Spatial Data Mining in Precision Agriculture

Martin-Luther-Universität Halle-Wittenberg

Georg Ruß, Otto-von-Guericke-Universität Magdeburg

01.03.2011

Data Mining

- Ausgangspunkt: große Datenmengen
  - Anwendung intelligenter Datenanalyse
  - Entwicklung und Einsatz neuer Algorithmen
  - Generierung von neuem Wissen bzw. neuen Informationen
- Einsatz im Precision Farming (PF):
  - große Datenmengen liegen vor (zunehmend)
  - Daten haben räumliche Struktur
  - Beantwortung praktischer Fragestellungen durch Data Mining

Feld F440

Figure: F440 bei Görzig, Quelle: Google Earth mit Overlay

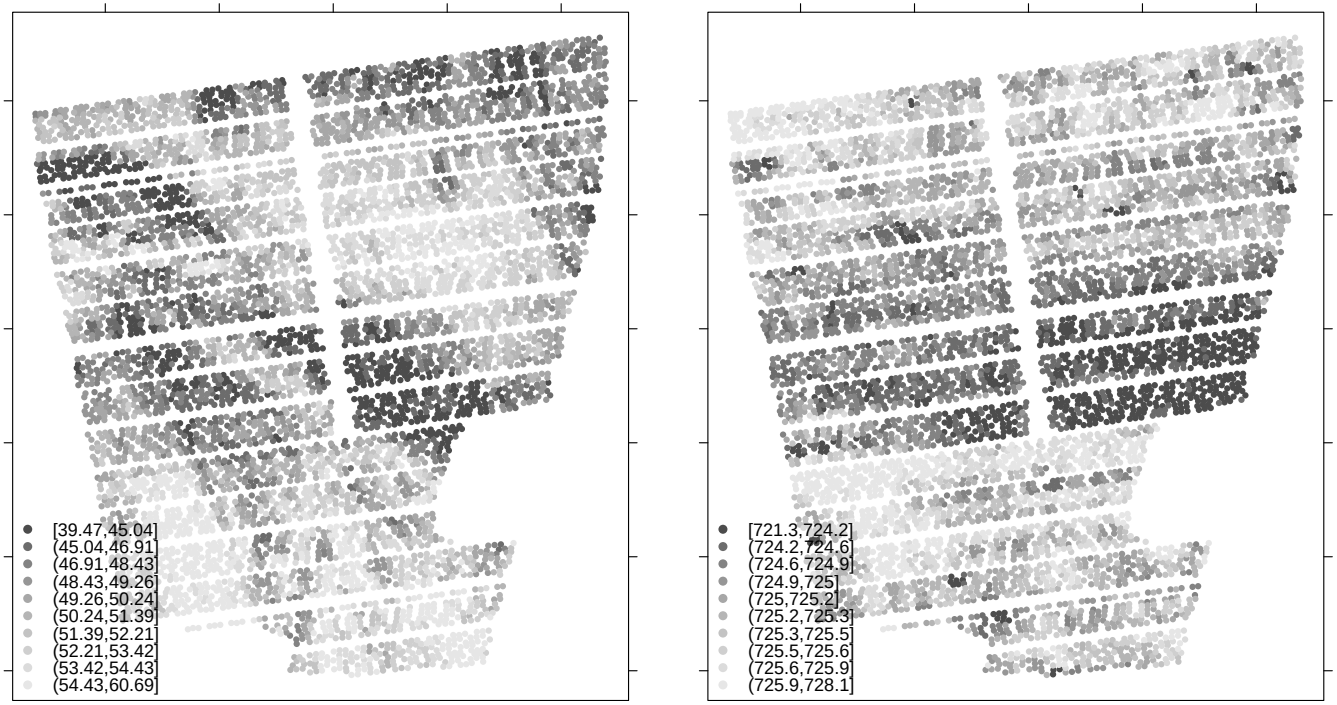
Fragestellungen in meiner Dissertation

- Ertragsvorhersage und Bedeutung einzelner Variablen
  - Betrachtung als Regressionsproblem
  - Notwendigkeit zur räumlichen Kreuzvalidierung bei Einsatz von Regressionsmodellen
  - Nutzung als Vehikel zur Bestimmung der Bedeutung einzelner Variablen (Sensoren, Datenquellen)
- Management-Zonen
  - Probleme existierender Algorithmen
  - Wünschenswerte Eigenschaften eines neuen Algorithmus
  - Hierarchisches Clustering als explorativer Ansatz

