5. Dresdner Flächennutzungssymposium

Dresden, 5. – 6. Juni 2013

Automatische Erkennung von Gebäudetypen auf Grundlage von Geobasisdaten

Robert Hecht

IÖR, Dresden

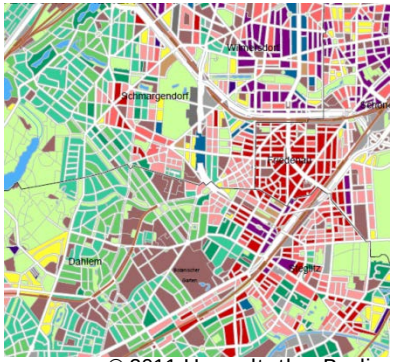
Gliederung

- **Hintergrund**
  (Motivation, Problemstellung, Ziel, Forschungsstand)
- **Methodik**
  (Daten, Typologie, Merkmale, Klassifikator, Umsetzung)
- **Ergebnisse**
  (Untersuchungsgebiete, Genauigkeit, Einflussfaktoren)
- **Fazit und Ausblick**
  (Bewertung, weiterer Forschungsbedarf)

Motivation

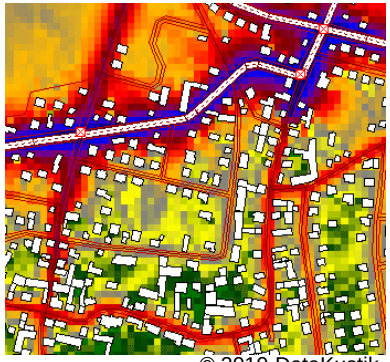
- **Wissenschaft und Praxis benötigt kleinräumige Informationen zur Siedlungsstruktur** (physisch, funktional, sozio-ökonomisch)

Stadt- und Regionalplanung



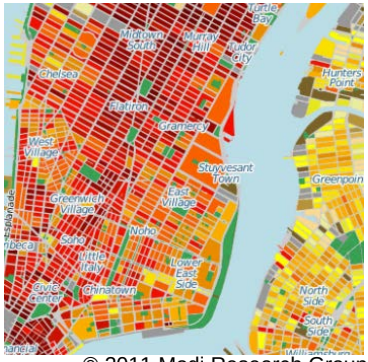
© 2011 Umweltatlas Berlin

Infrastrukturplanung (Lärm, Schadstoffe)



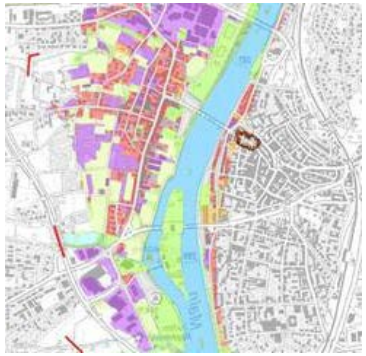
© 2010 DataKustik

Energiebedarfs -planung



© 2011 Modi Research Group

Risikoabschätzung (Hochwasser)



Tribet (2011)

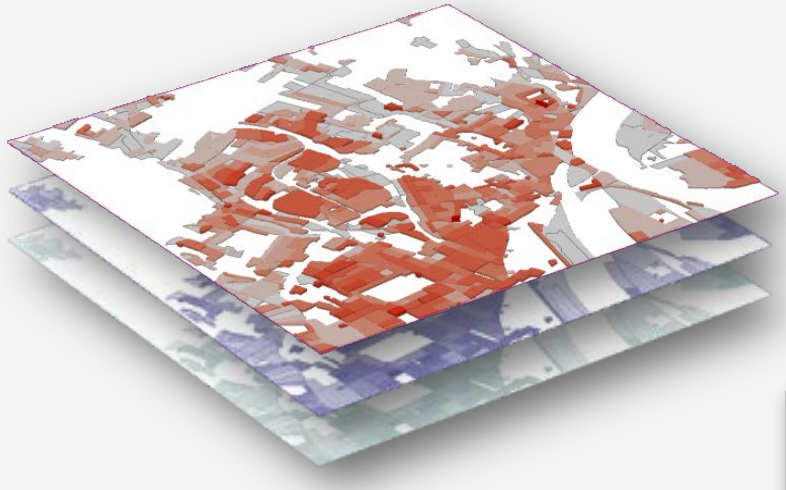
- **Anforderung:** hochauflösend (räumlich/inhaltlich), objektiv, flächendeckend, homogen

Problemstellung

- **Informationsbedarf nicht gedeckt** (insb. Gebäudestruktur, Bebauungsdichte, Geschossfläche, Wohnung- und Einwohnerzahl, Baualter, bauliche Aktivität)
  - **Amtliche Statistik:** geringe räumliche Auflösung (nur auf Gemeindeebene)
  - **Kommunalstatistik:** nicht flächendeckend, heterogen, teuer, erschwerter Zugriff
  - **Geobasisdaten:** geringe thematische Auflösung
  - **Kommerzielle Daten:** intransparent, teuer
  - **Eigene Erhebung:** zeit- und kostenintensiv

