

# gis.SCIENCE

DIE ZEITSCHRIFT FÜR GEOINFORMATIK

www.gis-biz.de 23. Jahrgang 4/2010



## URBAN DATA MINING

Mathematische Methoden in der urbanen Planung.

## OSM-DATEN-QUALITÄT

Vergleich der Straßennetze von OSM und NAVTEQ in Deutschland.

## KOOPERATION IN DER PLANUNG

Web-basierte Informationsplattform für die Raumplanung Taiwans.

## GEODATEN-FUSION IN GDIS

Informationsmehrwert durch Kombination verwandter Ressourcen.



## DIE ZEITEN ÄNDERN SICH, DIE GIS.SCIENCE AUCH! // TIMES ARE CHANGING, GIS.SCIENCE CHANGES TOO!

### EDITORIAL.

*Dirk Schmidbauer* **I**  
DIE ZEITEN ÄNDERN SICH,  
DIE GIS.SCIENCE AUCH!

### WISSENSCHAFTLICHE BEITRÄGE.

*Behnisch, Ultsch* **135**  
URBAN DATA MINING – EINE METHODIK ZUR  
RAUMBEZOGENEN WISSENSEXTRAKTION

*Ludwig, Voss, Krause-Traudes* **148**  
WIE GUT IST OPEN STREET MAP?  
ZUR METHODIK EINES AUTOMATISIERTEN OBJEKT-  
BASIERTEN VERGLEICHES DER STRASSENNETZE  
VON OSM UND NAVTEQ IN DEUTSCHLAND

*Lin, Streich* **159**  
DEZENTRALE KOOPERATION IN DER RAUM-  
PLANUNG DURCH EINSATZ EINER RAUM-  
INFORMATIONSPLATTFORM

*Wiemann* **166**  
KONZEPTION EINER FUSION VON GEODATEN  
UNTERSCHIEDLICHER QUELLEN IN GEODATEN-  
INFRASTRUKTUREN

### RUBRIKEN.

EDITORIAL **I**  
WISSENSCHAFTLICHE BEITRÄGE **135**  
AKTUELLES **172**

### //EDITORIAL.

*Dirk Schmidbauer* **I**  
TIMES ARE CHANGING,  
GIS.SCIENCE CHANGES TOO!

### //SCIENTIFIC PAPERS.

*Behnisch, Ultsch* **135**  
URBAN DATA MINING

*Ludwig, Voss, Krause-Traudes* **148**  
HOW GOOD IS OSM? – CONCERNING THE  
METHOD OF AN AUTOMATED AND OBJECT-  
WISE COMPARISON OF OSM AND NAVTEQ  
STREET NETWORK IN GERMANY

*Lin, Streich* **159**  
DECENTRALIZED COOPERATION IN SPATIAL  
PLANNING BY USING A SPATIAL INFORMATION  
PLATFORM

*Wiemann* **166**  
FUSION OF DIFFERENT GEODATA IN  
GEODATAINFRASTRUCTURES

### //RUBRICS.

EDITORIAL **I**  
SCIENTIFIC PAPERS **135**  
NEWS **172**

# URBAN DATA MINING – EINE METHODIK ZUR RAUMBEZOGENEN WISSENSEXTRAKTION

Martin Behnisch, Alfred Ultsch

**Zusammenfassung:** Durch den Fortschritt in der Informationstechnologie und das immer rapidere Anwachsen der Datenmengen steigen die Anforderungen an Systeme, die Wissen aus Daten extrahieren und darstellen. Urbanes Data Mining wird als Methodik zur Problemlösung verstanden, um logische oder mathematische, zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen mit Geobezug zu entdecken und Erkenntnisse (Wissen) aus numerischen Daten zu erzeugen. Mining impliziert einen aufwendigen Prozess der Suche nach verborgenen Zusammenhängen in einer großen Datenmenge. Eine für Menschen verständliche Erklärung des Zusammenhangs wird als Knowledge Discovery bezeichnet. Data Mining in Verbindung mit Techniken der Knowledge Discovery erhalten auch zunehmende Bedeutung für urbane Untersuchungs- und Planungsprozesse. Die Techniken des Urban Data Mining werden an einem raumbezogenen Datensatz in ihrer Anwendung beispielhaft vorgestellt. Dieser bezieht sich auf die Erfassung und Strukturierung von bereits erfolgten Agglomerations- und Verdichtungsprozessen in Deutschland. Ausgewählte raumstrukturelle Kenngrößen werden hierzu einer Einzeluntersuchung unterzogen und auf ihre Eignung zur Klassenbildung geprüft. Angestrebt wird eine sachlich-räumliche Differenzierung des deutschen Gemeindegemeinschaftensystems.

**Schlüsselwörter:** Urban Data Mining, Urban spatial monitoring, Knowledge Discovery

## // URBAN DATA MINING

// **Abstract:** Most of the large databases currently available have a strong geospatial component and contain potentially useful information that might be of value. »Urban Data Mining« is a methodological approach to discover logical, mathematical and partly complex descriptions of patterns and regularities inside a set of data. Data mining is defined as the inspection of data with the aim of discovering knowledge. Mining implies a laborious process of searching for hidden information in a large amount of data. Important requirements for 'knowledge discovery' are interpretability, novelty and usefulness of results. Data mining in connection with knowledge discovery techniques will be of increasing importance for the urban research and planning processes. The techniques of Urban Data Mining and their application will be presented using a dataset with reference to communities. The aim is to identify agglomeration and urbanization processes which already took place in Germany. Selected parameters will be subjected to an individual investigation and be tested on their ability for creating classes. It is intended to differentiate the system of German communities in a space-oriented way.

**Keywords:** Urban Data Mining, Urban spatial monitoring, Knowledge Discovery

### Anschrift der Autoren

Dr.-Ing. Martin Behnisch  
Limmattalstrasse 23  
CH-8049 Zürich  
E: Behnisch@urban-data-mining.de

Prof. Dr. Alfred Ultsch  
Philipps-Universität Marburg  
Datenbionik FB 12  
Hans-Meerwein-Straße  
D-35032 Marburg  
E: ultsch@informatik.uni-marburg.de

## 1. PLANUNG IM KONTEXT VON URBANEN ENTWICKLUNGSTRENDS

### 1.1 URBANITÄT & KONZENTRATION

In den nächsten zwei Generationen wird die urbane Lebensform weltweit dominant (Lichtensteiger 2006). Urbanität ist definiert als eine bestimmte Organisation des Politischen, als Demokratie, als Organisation des Ökonomischen und als eine bestimmte Art zu leben (Häussermann, Siebel 1999). Die vergangenen fünf Jahrzehnte haben gezeigt, dass eine beschleunigte Verschiebung menschlicher Siedlungen vom Ruralen ins Urbane stattgefunden hat (Baccini et al. 2002). Die Entfernung der Orte oder Räume ist in Anbetracht des technologischen Fortschritts von sekundärer Bedeutung (Sieverts 1997). Einerseits ist die Geschwindigkeitszunahme der Verkehrssysteme dafür verantwortlich, und andererseits beinhaltet der virtuelle Daten-Highway die Auflösung des geographischen Maßstabs (Prigge 1999). Es ergeben sich neue Anforderungen – überregionale Bezüge und Beziehungen der Bewohner gewinnen an Bedeutung – zur Messung von Zusammenhängen und Ähnlichkeiten von Objekten. In Deutschland besteht das Ziel der Raumentwicklungspolitik darin, räumliche/urbane Strukturen und Entwicklungen zu kennen (§ 18 Abs. 5 ROG, Raumordnungsgesetz), zu bewerten und zukünftige Entwicklungstrends in Strategien und Entscheidungen zu berücksichtigen. Allerdings haben zunehmende Kapitalkonzentration in den Zentren der wirtschaftlichen Entwicklung und gleichgerichtete Wanderungsbewegungen zu Wachstums- und Wohlstandsunterschieden zwischen den einzelnen Teilräumen geführt, die den gesellschaftspolitischen Zielvorstellungen – insbesondere dem Ziel gleichwertiger Lebensbedingungen für alle Teile der Bevölkerung (§ 1 ROG) – in eklatanter Weise widersprechen. Einerseits sind in Entleerungsgebieten weder ausreichende Erwerbsmöglichkeiten noch eine mit anderen Regionen vergleichbare Versorgung mit öffentlichen und privaten Gütern und Dienstleistungen gewährleistet, andererseits ist in einzelnen Verdichtungsräumen die Grenze der Belastbarkeit der natürlichen Umwelt bzw. Infrastruktureinrichtungen bereits überschritten. Es ist zu erwarten, dass sich Konzentrationstendenzen von Wirtschaft und Bevölkerung in Zukunft

fortsetzen, wodurch die Lebensqualität der Bevölkerung in Entleerungsgebieten und Verdichtungsräumen weiter beeinträchtigt wird. Diese Fehlentwicklungen lassen die Einsicht wachsen, dass eine gesellschaftspolitisch wünschenswerte Allokation der volkswirtschaftlichen Ressourcen durch die Selbststeuerungskräfte der Marktmechanismen nicht erreicht werden kann und deutlich stärker Maßnahmen der räumlichen Planung und regionalen Wirtschaftspolitik erforderlich werden. Als Forschungsgegenstand bietet die Raum- und Stadtstruktur seit Generationen für verschiedene wissenschaftliche Disziplinen interessante Arbeitsfelder, wobei der Erkenntnisgewinn in jüngster Zeit besonders durch inter- und transdisziplinäre Ansätze gelingt (Frühwald 1997). Es handelt sich um problemorientierte und von den Disziplinen unabhängige, sowohl praxis- als auch theoriebasierte Arbeitsweisen, die auf einer freien Wahl der Methodenanwendung und -entwicklung beruhen (Jaeger, Scheringer 1988).

### 1.2 DATENBESTAND UND LANGFRISTIGES RAUMVERSTÄNDNIS

Raumbezogene Daten wurden in der Vergangenheit in vielen Fällen unzugänglich, aber auch unzulänglich archiviert und oft nur für spezifische Zwecke zeitnah freigegeben oder lediglich in aggregierter Form routinemäßig aufbereitet. Erst im Verlauf der Zeit und aufgrund der Bestrebungen zur freien Verfügbarkeit dieser Daten (Open Geospatial Consortium 2010) erlangten einige ihren tatsächlichen Wert bzw. überhaupt ihre analytische Relevanz.

