Spatial Data Mining in Precision Agriculture

Application Lecture @UFZ

Georg Ruß

Otto-von-Guericke-Universität Magdeburg

13.01.2012

Introducing Myself

- Computer Science Degree (Diplom) in 2006
  - Work on Data Mining
  - Minor: Chemistry and Spectroscopy (MS, NMR)
- PhD work on *Spatial Data Mining in Precision Agriculture*
  - Interdisciplinary: computer science, geostatistics, precision agriculture
- Thesis submitted, PhD Defense: February 23rd
- PostDoc @UFZ?

Dissecting the (PhD Thesis) Title

$$\underbrace{\text{Spatial}}_{3}\,\underbrace{\text{Data Mining}}_{2}\text{ in }\underbrace{\text{Precision Agriculture}}_{1}$$

- 1 – Precision Agriculture
  - nowadays' technology applied to agriculture
  - small-scale, site-specific, data-based management

Dissecting the (PhD Thesis) Title

$$\underbrace{\text{Spatial}}_{3}\,\underbrace{\text{Data Mining}}_{2}\text{ in }\underbrace{\text{Precision Agriculture}}_{1}$$

- 1 – Precision Agriculture
  - nowadays' technology applied to agriculture
  - small-scale, site-specific, data-based management
- 2 – Data Mining
  - algorithms and ideas to *mine* data:
  - find novel, interesting and useful information in data [FPSS96]

Dissecting the (PhD Thesis) Title

$$\underbrace{\text{Spatial}}_{3}\,\underbrace{\text{Data Mining}}_{2}\,\text{in}\,\underbrace{\text{Precision Agriculture}}_{1}$$

- 1 – Precision Agriculture
  - nowadays' technology applied to agriculture
  - small-scale, site-specific, data-based management
- 2 – Data Mining
  - algorithms and ideas to *mine* data:
  - find novel, interesting and useful information in data [FPSS96]
- 3 – Spatial (Data)
  - result from most operations in environmental sciences
  - *must* consider spatial nature of data during data mining

Example Site F440

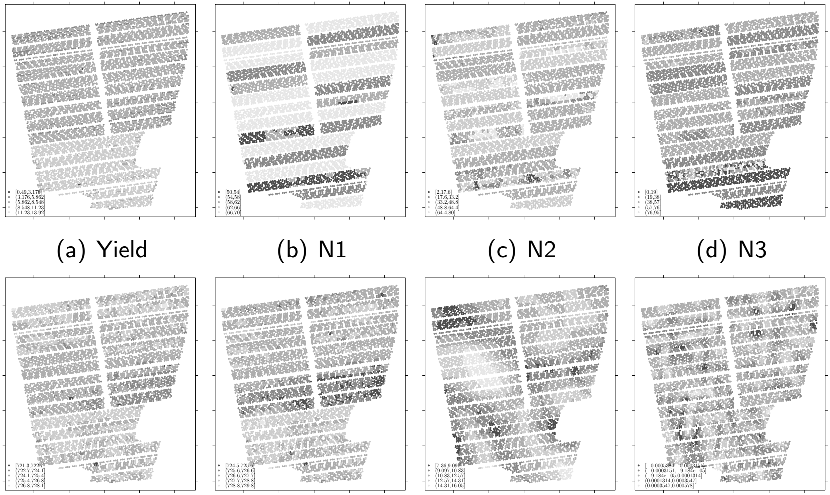
Figure: F440 near Köthen, Source: Google Earth w/ Overlay

Types of Data

- Yield and fertilizer
- Remote sensing (REIP32, REIP49, aerial/satellite imagery, ...)
- Geophysical data (apparent electrical conductivity EC25, ...)
- Soil sampling (pH, K, P, Mg, ...)
- Digital elevation models derivatives (slope, curvature, aspect, wetness index, ...)

- $\rightarrow$ High resolution spatial data sets
- $\rightarrow$ Use data mining on those sets for, e.g., optimization tasks

Example Site F440 (cont.)



(a) Yield (b) N1 (c) N2 (d) N3

(e) REIP32 (f) REIP49 (g) Wetness Index (h) Curvature

Figure: Yield and a few predictors (F440)

Task: Yield Prediction w/ Variable Importance

Yield Prediction = Regression

- Task: Predicting yield from other variables (ex-post)
  - based on an existing PhD thesis from 2006 [Wei06]
  - consider yield prediction as a (non-linear) regression task
  - issues with non-spatial models on spatial data (cp. [RB10a])