Ertragsvorhersage

- basiert auf Georg Weigerts Dissertation:
  - Einsatz neuronaler Netze (zur Regression)
  - Einsatz von Kreuzvalidierung
  - räumlicher Bezug der Daten nicht beachtet
  - Probleme mit systematischer Fehlerunterschätzung
- zugrundeliegender Umstand: räumliche Autokorrelation
  - räumlich benachbarte Datenpunkte sind häufig ähnlich
  - bei nicht-räumlicher Kreuzvalidierung landen räumlich benachbarte Punkte im Trainings- und Testdatensatz
  - führt zur Unterschätzung des Vorhersagefehlers, da das Regressionsmodell einige Punkte im Testdatensatz schon im Trainingsdatensatz "gesehen" hat
- Entwicklung einer räumlichen Kreuzvalidierung

Räumliche Autokorrelation



(a) EC25 (b) REIP32

Figure: F440, EC25/REIP32

Räumliche Kreuzvalidierung

Figure: Räumliche Zerlegung von F440 mit Hilfe von k-means, $k = 20$

Räumliche Variablenbedeutung

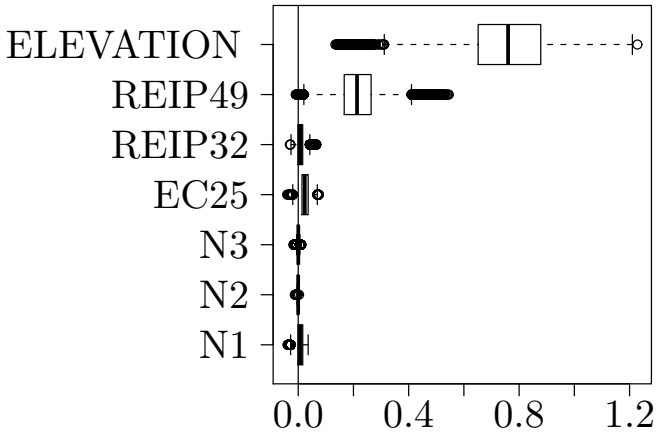
(engl. spatial variable importance)

- ausgehend vom Ertragsvorhersagemodell:
  - bestimme den Einfluß einzelner Variablen (Sensoren, Datenquellen)
  - Wie “gut” ist ein neuer Sensor zur Ertragsvorhersage geeignet?
  - Welche “neuen” Informationen liefert eine zusätzliche Datenquelle?

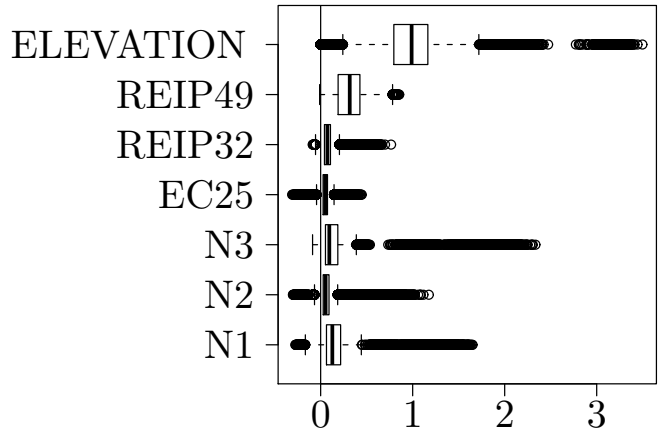
Räumliche Variablenbedeutung

- relativ einfache Grundidee:
  - wie bisher: räumliche Kreuzvalidierung
    - Modelltraining auf Trainingsdatensatz
    - Vorhersage auf Testdatensatz
    - Bestimmung des Fehlers als Differenz (RMSE) aus Modellvorhersage und tatsächlichem Wert
  - zusätzlich:
    - (wiederholte) Permutation einer Variable im Testdatensatz
    - Aufzeichnen des Einflusses dieser Permutation auf die Vorhersagegüte des trainierten Regressionsmodells
    - wenn Fehlerzunahme: Variable ist (im Modell) wichtig
    - keine/kaum Änderung: Variable ist eher unwichtig

