Spatial Data Mining in Precision Agriculture

NTNU, Trondheim

Georg Ruß, PhD candidate

August 13th, 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-referenced, 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)
  - pH, P, Mg, K: soil sampling data
- a few fields available, data records in up to $10 \times 10m$-resolution

Data Details – Temporal Aspects

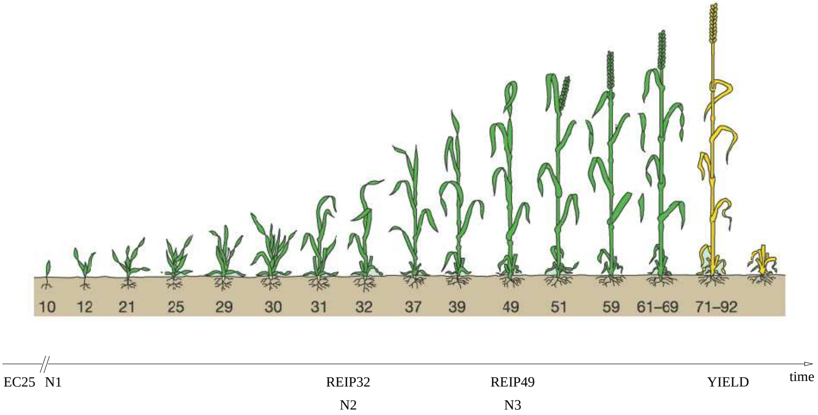


Figure: growing stages of cereals, source: adapted from BBCH

Data Details – Questions

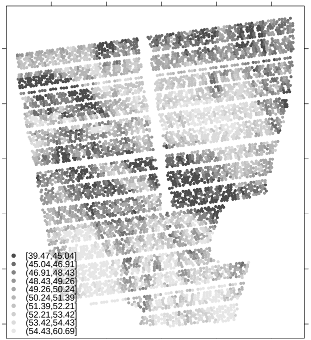
- Can the current year's yield be predicted from the available features? Which are the important variables for this task?
  - $\longrightarrow$ Spatial Regression
- How should the field be delineated into zones for basic fertilization?
  - $\longrightarrow$ Spatial Clustering

(Spatial) Regression – Basics

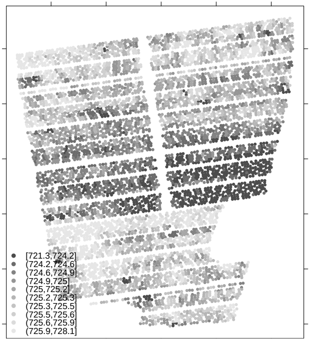
- multivariate regression: usually a cross-validation setup
  - divide data into training and test sets
  - train regression model on training set
  - report error on independent (!) test set
- linear model (usually as a baseline and with linear dependencies in data)
- support vector regression (support 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

- 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

Management Zone Delineation

- A common task in agriculture:
  - subdivide the field into smaller zones
  - zones are rather homogeneous
  - zones are spatially mostly contiguous
  - similarity between zones is low
- $\longrightarrow$ spatial clustering

Literature Approaches

- mostly non-spatial algorithms are used
  - no spatial contiguity
  - small islands, outliers, etc.
  - black-box models
  - fuzzy c-Means, k-Means, etc.
- spatial contiguity is not always required, but desirable
- spatial autocorrelation is usually neglected rather than exploited

Spatial Contiguity Constraint

- spatial clustering = clustering with a spatial contiguity constraint
- $\rightarrow$ constrained clustering
- Keep it simple and understandable:
  - hierarchical clustering
  - agglomerative clustering
- Idea:
  1. split field into small zones which are homogeneous
  2. iteratively merge these zones obeying similarity and spatial constraint

Spatial Tessellation

- k-Means clustering on the data points' coordinates
  - due to spatial autocorrelation, adjacent points are likely to be similar
  - this ensures homogeneity of these small zones
  - k is user-controllable and easy to understand
    - homogeneous field: smaller k
    - heterogeneous field: higher k

Spatial Tessellation

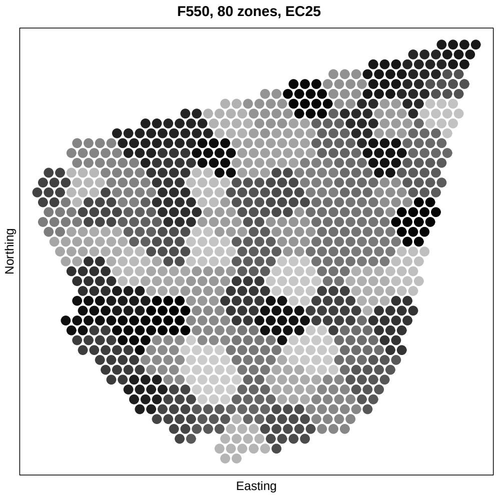
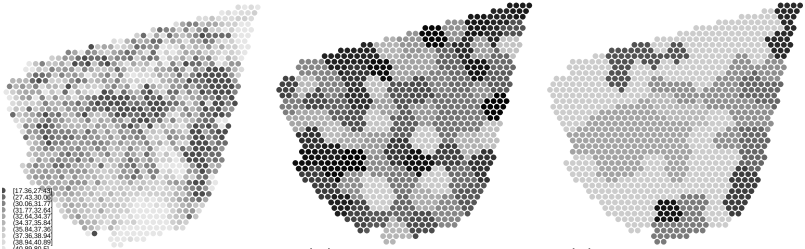


Figure: Tessellation of F550 using k-means, $k = 80$ (grey shades are for illustration only, no further meaning here)