- **Lösungsansatz:** Ableitung der siedlungsstrukturellen Informationen mithilfe **räumlicher Analysen im GIS**

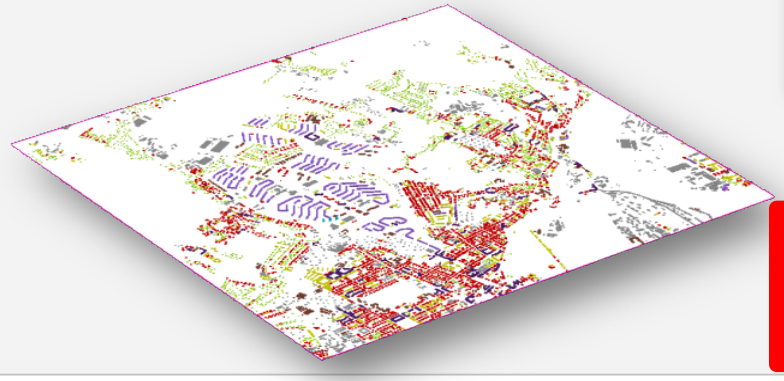
Lösungsansatz – Räumliche Analyse



Kleinräumige Information zur Siedlungsstruktur
(Gebäudedichte, Überbauungsgrad, Geschossflächendichte, Gebäudevolumen Wohnungs- und Einwohnerzahl, etc.)

Räumliche Analysen
(z.B. Verschneidung, Aggregation, Disaggregation)

Gebäudetypische Annahmen
(z.B. Geschosse, Wohnungsgröße, Haushaltsgröße)



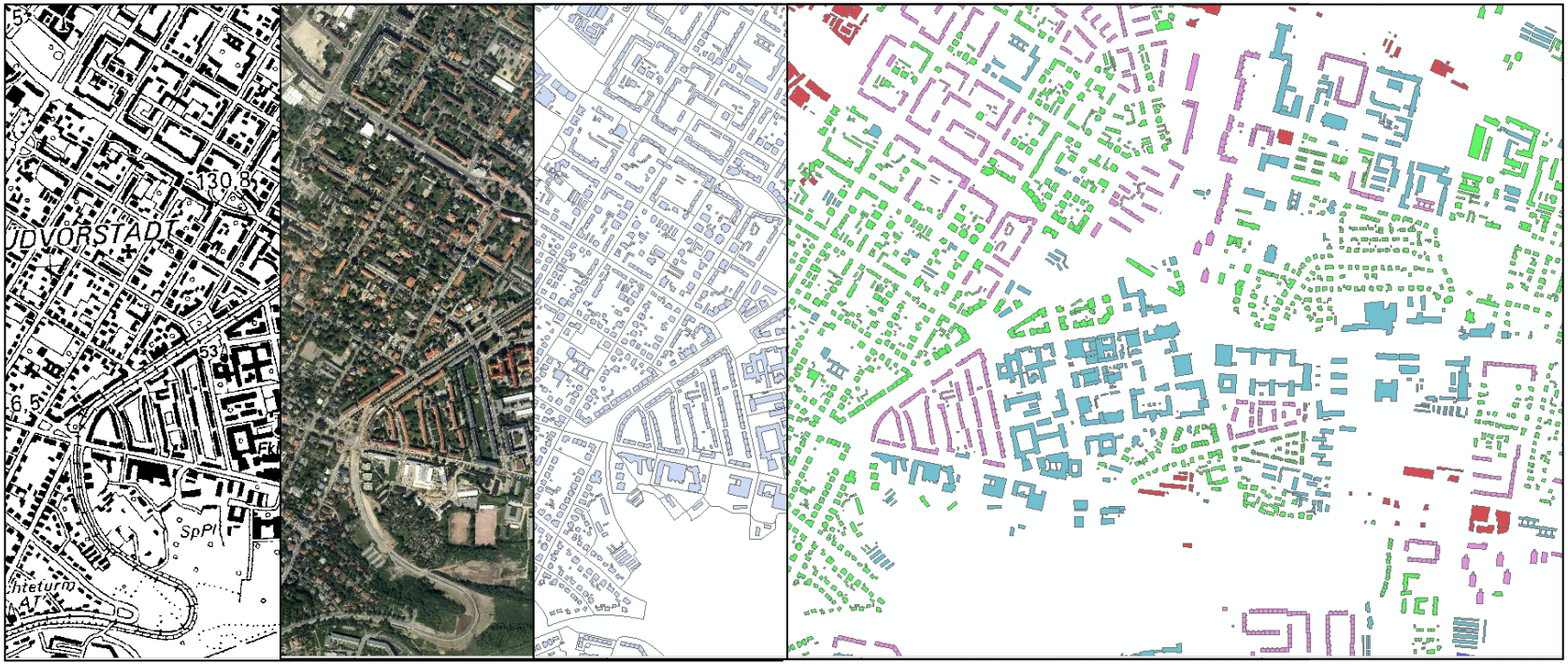
Zweckmäßig klassifiziertes Gebäudemodell
(z.B. Einfamilienhaus, Mehrfamilienhaus, Industriegebäude)

?

