Spatial Data Mining in Precision Agriculture

Dissertationsverteidigung

Georg Ruß

Otto-von-Guericke-Universität Magdeburg

23.02.2012

Gliederung

Titel und Themen

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

- 1 – Präzisionslandwirtschaft
  - Anwendung modernster Technologie in der Landwirtschaft
  - kleinräumige, teilflächenspezifische, datenbasierte Landwirtschaft

Titel und Themen

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

- 1 – Präzisionslandwirtschaft
  - Anwendung modernster Technologie in der Landwirtschaft
  - kleinräumige, teilflächenspezifische, datenbasierte Landwirtschaft
- 2 – Data Mining
  - Algorithmen, Methoden und Ideen, um in Daten zu schürfen:
  - finde neues, interessantes und nützliches Wissen in Daten

Titel und Themen

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

- 1 – Präzisionslandwirtschaft
  - Anwendung modernster Technologie in der Landwirtschaft
  - kleinräumige, teilflächenspezifische, datenbasierte Landwirtschaft
- 2 – Data Mining
  - Algorithmen, Methoden und Ideen, um in Daten zu schürfen:
  - finde neues, interessantes und nützliches Wissen in Daten
- 3 – Räumliche Daten
  - fallen zunehmend in Landwirtschaft und Umweltwissenschaften an
  - räumliche Komponente der Daten muss beachtet werden!

Beispielfeld F440

Abbildung: F440 bei Köthen, Quelle: Google Earth w/ Overlay

Daten, Daten, Daten

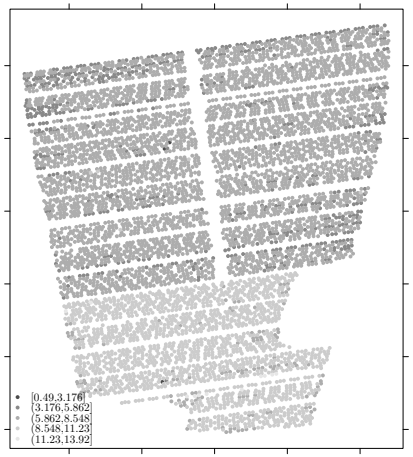
- Ertrag, Düngemengen
- Fernerkundung (Luftbilder, Satellitenfotos, versch. Spektralbereiche)
- Vegetationssensoren (REIP32, REIP49, etc.)
- Geophysikalische Daten (Bodenleitfähigkeit EC25, . . .)
- Bodenproben (pH, K, P, Mg, . . .)
- Digitale Höhenmodelle und Ableitungen daraus (Anstieg, Krümmung, Exposition, Feuchtigkeitsindizes, . . .)

Daten, Daten, Daten

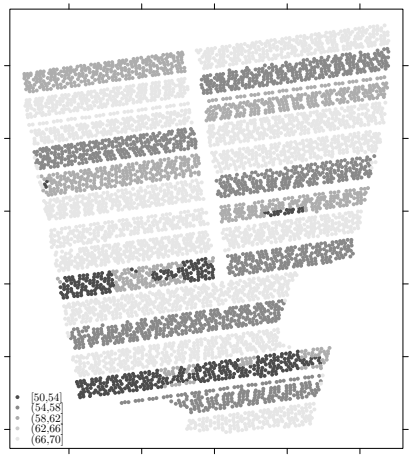
- Ertrag, Düngemengen
- Fernerkundung (Luftbilder, Satellitenfotos, versch. Spektralbereiche)
- Vegetationssensoren (REIP32, REIP49, etc.)
- Geophysikalische Daten (Bodenleitfähigkeit EC25, . . .)
- Bodenproben (pH, K, P, Mg, . . .)
- Digitale Höhenmodelle und Ableitungen daraus (Anstieg, Krümmung, Exposition, Feuchtigkeitsindizes, . . .)

- $\rightarrow$ Hochauflösende georeferenzierte Datensätze entstehen
- $\rightarrow$ Nutze Data Mining, zum Beispiel für Regressions-, Klassifizierungs- oder Optimierungsaufgaben.

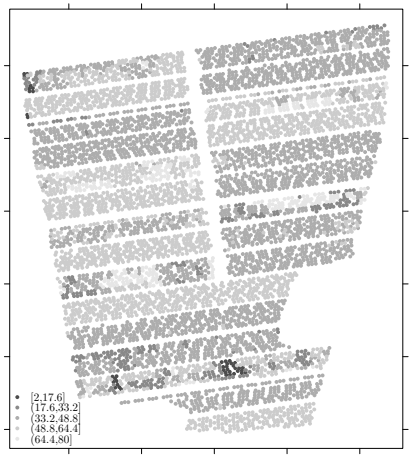
Beispielfeld F440 (fortg.)



