Feature Selection for Wheat Yield Prediction

SGAI AI-2009, Cambridge, Dec 16th, 2009

Georg Ruß, russ@iws.cs.uni-magdeburg.de

Dec 16th, 2009

Structure of this Talk

About me

- German computer scientist
- with interest in (spatial) data mining
- currently using mostly R for spatial data mining
- this talk is a part of what's going to be my PhD thesis
- my research blog: `http://research.georgruss.de/`
- my "Data Mining in Agriculture" workshop in 2010: `http://dma2010.de/`

Data Mining in Agriculture

basic idea

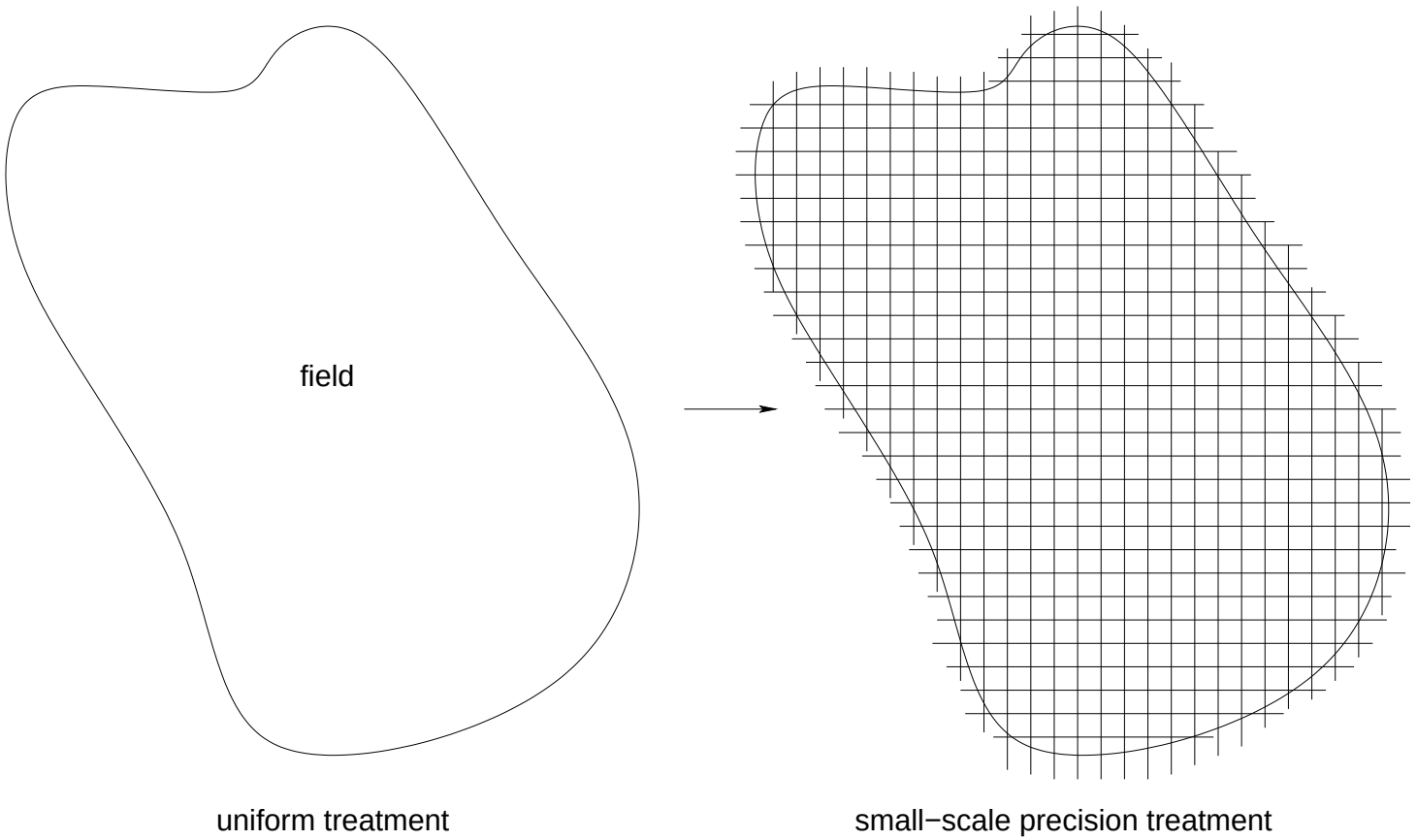


Figure: *Precision Agriculture* = data-driven approach to agriculture

Precision Agriculture

some more ideas

- precision agriculture
  - cheap data collection
  - GPS-based technology
  - divide field into small-scale parts
  - treat small parts independently instead of uniformly
- lots of data (sensors, imagery)
- use data mining to
  - improve efficiency
  - improve yield
  - identify useful sensors

Structure of this Talk

N Fertilizer and Yield

- Nitrogen fertilizer
  - easy to measure when manuring
  - three time points into the growing season when nitrogen fertilizer is applied
  - three attributes: N_1, N_2, N_3
- Yield 2003/2004
  - measure yield when harvesting
  - data from 2003 (previous year) and 2004 (current year)
  - two attributes: Yield03, Yield04

Vegetation Measuring and Electric Conductivity

- Red Edge Inflection Point
  - second derivative value along the spectrum's red edge region
  - aerial photography or tractor-mounted sensor
  - larger value means more vegetation
  - measured (chronologically) before N_2 and N_3
  - two attributes: $REIP_{32}$, $REIP_{49}$
- Electromagnetic Conductivity
  - measure apparent conductivity of soil down to 1.5m
  - uses commercial sensors
  - one attribute: EC_a

