Deutsche Gesellschaft für Photogrammetrie,
Fernerkundung und Geoinformation (DGPF) e.V.

Arbeitskreis „Auswertung von Fernerkundungsdaten“

26 Oktober, Halle (Saale)

Department
Geoinformation in Environmental Planning

Einsatz von Data Mining –
ein nichtparametrischer Klassifikator in der Umweltanalyse

Motivation für Data Mining

- Daten und Information
  - Daten → Information → Wissen
- Über die Umwelt sind große Datenmengen vorhanden
- Diese Daten enthalten implizit (nicht unmittelbar einsehbar) Informationen
- Problem:
  - Wie kann man diese Information explizit machen?
  - Wie kann man komplexe Umweltphänomene erkennen?

Data Mining: Definition

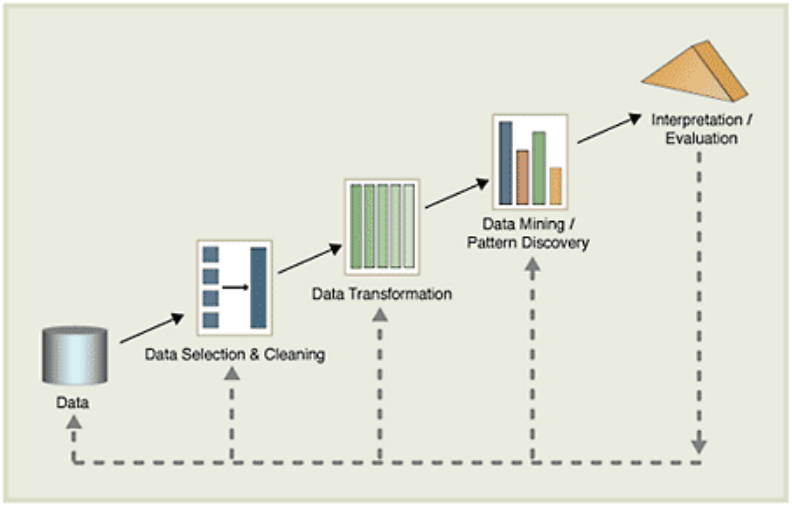
- Data Mining is a non-parametric method of identifying
  - valid
  - novel
  - potentially useful
  - ultimately understandable
- patterns in data.

(Fayyad et al. 1996)

- It employs techniques from
  - Machine learning
  - Statistics
  - Databases

Data Mining is a 5 step process

- The steps are not followed linearly, but in an iterative process



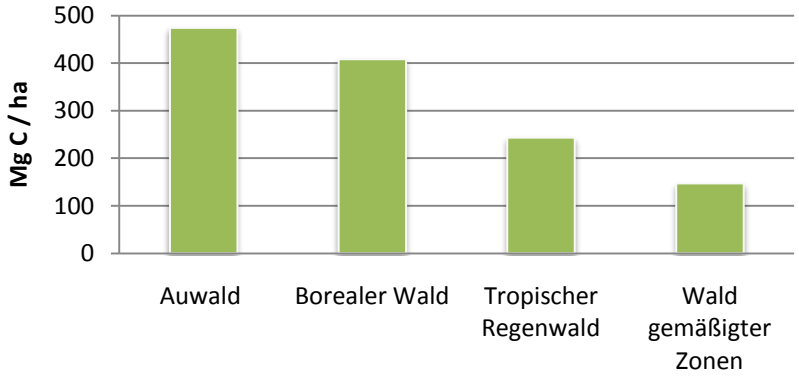
Source: http://alg.ncsa.uiuc.edu/tools/docs/d2k/manual/dataMining.html, after Fayyad, Piatetsky-Shapiro, Smyth, 1996