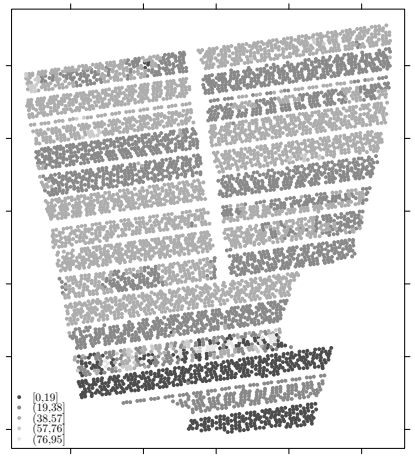
(a) Ertrag



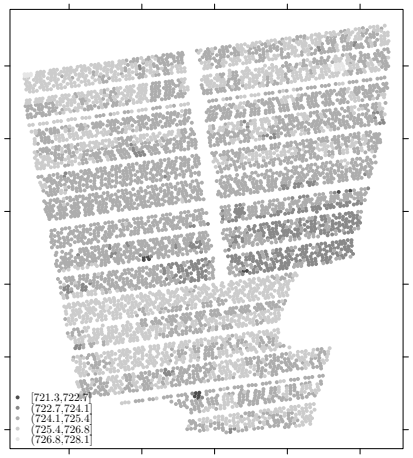
(b) N1



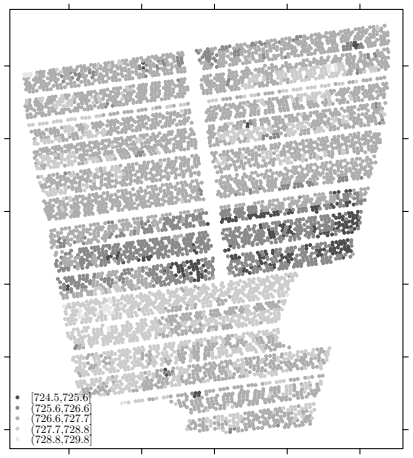
(c) N2



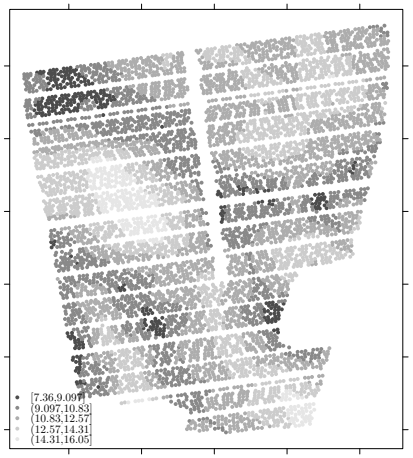
(d) N3



(e) REIP32



(f) REIP49



(g) Feuchteindex

(h) Krümmung

Abbildung: Ertrag und einige Variablen (F440)

Gliederung

Aufgabe: Ertragsvorhersage und Variablenwichtigkeit

Ertragsvorhersage mit Hilfe von Regression

- Abriss: Ertragsvorhersage aus anderen Variablen (ex-post)
  - knüpft an existierende Arbeit an
  - Ertragsvorhersage als (nicht-)lineare multivariate Regressionsaufgabe
  - Fortführung neuronaler Netze
  - Erweiterung um weitere Modelle
  - Probleme mit nicht-räumlicher Modellierung auf räumlichen Daten

Aufgabe: Ertragsvorhersage und Variablenwichtigkeit

Ertragsvorhersage mit Hilfe von Regression

- Abriss: Ertragsvorhersage aus anderen Variablen (ex-post)
  - knüpft an existierende Arbeit an
  - Ertragsvorhersage als (nicht-)lineare multivariate Regressionsaufgabe
  - Fortführung neuronaler Netze
  - Erweiterung um weitere Modelle
  - Probleme mit nicht-räumlicher Modellierung auf räumlichen Daten
- Regressionsmodelle
  - linear (lm), generalized additive (gam)
  - Regressionsbaum (rt), Bagging (bag)
  - Neuronales Netz (net)
  - Support Vector Regression (svr, e1071)
  - K-Nearest Neighbor (kknn)

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Ablauf: Regression und Kreuzvalidierung, generisch