(vgl. Meinel et al., 2008)

Gebäude in …

- Kartenwerken, Luftbildern, Vektordaten, …

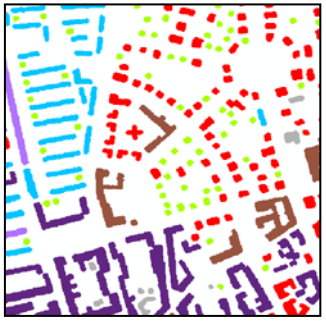


➔ **Daten bilden Gebäude ab**, besitzen jedoch häufig **keine explizite Information zum Gebäudetyp**

Zielstellung

- **Entwicklung methodischer Beiträge** zur automatischen Klassifizierung von Gebäudegrundrissen für die kleinräumige Beschreibung der Siedlungsstruktur
- **Bildanalyse**, **Geoinformatik** und des **Maschinellen Lernens**
- Berücksichtigung flächendeckend verfügbarer **Geobasisdaten**
- **Validierung** und **Bewertung**

Gebäudegrundrisse ohne semantischer Information



Klassifizierte Gebäudegrundrisse

Forschungsstand

- Automatische Ansätze mit unterschliedlicher/n
  - **Zielstellung und Typologie** (Stadtforschung, Generalisierung etc.)
  - **Interpretationsebene** (u.a. Gebäude, Gebäudegruppen, Baublöcke, Regionen)
  - **Eingangsdaten** (Fernerkundungsdaten, Katasterdaten, Topographische Karten etc.)
  - **Klassifikationsverfahren** (u.a. wissensbasierte Regelwerke, datengetriebene Verfahren)

- **Klassifizierung individueller Gebäude ein relativ junges Forschungsfeld**

Bisherige Ansätze

Ansätze zur Klassifizierung von Gebäudegrundrissen

top-down
(modellgetrieben)

Wissensbasierte Ansätze
(z.B. Regelwerke, Fuzzy-Regler)

u.a. Rengelink et al. (2000), Rainsford & Mackaness (2002), Orford & Radcliffe (2007), Meinel et al. (2008), Thomson (2009), Wurm et al. (2009) und Smith & Crooks (2010)

mixed
(kombiniert)

Kombinierte Verfahren
(z.B. Ontologie-gestützte Mustererkennung)

u.a. Lüscher (2009) und Triantakonstantis & Barr (2009)

bottom-up
(datengetrieben)

Überwachte Ansätze
(z.B. Entscheidungsbaum-Algorithmus, Künstliche Neuronale Netze, Support Vector Machines)

u.a. Sester (2000), Raheja (2004), Steiniger et al. (2008), Meng (2010), Römer & Plümer (2010) und Henn et al. (2011)

Unüberwachte Ansätze
(z.B. Clustering mit EM-Algorithmus)

u.a. Neidhart & Sester (2004), Werder et al. (2010) und Geiß et al. (2011)

Explizites Wissen, Transparenz, Aufwand beim Modellaufbau

Fokus: überwachte Erkennungsstrategie

Flexibilität, Automationsgrad, Komplexität

Forschungsstand

- **Defizite der bisherigen Ansätze**
  - geringe Klassenanzahl
  - geringe Anzahl an Merkmalen
  - kleine Untersuchungsgebiete
  - schwer adaptierbare Verfahren
  - unzureichende Validierung
  - wenig Erfahrungen mit Geobasisdaten in Dtld.

➔ Es steht ein breites Spektrum an Methoden der Mustererkennung und des Maschinellen Lernens zur Verfügung, deren Einsatz für die automatische Klassifizierung von Gebäudegrundrissen bislang nur in wenigen Arbeiten systematisch untersucht wurde.

Forschungsfragen

Wie können Gebäudegrundrisse automatisch klassifiziert werden?

- **Welche Anforderungen** werden an die Eingangsdaten gestellt?
- **Welche amtlichen Datengrundlagen** sind geeignet und wie lassen sich diese integrieren?
- **Welche Merkmale** können für die Beschreibung der Gebäude abgeleitet werden und welche Merkmale sind besonders wichtig?
- **Welche Klassifikationsverfahren** der Mustererkennung und des Maschinellen Lernens sind geeignet?
- **Welche Genauigkeit** kann unter Nutzung verschiedener Geobasisdaten erreicht werden?
- **Welche Faktoren** haben Einfluss auf die Klassifikationsgenauigkeit?

