Data Mining in der Landwirtschaft

Vortrag im DKE-Kolloquium

Georg Ruß, IWS

14. Januar 2010

Gliederung

Über mich

- Seit 2006 bei Prof. Kruse als Doktorand
- Interessen: Data Mining, Precision Agriculture
- Aktuelles Thema: seit Ende 2007
- Forschungs-Blog: `http://research.georgruss.de/`
- "Data Mining in Agriculture"-Workshop 2010: `http://dma2010.de/`

Data Mining in der Landwirtschaft

Grundidee

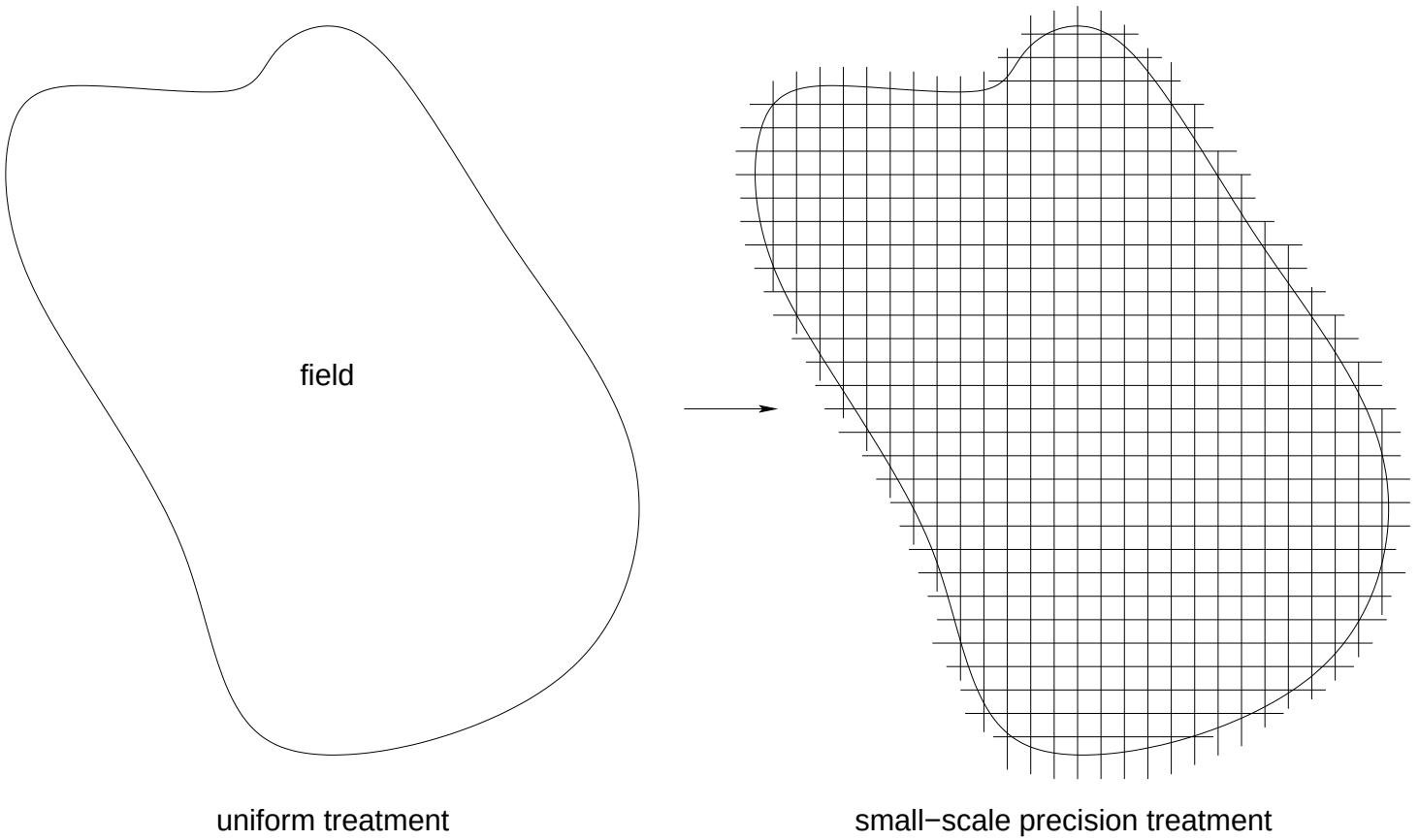


Abbildung: *Präzisionslandwirtschaft* = datengetriebene Herangehensweise

Präzisionslandwirtschaft

weitere Ideen

- Ursprünge:
  - günstige Datenquellen
  - GPS-Technologie für räumliche Daten
  - Feldunterteilung in Teilflächen
  - teilflächenspezifische Bewirtschaftung
- große Datenmengen
- Nutze Data Mining, um:
  - Effizienz zu erhöhen,
  - Ertrag zu erhöhen,
  - nützliche Daten herauszufiltern.

Gliederung

Räumliche vs. nicht-räumliche Daten

- First law of geography: *Everything is related to everything else, but near things are more related than distant things.* [5]
  - Landwirtschaftsdaten sind zwangsläufig räumliche Daten
  - Räumliche Autokorrelation liegt vor (Moran's I, Semivariogramme)
  - (benachbarte) Datenpunkte sind daher nicht unabhängig voneinander
  - eine natürliche Nachbarschaft liegt vor

Räumliche vs. nicht-räumliche Daten (*Fortsetzung*)

- Andererseits:
  - Klassische Data-Mining-Modelle vernachlässigen typischerweise die räumliche Komponente,
  - Datenpunkte werden als statistisch unabhängig voneinander angenommen,
  - daher treten Überlernen und Überanpassung auf.

Dateneinordnung