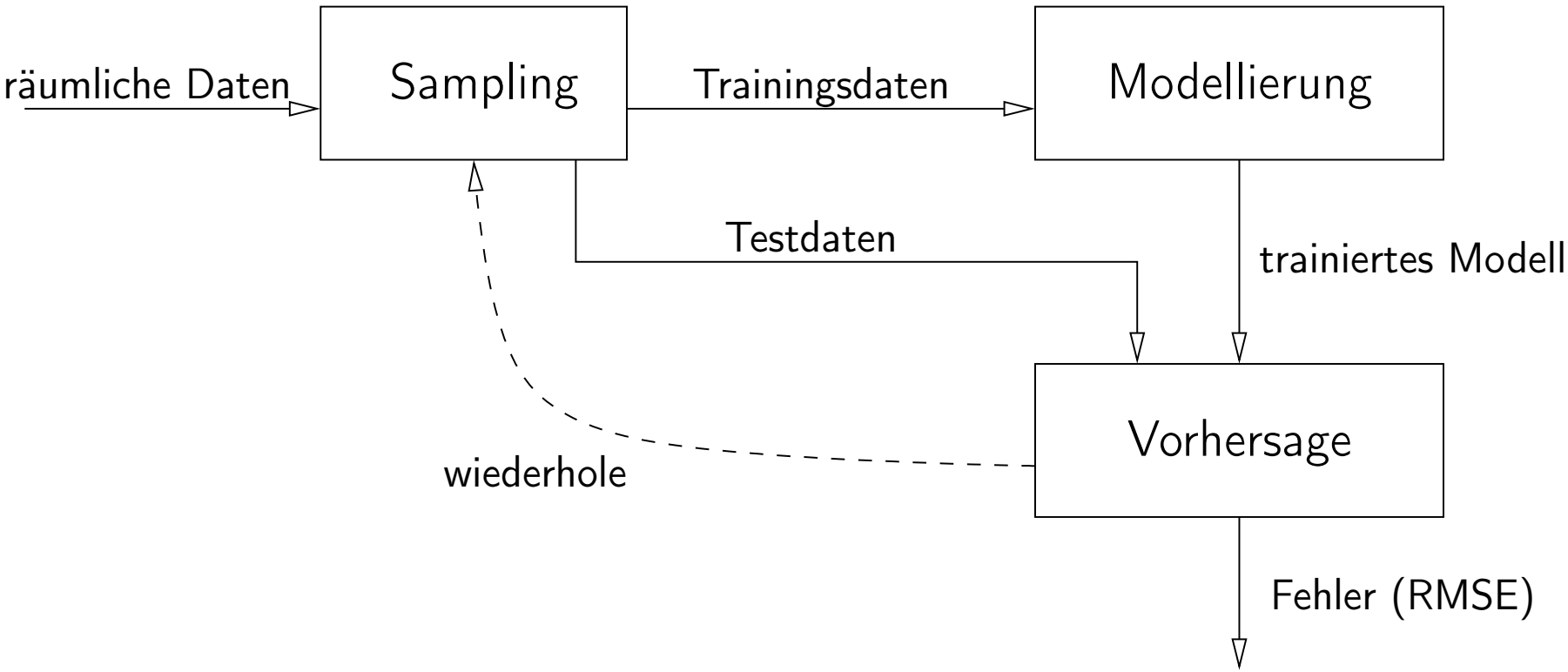


Abbildung: Kreuzvalidierung

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Nicht-räumliches Sampling auf räumlichen Daten

- Problem: räumliche Autokorrelation
- geographisch benachbarte (und damit wahrscheinlich ähnliche) Datenpunkte landen in Trainings- und Testdaten
- Verletzung der Unabhängigkeitsannahme der Kreuzvalidierung
- Systematische Fehlerunterschätzung

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Nicht-räumliches Sampling auf räumlichen Daten

- Problem: räumliche Autokorrelation
- geographisch benachbarte (und damit wahrscheinlich ähnliche) Datenpunkte landen in Trainings- und Testdaten
- Verletzung der Unabhängigkeitsannahme der Kreuzvalidierung
- Systematische Fehlerunterschätzung

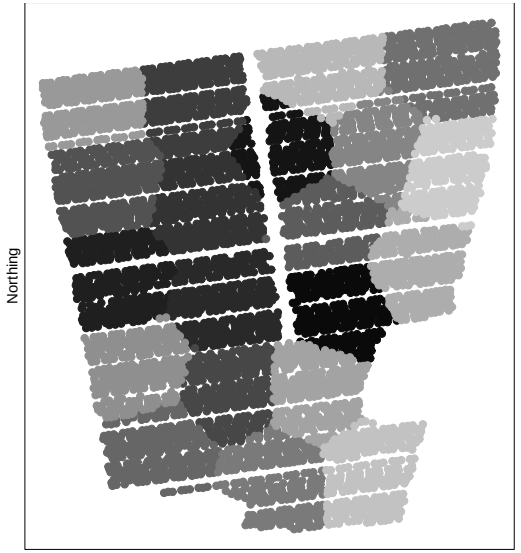


Abbildung: 20 räumliche Cluster

- $\rightarrow$ Räumliches Sampling mittels räumlichem Clustern:
  - k-Means-Clustering
  - wähle zufällig bspw. 90/10% der Cluster für Training/Test
  - mehr Cluster $\rightarrow$ Richtung nicht-räumliches Sampling

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Ablauf: räumliche Kreuzvalidierung