Durch das rapide Anwachsen von frei verfügbaren raumbezogenen Daten und den Fortschritt in der Informationstechnologie sind in jüngster Zeit die Anforderungen an Systeme deutlich gestiegen, die Wissen aus Daten mit Raumbezug extrahieren und abbilden können (Keim 2002). Bei vielen quantitativen Fragestellungen der Raumanalyse handelt es sich neuerdings um spezifische multifaktorielle (vieldimensionale) Probleme. Diese sind z.B. in kombinierter Form durch die Entwicklung von Bevölkerung, Arbeitsplätzen und Infrastruktur sowie deren räumlicher Verteilung und ihren Auswirkungen auf die Flächennutzung und den Verkehr geprägt (Siedentop 2003). Veränderte Rahmenbedingungen, wie der demographische Wandel, die Globalisierung oder die

weltweite Finanzkrise, vergrößern den Einfluss auf die Varianz oder Existenz von räumlichen und funktionalen Erscheinungsbildern zusätzlich. Vermutlich besteht die Herausforderung in Zukunft darin, den Überfluss an raumbezogenen Daten und Informationen überhaupt noch systematisch auswertbar zu gestalten und ihre Einbindung in Erfahrungszusammenhänge, durch welche erst Wissen geschaffen wird, dauerhaft abzusichern. Mit Blick auf die weiter fortschreitende Zunahme der Datenmenge und heute bereits erkennbare Vergrößerung der problembezogenen Dimensionen ist festzustellen, dass zur Analyse von derartig komplexen raumbezogenen Problemen und ebenso facettenreichen inhaltlichen Zusammenhängen nur selten die Integrationsmöglichkeiten von Konzepten des Data Mining und der Knowledge Discovery (in Databases) diskutiert werden (Miller, Han 2009): *„Due to the growth and wide availability of geo-referenced data in recent years, traditional spatial analysis tools are far from adequate at handling the huge volumes of data and the growing complexity of spatial analysis task. Geographic data mining and knowledge discovery represent important directions in the development of a new generation of spatial analysis tools in data-rich environment.“*

Aufgabe im Data Mining ist allgemein die Entdeckung von neuen und/oder verborgenen Zusammenhängen in hochdimensionalen Daten (Hand et al. 2001). Ausgehend von einer Menge von Zahlen, wird die Darstellung von bislang unbekanntem aber nützlichem Sachverhalten verfolgt. Die Wissensentdeckung – im Sinne der Disziplin Knowledge Discovery gekennzeichnet durch Syntax, Semantik und Pragmatik – unterstützt verfahrensbasiert die (natürlich-) sprachliche Interpretation und Ableitung von Wissen (Erkenntnissen) aus verschiedensten Datensammlungen und erlaubt darüber hinaus die Reproduktionen in maschineller Form (z.B. als wissensbasiertes System). Vor diesem Hintergrund unterliegen quantitative Messmodelle und automatisierte Beobachtungssysteme (Monitoring Systems) einem hohen Entwicklungsdruck, um sich anhand vielfältiger interoperabler Datenangebote (INSPIRE 2010) als Untersuchungsobjekt (urban/spatial unit) räumlich und zeitlich messen bzw. vergleichen zu können.

Es besteht ein Bedarf direkt verständliche Grundlagen für Planer und Entscheidungsträger zu erarbeiten (Arlt et al. 2001), die langfristig zum Aufbau und der Weiterentwicklung von umfassenderen und strategisch angemesseneren Planungswerkzeugen führen könnten. Vorstellbar werden – unter der Annahme eines auf diese Weise realisierten langfristig orientierten und wirkmächtigen Raumverständnisses – in ihrer Effizienz gesteigerte Förderinstrumente, die präzise und gezielt den Ausgleich oder Umgang mit interregionalen Unterschieden im jeweils erreichten Entwicklungsstand ermöglichen.

## 2. URBAN DATA MINING

### 2.1 VERFAHRENSSCHRITTE UND ZYKLISCHER ABLAUF

Der Begriff des ‚Urban Data Mining‘ charakterisiert die Erarbeitung einer für den urbanen Kontext entwickelten Methodik, die dazu dient, logische oder mathematische

und zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen zu entdecken sowie daraus Erkenntnisse abzuleiten und zu bewerten (Behnisch 2009). Urban Data Mining trägt unterstützend zum automatischen Generieren und Prüfen von Hypothesen bei. Verfolgt wird das Ziel der Wissensentdeckung, d.h. die planungsbezogene Übersetzung von Daten zu Wissen (Erkenntnis), welches zusätzlich zu den in Abschnitt 1.2 genannten Eigenschaften – Semantik, Syntax, Pragmatik – auch durch die Attribute „neu“, „nützlich“ und „überraschend“ charakterisierbar ist. Beschrieben wird ein zyklischer Arbeitsprozess (Abb. 1), so dass schrittweise gewonnene Erkenntnisse über raumbezogene Daten validiert und als Eingangsstufe der Folgeschritte verwendet werden. Wesentliche Verfahrensschritte sind dabei die Datensichtung, die Strukturerkennung, die Strukturbildung, die Strukturprüfung, die Operationalisierung und die Wissenskonzersion.

### 2.2 DATENBEISPIEL: „AGGLOMERATION UND VERDICHTUNG“

Die einzelnen Verfahrensschritte im Urban Data Mining sollen im Folgenden in ihrer Anwendung präzisiert werden. Der verwendete raumbezogene Datensatz bezieht sich auf die Gesamtmenge der 12.430 Gemeinden in Deutschland im Gebietsstand 2004. Sechs raumstrukturelle Untersuchungsvariablen werden auf ihre Eignung zur Klassenbildung geprüft. Angestrebt wird eine sachlich-räumliche Differenzierung des deutschen Gemeindegensystems im Sinne der Erfassung und Strukturierung von Agglomerations- und Verdichtungsprozessen. Im Vordergrund steht allerdings primär die Vorstellung der Arbeitsweisen, die zukünftig bei entsprechender Datenverfügbarkeit weitere Anwendungsmöglichkeiten für Untersuchungen von hochdimensionalen Daten bieten.

Die Variablen sind in Tabelle 1 durch die Messvorschrift charakterisiert.

Im Spannungsfeld einer geeigneten raumstrukturellen Abgrenzung zur Erfassung von Agglomerations- und Verdichtungsprozessen sei auf die sogenannte BOUSTEDT-Systematik verwiesen. Das dazugehörige Modell ist für Planungszwecke und als Instrument zur Beobachtung des Agglomerationsprozesses 1953 (Boustedt 1953) entwickelt und 1971 nochmals an Lebens- und Arbeitsverhältnisse angepasst worden (Boustedt 1975). Im Jahr 2000 wurden sogenannte BIK-Regionen für das wiedervereinigte Deutschland im Anschluss an eine bereits im Jahr 1987 erfolgte BOUSTEDT-Revision aufgebaut (Behrens, Marhenke 1997).

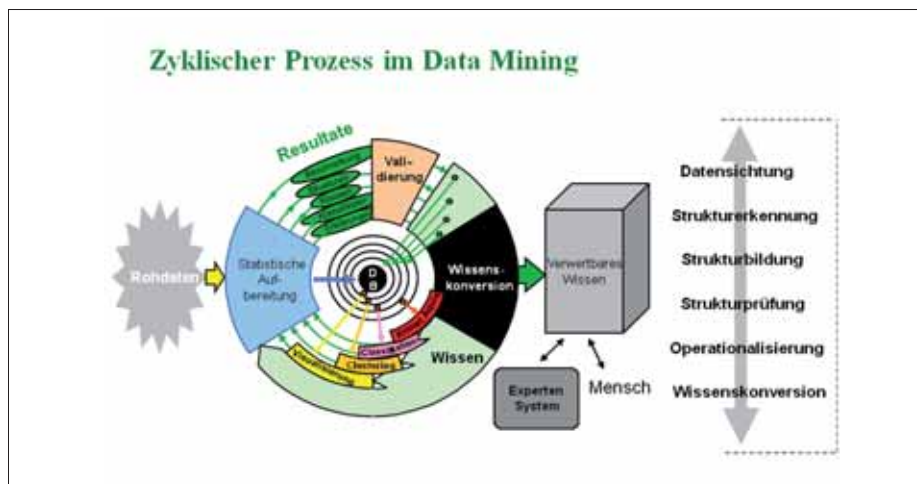


Abbildung 1: Zyklischer Prozess im ‚Urban Data Mining‘ (Behnisch, 2009).

Messgröße	Messvorschrift	Einheit
(1) Verstädterung	Anteil Siedlungs- und Verkehrsfläche an Katasterfläche	[%]
(2) Nutzungsproportion	Anteil Gebäude- und Freiflächen an Siedlungs- und Verkehrsfläche	[%]
(3) Konzentration	Einwohner und Arbeitsplätze je km <sup>2</sup> Gebäude- und Freifläche	[Personen je km <sup>2</sup> ]
(4) Entdichtung	Anteil Ein-/Zweifamilienhäuser am Wohnbaubestand	[%]
(5) Beschäftigungsdisparität	Quotient von sozialversicherungspflichtig Beschäftigten am Arbeitsort und sozialversicherungspflichtig Beschäftigten am Wohnort * 100	[dimensionslos]
(6) Erreichbarkeit	Fahrzeit zum nächsten Oberzentrum (PKW)	[Minuten]

Tabelle 1: Übersicht zu sechs raumstrukturellen Kenngrößen (Datengrundlage: Statistiken zur Raumordnung des Bundesinstituts für Bau-, Stadt- und Raumforschung (BBSR)).

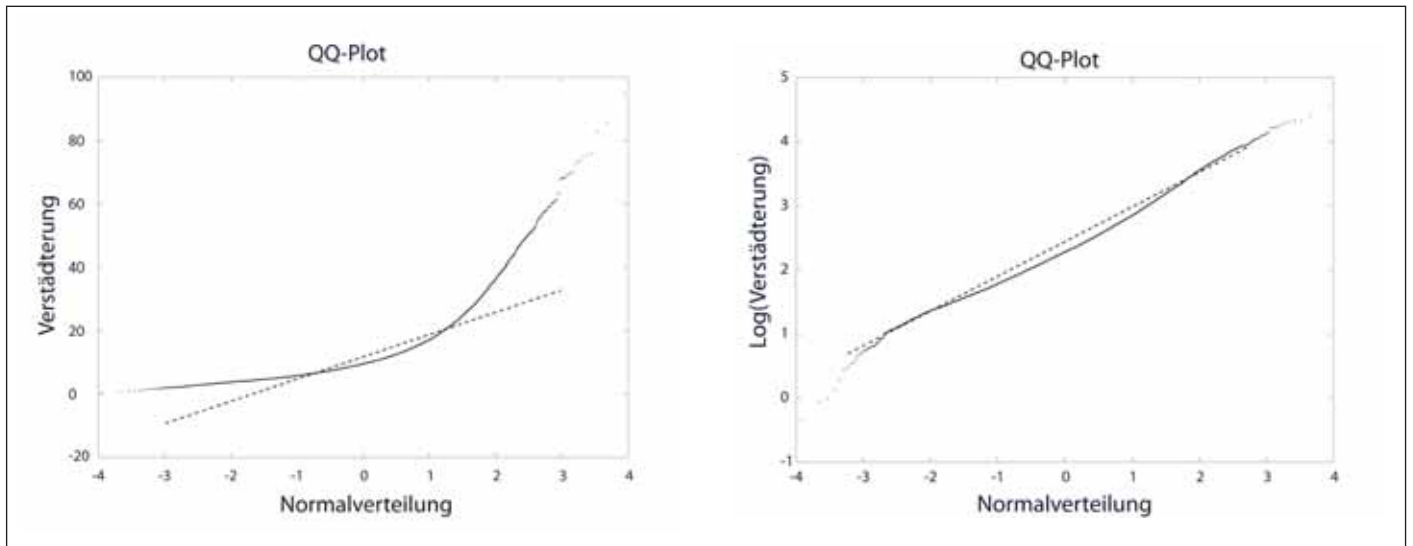


Abbildung 2: QQ-Plot, Messgröße Verstärderung (Eigene Bearbeitung).