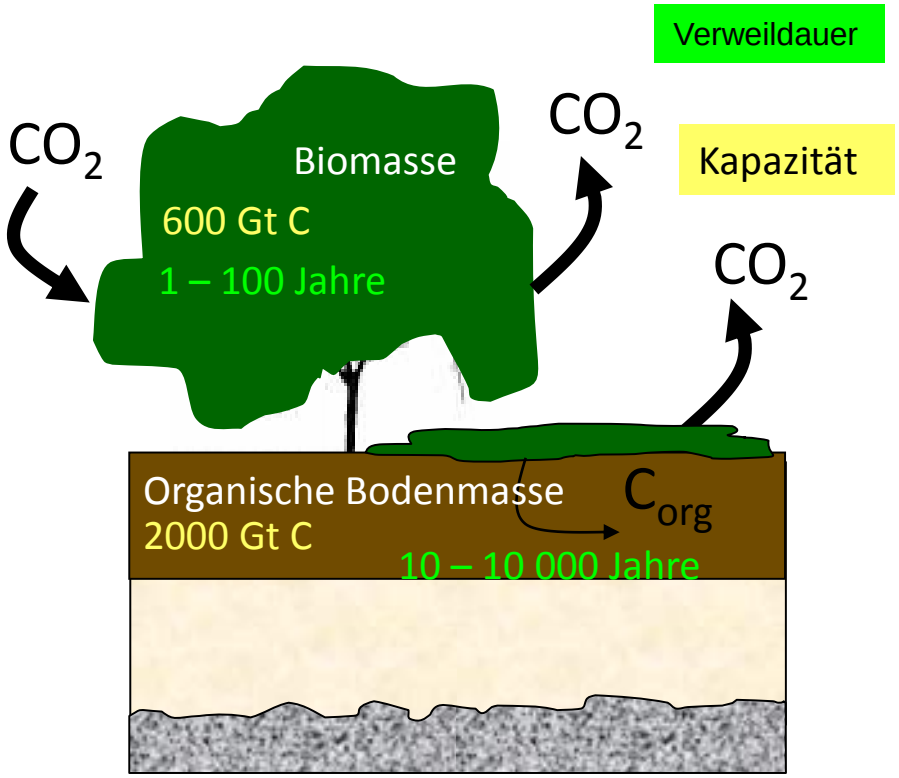
Data mining steps (Qi and Zhu 2003):

- Data selection
- Data preprocessing and transformation
- Pattern extraction
- Knowledge consolidation
  - Examination
  - Interpretation

Fallbeispiel: C_{org}-Modellierung in Auen

- Flussauen, insbesondere Auwälder, haben hohes Speicherpotential für organischen Kohlenstoff (C_{org}), bis zu 474 Mg C /ha (Cierjacks *et al.* 2010)
- Bisher nur wenig Daten zu Corg in Auwäldern, speziell Bodendaten