Spatial Variable Plots

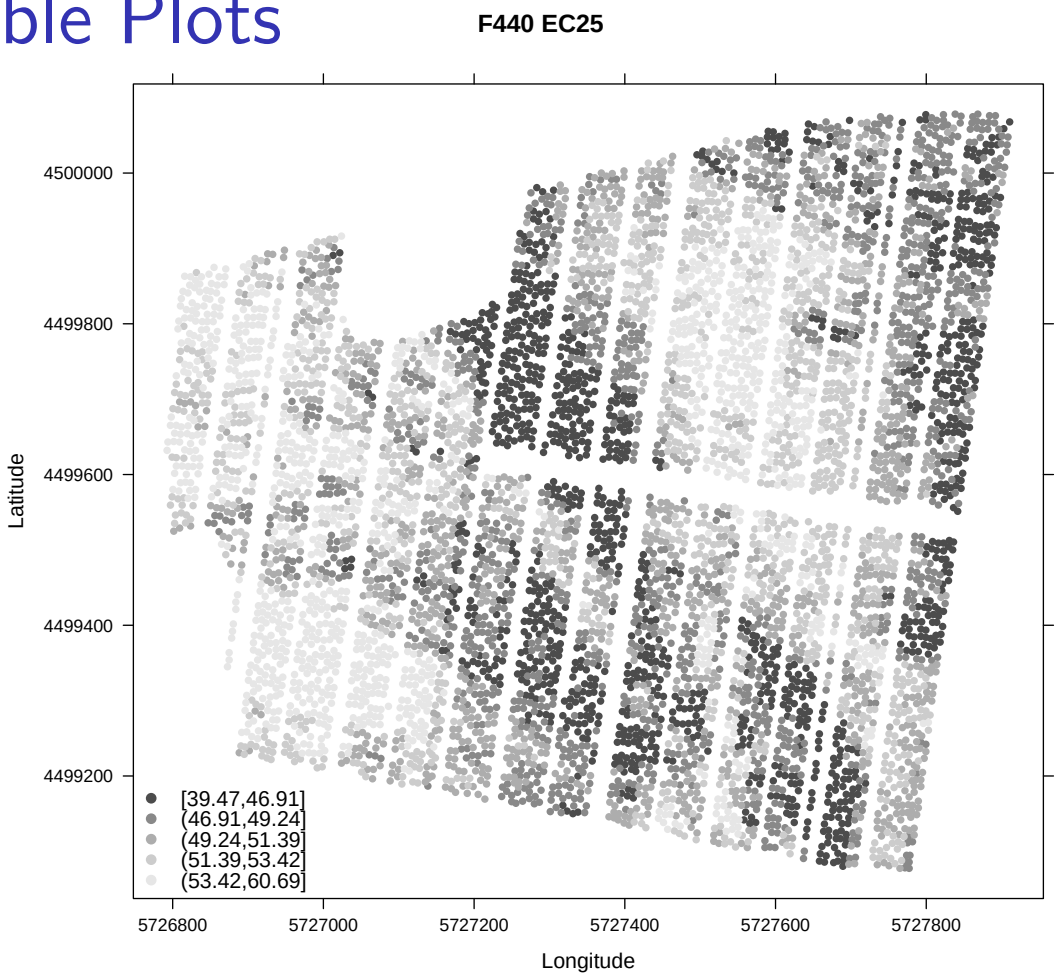


Figure: Apparent Electrical Conductivity for F440 field