### 2.3 DATENSICHTUNG

Eine umfassende Datensichtung einschließlich Vorverarbeitung erfüllt im Kontext der zyklischen Arbeitsweise des Urban Data Mining (Abb. 1) den Zweck den Grundvoraussetzungen üblicher Data Mining Methoden zu entsprechen. Die Datensichtung beginnt mit der Durchsicht der Objektdaten jeder einzelnen Variablen, indem man sich einen Überblick von Anzahl, Art, Wertebereichen und insbesondere der Verteilung verschafft. Klassische statistische Verfahren setzen – mehr oder weniger explizit – voraus, dass die Daten gewisse Eigenschaften (z.B. Normalverteilung, Unabhängigkeit, Linearität, Stetigkeit usw.) und Strukturen (z.B. ellipsoid oder kugelförmig) aufweisen. Diese gehen dann in die Modelle ein, die den Verfahren zur Strukturbildung (z.B. Clusterbildung) zugrunde liegen (Ultsch 2010). Da die Verteilung der Variablen üblicherweise nicht bekannt ist, besteht die Aufgabe darin, eine Hypothese über eine empirisch beobachtete Variable zu gewinnen. Geeignet ist die Visualisierung dazugehöriger Sachverhalte, die neben Lage- und Streumaßen eine Variablenbeschreibung ergänzen (z.B. Histogramme, Box-Plots, Quantil/Quantil-Plots (QQ-Plots), Pareto Density Estimation, PDE (Ultsch 2003), Modellierung mit Gauß-Mixturen). QQ-Plots dienen dazu, eine vorgelegte Verteilung graphisch mit einer standardisierten Verteilung, z.B. Normalverteilung oder Gleichverteilung, zu vergleichen. Bilden die so entstandenen Punkte annähernd eine Gerade, so kann davon ausgegangen werden, dass die beiden Verteilungen

gleich sind. Werden Abweichungen von Standardnormalverteilungen festgestellt, so gilt es entsprechende Umformungen (Transformationen) festzulegen, mit denen die Daten in eine bekannte Verteilung transformiert werden können. Die sogenannte ‚ladder of power‘ (Hartung 2005) ist eine Auflistung für die Größe (power) des Exponenten  $p$  zu  $y = x^p$ . Im Umkehrschluss kann aus dieser Transformation auf die empirische Verteilung geschlossen werden.

Für die Messgröße ‚Verstärderung‘ wird im Sinne des Arbeitsschrittes Datensichtung eine Verteilungsuntersuchung beispielhaft

dargelegt. Abbildung 2 zeigt QQ-Plots für diese Messgröße. Es werden die Quantile der Variable auf der Y-Achse aufgetragen.

Der QQ-Plot der Ausgangsgröße zeigt einen konkaven Bogen, wobei dies auf eine nichtnormale, „schiefe“ Verteilung hindeutet. Um eine schiefe Verteilung dennoch charakterisieren zu können, kann eine nicht-lineare Transformation angewendet werden. Eine Transformation auf annähernde Normalverteilung kann bei dieser Variablen durch Logarithmieren erreicht werden. In Teilbereichen folgen die logarithmierten Daten gemäß der rechten Abbildung in

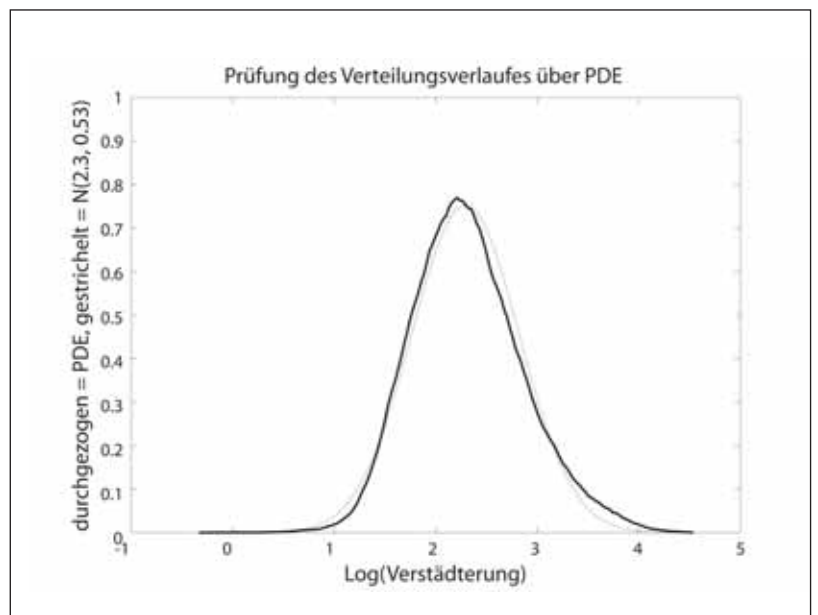


Abbildung 3: PDE, Messgröße LogVerstärderung (Quelle: Eigene Bearbeitung).

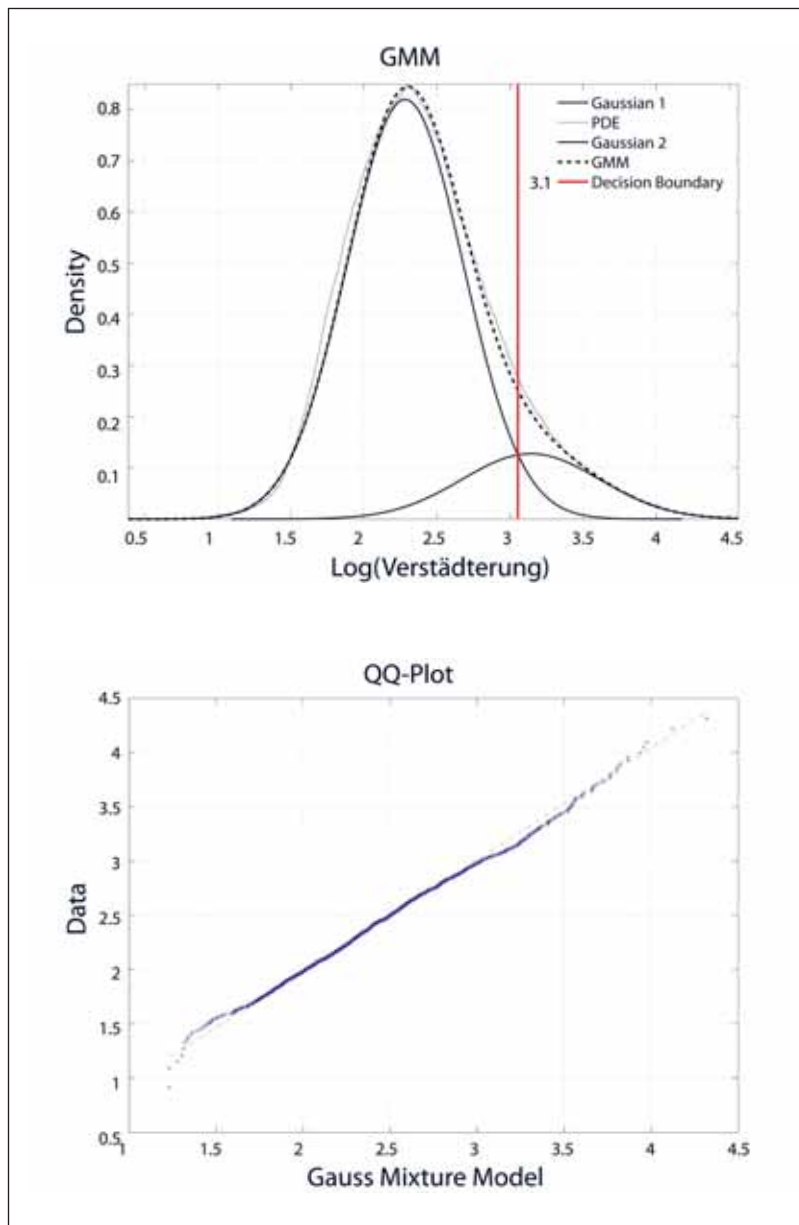


Abbildung 4: GMM und dazugehöriger QQ-Plot ‚Log(Verstädterung)‘.

Abbildung 2 einer Geraden, doch sind zusätzliche Unebenheiten vorhanden. Die Verteilungsuntersuchung der Messgröße ‚Verstädterung‘ wird ergänzt durch den PDE-Plot, Pareto Density Estimation (Ultsch 2003). Es wird die Datendichte an allen verschiedenen Datenpunkten der Datenmenge abgeschätzt. Abbildung 3 zeigt, dass der PDE der logarithmierten empirischen Messwerte (durchgezogene Kurve) und der Normalverteilung (gestrichelte Kurve) sich annähern, jedoch keine Überdeckung vorhanden ist.

Die Prüfung auf Log-Normalverteilung hat ergeben, dass für die Variable ‚Verstädterung‘ eine komplexere Modellierung sinnvoll wäre. Eine genauere Approximation ist

mit einer Gaußschen Mischverteilung (GMM) möglich. Die Verteilungsdichte eines Variablenvektors  $X$  lässt sich meistens nur ungenau mit Hilfe einer einzelnen Normal- bzw. Gaußverteilung beschreiben. Zur Schätzung der Parameter eines GMMs ist der Expectation-Maximization (EM) Algorithmus einsetzbar (Hand 2001). Die Lösungen sind besonders von den Initialisierungsparametern abhängig, so dass die Ergebnisse mehrfach zu berechnen sind. Als Gütekriterium eignet sich die Pareto-Dichteschätzung.

Gemäß Abbildung 4 (Vergleich von PDE und GMM) lässt sich die gegebene Verteilung mit zwei Gauß-Mixturen gut modellieren. Der QQ-Plot bestätigt die Verteilungsannahme – so folgt der Plot des GMMs und der empirischen Daten einer Geraden. Die Modellierung bildet die Grundlage, um Entscheidungsgrenzen (siehe Schnittpunkt der Kurvenverläufe) für Klassifizierungsvorgänge aufzubauen. Es wird die klassenbedingte Wahrscheinlichkeitsdichte berechnet, auf Basis derer ein Likelihood-Ratio-Klassifizierer ein gegebenes Muster einer Klasse zuweist (a posteriori Wahrscheinlichkeit). Diese liegt hier bei 20 Prozent (siehe  $\text{Log}(3, 1)$ ).

Ist die Beschreibung von einzelnen Variablen erfolgt (Tab. 2), beginnt nachfolgend die Untersuchung der Datensätze auf Zusammenhänge bzw. Abhängigkeiten zwischen zwei oder mehreren Variablen. Es wird zunächst geprüft, ob redundante Informationen in den Datensätzen existieren und Hinweise auf die Struktur des Datensatzes bzw. des Grundproblems zu erkennen sind. Hierzu sind einerseits visuelle Methoden wie Streu-Diagramme (Scatter-Plots) und andererseits statistische Maßzahlen wie Korrelationsmaße (z.B. Pearson, Spearmans oder Kendalls Rangkorrelationskoeffizient) einsetzbar, die den Abhängigkeitsgrad der Variablen messen. Korrelationen innerhalb der Merkmalsstruktur sind unvermeidbar (Vogel 1975), doch kann verschieden darauf reagiert werden, z.B. durch Variableneliminierung, Gewichtungsschema (Fischer 1982) oder Faktorenanalyse.