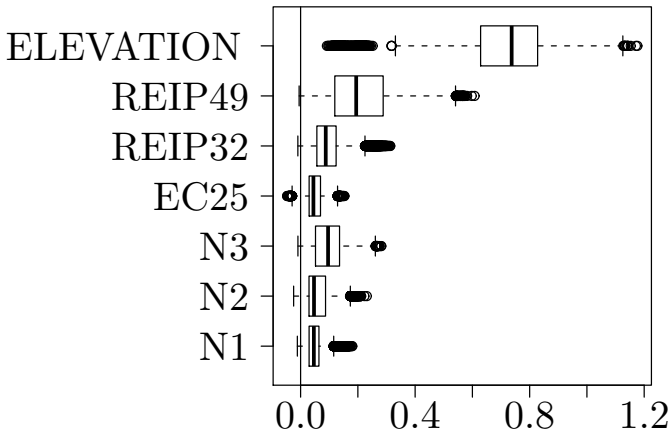
Räumliche Variablenbedeutung: vorläufige Ergebnisse



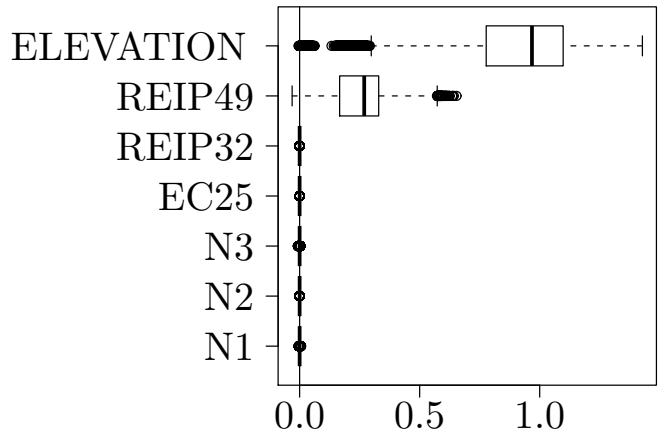
(a) lm



(b) neural network



(c) support vector regression, radial kernel



(d) bagging

Management-Zonen

engl. management zone delineation

- Grundidee: Bestimmung von Teilbereichen des Feldes, die unterschiedlich behandelt ("gemanagt") werden
- bisher eher heuristisch und ad-hoc behandelt, nicht unbedingt mit Rücksicht auf volle Ausnutzung der Daten
- PF-Behandlung etwa seit dem Jahr 2000, laut Literatur
- kein tatsächliches Qualitätsmaß vorhanden
- Zonen (zwangsläufig) abhängig vom Verwendungszweck
- **daher:** Entwicklung eines explorativen Ansatzes aus Datensicht mit Ausnutzung der räumlichen Struktur der Daten

Management-Zonen als Clusterproblem

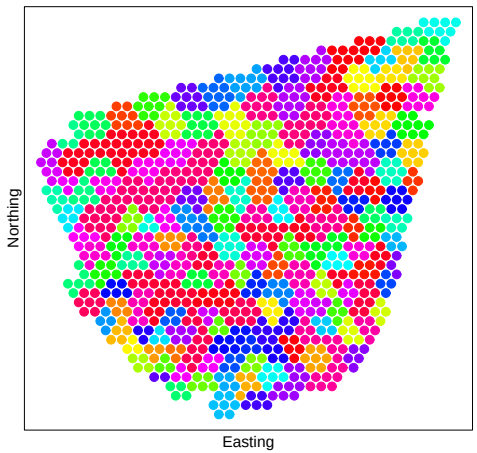
- aus Datensicht:
  - Zerlegung des Datenraums (des Feldes) in disjunkte Teilbereiche = Clusterproblem
  - Problem: zwei Datenräume
    - Geo-Raum: x-y-z-Koordinaten der Datenpunkte
    - Merkmalsraum: Sensordaten, Datenquellen, Meßwerte (EC, N, REIP, ...)
  - existierende Algorithmen sind immer auf nur einen der beiden Räume ausgelegt
  - Einbeziehung vieler Variablen (nicht nur EC/EM)
- zusätzliche wünschenswerte Eigenschaft: einstellbarer räumlicher Zusammenhang der Zonen
  - führt zu Erkenntnisgewinn über Ähnlichkeiten von Feldbereichen (Sinn des Data Mining)
  - manueller, explorativer Prozeß (automatisierbar?)