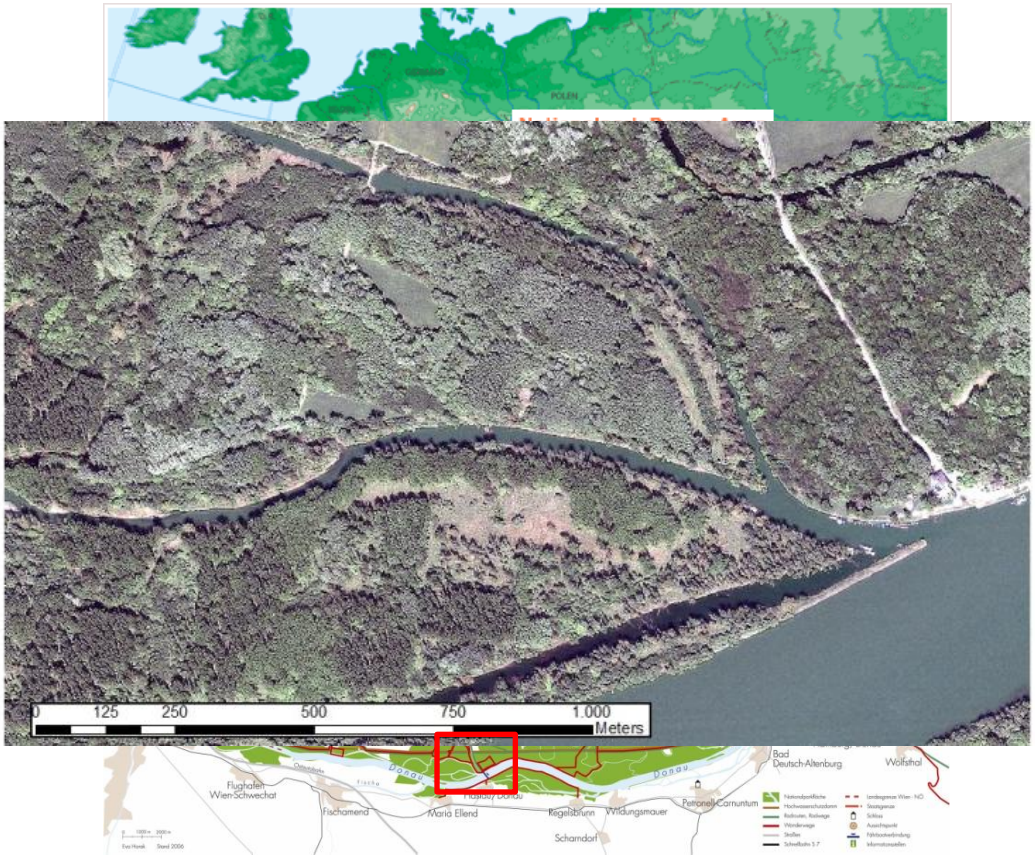




Swift 2001, Houghton 2005, Fontaine et al. 2007, Hamilton 2007, Hartmann und Kempe 2008