Im Sinne der Vergleichbarkeit von Daten ist eine Entscheidung zum Umgang mit Ausreißern, d.h. Objekte mit extremer Werteausprägung als auch der Fehlstellenbehandlung, zu treffen. Da in den meisten Fällen die Variablen nicht in gleicher Dimension vorliegen, sind die Variablen geeignet zu skalieren, z.B. durch Normierung, Standardisierung oder Lineare Transformation. Abbildung 5 vermittelt am gegebenen Untersuchungsdatensatz eine Vorstellung von linearen/nichtlinearen Zusammenhängen der Variablen (Scatter-Plot). Die Histogramme verdeutlichen den einheitlichen Verteilungsverlauf infolge zuvor durchgeführter Transformationen.

Um Objekte mehrdimensional miteinander vergleichen zu können, ist es notwendig, ein quantifizierbares Maß zu verwenden, welches Prinzipien zur Bestimmung der Gleichheit, Ähnlichkeit bzw. Verschiedenheit berücksichtigt. Bei der Festlegung einer geeigneten Metrik sind die Erkenntnisse der Datensichtung und damit die

Messgröße	grob inspiziert	GMM (Grenzen)	Regel	Gemeinden pro Klasse
(1) Verstädterung	Log(Data) folgt Normalverteilung	Bimodal, 2 Gaußverteilungen, Grenze: 20 %, Log (Data): 3.1	Klasse 1: Data ≤20 Klasse 2: Data >20	Klasse 1, [11029] Anteil: 88,73 % Klasse 2, [1401] Anteil: 11,27 %
(2) Nutzungs- proportion	Ohne Transformation	Bimodale Verteilung, 2 Gauß-Verteilungen, Grenze: 40 %	Klasse 1: Data ≤40 Klasse 2: Data >40	Klasse 1, [3938] Anteil: 31,68 % Klasse 2, [8492] Anteil: 68,32 %
(3) Konzentration	Log(Data) folgt Normalverteilung	Multimodal, 3 Gaußverteilungen, Grenzen: 2500, 4000 Log(Data): 7.9, 8.3	Klasse 1: Data ≤2500 Klasse 2: 2500<Data ≤4000 Klasse 3: Data >4000	Klasse 1, [5263] Anteil: 42,34 % Klasse 2, [4802] Anteil: 38,63 % Klasse 3, [2365] Anteil: 19,03 %
(4) Entdichtung	Umkehransatz der Variablendaten: $y = \log(100 - \text{Data})$ Log((100-Data)+1) folgt Normalverteilung	Multimodal, 3 Gaußverteilungen, Grenze: 2 %, 15 % Log(Data): 1,1 und 2,8 invertierte Grenze: 98 %, 80 %	Klasse 1: Data <85 Klasse 2: 85≤Data<98 Klasse 3: Data ≥98	Klasse 1, [1420] Anteil: 11,42 % Klasse 2, [8869] Anteil: 71,35 % Klasse 3, [2141] Anteil 17,22 % 100 % Einfamilienhäuser [518]
(5) Beschäfti- gungsdisparität	Log(Data+1) folgt Normalverteilung	Multimodal, Drei Gaußverteilun- gen, sachlogisch erzwungen: 100	Klasse 1: Data <100 Klasse 2: Data ≥100	Klasse 1, [10808] Anteil: 86,95 % Klasse 2, [1622] Anteil: 13,05 %
(6) Erreichbarkeit	Log(Data+1) folgt Normalverteilung	Bimodal, 2 Gauß-Verteilungen Grenze: 30 Min.. Log (Data): 3.3	Klasse 1: Data=0 Klasse 2: 0<Data ≤30 Klasse 3: Data > 30 Klasse 0: Data ='NAN'	Klasse 1, [133] Anteil: 1,07 % Klasse 2, [5092] Anteil: 40,96 % Klasse 3, [6964] Anteil: 56,02 % NAN: [241]

Tabelle 1: Übersicht zu sechs raumstrukturellen Kenngrößen.

Grundeigenschaften der Daten mit einzu-  
beziehen (Bock 1974). Die Wahl des Ähn-  
lichkeits- bzw. Distanzmaßes ist von ent-  
scheidender Bedeutung, da über Ähnlich-  
keit bzw. Unähnlichkeit von Objekten ent-  
schieden wird. Zum Zweck der Validierung  
sollte die Angabe von besonders ähnli-  
chen/unähnlichen Objekten erfolgen.

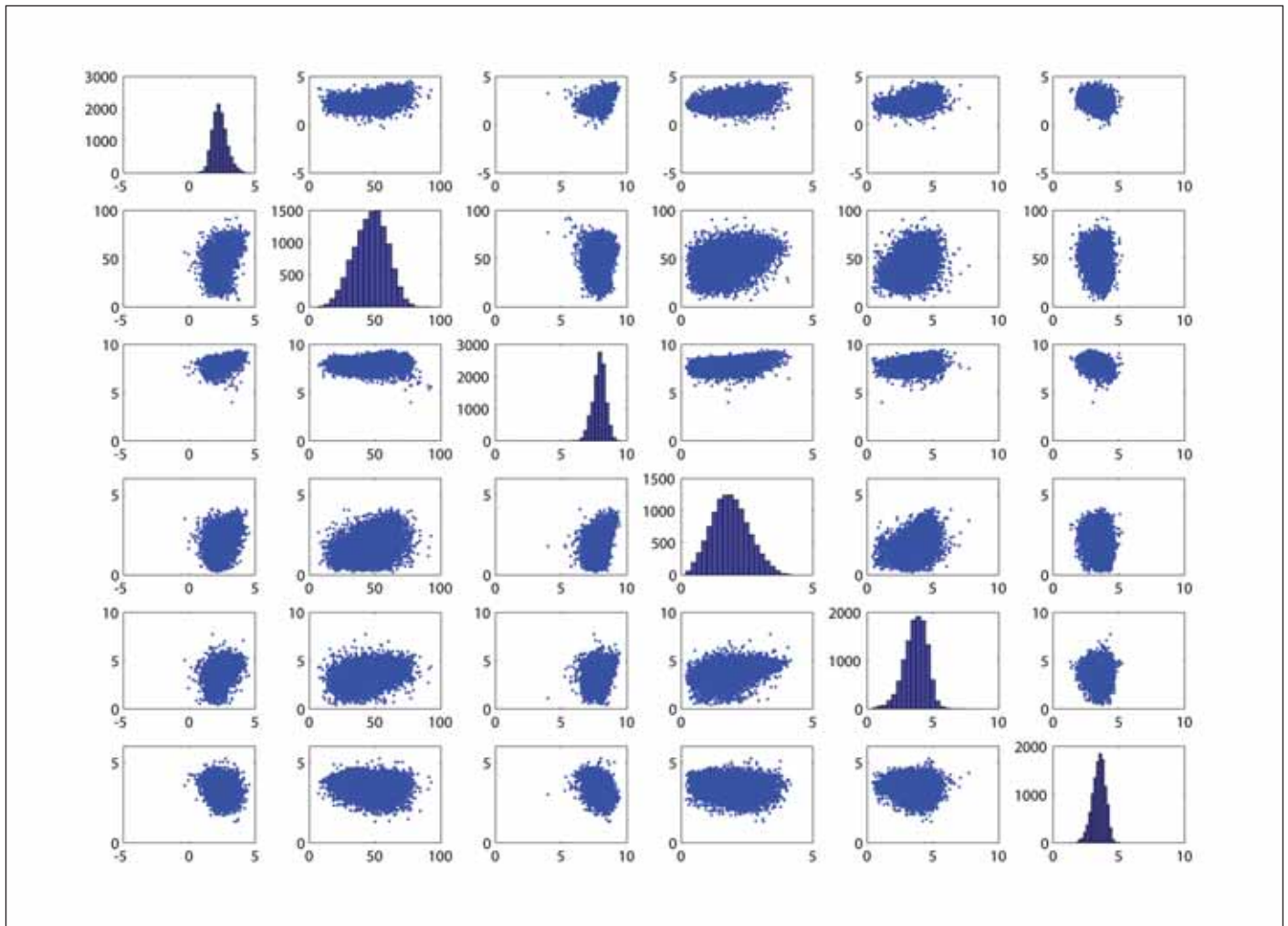
## 2.4 STRUKTURERKENNUNG

Durch die Strukturerkennung werden hoch-  
dimensionale Daten für einen menschli-  
chen Betrachter geeignet dargestellt, wo-  
bei dies in der Regel in graphischer Form  
geschieht. Einsetzbar sind Projektionsver-  
fahren, welche die wesentlichen Eigen-

schaften eines Datenbestandes aus dem  
zunächst hochdimensionalen und unein-  
sehbar Ursprungsraum in einen darstell-  
baren zwei- oder dreidimensionalen Re-  
präsentationsraum übertragen.

Es ist zwischen linearen und nichtlinea-  
ren Projektionsmethoden zu trennen, wo-  
bei gerade die nichtlinearen Projektionen  
sich dazu eignen, strukturelle Eigen-  
schaften (räumliche Beziehungen, Nachbar-  
schaftsverhältnisse) der Daten abzubild-  
en. Genannt seien die Multidimensiona-  
le Skalierung, Sammons-Abbildungen und  
insbesondere die Merkmalskarten, die  
den neuronalen Netzen zuzuordnen sind  
(Ultsch 2005). Übliche Selbstorganisie-

rende Merkmalskarten, SOM (Kohonen  
1982) sind durch eine geringe Anzahl  
Neuronen charakterisiert. Verwendet man  
eine sehr große Anzahl Neuronen, so ist  
es möglich, Strukturen in der Merkmalskar-  
te durch Emergenz abzubilden (jedem  
Neuron ist ein Prototyp im hochdimension-  
alen Datenraum zugeordnet). Die Emer-  
gente Selbstorganisierende Merkmalskar-  
te (ESOM) wurde entwickelt, um mit Hilfe  
einer dreidimensionalen Landschaftsdar-  
stellung eine Strukturerkennung in den Da-  
ten zu ermöglichen (Ultsch 1999). Zusam-  
mengehörnde Daten liegen in Tälern,  
während unterschiedliche Bereiche (Klas-  
sen) durch Mauern oder Gebirgszüge



Messgroe	Verstadterung	Nutzungsproportion	Konzentration	Entdichtung	Beschaftigungsdisparitat	Fahrzeit
MAXCORR (DATA)	0,55	0,40	0,55	0,52	0,37	0,31
MAXCORR (TRANSDATA)	0,52	0,41	0,52	0,41	0,52	0,41

Abbildung 5: Zusammenhange, Korrelation (Quelle, Eigene Bearbeitung).

getrennt werden. Die Abbildung 6 zeigt eine bliche Strukturerkennung in Form einer U\*-Map (Inseldarstellung). Es handelt sich um Daten, die eine hochdimensionale Struktur reprasentieren (Behnisch 2009) und zu klar unterscheidbaren Klassen (Gemeinden im Kontext von Schrumpfung und Wachstum) fhren.