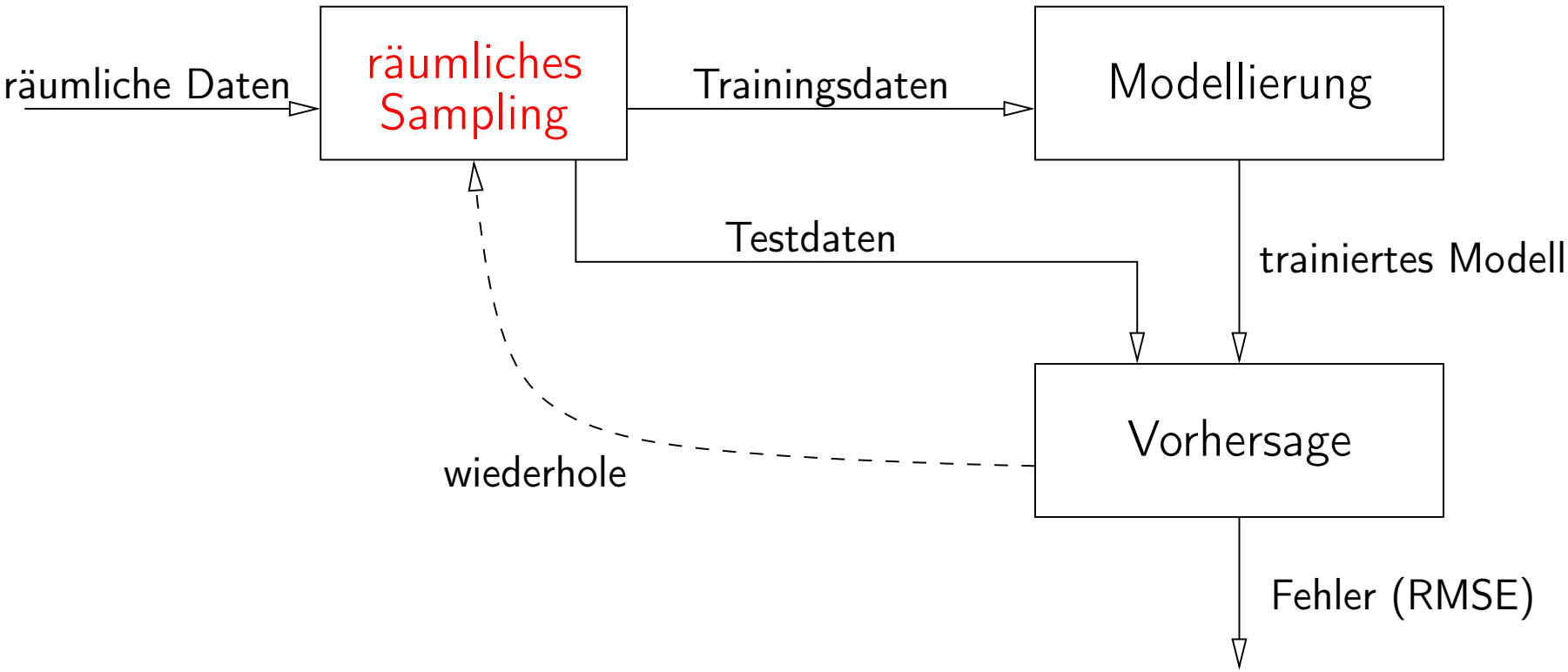


Abbildung: Kreuzvalidierung, räumlich

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Variablenwichtigkeit

- Frage:
  - Welchen Einfluss hat eine einzelne Variable auf die Vorhersagegüte?
  - = Trägt ein neuer Sensor wirklich neue Informationen bei?
  - = Sollte man weitere Datenquellen in Erwägung ziehen?

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Variablenwichtigkeit

- Frage:
  - Welchen Einfluss hat eine einzelne Variable auf die Vorhersagegüte?
  - = Trägt ein neuer Sensor wirklich neue Informationen bei?
  - = Sollte man weitere Datenquellen in Erwägung ziehen?
- Idee: Permutiere diese Variable im *Testdatensatz*!
  - Wenn der Fehler steigt, ist die Variable für dieses Modell wichtig
  - Wichtigkeit im Zusammenspiel mit anderen Variablen
  - unabhängig vom genutzten Regressionsmodell

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Variablenwichtigkeit

- Frage:
  - Welchen Einfluss hat eine einzelne Variable auf die Vorhersagegüte?
  - = Trägt ein neuer Sensor wirklich neue Informationen bei?
  - = Sollte man weitere Datenquellen in Erwägung ziehen?
- Idee: Permutiere diese Variable im *Testdatensatz*!
  - Wenn der Fehler steigt, ist die Variable für dieses Modell wichtig
  - Wichtigkeit im Zusammenspiel mit anderen Variablen
  - unabhängig vom genutzten Regressionsmodell
- Nebenbemerkung: wiederhole hinreichend oft ...
  - zufälliges räumliches Sampling
  - Modell-Training
  - Permutation

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Ablauf: Variablenwichtigkeit