Spatial Variable Plots

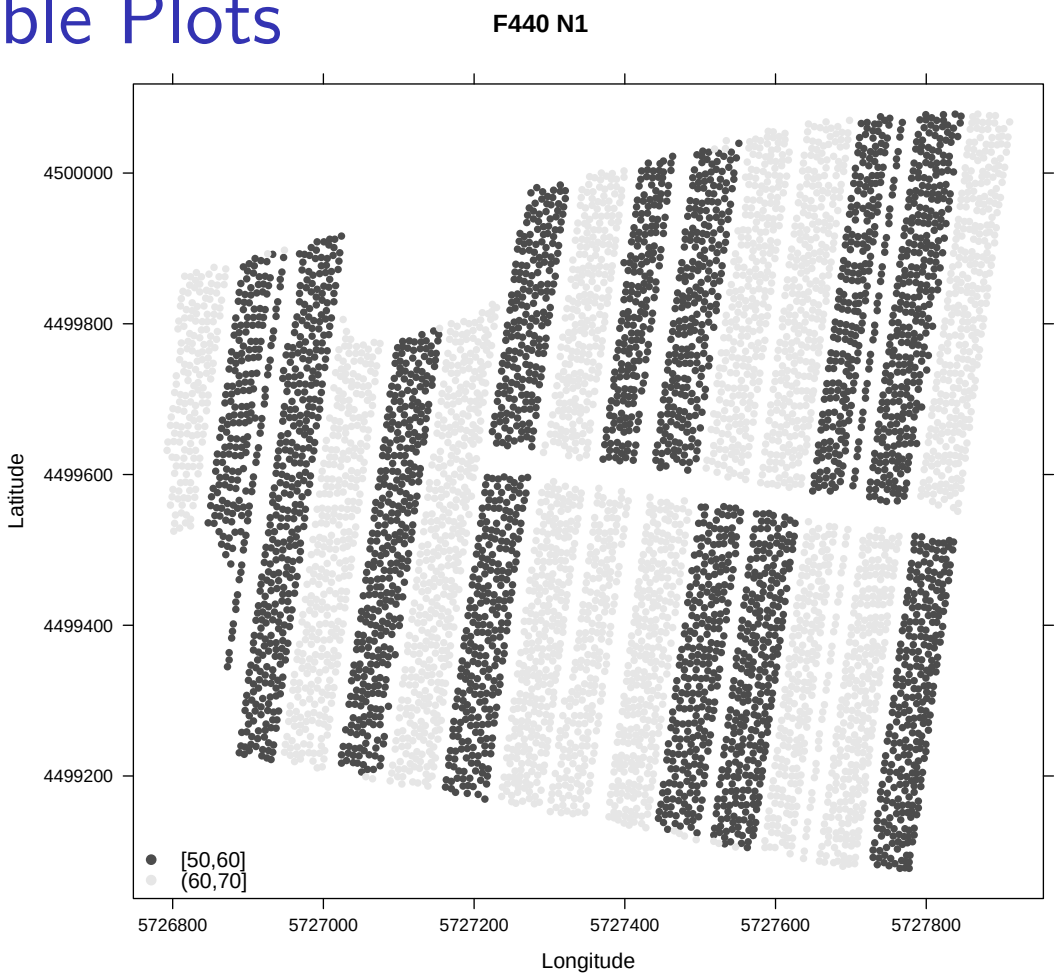
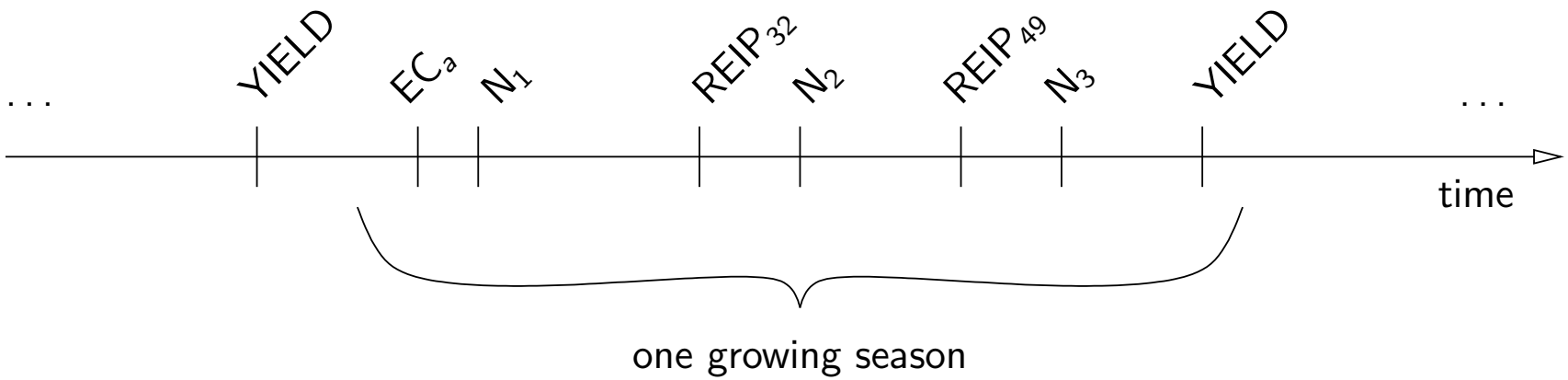