Task: Yield Prediction w/ Variable Importance

Yield Prediction = Regression

- Task: Predicting yield from other variables (ex-post)
  - based on an existing PhD thesis from 2006 [Wei06]
  - consider yield prediction as a (non-linear) regression task
  - issues with non-spatial models on spatial data (cp. [RB10a])
- Models
  - linear (lm), generalized additive (gam)
  - regression tree (rt), bagging (bag)
  - neural network (net)
  - support vector regression (svr)
  - k-nearest neighbor (kknn)

Task: Yield Prediction w/ Variable Importance

Regression Modeling, Process Flow

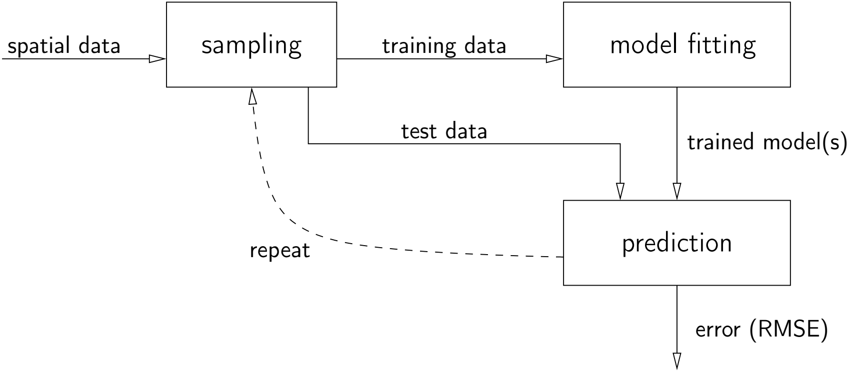


Figure: Generic cross-validation approach

Task: Yield Prediction w/ Variable Importance

Non-Spatial Sampling on Spatial Data

- Problem: spatial autocorrelation
- geographically adjacent data records likely to end up in training and test sets
- violates the independency assumption of cross-validation
- leads to systematic error underestimation

Task: Yield Prediction w/ Variable Importance

Non-Spatial Sampling on Spatial Data

- Problem: spatial autocorrelation
- geographically adjacent data records likely to end up in training and test sets
- violates the independency assumption of cross-validation
- leads to systematic error underestimation

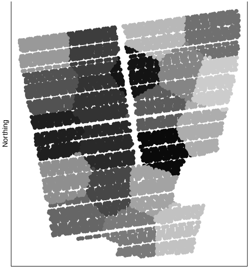


Figure: 20 spatial clusters

- → spatial sampling using spatial clustering:
  - k-means clustering
  - e.g., randomly choose 90/10% of clusters for training/test
  - more clusters → convergence towards non-spatial sampling (cp. [RB10b])

Task: Yield Prediction w/ Variable Importance

Spatial Regression Modeling, Process Flow

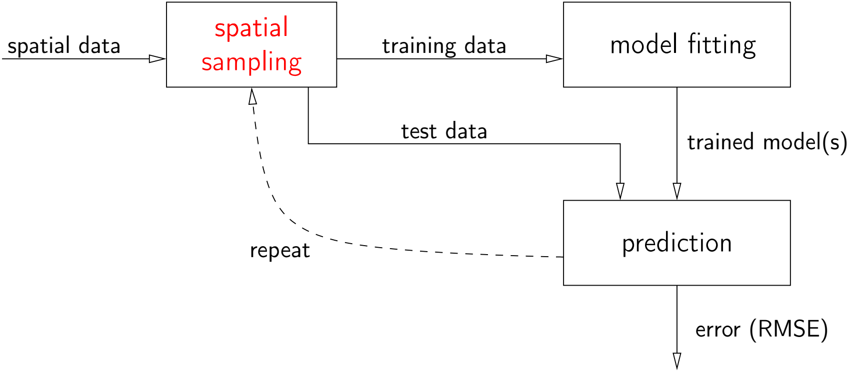


Figure: Generic spatial cross-validation approach

Task: Yield Prediction w/ Variable Importance

Variable Importance

- Question
  - What is the influence of a single variable on the model performance?
  - = Does a new sensor really contribute new information?
  - = Should we consider additional data sources?