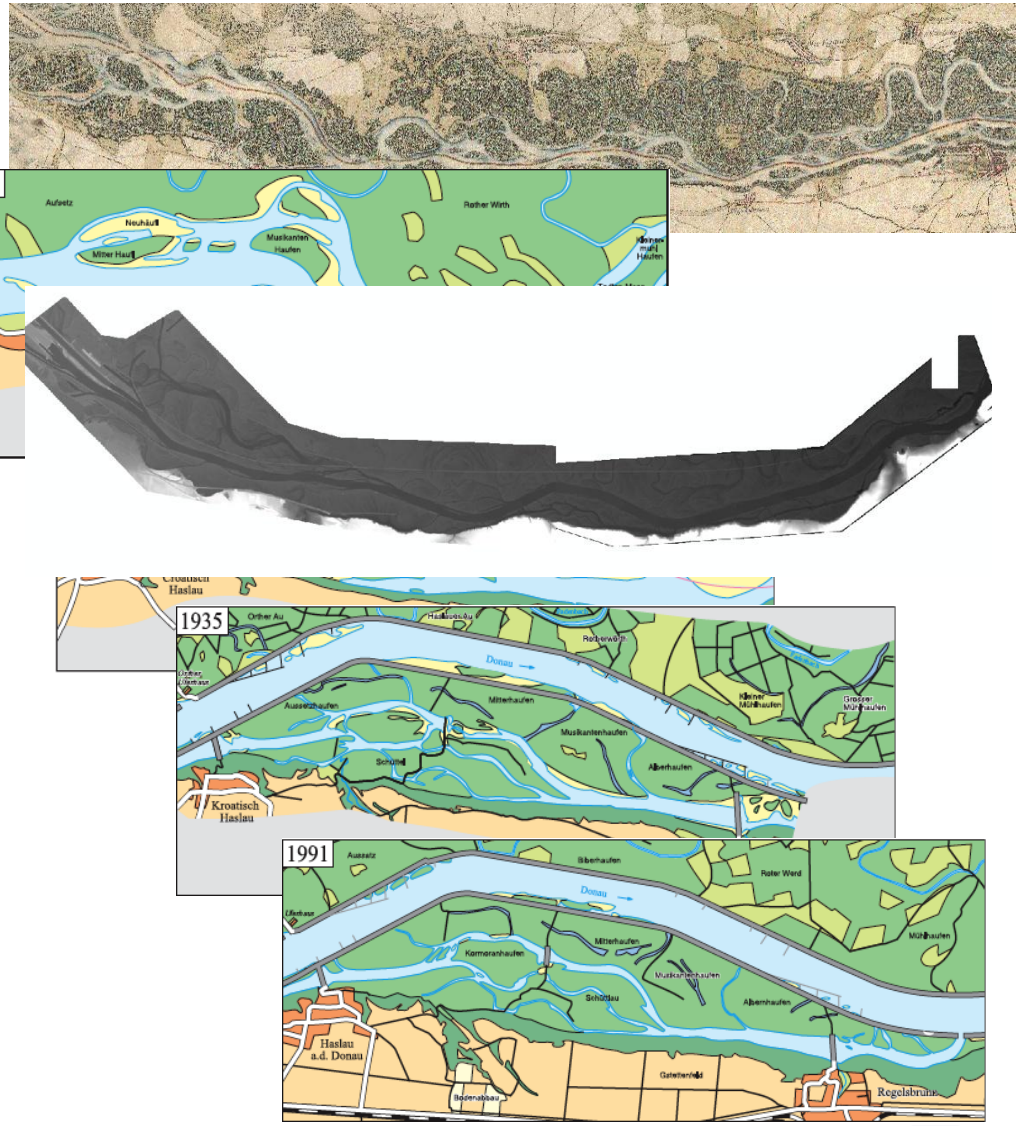
Untersuchungsgebiet

- Lage: im Donauauen - Nationalpark, Österreich
- Eine der letzten großen ursprünglichen Auengebiete in Mitteleuropa
- Hohe kleinräumige Variation

