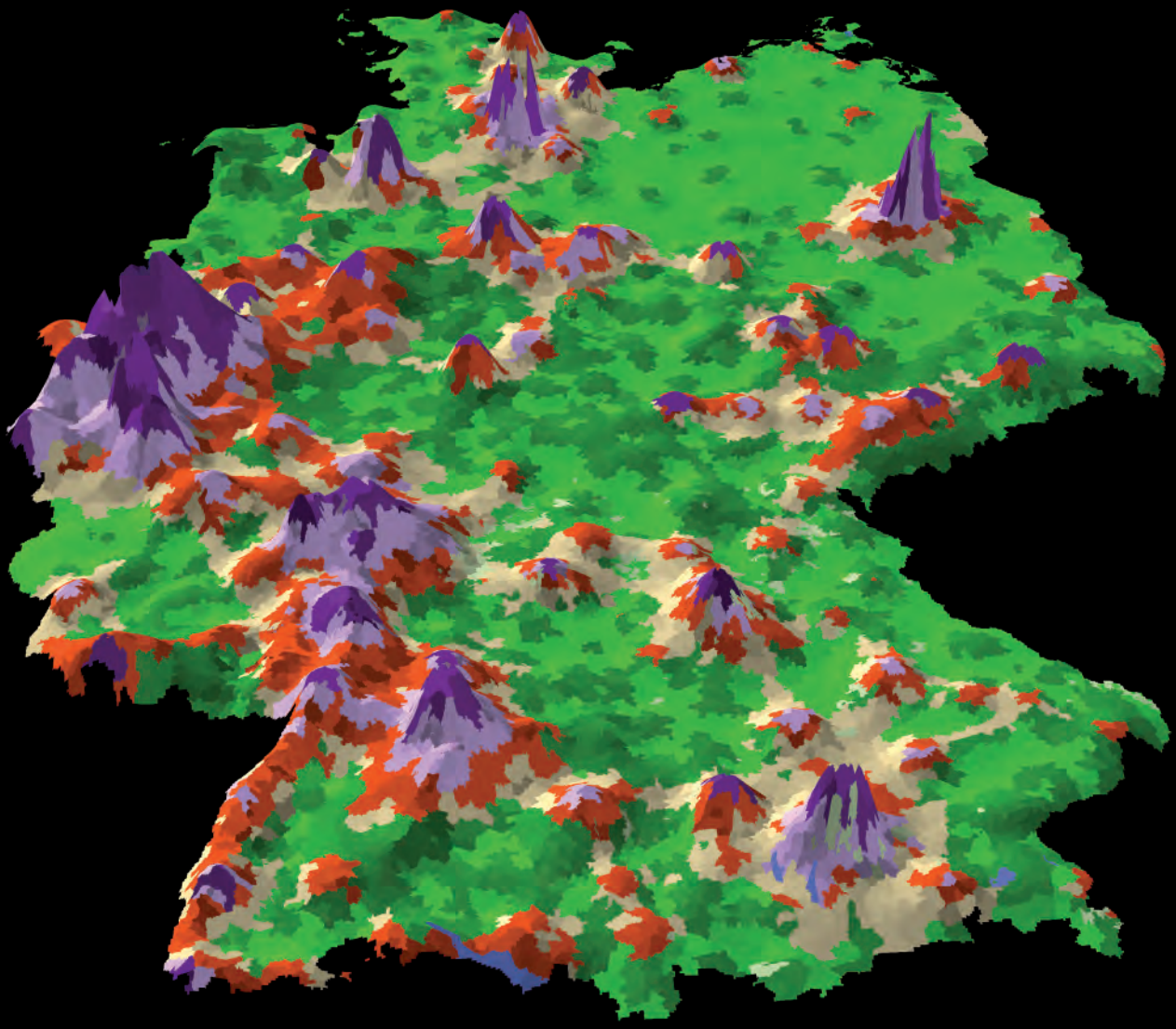
URBAN DATA MINING



Martin Behnisch

universitätsverlag karlsruhe

Martin Behnisch

URBAN DATA MINING

Operationalisierung der Strukturerkennung und Strukturbildung
von Ähnlichkeitsmustern über die gebaute Umwelt

Titelbild:
Dreidimensionale Visualisierung der Gebäudedichte in Deutschland, überlagert mit den bestehenden Raumstrukturtypen des Bundesamtes für Bauwesen und Raumordnung.