Fernerkundungsdaten Luftbilder, Satellitenbilder, abgeleitete Indizes wie NDVI, OSAVI, VARI, REIP, BIOMASS; nicht-invasiv, relativ günstig, hohe Auflösung

Bodenproben klassische Bodenuntersuchung auf Attribute wie OM (organisches Material), AN (verfügbarer Stickstoff), AP, AK, CEC, pH, Feuchtigkeit; invasive Methoden, sehr teuer, insbesondere bei höher Auflösung der Daten

Ertragskarten Ertrag wird bei der Ernte gemessen und geo-codiert; nicht-invasiv, günstig, hohe bis mittlere Auflösung je nach Ausrüstung

Topographie typischerweise aus GPS-Daten entnommen: Höhe, Anstieg, Senken, weitere abgeleitete Werte; nicht-invasiv, günstig, hohe Auflösung

Stickstoffdünger und Ertrag

- Stickstoffdünger
  - Menge kann beim Düngen relativ einfach gemessen werden
  - Dünger wird zu drei Zeitpunkten während der Vegetationsphase ausgebracht
  - Drei Attribute: N_1, N_2, N_3
- Ertrag 2007/2008
  - Ertrag wird bei der Ernte vollautomatisch geocodiert aufgezeichnet
  - Daten aus 2007 (Vorjahr) und 2008 (aktuelles Jahr)
  - Zwei Attribute: Yield07, Yield08

Vegetation und elektrische Leitfähigkeit

- REIP (Red Edge Inflection Point)
  - Wert der zweiten Ableitung des spektralen Rotbereichs eines Bildes
  - kann aus Luftbildern oder Satellitenfotos gewonnen werden
  - höherer Wert bedeutet mehr Vegetation
  - wird vor der Ausbringung von N_2 und N_3 gemessen
  - Zwei Attribute: $REIP_{32}$, $REIP_{49}$
- Elektromagnetische Leitfähigkeit
  - scheinbare Bodenleitfähigkeit wird per Sensor gemessen, bis zur Tiefe von etwa 1,50m
  - starke Korrelation mit verschiedenen Bodeneigenschaften wird erwartet
  - kommerzielle Sensoren sind verfügbar
  - Ein Attribut: EC_a