Methodik

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

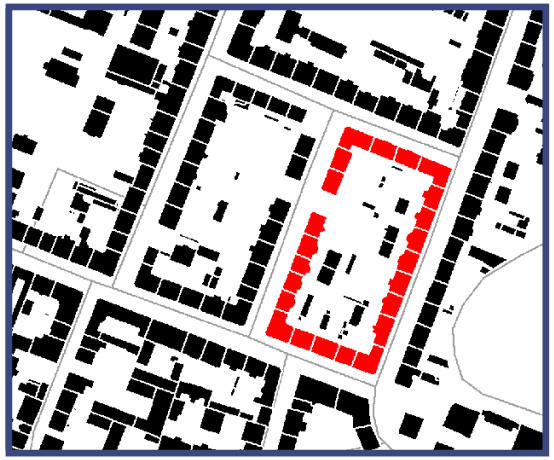
Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Umsetzung in ArcGIS/Python, R, HALCON-Prozeduren

Gebäudetypologie

- **Zweck:** Beschreibung der Siedlungsstruktur mit Fokus auf Wohnbebauung
- **Kriterien**: Nutzung, Wohnform, städtebauliche Struktur, Baualter/ Konstruktion



- Gebäude
  - Wohn-nutzung
    - Ein- und Zweifamilien-haus
      - **Einzelhaus**
      - **Doppelhaus**
      - **Reihenhaus**
    - Mehrfamilien-haus
      - Offene Bauweise
        - **Mehrfamilienhaus**
        - **Hochhaus**
        - **Villa**
      - Zeilenbau -weise
        - **Trad. Zeilenbauweise**
        - **Plattenbauweise**
      - Geschlossene Bauweise
        - **Blockrandbauweise**
        - **Historischer Ortskern**
    - **gemischt** (insb. Ländliche Hausformen)
  - Nichtwohn-nutzung
    - **Gebäude für Wirtschaft, Gewerbe und Landwirtschaft**
    - **Öffentliche Gebäude**
    - **Nebengebäude**

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

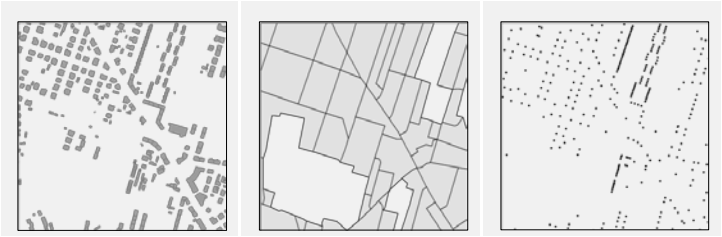
- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Eingangsdaten

- **Fokus auf Geobasisdaten** (flächendeckend, vollständig, homogen, in Fortführung gesichert)
- **Gebäudegrundrisse** (Kataster, Landesvermessung)
- **Baublöcke** (baulich geprägte Fläche, ATKIS®-Basis DLM)
- **Hauskoordinaten**



Eingangsdaten - Gebäudegrundrisse

Geobasisdaten mit Gebäudegrundriss

- **Rasterdaten** (mit grundrissähnlicher Gebäuderegion-Repräsentation)
  - Gebäude vereint mit anderen Elementen
  - Gebäude in separatem Layer
- **Vektordaten** (mit grundrisstreuer Einzelgebäude-Repräsentation)
  - ohne Nutzungsinformation
  - mit Nutzungsinformation
    - ohne Gebäudehöhe
    - mit Gebäudehöhe

Typ I (TK25 , DTK25-V)	**Typ II** (DTK25)	**Typ III** (HU, ATKIS)	**Typ IV** (ALK, ALKIS)	**Typ V** (3D-GS)

Zunahme von Informationsgehalt, Genauigkeit und Datenkosten

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

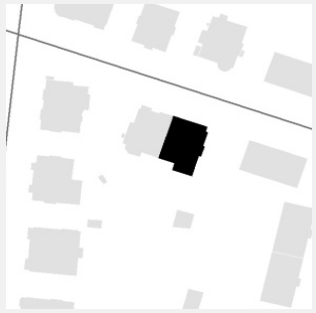
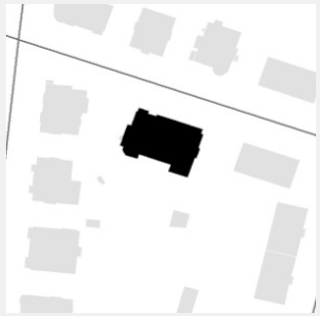
Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

- Festlegung gebäudebeschreibender Eigenschaften in Form **numerischer Merkmale**
- **Geometrische, Topologische, Statistische, Semantische Merkmale**
- Einsatz der **Räumlichen Analyse** und **digitalen Bildverarbeitung**
- **Nutzung bereits explizit modellierter Information** (z.B. Flächennutzung aus ATKIS Basis DLM, Gebäudenutzung)
- **Speicherung** der Merkmale