Figure: Nitrogen Fertilizer (first dressing) for F440 field

Structure of this Talk

Example: Yield Prediction



- try to predict current year's yield from fertilizer and soil status data (and maybe past year's yield?)
- classical regression problem?

Example: Yield Prediction

non-spatial case:

- standard cross-validation learning approach:
  - divide data into learning, validation and testing subsets
  - train model on learning subset
  - use validation subset to see when overfitting occurs (stop learning)
  - report error of model on testing subset
  - models tested: neural networks, support vector regression, regression tree, bagging etc.
- due to spatial autocorrelation very similar data records exist in the training, validation and testing subsets
- probably a violation of the *statistical independence* assumption
- $\longrightarrow$ use spatial cross-validation at a later stage

Structure of this Talk

Usefulness of Particular Data Attributes

Open Question

- Question:
  - How useful is a particular sensor (attribute)?
  - Is a new attribute related to existing ones?
  - Does a new attribute contribute much in terms of information content?
- Practical issues:
  - Question arises when developing new sensors
  - New sensors are evaluated in-season
  - (*current research and part of my PhD*)

Usefulness of Sensor Data

Open Question

- Ideas towards this issue:
  - Create a spatial (yield) prediction model and evaluate how much this is improved by adding new data attributes? (this talk)
  - Evaluate standard feature selection approaches for non-spatial data and adapt those? (future work)
  - Apply principal components analysis and check how the components change? (future work)
  - Check an attribute's importance by permutating its values and comparing models before and after the permutation? (future work)