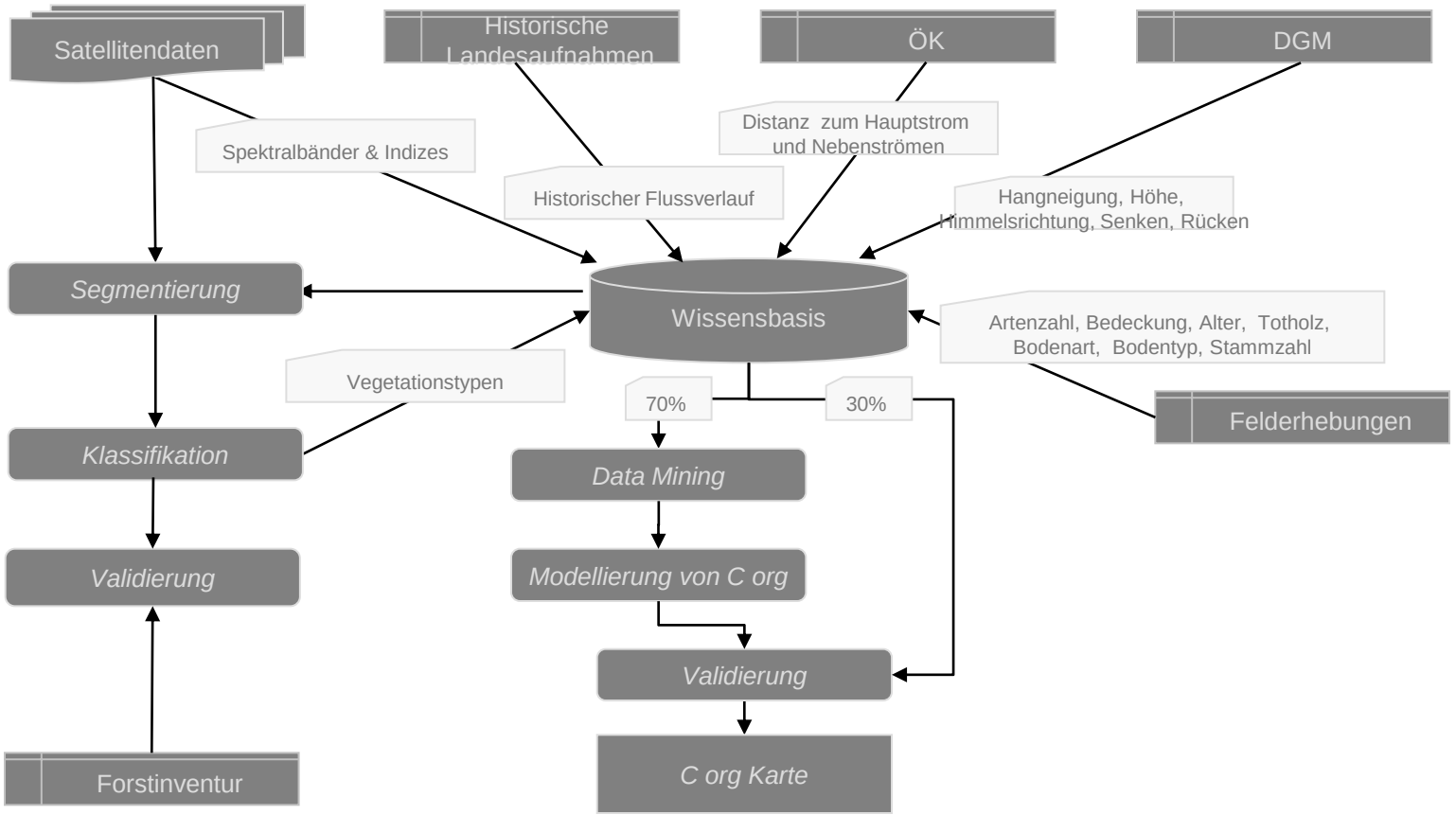


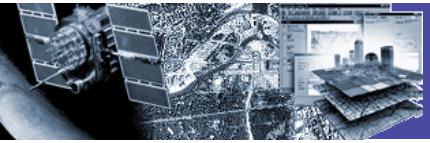
Daten

- Ikonos 2 (2009), RapidEye (2009)
- Historische Landesaufnahmen
- DGM
- Grundwassermodell
- Felderhebungen (2008; 68 Punkte) mit Daten zu Vegetation und Boden