Task: Yield Prediction w/ Variable Importance
Variable Importance

- Question
  - What is the influence of a single variable on the model performance?
  - = Does a new sensor really contribute new information?
  - = Should we consider additional data sources?
- Idea: Try permuting this variable in the test set!
  - if the RMSE increases, the variable is probably important
  - allows to assess the importance in the presence of other variables
  - allows to determine relationships between variables
  - works independently of the regression model used

Task: Yield Prediction w/ Variable Importance

Variable Importance

- Question
  - What is the influence of a single variable on the model performance?
  - = Does a new sensor really contribute new information?
  - = Should we consider additional data sources?
- Idea: Try permuting this variable in the test set!
  - if the RMSE increases, the variable is probably important
  - allows to assess the importance in the presence of other variables
  - allows to determine relationships between variables
  - works independently of the regression model used
- Side note: repeat steps below sufficiently often (statistics)
  - random spatial sampling
  - model fitting
  - test set permutation

Task: Yield Prediction w/ Variable Importance

Variable Importance, Process Flow

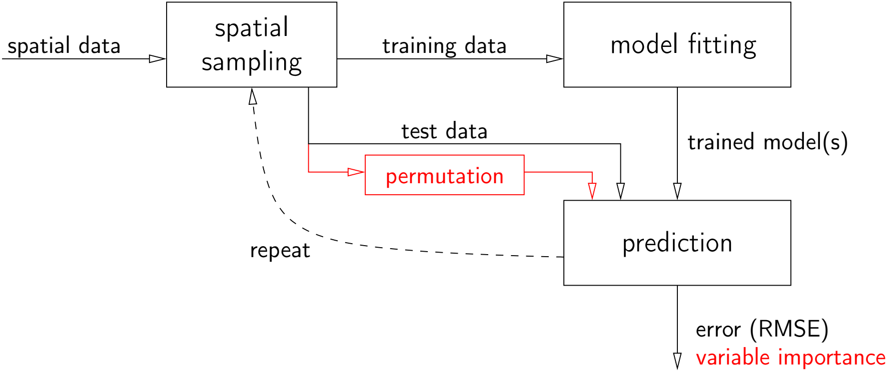


Figure: Spatial Variable Importance Approach

Task: Yield Prediction w/ Variable Importance

Results

- variable importance results:
  - vegetation indicators REIP32, REIP49 most important throughout models and sites
  - apparent electrical soil conductivity quite important (in conjunction with others)
  - digital elevation model variables quite important (wetness, curvature, slope)
  - further results vary between sites
- further relevant results:
  - comparison between different fertilization strategies made possible
  - comparison between models (winner: bagging/svr, loser: net)
- details in dissertation [Ruß12]
- in preparation: article for *Remote Sensing of Environment*

My Contribution to Data Integration/Inversion

- I'm not a geophysicist ...

My Contribution to Data Integration/Inversion

- I'm not a geophysicist ...
  - ... but we're having similar modeling and optimization problems in computer science!

My Contribution to Data Integration/Inversion

- I'm not a geophysicist ...
  - ... but we're having similar modeling and optimization problems in computer science!

- Specialties
  - preprocessing of spatial environmental data
  - computations on spatial environmental data (preferably in R)
  - parameter optimization of models (evolutionary, gradient descent, simulated annealing, PSO, ...)
  - exploitation of spatial heterogeneity
  - clustering, data and sensor fusion ($\approx$ integration)

My Contribution to Data Integration/Inversion

- I'm not a geophysicist …
  - …but we're having similar modeling and optimization problems in computer science!
- Specialties
  - preprocessing of spatial environmental data
  - computations on spatial environmental data (preferably in R)
  - parameter optimization of models (evolutionary, gradient descent, simulated annealing, PSO, …)
  - exploitation of spatial heterogeneity
  - clustering, data and sensor fusion ($\approx$ integration)
- Interests
  - work with further (geophysical?) data sets
  - work in conjunction with further disciplines
  - provide computer science and data mining expertise

Possible Research Profile

Environmental Data Mining

as the task of finding interesting, novel and potentially useful knowledge in spatial and temporal multi-layered data sets from environmental sciences.

(definition adapted from [FPSS96])

Acknowledgements

- Precision Agriculture Data: MLU Halle, Prof. Peter Wagner
- Digital Elevation Model Data: LVermGeo Sachsen-Anhalt, Magdeburg

- PhD blog at `http://research.georgruss.de`

Bibliography I

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth.
From data mining to knowledge discovery in databases.
AI Magazine, 17:37–54, 1996.

[RB10a] Georg Ruß and Alexander Brenning.
Data mining in precision agriculture: Management of spatial information.
In Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design*, volume 6178 of *LNAI*, pages 350–359, Berlin, Heidelberg, 2010. Springer.

[RB10b] Georg Ruß and Alexander Brenning.
Spatial variable importance assessment for yield prediction in precision agriculture.
In Paul R. Cohen, Niall M. Adams, and Michael R. Berthold, editors, *Proceedings of IDA2010*, volume 6065 of *LNCS*, pages 184–195, Heidelberg, 2010. Springer.

[Ruß12] Georg Ruß.
Spatial Data Mining in Precision Agriculture.
PhD thesis, Otto-von-Guericke-Universität Magdeburg, 2012.

[Wei06] Georg Weigert.
Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung.
PhD thesis, TU München, 2006.