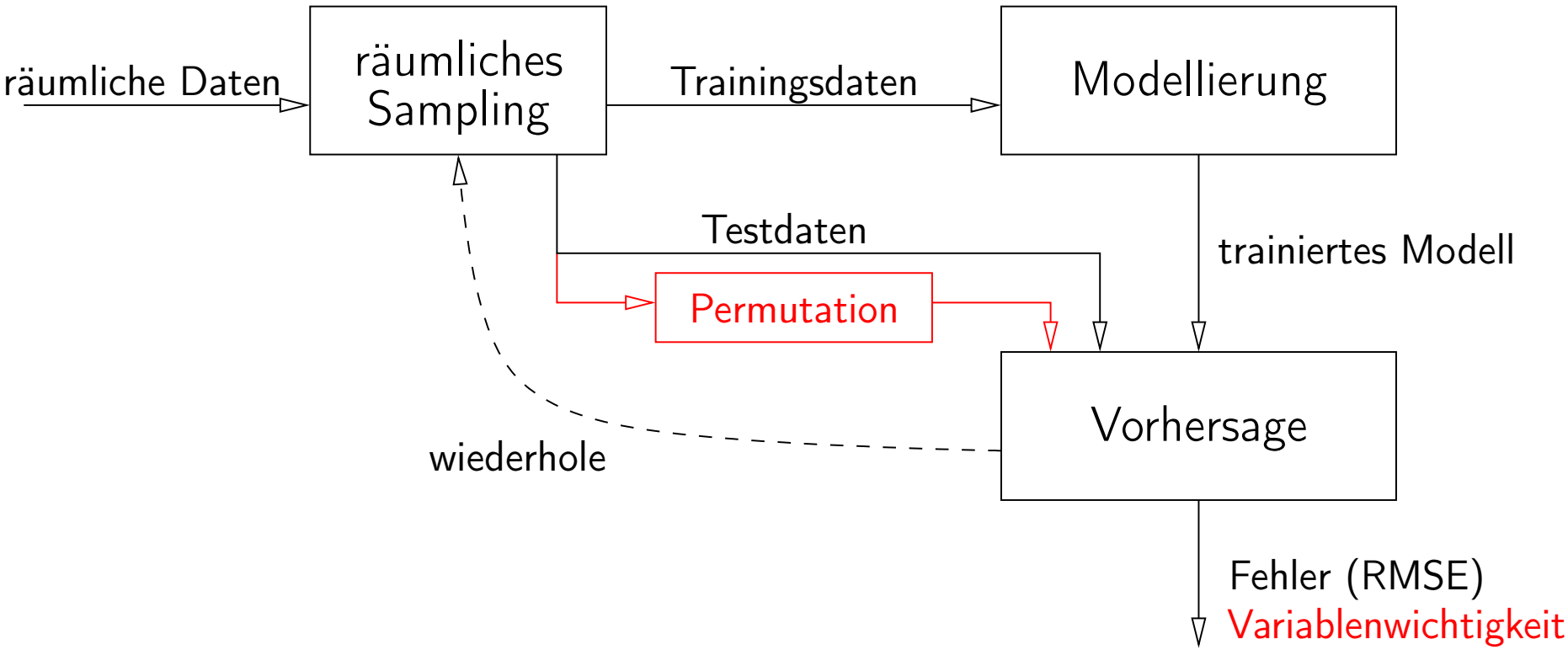


Abbildung: Räumliche Variablenwichtigkeit

Aufgabe: Variablenwichtigkeit und Ertragsvorhersage

Ergebnisse

- Variablenwichtigkeit:
  - Vegetationsindizes REIP32 & REIP49 am wichtigsten, unabhängig von Modell und Feld
  - Bodenleitfähigkeit relativ wichtig (im Zusammenspiel)
  - Digitales Höhenmodell potentiell wichtig (Feuchteindex, Krümmung, Anstieg)
  - Weitere Ergebnisse variieren zwischen Feldern
- weiteres Ergebnis:
  - Vergleich zwischen einzelnen Düngestrategien und deren Variablenwichtigkeit
  - Vergleich der Modelle (Gewinner: bagging/svr, Verlierer: net)

Gliederung

Management-Zonen

- 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
- Behandlung in der Präzisionslandwirtschaft etwa seit dem Jahr 2000, laut Literatur

Management-Zonen

- 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
- Behandlung in der Präzisionslandwirtschaft 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 exploratives Clusterproblem (1)

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, YIELD, ...)
- existierende Algorithmen sind fast ausschließlich auf nur einen der beiden Räume ausgelegt

Management-Zonen als exploratives Clusterproblem (2)

- zusätzliche wünschenswerte Eigenschaft: beeinflussbarer räumlicher Zusammenhang der Zonen
  - Kompromiss: Clusterähnlichkeit $\leftrightarrow$ räumlicher Zusammenhang
  - manueller, explorativer Prozess
  - führt zu Erkenntnisgewinn über Ähnlichkeiten von Feldbereichen (Sinn des Data Mining)

Hierarchisches Clustern

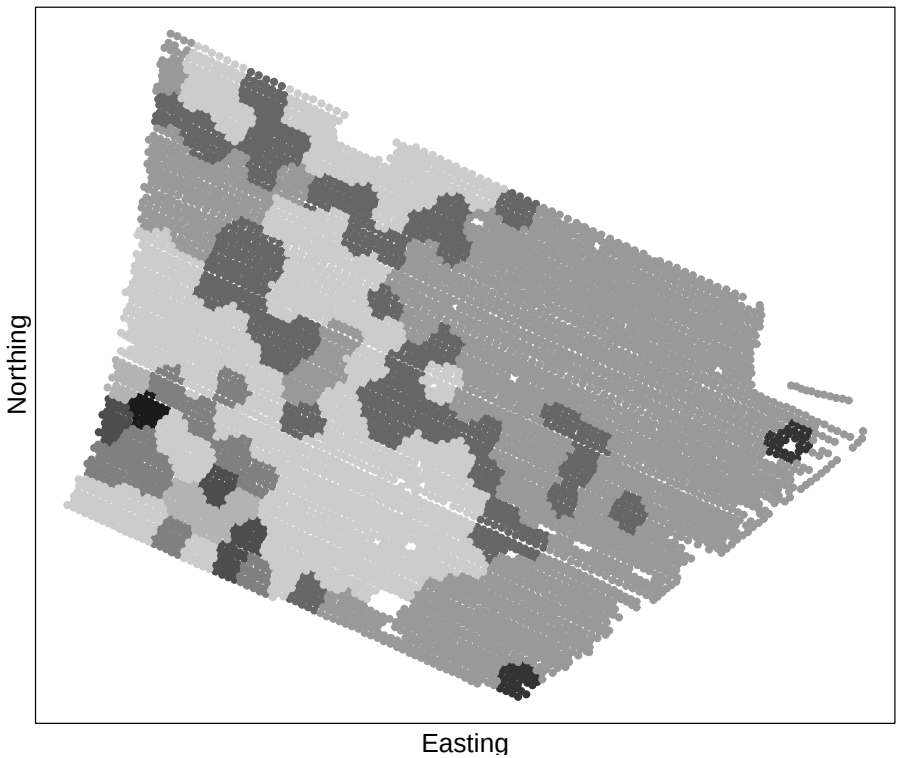
- Beginn: jeder Datenpunkt in einem einzelnen Cluster
- Zwischenschritte: Verschmelzen einzelner Datenpunkte/Cluster zu einem neuen Cluster
- Ende: alle Datenpunkte in einem Cluster