Data preparation, data cleaning and preprocessing





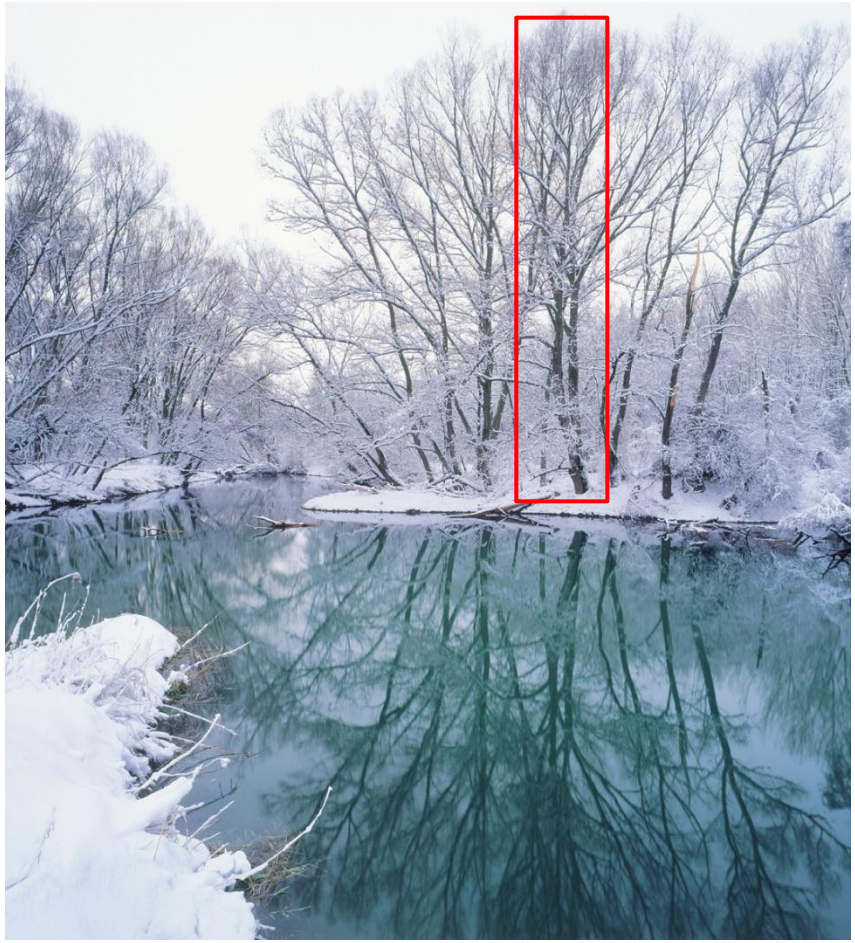
Input Daten für Data Mining

Satelliten-Daten	Zusatzdaten
Ikonos (22. April 2009):	**RÄUMLICHE LAGEPARAMETER** Absolute Höhe über NN
Blau (445-516 nm)	Relative Höhe über Fluss
Grün (506-595 nm)	Hangneigung
Rot (632-698 nm)	Distanz zum Hauptstrom
NIR (757-853 nm)	Distanz zu Seitenkanälen
Grün-blau	Existenz eines historischen Flussbetts bei der 1. Landesaufnahme
NDVI	Existenz eines historischen Flussbetts bei der 2. Landesaufnahme
NIR / grün	Existenz eines historischen Flussbetts bei der 3. Landesaufnahme
NIR-rot	**VEGETATIONSPARAMETER** Vegetationstyp
Rot/ blau	Artenzahl
Rot-blau	Bedeckungsgrad der Vegetation
Rot-grün	Baumalter
Rapid Eye (01. August 2009):	Totholzanteil
Blau (440-510 nm)	Stammzahl
Grün (520-590 nm)	**BODENPARAMETER** (Haupt-)Bodenart
Rot (630-685 nm)	Bodentyp
Red edge (690-730 nm)	Höhe – mittleres Grundwasser
NIR (760-850 nm)	Höhe – niederes Grundwasser

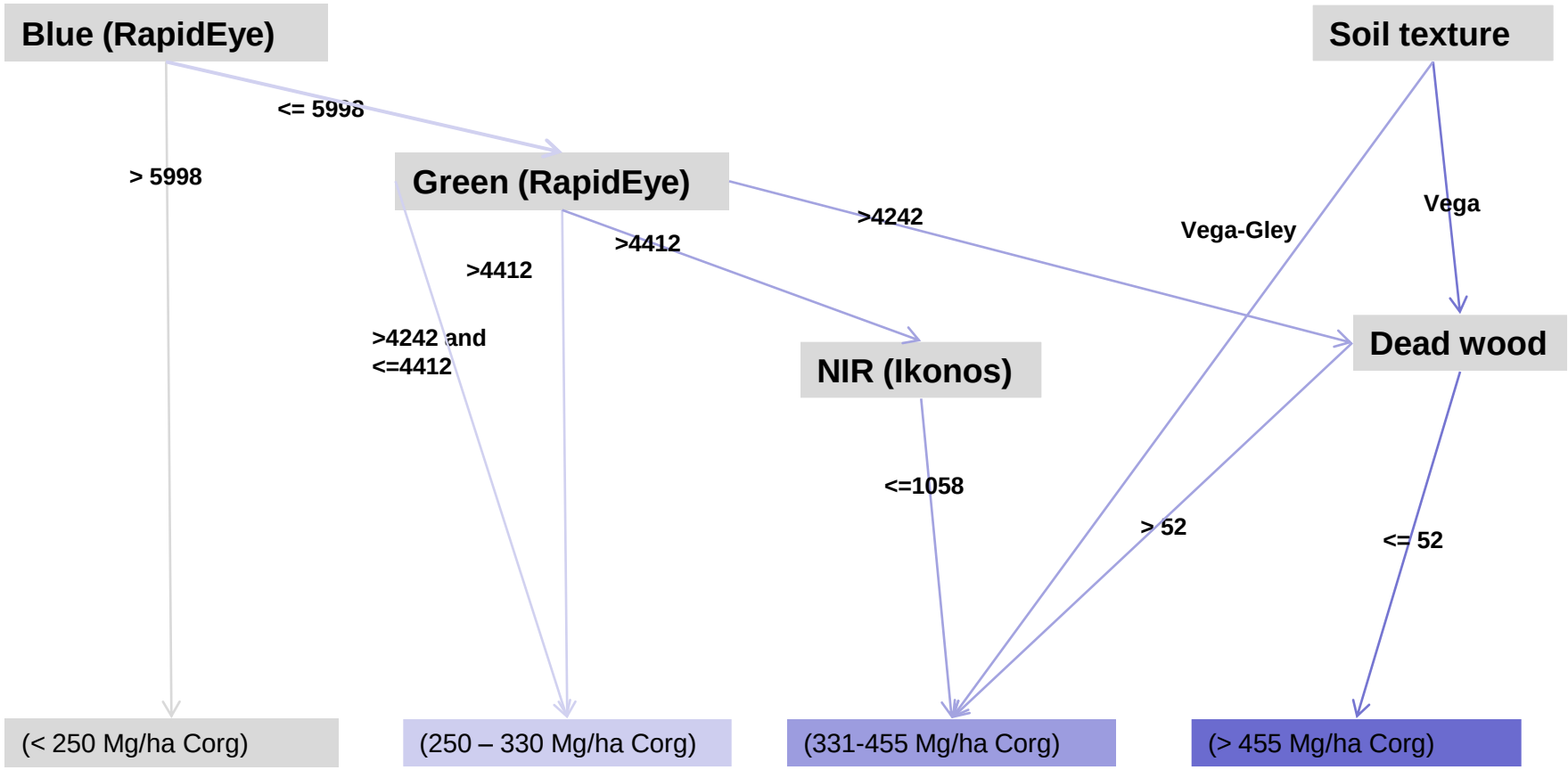
Data Mining / Knowledge Discovery durch Entscheidungsbäume

