Data Mining in Agriculture

ATO, Canberra, Dec 11th, 2009

Georg Ruß, georg.russ@ieee.org

Dec 11th, 2009

Structure of this Talk

About me

- German computer scientist
- with interest in (spatial) data mining
- currently using mostly R for spatial data mining
- most of this talk is about 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/`

Why I'm here

- German-Australian cooperation (DAAD/Go8)
  - two-year grant covering travel cost
  - cooperation between Universität Magdeburg (Prof Rudolf Kruse) and University of Melbourne (Prof Saman Halgamuge)
  - project about renewable energy distribution and optimization
- Invitation by Warwick Graco to talk about

"Data Mining in Agriculture"

Data Mining in Agriculture

basic idea

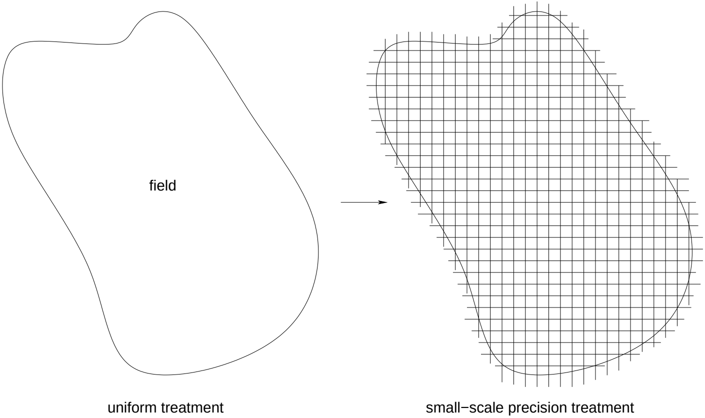


Figure: *Precision Agriculture* = data-driven approach to agriculture

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

Introduction / Motivation

Data Details

- N Fertilizer and Yield
- Vegetation and Electric Conductivity

Example Task: Yield Prediction

Research Questions

Summary

Spatial vs. Non-Spatial Data

- First law of geography: *Everything is related to everything else, but near things are more related than distant things.* [7]
  - agriculture data are spatial data
  - spatial autocorrelation exists (Moran's I, semivariograms)
  - data records are therefore *not* independent
  - natural neighborhood exists
- On the contrary:
  - classical data mining models often do not handle spatial data
  - data records are considered independent
  - overfitting and overlearning occur

Origin of Data

- invasive vs. non-invasive
- high-resolution vs. low-resolution
- cheap vs. expensive

Remote Sensing aerial images, satellite images, NDVI, OSAVI, VARI, REIP, BIOMASS; non-invasive, cheap, high-resolution

Soil Sampling EC_a, soil survey, OM, TN, AN, AP, AK, CEC, pH, water; mostly invasive, expensive, high resolution = expensive

Yield Mapping non-invasive, cheap, high to medium resolution

Topography often derived from GPS, elevation, slope, and derivatives; non-invasive, cheap, high-resolution

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 2007/2008
  - measure yield when harvesting
  - data from 2007 (previous year) and 2008 (current year)
  - two attributes: Yield07, Yield08