Yield Prediction

. . . as a tool for feature selection

- Create a standard non-spatial regression model setup
- Apply cross-validation to measure the models' error
- Change the learning data sets (add attributes) and evaluate how the models' error changes
- Good additional features should result in a lower prediction error

Forward Feature Selection

Algorithm

1: S = [ ], F ← features
2: **repeat**
3: $E \leftarrow [\,]$
4: **for** $j = 1 \ldots length(F)$ **do**
5: $f \leftarrow F[j]$ {select j-th feature}
6: $S_j \leftarrow S \cup f$ {add current feature to S_j}
7: $M_j \leftarrow model(S_j)$ {generate regression model from data}
8: $E_j \leftarrow evaluate(M_j)$ {calculate modeling error}
9: $E \leftarrow E||E_j$ {store error}
10: **end for**
11: $S \leftarrow S||F[min(E)]$ {add best feature to S}
12: $F \leftarrow F - F[min(E)]$ {remove best feature from F}
13: **until** min(E) $\leq$ threshold OR $F = [\,]$
14: **return** S {return list of features, best one first}

Results

Forward Feature Selection via Regression Tree

	F04		F131		F330	
step	error	feature	error	feature	error	feature
1	0.342	YIELD03	0.235	REIP49	0.279	N3
2	0.262	REIP49	0.136	YIELD05	0.246	REIP49
3	0.228	N3	0.104	EM38	0.223	EM38
4	0.215	EM38	0.104	REIP32	0.199	REIP32
5	0.210	REIP32	0.104	N1	0.189	N1
6	0.209	TRACFORCE	0.104	N2	0.187	N2
7	0.208	N1	0.104	N3	0.181	YIELD05
8	0.205	N2				

Results

Forward Feature Selection via Support Vector Regression

	F04		F131		F330	
step	error	feature	error	feature	error	feature
1	0.557	YIELD03	0.469	REIP49	0.519	N2
2	0.509	REIP49	0.356	YIELD05	0.508	YIELD05
3	0.483	REIP32	0.337	N2	0.493	REIP49
4	0.471	EM38	0.335	N1	0.491	N3
5	0.469	TRACFORCE	0.333	N3	0.469	EM38
6	0.466	N2	0.303	REIP32	0.454	N1
7	0.449	N1	0.285	EM38	0.439	REIP32
8	0.444	N3				

Structure of this Talk

Summary

- Core Points:
  - a yield prediction task from agriculture, a non-spatial regression approach
  - a forward feature selection approach to find the most important attributes
- Future Work:
  - Agriculture will become ever more data-driven.
  - Important difference between non-spatial and spatial data.
  - Standard data mining techniques can not be copied one-to-one to spatial data, but must be adapted
- Overall: successful application of regression and feature selection in an agriculture context

Bibliography I

Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner.

Estimation of neural network parameters for wheat yield prediction.
In Max Bramer, editor, *Artificial Intelligence in Theory and Practice II*, volume 276 of *IFIP International Federation for Information Processing*, pages 109–118. Springer, July 2008.

Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner.

Optimizing wheat yield prediction using different topologies of neural networks.
In José Luis Verdegay, Manuel Ojeda-Aciego, and Luis Magdalena, editors, *Proceedings of IPMU-08*, pages 576–582. University of Málaga, June 2008.

Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner.

Visualization of agriculture data using self-organizing maps.
In Tony Allen, Richard Ellis, and Miltos Petridis, editors, *Applications and Innovations in Intelligent Systems*, volume 16 of *Proceedings of AI-2008*, pages 47–60. BCS SGAI, Springer, January 2009.

Georg Ruß, Rudolf Kruse, Peter Wagner, and Martin Schneider.

Data mining with neural networks for wheat yield prediction.
In Petra Perner, editor, *Advances in Data Mining (Proc. ICDM 2008)*, pages 47–56, Berlin, Heidelberg, July 2008. Springer Verlag.

Finally . . .

Questions & Answers

- my research blog: `http://research.georgruss.de/`
- my "Data Mining in Agriculture" workshop in 2010: `http://dma2010.de/`