Hierarchisches Clustern

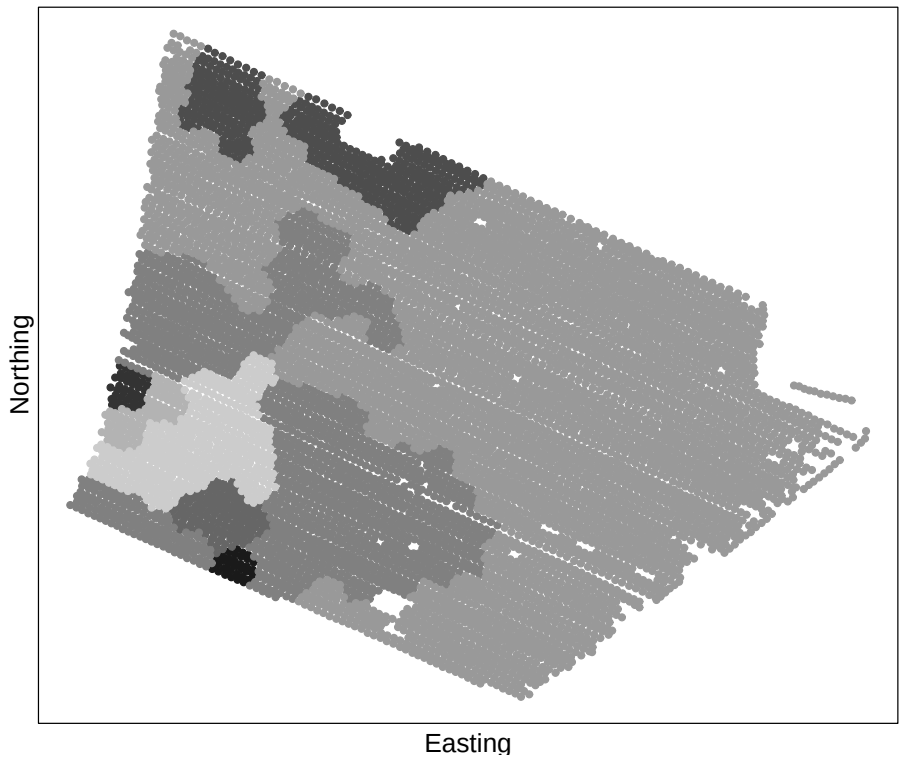
- Beginn: jeder Datenpunkt in einem einzelnen Cluster
- Zwischenschritte: Verschmelzen einzelner Datenpunkte/Cluster zu einem neuen Cluster
- Ende: alle Datenpunkte in einem Cluster
- Kriterien zum Verschmelzen:
  - a) Ähnlichkeit im Geo-Raum (Cluster nah beieinander)
  - b) Ähnlichkeit im Merkmalsraum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

Management-Zonen, Beispiel

Vergleich niedriger/hoher räumlicher Zusammenhang



(a) F631, HACC-spatial angewandt auf EC25, niedrige Kontiguität

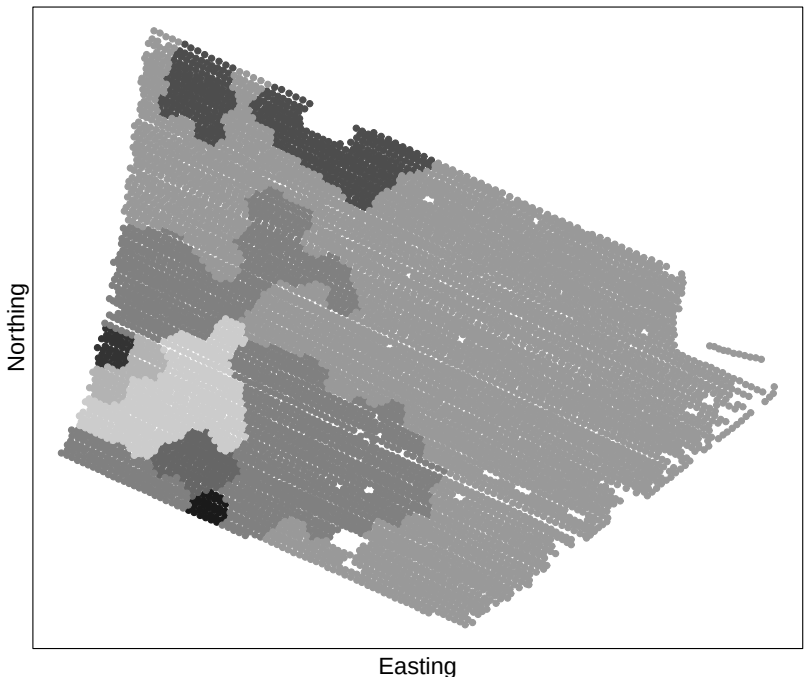


(b) F631, HACC-spatial angewandt auf EC25, hohe Kontiguität

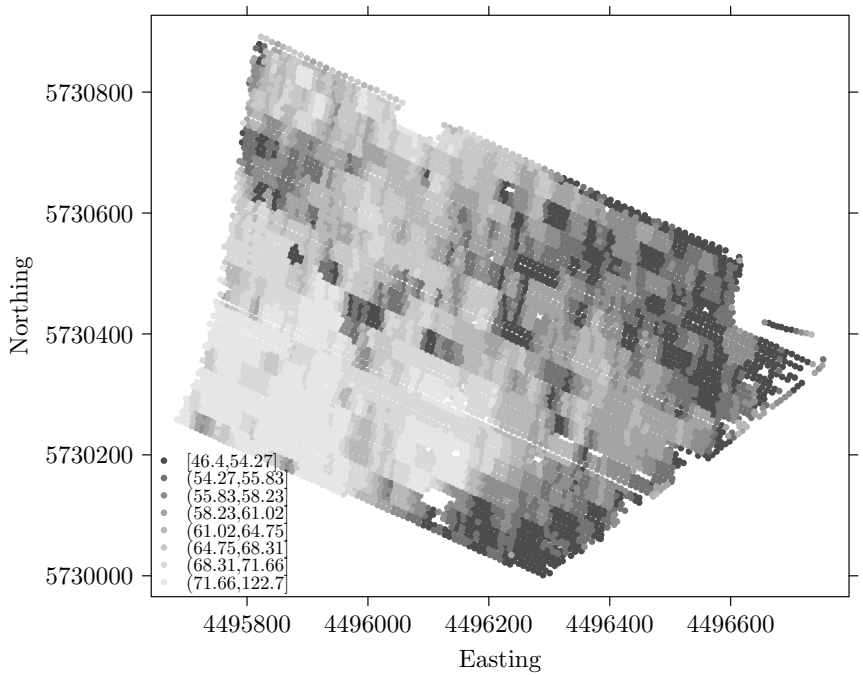
Abbildung: Vergleich zwischen niedrigem/hohen räumlichen Zusammenhang