Merkmalsberechnung

- **Ebenen**

Mikro-Ebene		Meso-Ebene		Makro-Ebene
Einzelgebäude	Gebäuderegion	Umgebung (100m Suchradius)	Baublock	Siedlungskörper
				
z. B. Fläche, Umfang, Gebäudehöhe, Nutzung	z. B. Fläche, Umfang, Anzahl adjazenter Gebäude	z. B. Gebäudeanzahl, Mittlerer Abstand der Gebäude	z. B. Gebäudeanzahl, Überbauungsgrad, ATKIS®-Nutzungsart	z. B. Größe des Siedlungskörpers

➔ **Merkmalsset pro Eingangsdatentyp** (72 bis 90 Merkmale)

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Vorverarbeitung

- **Konsistenzprüfung und Datenbereinigung** (z.B. Eliminierung konstanter Merkmale)
- **Skalierung** der Merkmale (Z-Transformation)
- **Transformation kategorialer Merkmale** in nummerische Merkmale
- **Ersetzen fehlender Werte** mit k-Nächste-Nachbar-Imputation (Hastie et al., 1999)

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

Modellanwendung und -bewertung

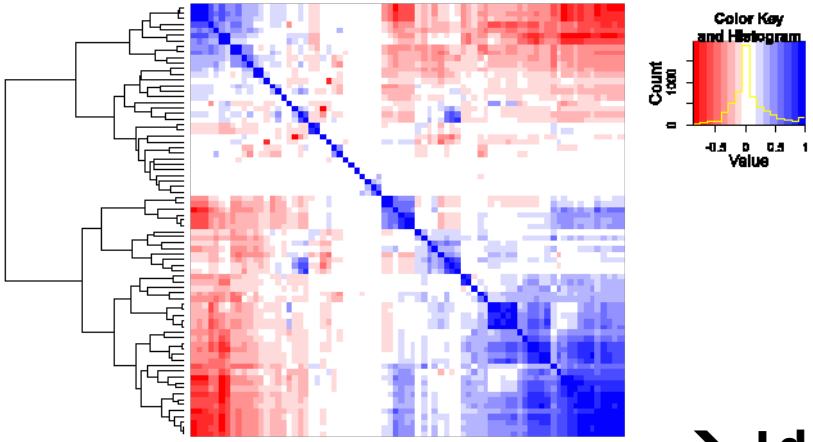
- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Einsatz zweier Verfahren

- **Filterbasierte Merkmalsreduktion** während der Modellentwicklung
- **Wrapperverfahren** nach Wahl des besten Klassifikators

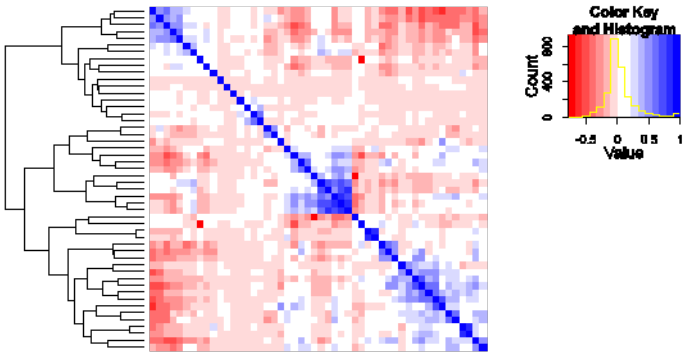
Bsp. Filterbasierte Merkmalsreduktion

Korrelation der originalen Merkmale (77 Merkmale)



Beispieldatensatz Halle (DTK25-V)
77 Merkmale
32 hoch korrelierte Merkmale mit (|ρ|>0,9)

Korrelation nach Merkmalsreduktion (50 Merkmale)



Reduktion auf 50 Merkmale
(nicht korr. Merkmale + 5 PCA-Komponenten)

- ➔ **Identifikation stark korrelierter Merkmale**
- ➔ **PCA auf stark korrelierte Merkmale**

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Entwicklung des Klassifikators

- **Test 16 verschiedener Lernverfahren** an 5 Eingangsdatentypen
- **Wahl des besten Klassifikationsverfahrens** anhand Klassifikationsgenauigkeit und Laufzeit
- **Klassische** (KNN, LDA) und **moderne Algorithmen** (SVM, Ensemble-Methoden)
- **R-Pakete stellen Algorithmen** (GIS-Kopplung möglich)

Methodisches Design

Rahmenbedingungen

- **Festlegung einer Typologie**
- **Wahl der Eingangsdaten**

Methodische Bausteine von BFClassTool

- **Datenaufbereitung** (Datenintegration, Klassenlabels)
- **Merkmalsberechnung** (Merkmalstypen, Umsetzung, Eliminierung kleiner Nebengebäude)
- **Vorverarbeitung** (Bereinigung, Skalierung, Transformation)
- **Merkmalsselektion** (Filterbasierte Vorselektion, Wrapper-Ansatz)
- **Entwicklung des Klassifikators** (Klassifikationsverfahren, Tuning, Modellselektion)