Auf Grundlage der sechs raumstrukturellen Variablen wurde analog dazu geprft, ob und in welcher Weise eine Klassenbildung plausibel und sinnvoll ist. Abbildung 7 zeigt die Anwendung einer mehrdimensionalen Skalierung. Die Vermutung, dass es nur schwer mglich ist, unterscheidbare Klassen mit Hilfe eines Clusteralgorithmus

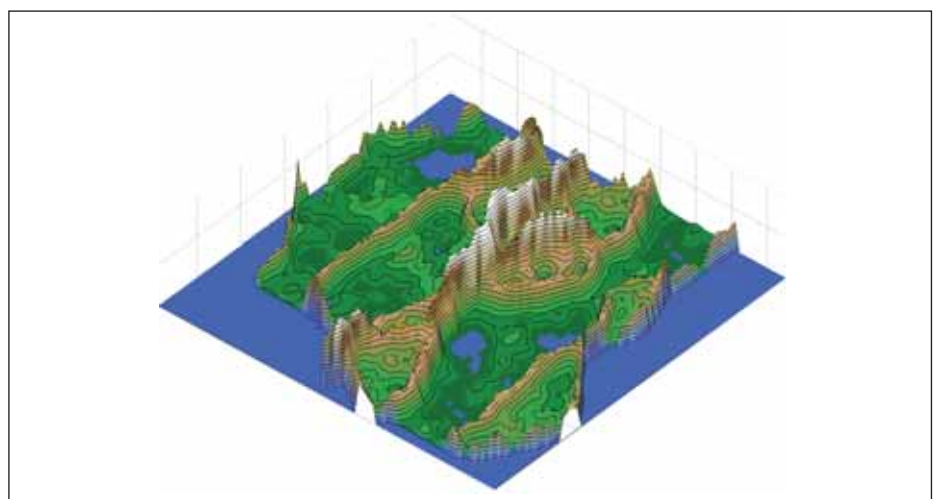


Abbildung 6: U\*-Map (Inseldarstellung) (Quelle: Behnisch 2009).

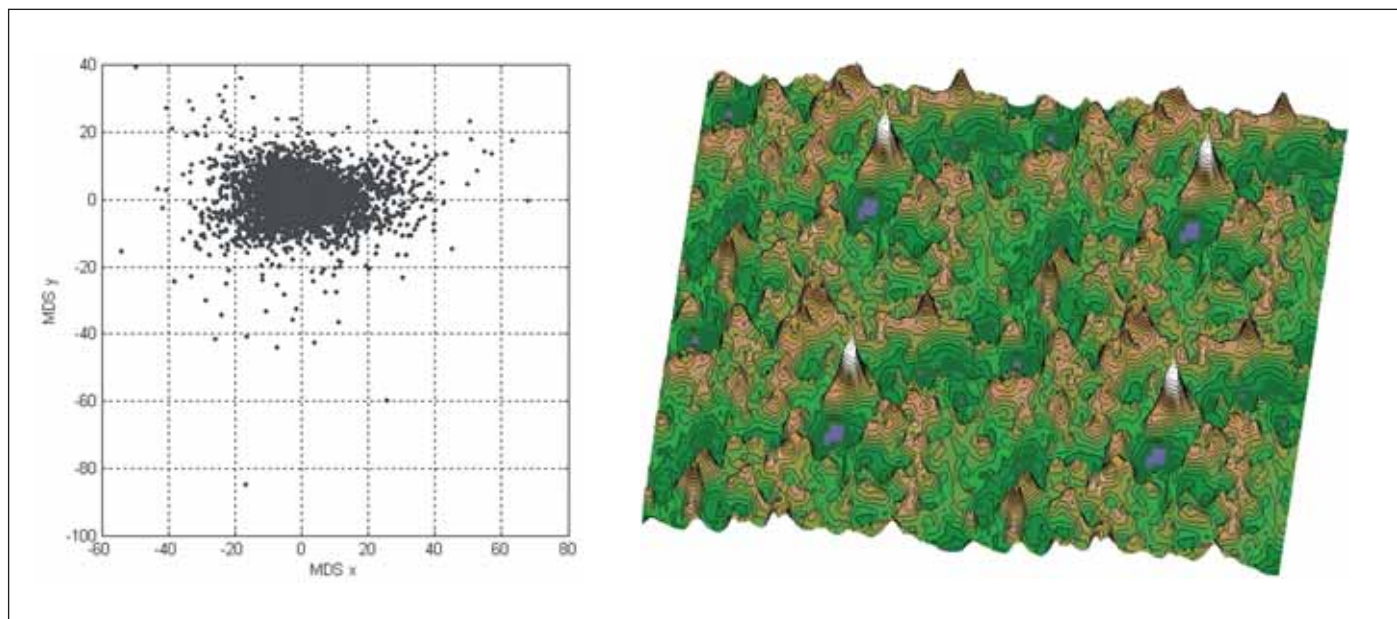


Abbildung 7: MDS und U\*-Matrix (Quelle: Behnisch 2009).

bilden zu können, wird durch die Emergente SOM aus Abbildung 7 ohne deutliche Struktur bestätigt. Aufgrund dieser Abbildungen wird die Notwendigkeit einer Strukturerkennung vor einer geplanten Strukturbildung betont.

## 2.5 STRUKTURBILDUNG

Die Klassifikation bildet ein wesentliches Instrument im ‚Urban Data Mining‘. Es sind drei wesentliche Klassifizierungsprobleme zu benennen (Deichsel, Trampisch 1985):

- 1 **Aufgliederungsproblem (Klassifikation, im engeren Sinn):** Es liegen keine Informationen über Gruppen in der Gesamtheit vor. Die Aufgabe besteht darin, die Gesamtheit oder Stichproben aus der Gesamtheit in eine zunächst unbekannt Anzahl möglichst homogener und einander möglichst ungleichartiger Gruppen zu zerlegen. Die zu lösenden Hauptprobleme bestehen in der Bestimmung der Anzahl der Gruppen und in der Zuordnung der Einheiten zu diesen Gruppen. Die Datenmatrix bildet die Informationsquelle.
- 2 **Schichtungsproblem (Mischform):** Zu bestimmen ist eine Untergliederung der Gesamtheit in Teilgesamtheiten (Schichten) mit a priori vorbestimmter Anzahl. Die Datenmatrix ermöglicht eine Bearbeitung und wird zusätzlich durch das Wissen über Anzahl und Eigenschaften der Schichten ergänzt.
- 3 **Zuordnungsproblem (Diskriminanzproblem):** Es liegen Informationen über An-

zahl und Eigenschaften von Teilgesamtheiten a priori vor. Die Aufgabe besteht darin, den schon bekannten Teilgesamtheiten die aus einer Grundgesamtheit entnommenen Einheiten anhand ihrer Variablenwerte mit möglichst großer Sicherheit zuzuordnen. Die bereits definierten Teilgesamtheiten und die Datenmatrix unterstützen die Vorgehensweise. Die Clusteranalyse folgt dem Wunsch, den Klassifikationsprozess systematisch und quantitativ erfassen zu wollen und durch Berücksichtigung numerischer Kriterien die Güte von Gruppierungen „objektiv“ zu vergleichen. Es erfolgt dabei die Durchmusterung und Auswertung von entweder Ähnlichkeits- oder Distanzmatrizen. Die Verfahren der Strukturbildung sind außerordentlich zahlreich und können nach verschiedenen Kriterien (z.B. Form, Dichte, Vagheit) systematisiert werden. Im Kontext der Klassifizierung sei zusätzlich auf Gaußsche Mixture-Modelle (Gaussian Mixture Models, GMM) verwiesen. GMMs sind eng mit dem Bayes'schen Klassifizierer verwandt und gelten als ein probabilistisches Modell für multivariate Wahrscheinlichkeitsdichten.

Am konkreten Untersuchungsdatensatz nimmt der Arbeitsschritt Strukturbildung Bezug auf die bereits erkannten Gruppierungsschwierigkeiten gemäß MDS und U\*-Matrix (siehe Abb. 7). Es wird deshalb entschieden, dass für eine mehrdimensionale Betrachtung die bereits gewonnene Einzelklassifizierung im Sinne der Verteilungsuntersuchung (siehe Tab. 2) besser geeignet ist, um

auf Basis dieser ermittelten Klassen eine Gesamtbetrachtung durchzuführen.

Es ist anzumerken, dass bei einer Klassifizierung anhand von Entscheidungsgrenzen mit steigender Anzahl der Untersuchungsvariablen sehr große Klassenzahlen entstehen können. Im Hinblick auf eine sinnvolle Interpretierbarkeit von Klassenergebnissen werden drei der sechs Variablen (‚Verstädterungsgrad‘, ‚Konzentration‘, ‚Fahrzeit‘) gewählt. Die Variablen werden nicht nur aus inhaltlichen Aspekten ausgesucht, sondern auch unter dem Gesichtspunkt einer guten Objektrennung (3D-Scatter-Plot).

Mit drei Variablen wurden 22 Klassen aus den Gemeindedaten ermittelt. Das Ergebnis kommt der theoretisch zu erwartenden Klassenzahl von 24 sehr nahe. Tabelle 3 beschreibt die Klassen und Variableneigenschaften. Verwiesen sei auf den Arbeitsschritt der Wissenskonversion, der die Vergabe der Semantik und Interpretation der Klassen maßgeblich beeinflusst.

## 2.6 STRUKTURPRÜFUNG

Die Strukturprüfung ermöglicht die Validierung einer gefundenen Clusterung. Die recht ungenau beschriebene Zielsetzung der Clusteranalyse, Cluster möglichst ähnlicher Objekte zu bilden, lässt unterschiedliche Interpretationen zu. Es ist oftmals nicht zwischen richtigen oder falschen Gruppierungen zu unterscheiden, sondern vielmehr im Sinne der jeweiligen Anwendung nach brauchbaren bzw. unbrauchbaren Lösungen.

Semantik	Verstädterungsgrad			Konzentration			Erreichbarkeit			Anteil [in %]
	M	$\varnothing$	$\sigma$	M	$\varnothing$	$\sigma$	M	$\varnothing$	$\sigma$	
Peripherie, geringe Dichte	7,30	7,82	3,1	1736	1709	496	46	49	15	28,02
Peripherie, mittlere Dichte	9,70	10,1	3,8	3064	3120	411	41	44	12	19,66
Peripherie, hohe Dichte	10,97	11,2	4,2	4565	4781	730	39	43	11	4,72
Regionalzentrum, geringe Dichte	25,17	27,1	7,4	2033	1843	618	41	47	16	0,59
Regionalzentrum, mittlere Dichte	24,19	27,6	9,6	3198	3244	430	40	44	14	1,54
Regionalzentrum, hohe Dichte	25,71	28,2	8,5	4942	5229	998	36	42	13	1,49
Inter-Agglomerativ, geringe Dichte	7,60	8,2	3,1	1948	1863	457	22	22	6	12,00
Inter-Agglomerativ, mittlere Dichte	10,59	10,9	3,7	3147	3183	417	22	21	6	15,58
Inter-Agglomerativ, hohe Dichte	14,11	13,6	4,1	4665	4891	797	21	20	6	6,75
Agglomerationsseinheit, mittlere Dichte	23,47	26,3	7,9	3432	3366	403	20	19	6	1,45
Agglomerationsseinheit, hohe Dichte	27,68	30,9	10,2	5623	5880	1445	18	18	6	4,85
Agglomerationszentrum	37,57	40,0	12,6	7558	7874	2074	0	0	0	0,90
Sonderfälle: Agglomerationsseinheit	27,65	30,0	8,8	1711	1624	660	19	20	6	0,35
Sonderfälle: Agglomerationszentrum	17,48	15,9	3,6	4859	5057	846	0	0	0	0,11
Sonderfälle: Agglomerationszentrum	9,64	12,2	4,7	3439	3504	375	0	0	0	0,05
Sonderfälle: Agglomerationszentrum	22,07	22,0		3899	3899		0	0		0,01

**Tabelle 3:** Eigenschaften (M=Median, Griechisch Phi=Mittelwert, Griechisch Sigma=Standardabweichung) (Quelle: Eigene Bearbeitung).