Management-Zonen, Beispiel

Vergleich der Zonen mit EC25 auf F631



(a) F631, HACC-spatial angewandt auf EC25, hohe Kontiguität



(b) F631, Variable EC25

Abbildung: Vergleich zwischen Zonen und tatsächlicher Variable

Gliederung

Zusammenfassung (1)

- *precision agriculture* ist ein notwendiger Ansatz für zukünftige datengetriebene Landwirtschaft
- räumliche, hochaufgelöste, georeferenzierte große Datenmengen
- 1. Thema: Ertragsvorhersage als Vehikel zur Bestimmung der räumlichen Variablenwichtigkeit
  - Ertragsvorhersage als Regressionsmodell
  - Permutationsbasierte Variablenwichtigkeit
  - Anpassung der Vorgehensweise für räumliche Daten

Zusammenfassung (2)

- 2. Thema: Management-Zonierung per räumlichem Clusteralgorithmus HACC-spatial
  - hierarchisches agglomeratives Clustering für räumliche Daten
  - Einführung eines Kompromiss-Parameters zwischen Clusterähnlichkeit und räumlichem Zusammenhang
  - Derzeit experimenteller Einsatz von HACC-spatial:
    - am CIRAD (*Centre de Cooperation Internationale en Recherche Agronomique pour le Developpement*, Kontakt via R-sig-geo-Mailingliste)
    - am Department of Ecosystem Analysis, School of Forest Resources, University of Washington, Seattle (Research-Blog-Kontakt)

Ausblick

- Verfeinerung der Methodik
- Erweiterung um zusätzliche Daten
- Anwendung auf andere Umweltdaten:

Ausblick

- Verfeinerung der Methodik
- Erweiterung um zusätzliche Daten
- Anwendung auf andere Umweltdaten:

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.

Danksagung

- Felddaten: MLU Halle, Prof. Peter Wagner
- Digitales Höhenmodell: LVermGeo Sachsen-Anhalt, Magdeburg

Und jetzt: Das Wetter

- Das Wetter hätte natürlich einen entscheidenden Einfluss auf die Ertragsvorhersage!

Und jetzt: Das Wetter

- Das Wetter hätte natürlich einen entscheidenden Einfluss auf die Ertragsvorhersage!
- aber:
  1. Es kann nicht für einen langen Zeitraum mit hoher Sicherheit vorhergesagt werden.
  2. Es wirkt nicht direkt teilflächenspezifisch.
  3. Es ist indirekt in den Vegetationsindikatoren zu bestimmten Zeitpunkten enthalten und geht daher auch in die Regressionsmodelle mit ein.