Hierarchisches Clustern

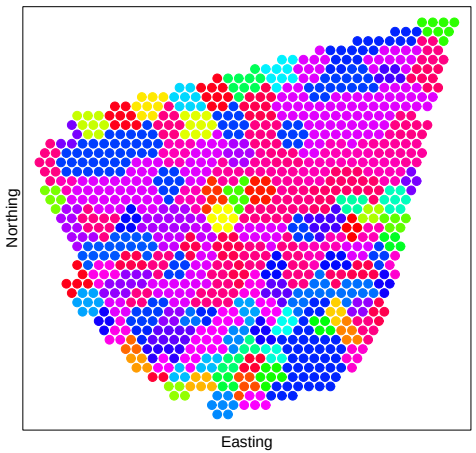
- Beginn: jeder Datenpunkt in einem einzelnen Cluster
- Zwischenschritte: Verschmelzen einzelner Datenpunkte zu einem neuen Cluster
- Ende: alle Datenpunkte in einem Cluster
- Kriterien zum Verschmelzen:
  - a) Ähnlichkeit im Geo-Raum (Cluster benachbart)
  - b) Ähnlichkeit im Merkmalsraum (ähnliche Feldeigenschaften)
- Parameter zur Einstellung der räumlichen Kontiguität

Management-Zonen, Beispiel

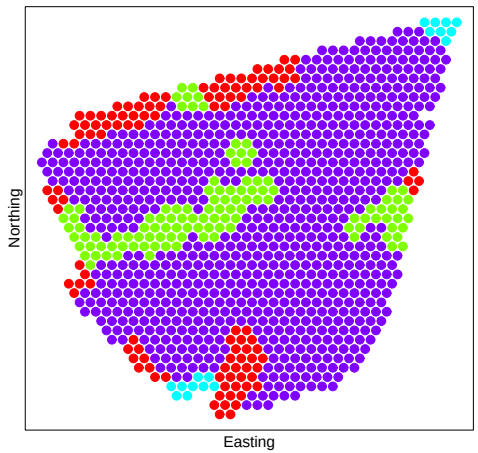
niedrige Kontiguität



(e) F550, 150 Cluster



(f) F550, 60 Cluster

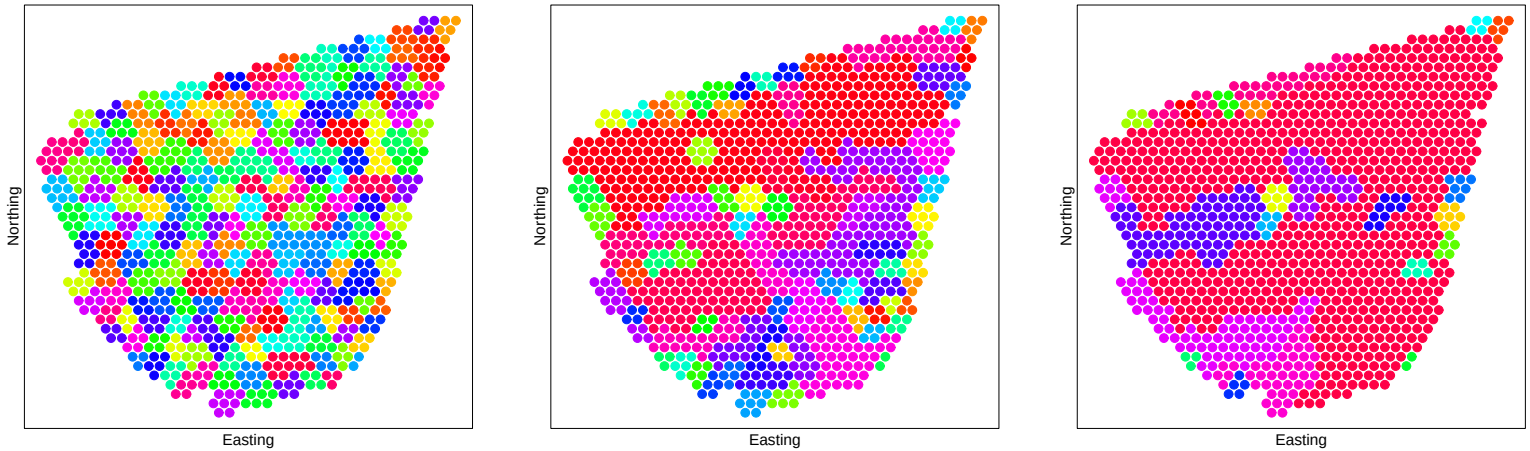


(g) F550, 4 Cluster

Figure: F550, Clustern mit niedriger Kontiguität, Variablen: pH, K, Mg, P (Farben nur zur Abgrenzung der Cluster, keine weitere Bedeutung)