Eine Möglichkeit der Strukturprüfung basiert auf einem Vergleich der Ergebnisdaten einer Clusterung mit einer bereits a priori bekannten Klassifizierung. Als Maßzahlen für die Qualitätsprüfung im Sinne einer Übereinstimmung zwischen einer bekannten Klassifikation und zusätzlich ermittelten Clusterung eignen sich die Sensitivität, Spezifität, Akkuratheit und die sogenannte ROC-Kurve (Receiver-Operation-Characteristic).

Eine weitere Möglichkeit der Strukturprüfung besteht darin, unabhängig von einer bereits a priori bekannten Klassifizierung die Homogenität bzw. Heterogenität innerhalb des Clusters und über die Cluster Grenzen hinweg zu bestimmen. Mit Blick auf andere multivariate Verfahren eignet sich die Diskriminanzanalyse dazu, anhand einer Clusterstruktur z.B. die Unterschiede zwischen Clustern hinsichtlich vorgegebener Variablen zu analysieren oder die Trennkraft der Variablen zu ermitteln. Die Regressionsanalyse verfolgt das Ziel,

ein funktionales Modell zwischen einer abhängigen (erklärte Variable) und einer oder mehreren unabhängigen Variablen (erklärende Variablen) zu finden.

## 2.7 OPERATIONALISIERUNG

Die Operationalisierung bildet die Grundlage für die nachfolgende Wissenskonversion und ermöglicht insbesondere die Wissensgewinnung aus bestehenden Daten unter Einbeziehung bereits entdeckter Strukturen. Es werden Zuordnungsvorschriften gesucht, die die gewonnenen Klassifikationen charakterisieren und darüber hinaus eine nachträgliche Zuordnung von nicht klassifizierten Daten realisieren.

Im Kontext des Data Mining bilden sogenannte Klassifikatoren eine Quelle zur Wissensdarstellung. Es ist zwischen symbolischen und symbolischen Klassifikatoren zu unterscheiden. Während ein symbolischer Klassifikator die Aufgabe ohne ein genaues Verständnis der Klassen erledigt (z.B. k-Nearest Neighbor, Neuronale

Netze), stellt ein symbolischer Klassifikator die Anforderung einer nahezu natürlich-sprachigen Beschreibungsform an die einzusetzenden Algorithmen. Gerade symbolische Klassifikatoren tragen dazu bei, dass der Mensch ein Verständnis für Klassen gewinnt und eine Abstraktion von Klassen möglich wird (z.B. Bayes'sche Klassifikatoren, Entscheidungsbäume/-regeln). Der Vorteil der Algorithmen CART (Classification and Regression Trees) nach Breiman (1984) und ID3 (Iterative Dichotomiser 3) nach Quinlan (1986) besteht in der Möglichkeit, die Entscheidungen für eine Klasse graphisch als Entscheidungsbaum darzustellen. Der Algorithmus sig\* (Ultsch 1991) zielt auf das Verstehen eines Clusters anhand ausgewiesener, signifikanter Regeln ab (Diagnose) und legt dabei weniger Wert auf die Hierarchie der Entscheidungen.

Bei der Bestimmung der Güte von Klassifikatoren ist nicht nur die Klassifikationsleistung im Vergleich zu gegebenen Klassifikationen zu untersuchen, sondern auch die Fähigkeit der Klassifikatoren, neue Datensätze einzuordnen. In der Regel werden drei Datenmengen gebildet: Lern- und Testdatensatz sowie ein Validierungsdatensatz.

Die Techniken aus dem Gebiet der künstlichen Intelligenz fördern die Wissensdarstellung und unterstützen bei Bedarf zusätzlich die Nutzung von Wissen in maschinellen Systemen und eignen sich z.B. für Monitoring- bzw. Diagnosesysteme. Diese müssen in der Lage sein, die von ihnen getroffenen Diagnosen nicht nur zu treffen, sondern sie auch zu begründen. Es werden Schätzkalküle verwendet, die zur Abbildung von unvollständigem, widersprüchlichem oder annäherndem Wissen verwendet werden.

Auch für das Beispiel ist eine Klassenerklärung denkbar. Es lassen sich Erkenntnisse über die von Agglomerations- und Verdichtungsprozessen betroffenen Gemeinden gewinnen und weitere Variablen zur Erklärung finden.

## 2.8 WISSENSKONVERSION

In Zusammenhang mit entwickelten Werkzeugen im Bereich des Data Mining ist festzustellen, dass diese nicht umfassend über eine Möglichkeit der Wissenskonversion verfügen (Gaul, Säuberlich 1998, S. 145). In Tabelle 4 sind Arten der Wissenskonversion und dazugehörige Methoden zusammenfassend aufgeführt.

<b>Sozialisation</b>	<b>Observieren</b> (Beobachten z.B. eines Experten), <b>Imitieren</b> (Nachahmung der Handlung z.B. eines Experten), <b>Praktizieren</b> (Überführung der theoretischen Grundlagen in praktische Erfahrung) und <b>Kommunizieren</b> (direkte verbale Vermittlung von Wissen)
<b>Externalisierung</b>	<b>Reflektieren</b> (individuelle Konzentration auf Ideen, die Arbeit und die damit verbundene Explizierung des Wissens), <b>Metapherbildung</b> (lebendige, anschauliche Versprachlichung von Zusammenhängen), <b>Analogiebildung</b> (Aufzeigen funktionaler Gemeinsamkeiten zwischen getrennten Wissensgebieten und direkter Transfer) und <b>Modellbildung</b> (komplexe Zusammenhänge problemspezifisch vereinfachen und strukturieren)
<b>Kombination</b>	<b>Sortieren</b> (Neuanordnung), <b>Hinzufügen</b> (Entstehung), <b>Vereinigen</b> (Zusammenfügen), <b>Aggregieren</b> (Ansammlung), <b>Selektieren</b> (Auswahl), <b>Kategorisieren / Klassifizieren</b> (Generierung) und <b>Rekombinieren</b> (Erzeugung aus dem Bestand)
<b>Internalisierung</b>	Prüfendes und vergleichendes Nachdenken über <b>Lesen</b> (Texten), <b>Sehen</b> (Bilder bzw. Grafiken) und <b>Hören</b> sowie vereinzelt <b>Tasten</b> und <b>Riechen</b>

Tabelle 4: Art der Wissenskonversion und ihre Methoden (Quelle: Eigene Bearbeitung unter Verwendung von Nonaka/Takeuchi (1997), Preece et al. (1994), Schreiber et al. (2000), Hyttinen (2004, S.14 ff.).)

Die Unterscheidung von Daten, Information und Wissen unterliegt einer großen Zahl von Definitionen, wobei viele zusätzlich hilfreich sind, um unterstützend den Übergang von Daten zu Wissen zu realisieren (Streich 2005, S.17): „Wissen ist die intellektuelle Vernetzung von Informationsatomen bzw. Einzeltatsachen zu komplexen Kenntnisstrukturen auf der Grundlage von Erfahrungstatbeständen und/oder Lernvorgängen von Einzelsubjekten oder Gruppen. Informationen bestehen aus sinnvoll strukturierten Daten, Daten wiederum sind die ‚atomaren‘ Bausteine für Informationen.“

Für das vorgestellte Datenbeispiel und die daraus resultierende Klassenbildung ist festzustellen, dass eine Raubeobachtung erreicht wird, die der Grundidee folgt, die heutige Situation der Verdichtung bzw. Agglomerationen abzubilden. Abbildung 8 zeigt schematisch die im Zuge der Wissenskonversion abgeleitete Klassengrundstruktur.

In Bezug auf ein gewähltes Agglomerationszentrum wird ein Verflechtungsgebiet untersucht, wobei als Verflechtungsmaß die Fahrzeit im motorisierten Individualverkehr zu den bestehenden Oberzentren verwendet wird. Die Oberzentren werden bei dieser Klassifizierung unter dem Begriff des Agglomerationszentrums beschrieben, und ein sogenannter Interaktionsraum entsteht zwischen den Gemeinden, die eine Fahrzeit von

höchstens 30 Minuten zum Agglomerationszentrum aufweisen. Außerhalb dieses sogenannten Interaktionsraumes befinden sich die Gemeinden, welche der Peripherie zugeordnet werden. Die Messgröße ‚Konzentration‘ als modifizierte Siedlungsdichte (Auslastung der Gebäude- und Freifläche durch Einwohner und Beschäftigte) dient zum Aufbau von drei Dichteklassen.

Der Grad der Verstärkung als Anteil der Siedlungs- und Verkehrsfläche an der Gemeindefläche insgesamt ermöglicht die Identifizierung von besonders verstärkten Gemeinden. In diesem Ansatz werden die

Gemeinden, welche innerhalb des Interaktionsraums liegen, als Agglomerationseinheiten bezeichnet. Diese tragen bei ähnlichen Verstärkungseigenschaften der Nachbargemeinden zum Wachstum der Agglomeration insgesamt bei. Gemeinden, die in der Peripherie liegen und einen besonders großen Verstärkungsgrad aufweisen, werden als Regionalzentrum bezeichnet. Eine Gemeinde gilt als inter-agglomerativ, wenn diese sich im Interaktionsraum befindet, jedoch nicht über einen Verstärkungsgrad von mehr als 20 Prozent verfügt. Die einzelnen Agglomerationszentren werden untereinander mit ihrem dazugehörigen Interaktionsraum vergleichbar. Hinsichtlich der Einwohner und Arbeitsplätze je km<sup>2</sup> Gebäude- und Freifläche und dem Verstärkungsgrad wurden verschiedene Verflechtungsmuster erfasst. Die Regionen Rhein-Main, Rhein-Ruhr sowie die Region Stuttgart repräsentieren hochverdichtete und deutlich verstärkte Agglomerationen. Es ist zu vermuten, dass in Zukunft bei weiter fortschreitendem regionalem Wachstum von Bevölkerung und Beschäftigten eine noch stärkere Agglomeration entstehen wird und damit die Übergänge zwischen Verflechtungsgebieten noch schwieriger abgrenzbar sein werden. In Bezug auf einige ostdeutsche Agglomerationszentren (z.B. Chemnitz, Neubrandenburg oder Schwerin) lassen sich deutlich geringere Dichte- und Verstärkungseigenschaften bei den Gemeinden im Verflechtungsgebiet erkennen. Hier kann nicht von einer homogen dichten städtischen Agglomeration gesprochen werden, sondern eher von einem Kerngebiet, welches

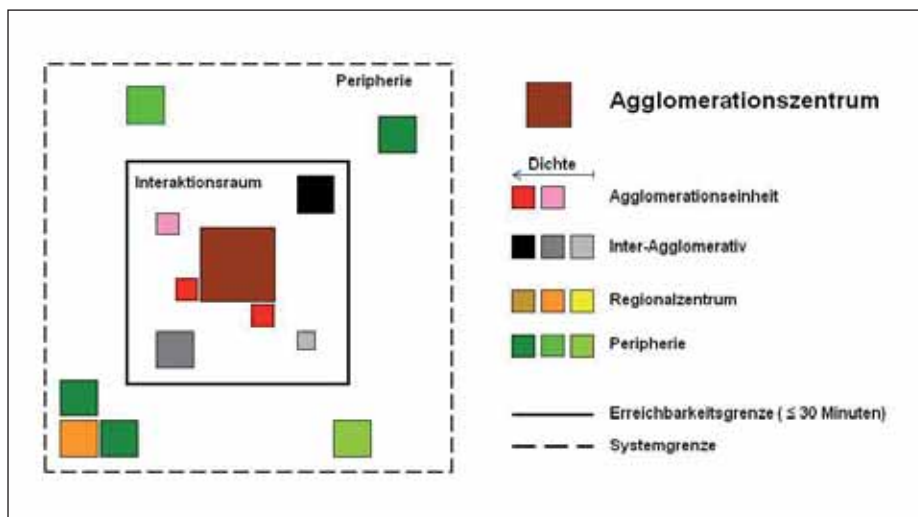


Abbildung 8: Klassengrundstruktur (Quelle: Behnisch, 2009).