- Entscheidungsbaum als ein Unterstützungswerkzeug in Form eines Graphen mit Entscheidungsästen und möglichen Konsequenzen
- Einfache Form die kompakt gespeichert werden kann und effizient neue Daten klassifiziert
- Führt automatische Feature-Auswahl und Komplexitätsreduzierung durch, Baumstruktur gibt leicht verständliche und interpretierbare Information

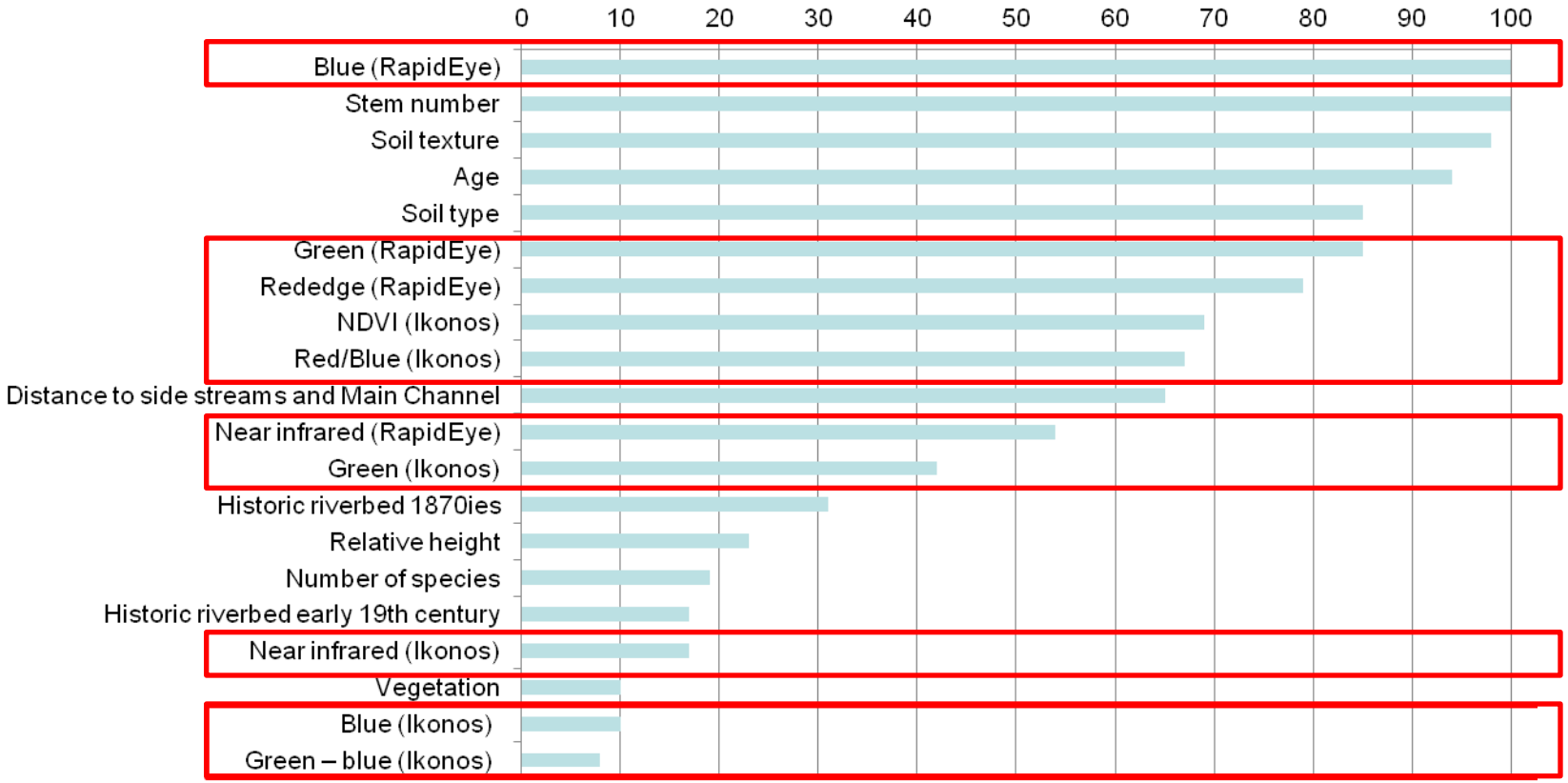
Data Mining Regelsätze



Beispiel Regelwerk (Entscheidungsbaum)