URBAN DATA MINING

Operationalisierung der Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt

von
Martin Behnisch

universitätsverlag karlsruhe

Dissertation, Universität Karlsruhe (TH)
Fakultät für Architektur, 2007

Impressum

Universitätsverlag Karlsruhe
c/o Universitätsbibliothek
Straße am Forum 2
D-76131 Karlsruhe
www.uvka.de

Universitätsverlag Karlsruhe 2008
Print on Demand

ISBN: 978-3-86644-249-8

Urban Data Mining

Operationalisierung der Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt.

Zur Erlangung des akademischen Grades
eines
Doktor-Ingenieurs

der Fakultät für Architektur
der Universität Karlsruhe (TH)

eingereichte Dissertation

von
Dipl. Ing. Martin Behnisch
aus Aachen, 2007

Tag der mündlichen Prüfung: 13.11.2007

Referent: Professor Dr. Niklaus Kohler

Korreferenten: Professor Dr. Alfred Ultsch
Professor Markus Neppl
PD Dr. Nguyen Xuan Thinh

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Stipendiat der Landesgraduiertenförderung (LGF) Baden-Württemberg am Institut für industrielle Bauproduktion (ifib) der Universität Karlsruhe (TH).

Dem Leiter des Instituts, Herrn Prof. Dr. ès. sc. techn. **Niklaus Kohler** möchte ich an dieser Stelle meinen innigen Dank für seine großzügige Unterstützung, sein konsequentes motivierendes Vertrauen und die Übernahme des Hauptreferats aussprechen. Mein aufrichtiger Dank gilt ebenso Herrn Prof. Dipl.-Ing. **Markus Neppl**, dem Leiter des Lehrstuhls für Stadtquartiersplanung und Entwerfen am Institut für Städtebau und Entwerfen der Universität Karlsruhe (TH). Ich bedanke mich für seine konstruktiven Anmerkungen, seine hilfreichen Anregungen und die Übernahme des Korreferats.

Bei Herrn Professor Dr. **Alfred Ultsch**, dem Leiter der Arbeitsgruppe Datenbionik des Fachbereichs Mathematik und Informatik an der Phillipps-Universität Marburg, bedanke ich mich für die Möglichkeit, dass ich als ausgebildeter Architekt die Verfahren und Arbeitsweisen des Data Mining erlernen konnte. Ergänzt wurde diese große Unterstützung durch die überaus ungewöhnliche Hilfsbereitschaft von Herrn Dr.-Ing. **Steffen Bocklisch,** dem Lehrstuhlinhaber der Professur für Systemtheorie der Fakultät für Elektrotechnik an der Technischen Universität Chemnitz. Ihm danke ich für die Diskussionen am Wochenende in Chemnitz über das Prinzip der Klassifikatoren und des Konzepts der Fuzzy-Pattern-Klassifikation.

Bei Herrn Dr. rer. nat. habil. **Nguyen Xuan Thinh**, dem Wissenschaftlichen Mitarbeiter des Leibniz-Instituts für ökologische Raumentwicklung e.V. (IÖR), bedanke ich mich ganz herzlich für die Bereitstellung experimenteller Daten zur räumlichen Analyse von Flächennutzungsmustern und seine Offenheit, die langjährige Forschungserfahrung weiterzugeben. Gedankt sei auch seinem Kollegen Dr.-Ing. **Stefan Siedentop** für die Möglichkeit, jederzeit fachliche Fragen klären zu können.

Der Aufbau einer statistischen Datenbasis mit Raumbezug wäre ohne die Mitarbeiter des **Bundesamtes für Bauwesen und Raumordnung (BBR)** nicht möglich gewesen. Gedankt sei ebenso den Mitarbeitern des **Statistischen Bundesamtes** und der **Statistischen Landesämter** sowie der **amtlichen Vermessung** in Deutschland, die umfangreiches Datenmaterial bereitgestellt haben.

Ferner bedanke ich mich bei den Herren **Claudio Ferrara** und **Patrick Erik Bradley** für die motivierenden Worte, den wissenschaftlichen Arbeitsweg zu suchen sowie bei meinen lieben Kollegen am ifib, die gerade durch kollegialen Umgang zum Gelingen dieser Arbeit beitrugen.

Schließlich bedanke ich mich für die Unterstützung des **Landes Baden-Württemberg**, ohne dessen Finanzierung diese Arbeit nicht möglich gewesen wäre. Der **Ruth und Erich Rossmann-Stiftung** danke ich für die Finanzierung eines Auslandsaufenthaltes an der Universität von Tianjin in China.