Modellanwendung und -bewertung

- **Genauigkeitsanalyse** (Genauigkeit nach Untersuchungsgebiet, Eingangsdatentyp und Trainingsmenge, Generalisierungsfähigkeit)

Genauigkeitsanalyse

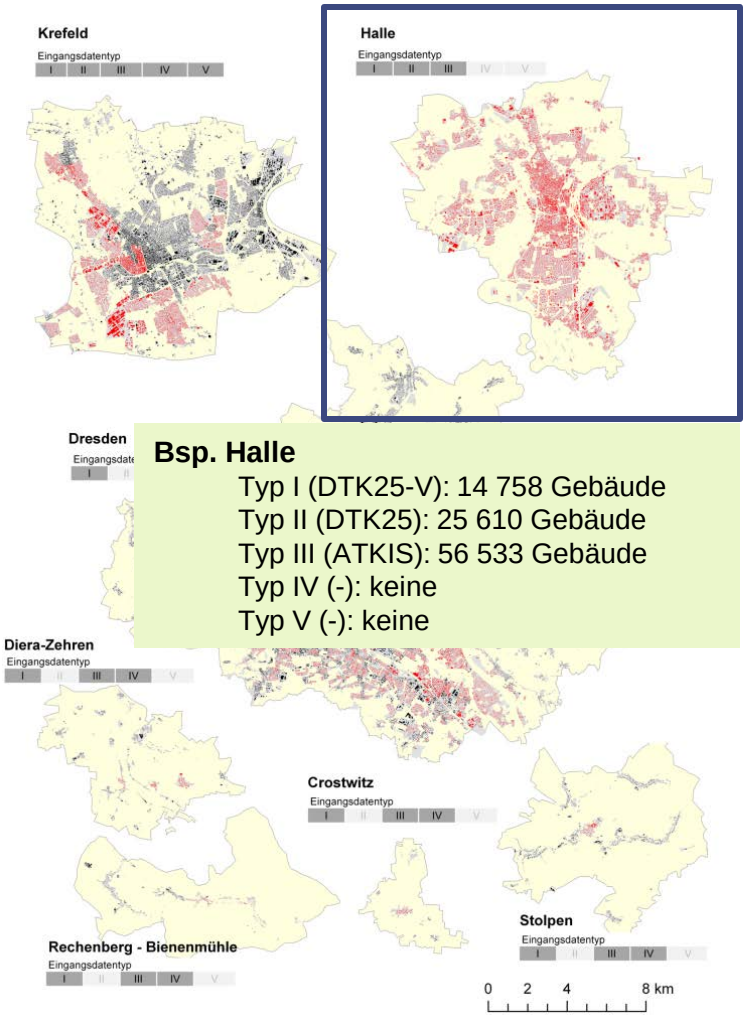
- Anwendung auf alle Datensätze
- Bewertung der Genauigkeit in Abhängigkeit vom
  - Eingangsdatentyp
  - Siedlungstyp
  - Trainingsdatenmenge
- Untersuchung der Generalisierungsfähigkeit (kreuzweises Trainieren und Testen verschiedener Städte)
- Einfluss der Merkmalsselektion (Wrapper – Ansatz)

Ergebnisse

Untersuchungsgebiete

- **Siedlungstypen**
  - **Städtisch** (Dresden, Halle, Krefeld)
  - **Kleinstädtisch** (Stolpen)
  - **Ländliche Gebiete** (sächs. Gemeinden)