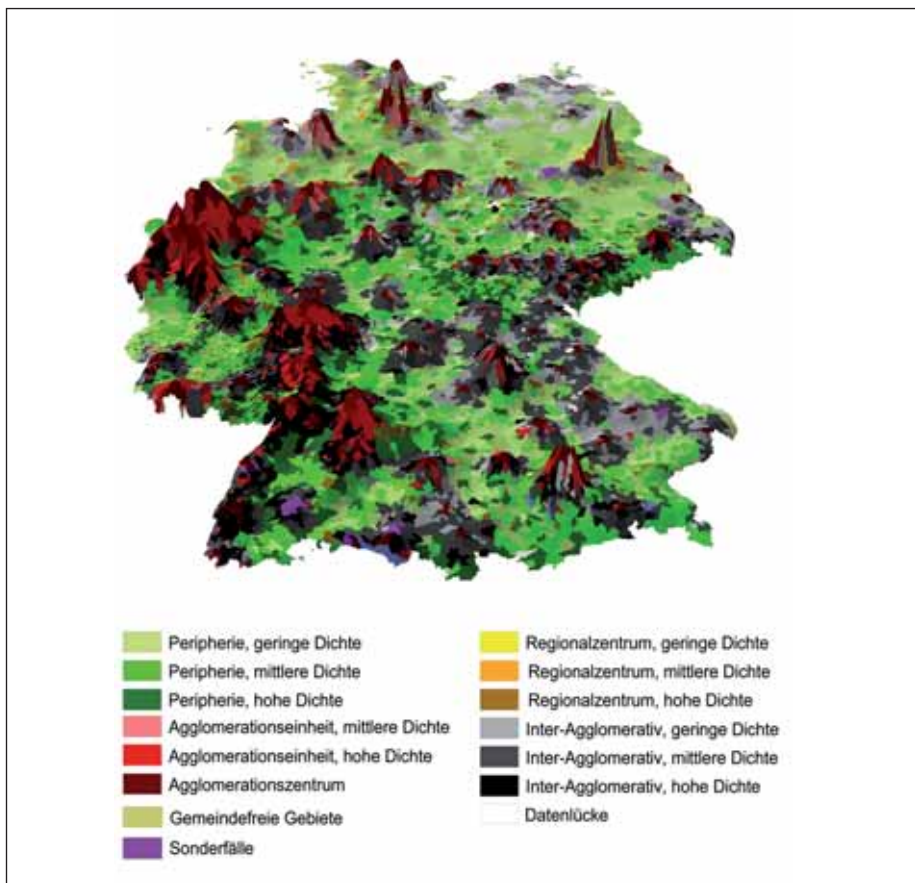


Abbildung 9: Klassen und Gebäudedichte (Quelle: Behnisch, 2009).



Abbildung 10: Klassen und Gebäudedichte (Quelle: Behnisch, 2009).

als Oberzentrum auf das Umland einen Einfluss ausübt.

Gerade in den Gebieten mit einer geringeren Verdichtung in den Verflechtungsgebieten fallen entweder Gemeinden mit hoher Dichte oder sogar Gemeinden mit hoher Dichte und großer Verstädterung in einem gewissen Erreichbarkeitsabstand zum Agglomerationszentrum auf. Exemplarisch herausgegriffen sei das Agglomerationszentrum Kempten und die speziellen Gemeinden

mit einer hohen Dichte: Marktobersdorf, Pfronten und Immenstadt im Allgäu.

Innerhalb der Peripherie sind zwischen den dazugehörigen Gemeinden ebenfalls Besonderheiten im Hinblick auf die Dichte und den Verstädterungsgrad messbar. Die als Regionalzentren definierten Gemeinden sind durch einen höheren Verstädterungsgrad gekennzeichnet. Es ist zu erwarten, dass diese Gemeinden weitere besondere Eigenschaften haben und auch

einen stärkeren Einfluss auf die Nachbargemeinden ausüben. Zu bemerken ist, dass es sich vielfach um Gemeinden mit Stadtrecht handelt. Darüber hinaus werden Gemeinden in der Peripherie mit einer hohen Dichte identifiziert, die möglicherweise weitere regionalspezifische Besonderheiten aufweisen.

Abbildung 9 vermittelt einen Eindruck von einer eindimensionalen Informationsüberlagerung. Dargestellt werden die durch Klassenbildung erzeugten Verstädterungseigenschaften und die Gebäudedichte, wobei große Höhenwerte Ausdruck einer größeren Gebäudedichte mit Bezug auf die Katasterfläche der Gemeinde repräsentieren. In Behnisch (2009) wird ein Schätzansatz vorgestellt, der darauf hindeutet, dass 20 Prozent der Gemeinden bereits 80 Prozent des Gebäudebestandes enthalten.

In Abbildung 10 sind Regionalstrukturen in einem größeren Detailausschnitt ersichtlich. Es handelt sich um die Region Stuttgart und den Großraum Berlin.

In der Region Stuttgart kann von flächenhaften Verstädterungsausprägungen gesprochen werden. Im Interaktionsraum von Stuttgart werden hoch verdichtete und auch hoch verstädterte Gemeinden erfasst. Die ermittelte Erreichbarkeitsgrenze von maximal 30 Minuten Fahrzeit zum nächsten Agglomerationszentrum führt zur Identifizierung der Interaktionsräume. In der Region Stuttgart und den benachbarten Agglomerationszentren Pforzheim, Karlsruhe und Heilbronn sowie Tübingen grenzen diese aneinander und überlagern sich in Teilbereichen. Im Großraum Berlin wird in direkter Nachbarschaft zum Agglomerationszentrum eine größere Anzahl Gemeinden als Regionalzentrum erkannt. Diese zeigen oftmals eine geringe bis mittlere Dichte. Es handelt sich um ein frühes Suburbanisierungsstadium. Das sich in der Industrialisierung heraus gebildete Siedlungssystem mit einem Ring größerer Mittelstädte um das Zentrum ist weitgehend noch erhalten.

#### 4. FAZIT UND AUSBLICK

Der Untersuchungsansatz (Urban Data Mining) dokumentiert die Auseinandersetzung mit methodischen Vorgehensweisen des Data Mining und der Knowledge Discovery. Auf der Grundlage dieses explizit für den urbanen Kontext ausgewählten Methodenrepertoires kann in Zukunft ein umfassenderes Regelwerk

aufgebaut werden, welches sich zur dauerhaften Bewertung von räumlichen Strukturen und Entwicklungen eignet.

Im methodischen Sinne hat die Untersuchung der sechs raumstrukturellen Variablen eine problemorientierte Sicht auf die Notwendigkeit einer Verteilungsuntersuchung geliefert. Diese unterstützt den Aufbau von Entscheidungsgrenzen und fördert die Identifizierung von Schwellwerten im urbanen Forschungsfeld. Die Verortung von einzelnen Messgrößen beruht nicht mehr nur auf subjektiv festgelegten Klassengrenzen oder reinem Expertenwissen. Vielmehr bilden Gauß-Mixtur-Modelle eine Charakteristik in den Daten ab und liefern eine Entscheidungsgrundlage.

Mit Hilfe einer Emergenten SOM (Ultsch 2005) wurde die Wichtigkeit einer Strukturerkennung betont. Es zeigte sich, dass keine deutliche Struktur in den Daten der sechs Messgrößen vorhanden ist. Eine unkontrollierte Anwendung eines Clusteralgorithmus hätte somit nicht zu einer klar unterscheidbaren Gruppenstruktur geführt. Mit Blick auf die Eindeutigkeit einer Gruppenstruktur wurde eine mehrdimensionale Gemeindebetrachtung schließlich auf Basis der zuvor bereits gefundenen Entscheidungsgrenzen umgesetzt.

Raumstrukturell vermittelt das Klassifizierungsergebnis einen Eindruck der Verstädterung, der Erreichbarkeit der nächstgelegenen Oberzentren und der Konzentration von Einwohnern und Beschäftigten. Identifiziert werden polyzentrische Regionalstrukturen im Verflechtungsbereich von Oberzentren und monozentrisch geprägte geringer besiedelte Regionalstrukturen.

In der Zukunft ist die Einbeziehung von grenznahen Gemeinden im benachbarten Ausland zu befürworten, um gegenseitige Grenzbeziehungen bzw. regional spezifische Einflüsse und Wirkungsbeziehungen aufzudecken. Eine zeitliche und räumliche Vergleichbarkeit von Gebäudenutzungsdaten ist aufgrund der ALKIS-Einführung (Bill et al. 2002) ebenso relevant und für die Klassenerklärung (im Sinne des Arbeitsschrittes der Operationalisierung) besonders wertvoll, um weitere inhaltliche Schwerpunkte setzen zu können und neue Regeln für eine gefundene raumstrukturelle Klasse zu generieren.

Viele Ansätze der Stadtklassifikation repräsentieren eher unifaktorielle oder deskriptive Arbeitsweisen (Behnisch 2009). Im Data Mining wurden jedoch effiziente Algorithmen entwickelt, die die Analyse vollständiger

Populationen ermöglichen. Im engeren Sinne kann also zukünftig der nichttriviale Prozess der Identifizierung gültiger, neuartiger, potenziell nützlicher und verständlicher Muster in (großen) kontextbezogenen statistischen Datenbeständen sichergestellt werden. In diesem Beitrag wurden dazu exemplarisch einige Techniken anhand von sechs Variablen vorgestellt. Die Durchdringung von weitaus umfangreicheren Variablenmengen fördert die Auseinandersetzung bzw. den Vergleich mit bestehenden Raumtypen und Raumeigenschaften.