Kartierung

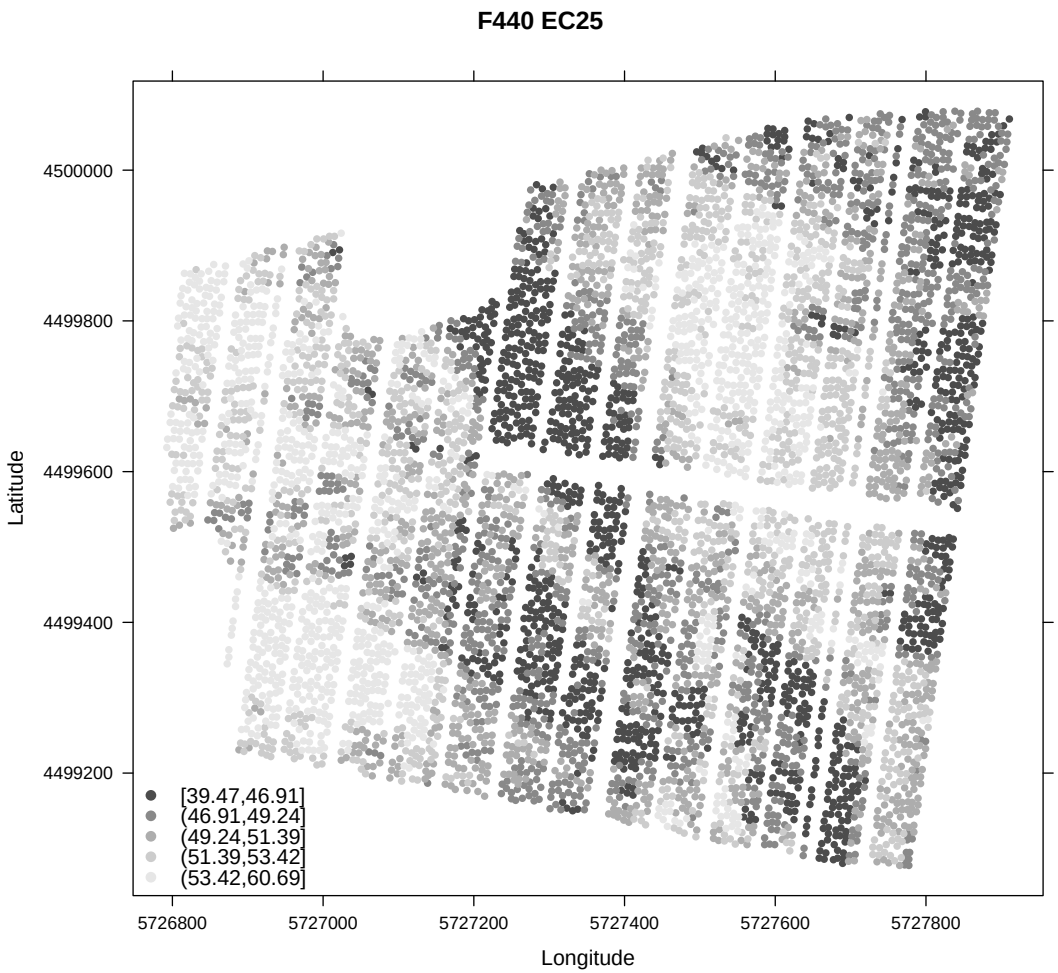


Abbildung: Scheinbare Bodenleitfähigkeit für Feld F440