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

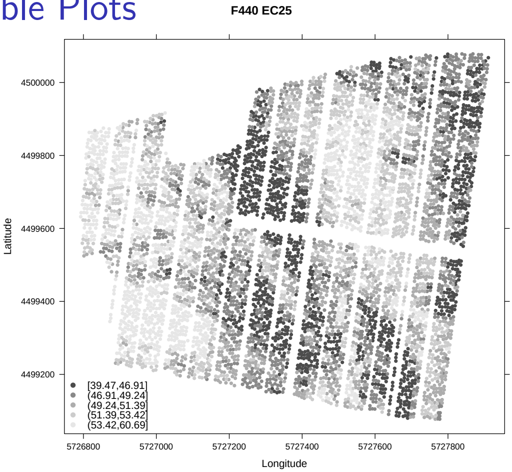


Figure: Apparent Electrical Conductivity for F440 field

Spatial Variable Plots

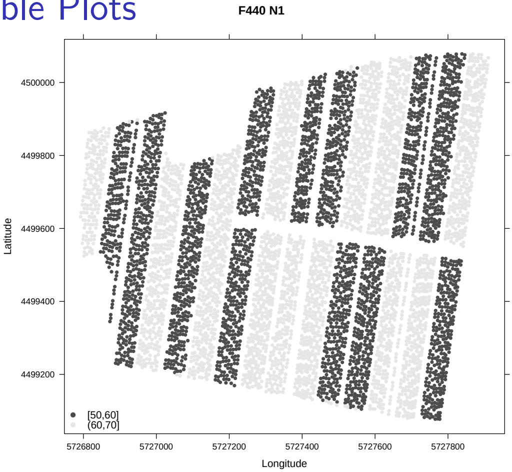
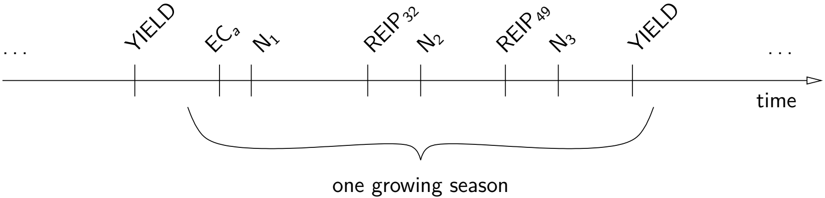


Figure: Nitrogen Fertilizer (first dressing) for F440 field

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
- violation of the *statistical independence* assumption
- $\longrightarrow$ use spatial cross-validation

Example: Yield Prediction

spatial case:

- extend standard approach to spatial data
  - divide data into spatial subsets (contiguous parts of the field)
  - train standard regression model on learning subset
  - use validation subset to stop training
  - report error of model on testing subset
  - spatial subset generation: via k-means (simple approach)
  - regression models: as before
- more statistically valid way of yield prediction
- generate map from prediction errors
  - find extraordinary parts in the field
  - uncover hidden relationships in the data

Spatial Prediction Plot

Figure: Spatial Cross-Validation Approach, 50 Clusters, SVR

Structure of this Talk

Introduction / Motivation

Data Details

Example Task: Yield Prediction

Research Questions
- Interestingness of Subfields
- Usefulness of Particular Data Attributes

Summary

Research Questions

- How well can the current year's yield be predicted? (*solved*)
- Are there any subparts of the field which differ considerably from the rest? Can we uncover hidden relationships from a data mining perspective? (*current work*)
- How useful are the additional sensor data that were introduced? (*current work*)
  - EC_a, $REIP_{32}$, $REIP_{49}$ et al.

Interestingness of Subfields

Literature

- Classical data-driven approach in agriculture: delineation into management zones
  - classic: yield mapping
  - often: expert knowledge
  - recently: fuzzy clustering on non-spatial data, PCA
- Drawbacks
  - no dynamics – zones are static throughout season
  - incontiguity of zones – no consideration of spatial relationships

Interestingness of Subfields

Literature

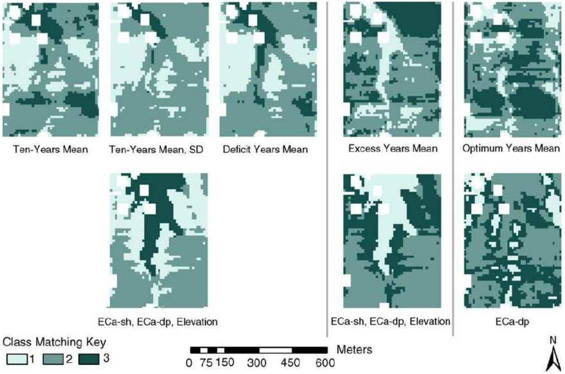


Figure: taken from [2], showing different field zones based on different measurements of soil apparent electrical conductivity (EC_a)

Interestingness of Subfields

Literature

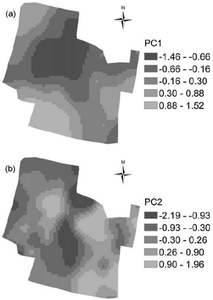


Fig. 3 – Contour maps for (a) first and (b) second principal component (PC).

(a) Principal Components

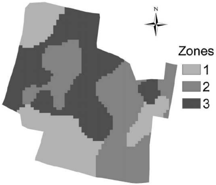


Fig. 5 – Management zones (MZs) map for optimum clusters in study area.