Hierarchical Agglomerative Constrained Clustering

- principle: merge only adjacent zones, if they are similar enough
  - this ensures spatial contiguity
  - $\longrightarrow$ spatial constraint, non-adjacent zones *cannot link*
- once non-adjacent zones become much more similar than adjacent ones, they may be merged
  - introduce a user-controllable *contiguity factor cf*
  - $cf \geq 2$: high contiguity
  - $cf \in [1, 2]$: low contiguity
  - $cf \leq 1$: no contiguity

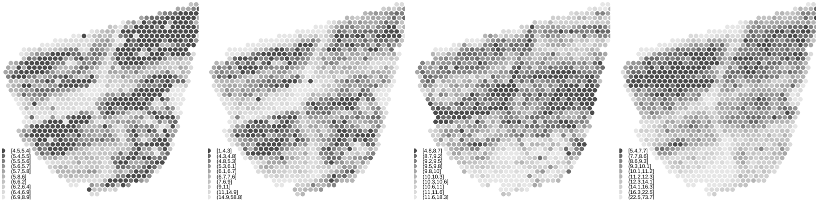
HACC – 1D example



(a) EC25 plot (b) EC25, 80 zones (c) EC25, final zones

Figure: F550, EC25 clustering

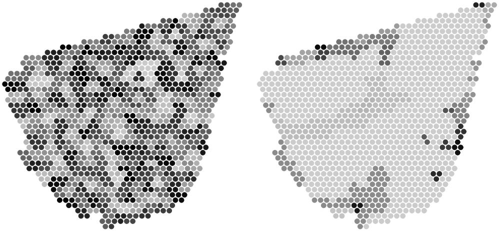
HACC – 4D example



(a) pH plot (b) P plot (c) Mg plot (d) K plot

Figure: F550, four attributes

HACC – 4D example (cont.)



(a) F550-4D, beginning (b) F550-4D, ten zones

Figure: F550, management zones

- actually, 3 zones (when comparing attribute values)
  - low pH, low P, low Mg, low K (largest zone)
  - high pH, high P, high Mg, high K (border zones)
  - high pH, high P, low Mg, high K (middle, from left)

Summary

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

Time for . . .

Questions?

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

Spatial Variable Importance – Results

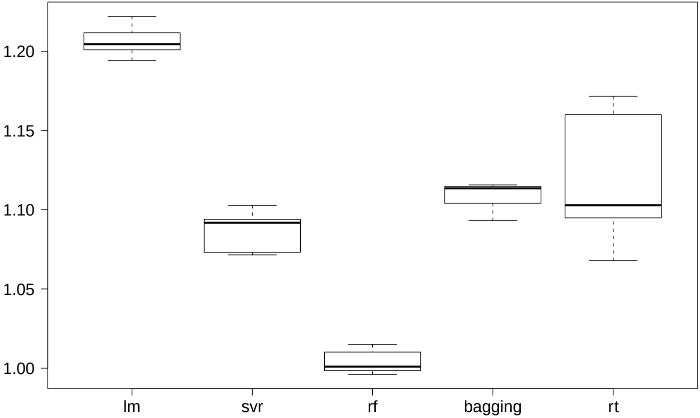
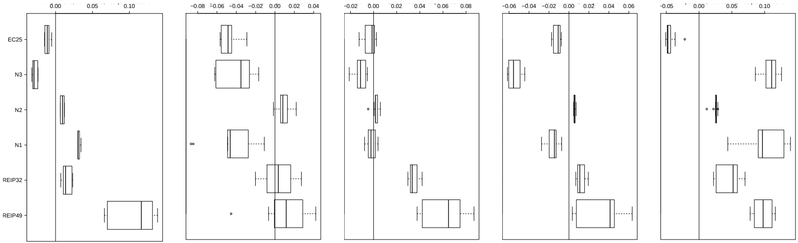


Figure: F440, RMSE of models

Spatial Variable Importance – Results



(a) lm (b) rt (c) rf (d) bagging (e) svr

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

Spatial Variable Importance – Results

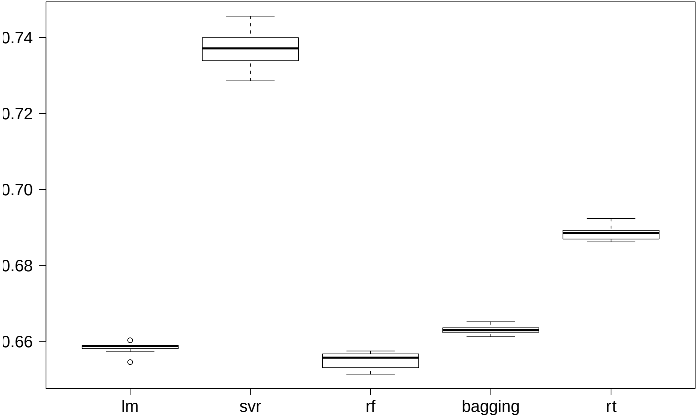
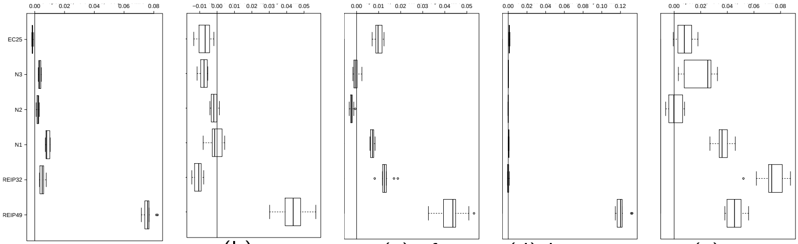


Figure: F611, RMSE of models

Spatial Variable Importance – Results



(a) lm (b) rt (c) rf (d) bagging (e) svr

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