Häufigkeit Attributverwendung



Evaluierung mit Trainingsdaten (70%)

Regelwerk

	Regeln	Fehler (%)
0	8	8 (16.0)
1	7	13 (27.0)
2	9	14 (29.2)
3	8	11 (22.0)
4	7	11 (22.9)
5	7	15 (31.9)
6	8	11 (22.9)
7	7	13 (27.1)
8	8	13 (27.1)
9	6	12 (25.0)
boost		0 (0.0) <<

(a)	(b)	(c)	(d)	← klassifiziert als
12				(a): < 250 Mg/ha Corg
	11			(b): 250 – 330 Mg/ha Corg
		11		(c): 330-455 Mg/ha Corg
			14	(d): > 455 Mg/ha Corg

Evaluierung mit Testdaten (30%)

Regelwerk

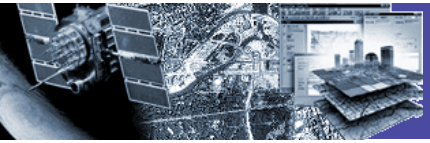
	Regeln	Fehler (%)
0	8	14 (70.0)
1	7	14 (70.0)
2	9	12 (60.0)
3	8	13 (65.0)
4	7	13 (65.0)
5	7	12 (60.0)
6	8	15 (75.0)
7	7	14 (70.0)
8	8	11 (55.0)
9	6	13 (65.0)
boost		10 (50.0)

<<

(a)	(b)	(c)	(d)	← klassifiziert als
2	3			(a): < 250 Mg/ha Corg
	4	1	1	(b): 250 – 330 Mg/ha Corg
1	1	2	2	(c): 330-455 Mg/ha Corg
	1		2	(d): > 455 Mg/ha Corg

Diskussion

- Vorteile
  - Automatisierter Prozess statt Wissensbasierter Auswertung
  - Erkennung bisher unbekannter Muster
- Nachteile
  - Einzelne Regeln nicht nachvollziehbar (Gefahr des Data Dredging)
  - Geringe Flexibilität der Software



Zusammenfassung & Ausblick

Zusammenfassung:

- Kohlenstoffverteilung kann mittels Fernerkundung und zusätzlichen Geodaten dargestellt werden
- Entscheidungsbaum kann Beziehungen abbilden
- Ergebnisse haben hohe Fehlerquote
  - Aufgrund hoher Inhomogenität der Kohlenstoffverteilung
  - Auswahl der Zusatzparameter muss evaluiert werden
  - Bisher geringe Anzahl an Stichproben (68 Stichproben)

Ausblick:

- Ausweitung des Data Mining auf Textur-Parameter und eine höhere Stichprobenzahl
- Test auf Übertragbarkeit

Referenzen

- leonhard.suchenwirth@tu-berlin.de
- michael.foerster@tu-berrlin.de
- birgit.kleinschmit@tu-berlin.de

Gefördert durch die DFG

- Cierjacks, A., B. Kleinschmit, M. Babinsky, F. Kleinschroth, A. Markert, M. Menzel, U. Ziechmann, T. Schiller, M. Graf, and F. Lang. 2010. Carbon stocks of soil and vegetation on Danubian floodplains. *Journal of Plant Nutrition and Soil Science* 173(5) : 644-653.
- Cierjacks, A., B. Kleinschmit, I. Kowarik, M. Graf, and F. Lang. 2011. Organic matter distribution in floodplains can be predicted using spatial and vegetation structure data. *River Research and Applications* 27: 1048-1057.
- Suchenwirth, L., Förster, M., Cierjacks, A., Lang, F. & Kleinschmit B. (submitted): Knowledge-based classification of Remote Sensing data for the estimation of below- and above-ground organic carbon stocks in riparian forests.