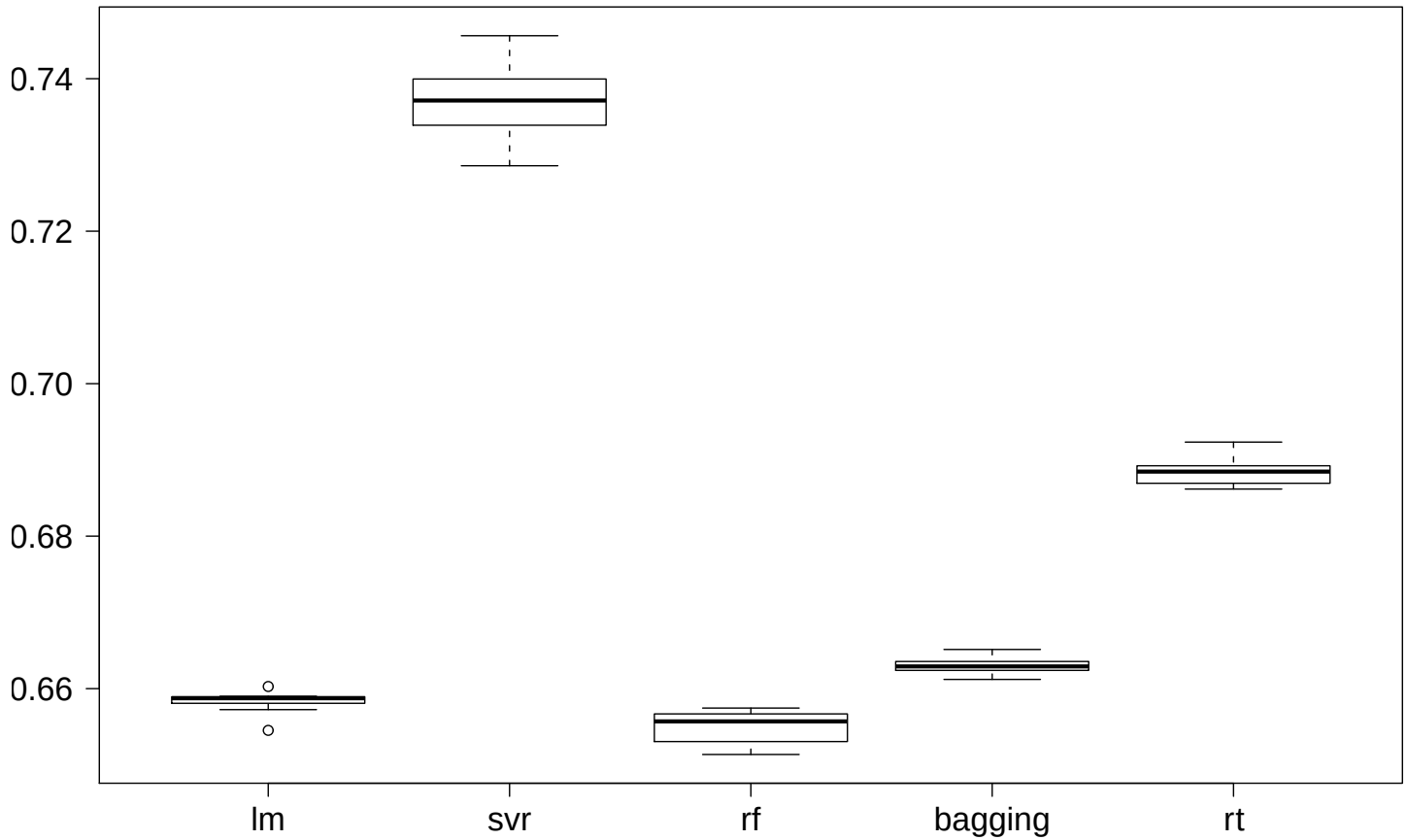


Figure: F611, RMSE of models

Spatial Variable Importance – Results

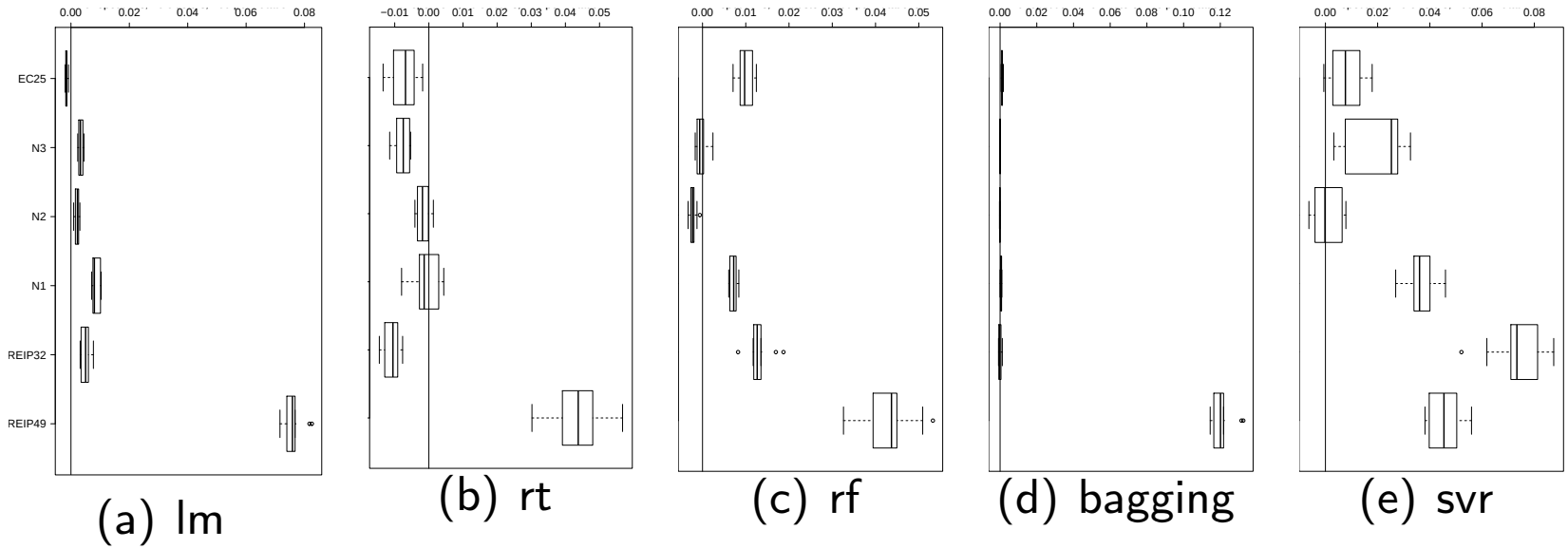


Figure: F611, RMSE increase of models after permuting one variable

Spatial Variable Importance – Conclusions

- REIP49 most important for yield prediction
  - obvious, since it shows the biomass amount close to harvest
- F440: REIP32 close second
- F611: likely linear relationships in data (*lm* best)
- issues with different numbers of levels for variables occur (4 levels for N1, 45/50 for N2/N3, 367/397 for REIP32/49)
- difference in modeling (linear vs. tree-based vs. support vector regression) can be seen

Summary

- precision agriculture as a data-driven approach
- spatial, geo-coded data in large amounts
- yield prediction solved as spatial cross-validation (regression)
- novel approach to assessing spatial variable importance

Time for . . .

Questions?

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