Kartierung

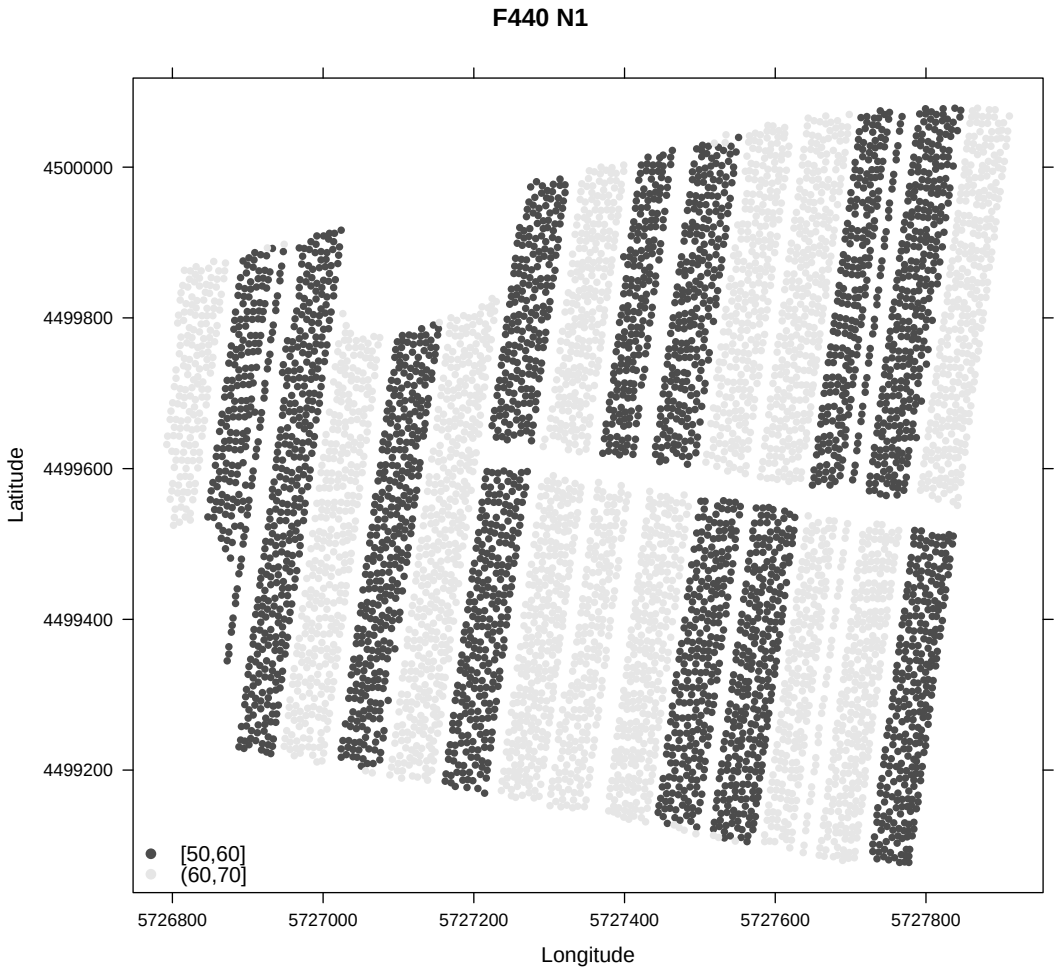
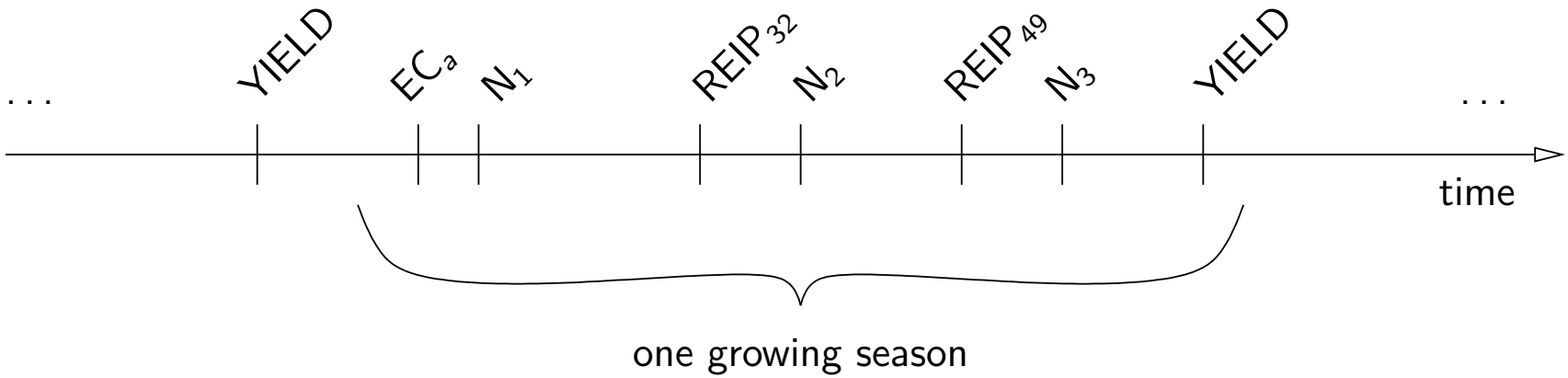