Wie zu Beginn bereits erwähnt, werden aber auch Urbane Raumbewachungssysteme (Urban Monitoring Systems) langfristig unabdingbar, da diese in der Lage sind, automatisiert Entwicklungen in einem bestimmten Gebiet über die Zeit zu beobachten bzw. nach zu verfolgen. Der Benutzer dieser Systeme ist also gewarnt (oder sensibilisiert), wenn eine Entwicklung nicht wie gewünscht verläuft oder sich eine negative Veränderung einzustellen droht. Allerdings bedarf es zuvor erst noch einer wesentlich genaueren Kenntnis der Definition und Aussagekraft von Indikatoren und Ihrer Zusammenhänge, so dass Verfahren des Data Mining und der Knowledge Discovery wichtige Beiträge leisten können. ◀

**Literatur**

Arlt, G.; Gössel, J.; Heber, B.; Hennersdorf, J.; Lehmann, I.; Thinh, N. X. (2001): Auswirkungen städtischer Nutzungsstrukturen auf Bodenversiegelung und Bodenpreis. In: IÖR Schriften (IÖR, Leibniz-Institut für ökologische Raumentwicklung e. V., Dresden) Nr. 34.

Baccini, P.; Kytzia S.; Oswald, F. (2002): Restructuring Urban Systems. In: Moavenzadeh, F.; Hanaki, K.; Baccini, P. (Hrsg.): Future Cities: Dynamics and Sustainability. Springer, S. 17-43.

Behnisch, M. (2009): Urban Data Mining. Universitätsverlag Karlsruhe.

Behrens, K.; Marhenke, W. (1997): Die Abgrenzung von Stadtregionen und Verflechtungsgebieten in der Bundesrepublik Deutschland. In: Statistisches Landesamt Baden-Württemberg (Hrsg.): Jahrbuch für Statistik und Landeskunde. Stuttgart, S. 165-186.

Bill, R.; Seuß, R.; Schilcher, M., Hrsg. (2002): Kommunale Geo-Informationssysteme: Basis-

wissen, Praxisberichte und Trends. Wichmann.

Bock, H. H. (1974): Automatische Klassifikation. Vandenhoeck & Ruprecht, S. 44 ff. und S. 77.

Boustedt, O. (1953): Die Stadtregion. Ein Beitrag zur Abgrenzung städtischer Agglomerationen. In: Allgemeines statistisches Archiv, Nr. 37, S.13-26.

Boustedt, O. (1975 a): Grundriß der empirischen Raumforschung. Teil 3: Siedlungsstrukturen. Hermann Schroedel Verlag, Hannover.

Boustedt, O. (1975 b): Grundriß der empirischen Raumforschung. Teil 4: Regionalstatistik. Hermann Schroedel Verlag, Hannover.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984): Classification and Regression Trees. Wadsworth, California.

Deichsel, G.; Trampisch, H. J. (1985): Clusteranalyse und Diskriminanzanalyse. Gustav Fischer, S. VII.

Fischer, M. M. (1982): Eine Methodologie der Regionaltaxonomie: Probleme und Verfahren der Klassifikation und Regionalisierung in der Geographie und Regionalforschung. In: Bahrenberg, G.; Barth, H.-K.; Leuze, E.; Taubmann, W.(Hrsg.): Bremer Beiträge zur Geographie und Raumplanung. Presse- und Informationsdienst, Universität Bremen, Heft 3.

Frühwald, W. (1997): Neue Perspektiven in der Wissenschaft. In: Zeit Punkte, Nr. 6, Steiner, S. 62-64.

Gaul, W. G.; Säuberlich, F. (1998): Classification and Positioning of Data Mining Tools. In: Gaul, W.; Locarek-Junge, H. (eds.): Classification in the Information Age. Proc. 22th Annual Conference of the GfKI, University of Dresden, March 4-6, 1998. Springer-Verlag, S. 145 ff.

Hand, D. J.; Mannila, H.; Smyth, P. (2001): Principles of Data Mining, Bradford, S. 281.

- Hartung, J. (2005): Statistik – Lehr- und Handbuch der angewandten Statistik. Oldenbourg, S. 833.
- Häussermann, H.; Siebel, W. (1999): Neue Urbanität. In: Raum Journal I. Schrift der Kulturregion Stuttgart e.V. (Hrsg.), Stuttgart, S. 19-21.
- Hytinen, L. (2004): Knowledge conversions in knowledge work – a descriptive case study. Licentiate Thesis. Espoo: Helsinki University of Technology.
- INSPIRE (2010): Infrastructure for Spatial Information in Europe - History. (online: <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/4>, Zugriff 03/2010).
- Jaeger, J.; Scheringer, M. (1998): Transdisziplinarität: Problemorientierung ohne Methodenzwang. In: GAIA, Vol. 7, Nr. 1, 1998, oekom, S. 10-25.
- Keim, D. A. (2002): Datenvisualisierung und Data Mining. In: Datenbank-Spektrum – Zeitschrift für Datenbanktechnologie. 2. Jahrgang, Heft 2, dpunkt, S. 31.
- Kohonen, T. (1982): Self-Organized Formation of Topologically Correct Feature Maps. In: Biological Cybernetics. Vol. 43, Springer, pp. 59-69.
- Lichtensteiger, T. (2006): Bauwerke als Ressourcennutzer und Ressourcenspender in der langfristigen Entwicklung urbaner Systeme. vdf, S. 1.
- Miller, H. J.; Han, J. (2009): Geographic Data Mining and Knowledge Discovery. Chapman & Hall, S. 21.
- Nonaka, I.; Takeuchi, H. (1997): Die Organisation des Wissens. Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen. Campus Verlag.
- Open Geospatial Consortium (2010): OGC History. (online: <http://www.opengeospatial.org/ogc/history>, Zugriff 03/2010).
- Preece, J.; Rogers, Y.; Sharp, H.; Benyon, D.; Holland, S.; Carey, T. (1994): Human-Computer Interaction. Addison Wesley.
- Prigge, W. (1999): Wie urban ist das digitale Zeitalter?, In: Raum Journal I, Schrift der Kulturregion Stuttgart e.V. (Hrsg.), S. 54.
- Quinlan, J. R. (1986): Introduction of Decision Trees. In: Machine Learning, Vol. 1, Springer, S. 81-106.
- Schreiber, G.; Akkermans, H.; Anjewierden, A.; De Hoog, R.; Shadbolt, N.; Van de Velde, W.; Wielinga, B. (2000): Knowledge Engineering and Management – The Common KADS Methodology. MIT Press.
- Siedentop, St.; Kausch, St.; Einig, K.; Gössel, J. (2003): Siedlungsstrukturelle Veränderungen im Umland der Agglomerationsräume. Bundesinstitut für Bau-, Stadt- und Raumforschung (Hrsg.), Heft 114, S. 190.
- Sieverts, T. (1997): Zwischenstadt zwischen Ort und Welt, Raum und Zeit, Stadt und Land. Birkhäuser, S. 35.
- Streich, B. (2005): Stadtplanung in der Wissensgesellschaft: ein Handbuch. VS Verlag für Sozialwissenschaften, Wiesbaden.
- Ultsch, A. (1991) Konnektionistische Modelle und ihre Integration mit wissensbasierten Systemen. Habilitationsschrift, Fakultät für Informatik, Universität Dortmund.
- Ultsch, A. (1999): Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. In: Kaski, S.; Oja, E. (eds.): Kohonen Maps. Elsevier, pp. 33-46.
- Ultsch, A. (2003): Pareto Density Estimation: A Density Estimation for Knowledge Discovery. In: Baier, D.; Wernecke, K. D. (Eds): Innovations in Classification, Data Science, and Information Systems. Proceedings 27th Annual Conference of the German Classification Society (GfKl), Springer, pp. 91-100.
- Ultsch, A. (2005): Clustering with SOMU\*C. In: Proceedings of the 5th Workshop on Self-Organizing Maps (WSOM 2005), Paris, September 5-8, pp. 75-82.
- Ultsch, A. (2010): Knowledge Discovery. Vorlesungsunterlagen, Fachbereich Mathematik und Informatik, Philipps-Universität Marburg.
- Vogel, F. (1975): Probleme und Verfahren der numerischen Klassifikation. Vandenhoeck & Ruprecht, S. 52 ff.

## IMPRESSUM // PUBLICATION INFORMATION:

GIS.SCIENCE – Die Zeitschrift für Geoinformatik ISSN 1430-3663 // Herausgeber: Dirk Schmidbauer // Redaktion: Alexandra Wojtanowska M.A., [alexandra.wojtanowska@abcverlag.de](mailto:alexandra.wojtanowska@abcverlag.de), Tel. +49 (0)221-96 08 536 // Hauptschriftleiter: Prof. Dr.-Ing. Ralf Bill, [ralf.bill@uni-rostock.de](mailto:ralf.bill@uni-rostock.de), Tel. +49(0)0381498-3200 // Editorial-Board: Prof. Dr. Lars Bernard, TU Dresden; Dr. Andreas Donaubauer, ETH Zürich; Prof. Dr. Max Egenhofer, University of Maine Orono; Prof. Dr. Manfred Ehlers, Universität Osnabrück; Prof. Dr. Klaus Greve, Universität Bonn; Dr. Stefan Lang, Universität Salzburg, Prof. Dr. Stephan Nebiker, Fachhochschule Nordwestschweiz, Prof. Dr. Matthäus Schilcher, TU München, Prof. Dr. Josef Strobl, Universität Salzburg // Konzeption und Layout: Dipl. Des. Birgit Speiser // Leserservice: Ingrid Gimbel, [ingrid.gimbel@abcverlag.de](mailto:ingrid.gimbel@abcverlag.de), +49(0)06221/75704-100 // GIS.SCIENCE erscheint im abcverlag GmbH, Waldhofer Str. 19, 69123 Heidelberg, Tel. +49(0)6221/75704-100, Fax +49(0)6221/75704-109, [info@abcverlag.de](mailto:info@abcverlag.de) // Geschäftsführung: Dirk Schmidbauer, HRB 337388, Ust-ID: DE 227 235 728 // Druck: abcdruck, Heidelberg // Erscheinungsweise: 12 x jährlich, davon 4 Ausgaben GIS.SCIENCE plus 2 Sonderthemenhefte GIS.TRENDS+MARKETS // Jahresabonnement (12 Hefte): Inland 157,25 EUR inkl. Versandkosten, Ausland 166 EUR inkl. Versandkosten, Studenten/Auszubildende 89,- EUR inkl. Versandkosten, Mitglieder des Deutschen Dachverbandes für Geoinformation e.V. (DDGI) erhalten das Abo im Rahmen ihrer Mitgliedschaft // Bezugszeitraum: Das Abonnement läuft zunächst für 12 Monate. Zum Ablauf des ersten Bezugsjahres kann das Abonnement zum Ende des Kalenderjahres mit einer Kündigungsfrist von 3 Monaten gekündigt werden. Bei Nichterscheinen aus technischen Gründen oder höherer Gewalt entsteht kein Anspruch auf Ersatz. // Alle in GIS.BUSINESS und GIS.SCIENCE und GIS.TRENDS+MARKETS erscheinenden Beiträge, Abbildungen und Fotos sind urheberrechtlich geschützt. Reproduktion, gleich welcher Art, kann nur nach schriftlicher Genehmigung des Verlags erfolgen. © 2010 abcverlag GmbH, Heidelberg