Management-Zonen, Beispiel

hohe Kontiguität



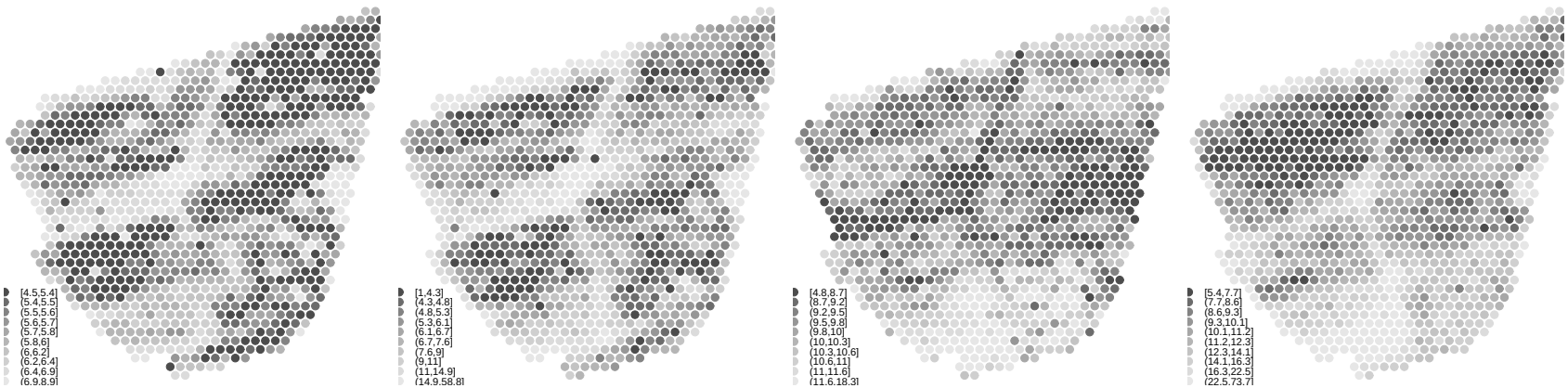
(a) F550, Beginn (b) F550, 75 Cluster (c) F550, 22 Cluster

Figure: F550, Clustern mit hoher Kontiguität, Variablen: pH, K, Mg, P
(Farben nur zur Abgrenzung der Cluster, keine weitere Bedeutung)

(Demo: Videovergleich)

Management-Zonen

Vergleich der Zonen mit Bodenattributen



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

Figure: F550, Variablen

Management-Zonen

Vergleich der Zonen mit Bodenattributen

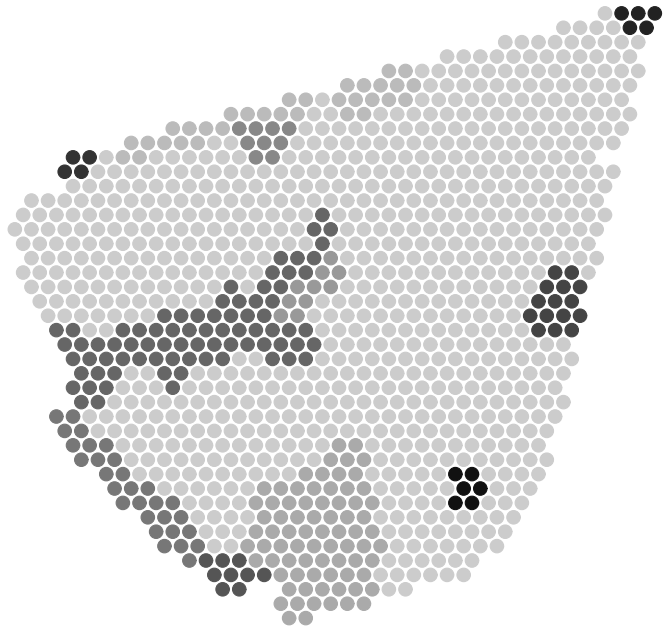


Figure: F550, elf Management-Zonen laut Algorithmus

- letztendlich nur 3 Zonen:
  - pH niedrig, P niedrig, Mg niedrig, K niedrig (größte Zone)
  - pH hoch, P hoch, Mg hoch, K hoch (Ränder)
  - pH hoch, P hoch, Mg niedrig, K hoch (Mitte links)

Zusammenfassung

- *precision farming* ist ein datengetriebener Ansatz für zukünftige Landwirtschaft
- räumliche, georeferenzierte, große Datenmengen
- Ertragsvorhersage als Vehikel zur Bestimmung der räumlichen Variablenbedeutung
- Management-Zonierung per räumlichem Clusteralgorithmus

Zeit für . . .

Fragen?

- georg.russ@ovgu.de
- Folien, Veröffentlichungen, Skripte unter http://research.georgruss.de