Abbildung: Stickstoffdünger (N1) für Feld F440

Gliederung

Ertragsvorhersage



- Versuche, den diesjährigen Ertrag basierend auf den vorhandenen Datenattributen vorherzusagen!
- Ist das ein klassisches Regressionsproblem?

Ertragsvorhersage:

nicht-räumlicher Fall

- Standardansatz, Lernen mit Kreuzvalidierung:
  - Datenaufteilung in Lern-, (Validierungs-), Testdatensatz
  - Modell wird auf Lernstichprobe trainiert
  - Validierungsstichprobe bestimmt das Ende des Lernprozesses
  - Modellfehler auf Testdatensatz wird berechnet
  - benutzte Modelle: Neuronale Netze, Support Vector Regression, Regressionsbaum, Random Forests, etc.

Ertragsvorhersage:

nicht-räumlicher Fall (Fortsetzung)

- allerdings: wegen des Vorliegens räumlicher Autokorrelation existieren sehr wahrscheinlich (bzw. zwangsläufig) sehr ähnliche Datenpunkte in Lern- und Testdatensatz
- das verletzt die Annahme der statistischen Unabhängigkeit der Datenpunkte
- Modellfehler wird (stark) unterschätzt
- $\longrightarrow$ benutze räumliche Kreuzvalidierung

Ertragsvorhersage

räumlicher Fall:

- Erweitere den Standardansatz auf räumliche Daten:
  - teile Daten in räumliche Teilmengen (Teilflächen)
  - benutze Standard-Ansatz auf diesen räumlichen Teilflächen
  - Methode zur Teilflächengenerierung: k-Means auf Punktkoordinaten
  - Regressionsmodelle: wie vorher
- statistisch gültiger Ansatz zur Ertragsvorhersage
- zusätzlich: räumliche Kartierung des Vorhersagefehlers ist möglich
  - finde ungewöhnliche Teilflächen des Feldes
  - finde bisher unbekannte Zusammenhänge in den Daten

Kartierung

Abbildung: Räumliche Kreuzvalidierung, 50 Teilflächen, Support Vector Regression

Ergebnisse: Unterschätzung des Modellfehlers

		F440		F611	
	k	spatial	non-spatial	spatial	non-spatial
SVR	10	1.06	0.54	0.73	0.40
	20	1.00	0.54	0.71	0.40
	50	0.91	0.53	0.67	0.38
RegTree	10	1.09	0.56	0.69	0.40
	20	0.99	0.56	0.68	0.42
	50	0.91	0.55	0.66	0.40
RandForest	10	0.99	0.50	0.65	0.41
	20	0.92	0.50	0.64	0.41
	50	0.85	0.48	0.63	0.39
Bagging	10	1.09	0.59	0.66	0.42
	20	1.01	0.59	0.66	0.42
	50	0.94	0.58	0.65	0.41

Tabelle: Regressionsvorhersagefehler, Vergleich zwischen räumlicher (spatial) und nicht-räumlicher (non-spatial) Behandlung der Daten

Gliederung

Interessantheit von Attributen

Wie nützlich sind zusätzlich eingeführte Sensordaten im Hinblick auf die Ertragsvorhersage?