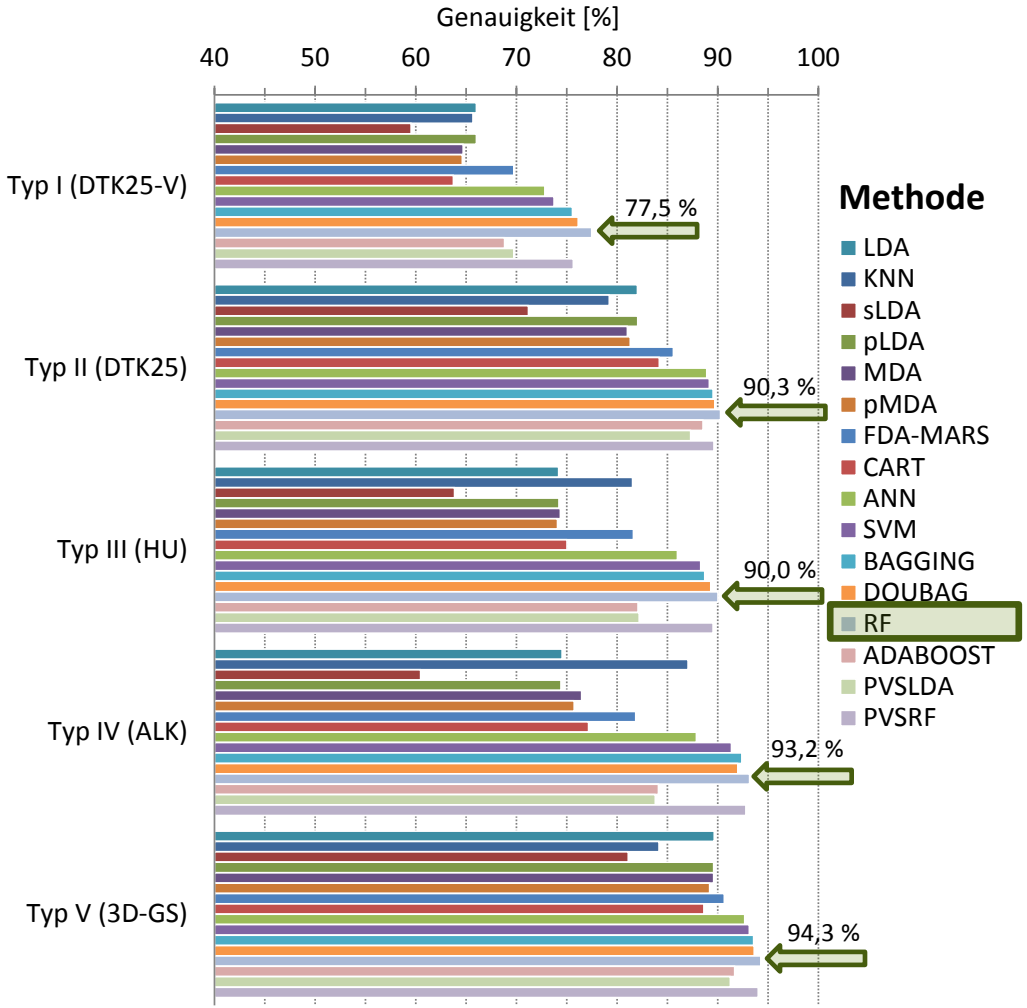
- **Referenzdatenbasis**
  - **Nutzung bereits vorhandener Kartierungen** (Umkodierung, räumliche Zuordnung, Transformation etc.)
  - **11 Gebäudetypen**
  - **5 Eingangsdatentypen**



Bsp. Halle

Typ I (DTK25-V): 14 758 Gebäude
Typ II (DTK25): 25 610 Gebäude
Typ III (ATKIS): 56 533 Gebäude
Typ IV (-): keine
Typ V (-): keine

Wahl des Klassifikators - Genauigkeit



- Genauigkeitsabschätzung mittels 10-facher Kreuzvalidierung
- höhere Klassifikationsleistung mit nichtlinearen Klassifikationsverfahren
- höchste Klassifikations-genauigkeit mit Random Forest (RF)
- PVSRF, BAGGING, DOUBAG, SVM und ANN nicht signifikant schlechter als RF
- Laufzeitvergleich identifiziert Random Forest als besten Klassifikator

Genauigkeitsanalyse

Gesamtgüte

- **10-fache Kreuzvalidierung** an allen Datensätze
- Gütemaße: *Overall Accuracy* und *Cohens Kappa*

Bsp. Overall Accuracy

Eingangsdaten	Untersuchungsgebiete: Stadt Dresden	Stadt Halle	Stadt Krefeld	Kleinstadt Stolpen	Dörfl. Gebiete
Typ I DTK25-V	78,9 %	76,6 %	76, 6 %	67,4 %	56,4 %
Typ II DTK25		88,9 %	78,2 %		
Typ III HU	90,4 %	91,8 %	93,6 %	72,2 %	62,4 %
Typ IV ALK	94,3 %		94,3 %	84,2 %	78,7 %
Typ V 3D-GS			94,9 %		

Klassenspezifische Genauigkeit

- Konfusionsmatrix aus 10-facher KV für jeden Datensatz
- Gütemaße: *Producer´s Accuracy* und *User´s Accuracy*

Einflussfaktoren auf Genauigkeit

- **Datengrundlage:**
  - von Eingangsdatentyp I bis V zunehmend
  - 90 - 95 % für Gebäude aus Liegenschaftskataster (z.B. HU, ALK, 3D-Gebäude), 76 - 88 % für Gebäude aus topographischen Karten
- **Siedlungstyp:**
  - Genauigkeit in ländlichen Gebieten ca. 20 Prozentpunkte geringer als in städtischen,
  - Große Unsicherheiten in Gebäudenutzung
- **Trainingsdatenmenge:**
  - Genauigkeit nimmt mit Trainingsdatenmenge zu
  - Bsp. Halle (Trainingsobjekte/Genauigkeit):
    - 625/>80%, 2 500/>85 %, 10.000/>90 %
    - 350 Gebäude pro Typ/ PA für alle Klassen > 70%
- **weitere** (Datenqualität, Generalisierungsgrad)