(b) Resulting Zones

Figure: taken from [8], PCA is run on whole data set and management zones are generated from the first principal components

Interestingness of Subfields

Literature

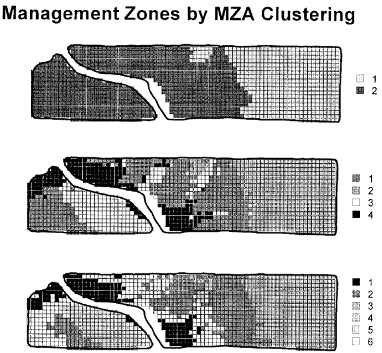


Fig. 3. (Top) Apparent soil electrical conductivity (EC_a), elevation, and slope as MZA clustering variables for Field 2 and (bottom) MZA output for two, four, and six clusters.

Figure: taken from [1], Management Zone Analyst Software, grid-based clustering based on different subsets of available data

Interestingness of Subfields

Current Research

- Improve the existing approach of management zone delineation:
  - Use spatial data.
  - Find contiguous subparts.
  - Adapt management zones throughout the season.
- Approach (split-and-merge):
  - Cluster the field (spatially) using k-means into an appropriate number of zones.
  - Merge neighboring zones according to some non-spatial criterion (similarity, distance, etc.).
  - Repeat this process in-season with available in-season vegetation data ($REIP_{32}$, $REIP_{49}$).
  - Investigate changes in zones.

Interestingness of Subfields

Split into Spatial Clusters (k-means)

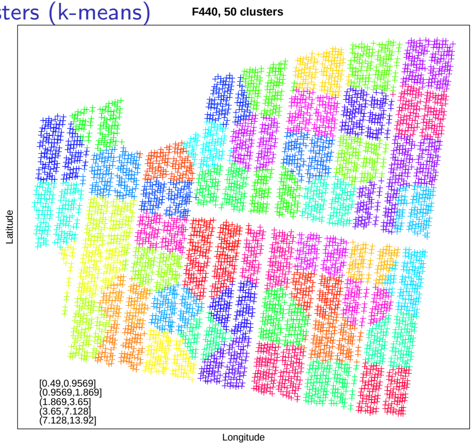


Figure: split of F440 field into 50 spatial clusters using k-means on the data points' coordinates

Interestingness of Subfields

Merging Clusters

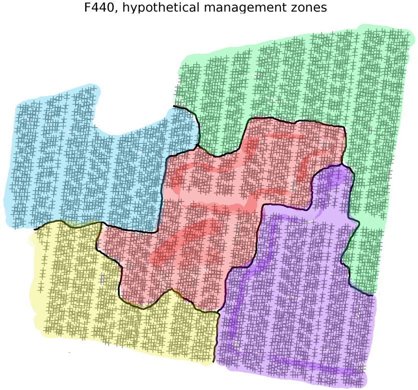


Figure: suggested idea: merging the previous clusters into a fixed number of management zones

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?
  - Apply principal components analysis and check the components?
  - Check an attribute's importance by permutating its values and comparing models before and after the permutation?
  - Evaluate standard feature selection approaches for non-spatial data and adapt those?

Structure of this Talk

Summary

- There's a difference between non-spatial and spatial data.
- Agriculture will become ever more data-driven.
- Standard data mining techniques can not be copied one-to-one to spatial data, but may be adapted:
  - Clustering
  - Regression
  - Feature Selection
  - Principal Components Analysis
  - etc.
- overall: successful application of data mining ideas in agriculture

Bibliography I

Jon J. Fridgen, Newell R. Kitchen, Kenneth A. Sudduth, Scott T. Drummond, William J. Wiebold, and Clyde W. Fraisse.
Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation.
Agronomy Journal, 96(1):100–108, 2004.

N.R. Kitchen, K.A. Sudduth, D.B. Myers, S.T. Drummond, and S.Y. Hong.
Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity.
Computers and Electronics in Agriculture, 46(1-3):285 – 308, 2005.
Applications of Apparent Soil Electrical Conductivity in Precision Agriculture.

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.

Bibliography II

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.

Waldo Tobler.
A computer model simulation of urban growth in the detroit region.
Economic Geography, pages 234–240, 1970.

Wang Xin-Zhong, Liu Guo-Shun, Hu Hong-Chao, Wang Zhen-Hai, Liu Qing-Hua, Liu Xu-Feng, Hao Wei-Hong, and Li Yan-Tao.
Determination of management zones for a tobacco field based on soil fertility.
Computers and Electronics in Agriculture, 65(2):168 – 175, 2009.

Finally ...

Questions & Answers

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