- Fragestellung taucht bei der Entwicklung neuer Sensoren oder Datenquellen auf
- neue Datenquellen sollen so schnell wie möglich evaluiert werden

Interessantheit von Attributen

- nutze das entwickelte räumliche Ertragsvorhersagemodell:
  - im Zusammenhang mit Standard-Ansätzen zur Merkmalsauswahl (*feature selection*)
  - im Zusammenhang mit Ansätzen zur Variablen-Vertauschung (*variable permutation*)

Management-Zonen

Gibt es Teilflächen, die aufgrund der Datenlage besonders hervorstechen? Gibt es dort bisher unbekannte Zusammenhänge zwischen Datenattributen?

- enger Zusammenhang mit dem Begriff der "Management-Zone", der in der Landwirtschaft lange genutzt wird
- existierende Ansätze vernachlässigen räumliche Datenkomponente (keine zusammenhängenden Teilflächen)
- bisherige Zonen sind typischerweise statisch – Änderungen während der Vegetationsphase werden nicht berücksichtigt.

Management-Zonen

- Schritt 1: schrittweiser Aufbau von zusammenhängenden Teilflächen auf der Grundlage der räumlichen Cluster
- Schritt 2: Beobachtung der Veränderung von Zonen während der Saison

Management-Zonen

- teile Feld in Teilflächen (k-means, wie vorher)
- bestimme Nachbarschaft von Teilflächen
- verschmilz benachbarte (ähnliche) Teilflächen
- stoppe, wenn festgelegte Anzahl von Zonen erreicht ist
- werte entstandene Zonen aus

Management-Zonen

Aufteilung in Teilflächen (k-means)