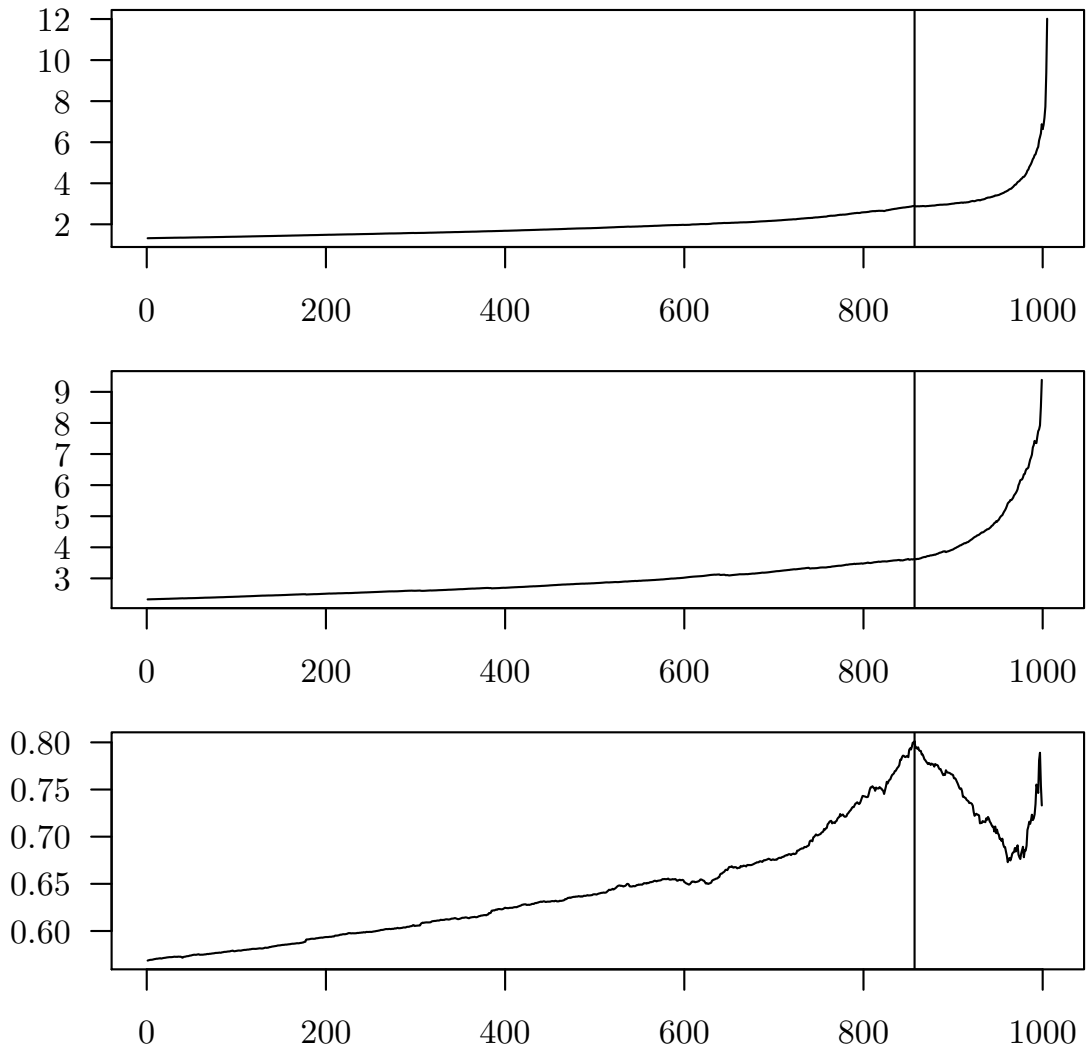
Räumlich vs. nicht-räumlich in Zahlen

exemplarisch an zwei Datensätzen

		F440		F611	
	k	spatial	non-spatial	spatial	non-spatial
Regression Tree	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
Supp. Vec. Regression	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
Bagging	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

Tabelle: Vergleich zwischen räumlicher und nicht-räumlicher Regression, RMSE-Werte. Der zahlenmäßige Unterschied zwischen *räumlich* und *nicht-räumlich* überwiegt bei weitem die Unterschiede zwischen einzelnen Modellen und verschiedenen Parametereinstellungen.

Kompromiss-Parameter bei HACC-spatial



Mittlere Distanzen im Merkmalsraum beim Clustering. Oben: nicht räumlich benachbarte Cluster, Mitte: räumlich benachbarte Cluster, Unten: Verhältnis der beiden Werte.

contiguity threshold: 0,8.
Bis zum Erreichen des Schwellwertes werden nur die ähnlichsten räumlich benachbarten Cluster verschmolzen. Danach werden nur die ähnlichsten verschmolzen, ohne die räumliche Einschränkung.

Veröffentlichungen

- Ertragsvorhersage mit NN: [RKWS08], [RKSW08a], [RKSW08b]
- Ertragsvorhersage weitere Regressionsmodelle: [RKSW10], [Ruß09], [RK10b]
- räumliche Variablenwichtigkeit: [RK10a], [RB10a], [RB10b]
- räumliches Clustering: [RKS10], [RS10], [RK11b], [RK11a], [Ruß11]

Veröffentlichungen I

[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.

[RK10a] Georg Ruß and Rudolf Kruse.
Feature selection for wheat yield prediction.
In Tony Allen, Richard Ellis, and Miltos Petridis, editors, *Research and Development in Intelligent Systems XXVI, Incorporating Applications and Innovations in Intelligent Systems XVII*, volume 26 of *Proceedings of AI-2009*, pages 465–478, London, January 2010. BCS SGAI, Springer.

[RK10b] Georg Ruß and Rudolf Kruse.
Regression models for spatial data: An example from precision agriculture.
In Petra Perner, editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 6171 of *LNAI*, pages 450–463, Berlin, Heidelberg, July 2010. Springer.

[RK11a] Georg Ruß and Rudolf Kruse.
Exploratory hierarchical clustering for management zone delineation in precision agriculture.
In Petra Perner, editor, *Proceedings of ICDM 2011*, volume 6870 of *LNAI*, pages 161–173, Berlin, Heidelberg, August 2011. Springer.

Veröffentlichungen II

[RK11b] Georg Ruß and Rudolf Kruse.
Machine learning methods for spatial clustering on precision agriculture data.
In Anders Kofod-Petersen, Fredrik Heintz, and Helge Langseth, editors, *Eleventh Scandinavian Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications, pages 40–49, Amsterdam, Netherlands, May 2011. IOS Press.

[RKS10] Georg Ruß, Rudolf Kruse, and Martin Schneider.
A clustering approach for management zone delineation in precision agriculture.
In Rajiv Khosla, editor, *Proceedings of ICPA 2010*, Denver, July 2010. International Society of Precision Agriculture.

[RKSW08a] 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 Boston, July 2008.

[RKSW08b] 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 the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-08)*, pages 576–582. University of Málaga, June 2008.

[RKSW10] Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner.
Using advanced regression models for determining optimal soil heterogeneity indicators.
In Hermann Locarek-Junge and Claus Weihs, editors, *Classification as a Tool for Research, Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 463–471, Berlin, Heidelberg, New York, June 2010. Springer.

Veröffentlichungen III

[RKWS08] 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 – Medical Applications, E-Commerce, Marketing, and Theoretical Aspects*, volume 5077 of *LNAI*, pages 47–56, Berlin, Heidelberg, July 2008. Springer Verlag.

[RS10] Georg Ruß and Martin Schneider.
Hierarchical spatial clustering for management zone delineation in precision agriculture.
In Isabelle Bichindaritz, Petra Perner, and Georg Ruß, editors, *Advances in Data Mining*, pages 95–104, Leipzig, July 2010. IBaI Publishing.

[Ruß09] Georg Ruß.
Data mining of agricultural yield data: A comparison of regression models.
In Petra Perner, editor, *Advances in Data Mining – Applications and Theoretical Aspects*, volume 5633 of *LNAI*, pages 24–37. Springer, July 2009.

[Ruß11] Georg Ruß.
Hacc-spatial: Hierarchical agglomerative spatially constrained clustering.
In Isabelle Bichindaritz, Petra Perner, and Georg Ruß, editors, *11th ICDM Conference, New York, USA, Workshop Proceedings*, Leipzig, September 2011. IBaI Publishing.