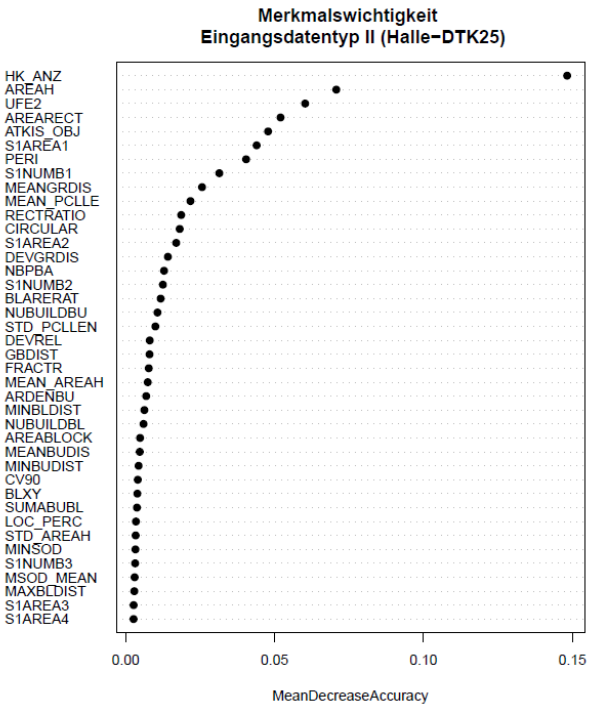
Weitere Erkenntnisse

Merkmalswichtigkeit

- Random Forest liefert Maße zur Messung der Merkmalswichtigkeit
- Merkmalsselektion durch Rückwärtseliminierung (Díaz-Uriarte & Alvarez de Andrés, 2006)
- Einsparpotenzial an Rechenressourcen

Übertragbarkeit

- Genauigkeitsverlust durch regionale Unterschiede (baukultur-historisch, Modellierungsunterschiede)
- Bsp: Klassifizierung von Dresden mit Trainingsdaten von Halle: 73 % (Verlust von 17 %)



Fazit

- **Kombination** der räumlichen Analysen mit Methoden der Mustererkennung und des maschinellen Lernens ermöglicht **automatische Klassifizierung**
- Ausschließliche **Nutzung von Geobasisdaten**
- Beschreibung der Gebäude über **geometrische, topologische, statistische und semantische Merkmale**
- **Nichtlineare Klassifikatoren** besser als lineare (besonders Ensemble-Methoden, wie Random Forest)
- **Klassifikationsgenauigkeit:** 90 – 95 % (Vektordaten), 76 - 88 % (Rasterkarten)
- **Einflussfaktoren:** Eingangsdaten, Siedlungstyp, Trainingsdatenmenge

Ausblick

- **Verschiedene Anwendungsszenarien denkbar** (Quantitativen Stadtforschung, Planung, Geomarketing, Energiemodellierung, Risikoabschätzung, Monitoring etc.)
- **Berücksichtigung weiterer Daten** (LoD2, Flurstücke, Bilddaten, semantische Information aus OSM, Firmendaten)
- Untersuchungen zur **Adaptierbarkeit** auf andere Probleme (Gebäudetypologie, Daten, Untersuchungsgebiete)
- **Evaluierung** projizierter sozioökonomischer Zielgrößen (z.B. Bevölkerung, Haushalte, Energieverbrauch etc.)
- Überwindung **regionaler Unterschiede**

Vielen Dank!

Robert Hecht
Leibniz-Institut für ökologische Raumentwicklung
Weberplatz 1
01217 Dresden
r.hecht@ioer.de

Dank
Jörg Hennersdorf, Anne Bräuer, Benjamin Richter und Michel Kaden für die Unterstützung bei der Erhebung der Referenzdaten

Referenzen

- Breimann (2001): L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- Díaz-Uriarte und Alvarez (2006): R. Díaz-Uriarte and S. Alvarez de Andrés, “Gene selection and classification of microarray data using random forest,” BMC Bioinformatics, vol. 7, p. 3, Jan. 2006.
- Herold et al (2012): H. Herold, G. Meinel, R. Hecht, E. Csaplovics: A GEOBIA Approach to Map Interpretation - Multitemporal Building Footprint Retrieval for High Resolution Monitoring of Spatial Urban Dynamics. In: International Conference on Geographic Object-Based Image Analysis, 2012, Rio de Janeiro, Brazil. Proceedings of the 4th GEOBIA. São José dos Campos : INPE, 2012, S.252-256
- Meinel et al (2009): G. Meinel, R. Hecht, and H. Herold, “Analyzing building stock using topographic maps and GIS,” Building Research & Information, vol. 37, no. 5–6, pp. 468–482, Sep. 2009.