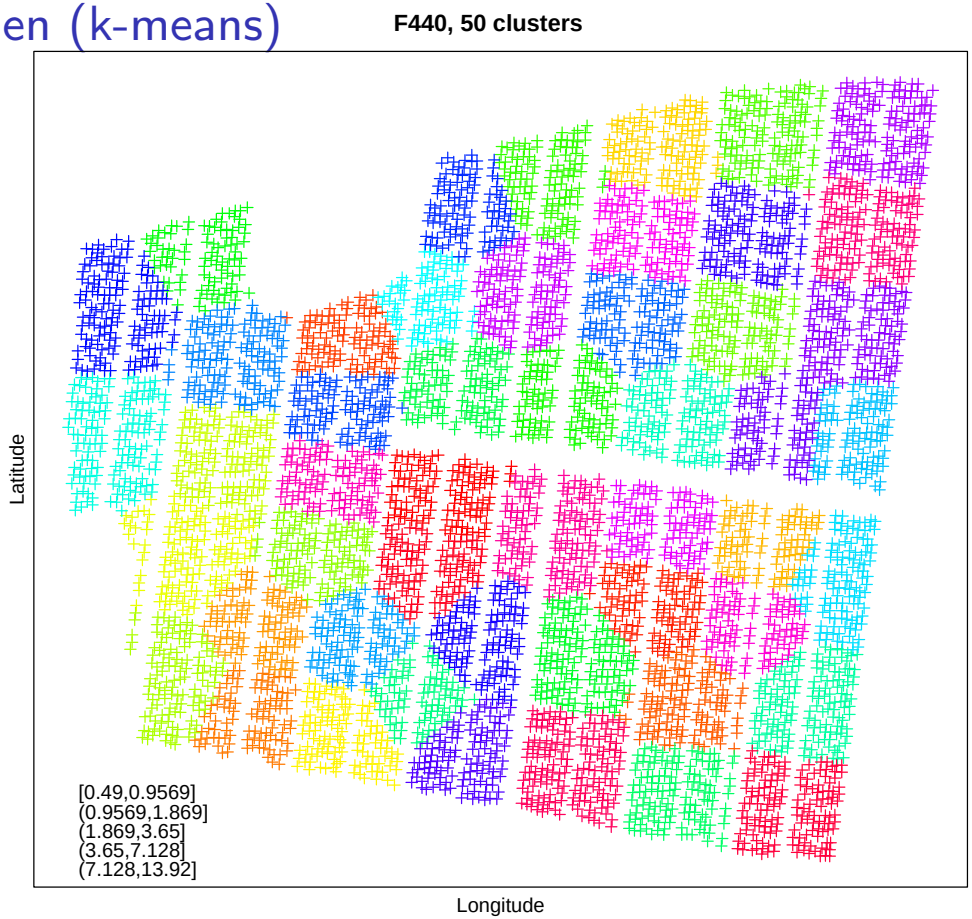


Abbildung: Aufteilung von F440 in 50 räumliche Cluster (k-means auf Datenpunktkoordinaten)

Management-Zonen

Verschmelzen

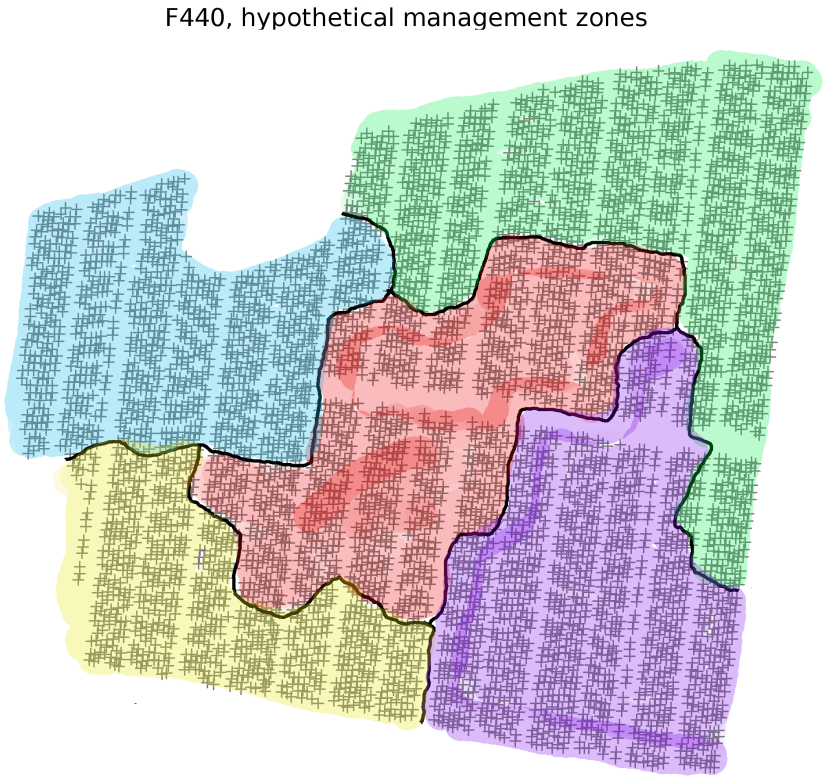


Abbildung: Vorschlag: konsekutives Verschmelzen der Teilflächen zu einer festgelegten Anzahl von Zonen

Gliederung

Zusammenfassung

- Die Unterscheidung zwischen räumlichen und nicht-räumlichen Daten ist sehr wichtig.
- schlecht: Standard-Methoden zur Datenanalyse können nicht ohne weiteres auf räumlichen Daten angewandt werden
- gut: Modifikationen in der Datenauswahl (*sampling*) ermöglichen die Beibehaltung klassischer Methoden.
- Landwirtschaft wird mehr und mehr datengetrieben sein.
- für Data Miner gibt's hier jede Menge unbeackerter Felder

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.

Waldo Tobler.

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

Abschluß...

Fragen & Antworten

- Research Blog: `http://research.georgruss.de/`
- "Data Mining in Agriculture" Workshop 2010: `http://dma2010.de/`