Der größte Dank gilt meinen **Eltern**, die mich zu jeder Zeit unterstützt und motiviert haben. Ihnen möchte ich diese Arbeit widmen.

Karlsruhe, im April 2007 Martin Behnisch

Autorenreferat

BEHNISCH, Martin: Urban Data Mining – Operationalisierung der Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt. Universität Karlsruhe, Fakultät Architektur, Dissertation, 2007.

Hauptteil: 303 Seiten, 578 Literaturquellen, 100 Abbildungen, 64 Tabellen
Nebenteil A – Räumliche Dateninspektion und Ergebnisvisualisierung: 98 Karten, 100 Seiten
Nebenteil B – Theoretische Ergänzungen - Berechnungen: 22 Anlagen, 115 Seiten

Nebenteil A und Nebenteil B sind auf einer CD-ROM in der Buchausgabe enthalten.

Zusammenfassung

Durch den schnellen Fortschritt in der Informationstechnologie und das immer schnellere 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 zu entdecken. Der empirische Teil dieser Arbeit bezieht sich auf die 12504 Gemeinden in Deutschland im Jahre 2004. Vor dem Hintergrund räumlicher Untersuchungsaufgaben wird die Emergente SOM zur Strukturerkennung eingesetzt. Die Klassifikation ist ein wichtiges Instrument, das zu einer Entwicklung von Maßstäben und Bewertungsskalen der städtischen Phänomene beiträgt. Die Konstruktion von Klassifikatoren findet Zuordnungsvorschriften, welche die gewonnenen Klassifikationen nachvollziehen und bislang nicht klassifizierte Daten richtig einordnen. Die Daten beziehen sich auf die Verstädterung und Zentralität von Gemeinden und auf zeitliche Entwicklungen. Die Daten zur Beschäftigung ermöglichen eine Beschreibung der Diversität von Gemeinden. Darüber hinaus wurden methodische Möglichkeiten gesucht, um räumliche Informationen zwischen verschiedenen Objekten zu übertragen. Die gemeindescharfe Schätzung des deutschen Gebäudebestandes wird als besondere Herausforderung angesehen, da bisher keine amtlichen Statistiken in Deutschland vorliegen, die den Gesamtbestand auf Gemeindeebene ausweisen. Zum Aufbau einer digitalen Datenbasis ist der Einsatz von GIS unverzichtbar. Da die gebaute Umwelt von hoher Komplexität und dynamischer Entwicklung geprägt ist, werden interdisziplinäre Forschungsansätze verfolgt.

Abstract

The term of Urban Data Mining is defined to describe a methodological approach that discovers logical or mathematical and partly complex descriptions of urban patterns and regularities inside the data. The concept of data ining in connection with knowledge discovery techniques plays an important role for the empirical examination of high dimensional data in the field of urban research. The procedures on the basis of knowledge discovery systems are currently not exactly scrutinised for a meaningful integration into the regional and urban planning and development process. Emergent SOM and classification are used to analyse 12504 communes in Germany. The data deals mainly with the question of centrality and substance of cities and also with the development of urban settlements. Especially data about employment and the building stock support the description of communes according to the diversity of urban functions. A method is integrated that allows an estimation of the German building stock. Moreover the construction of classifiers enables the allocation of unclassified objects to an already existing structure of clusters. In the future it might be possible to establish an instrument that defines objective criteria for the benchmark process about urban phenomena. The use of GIS supplements the process of knowledge conversion and communication. The approach is based on interdisciplinary research because of its complex and dynamic behaviour.

Es ist aber bisweilen schwer zu beurteilen, für welche von zwei Möglichkeiten man sich entscheiden oder welches von zwei Übeln man über sich ergehen lassen soll; und oft ist es noch schwerer, bei dem gefassten Entschluss zu bleiben.

Aristoteles, Nicomachische Ethik, 3,1, 4. Jahrhundert v. Chr.

Es zeichnet einen gebildeten Geist aus, sich mit jenem Grad an Genauigkeit zufrieden zu geben, den die Natur der Dinge zulässt, und nicht dort Exaktheit zu suchen, wo nur Annäherung möglich ist.

Aristoteles, Nikomachische Ethik,

Präzision ist nicht Wahrheit.

Henri Matisse, aus Demant, Bernd [1993, S.1]

"… one of the most important facts of human thinking is the ability to summarize information into labels of fuzzy-sets which bear an approximate relation to the primary data …"

Lofti Zadeh, 1973

"It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it."

Tukey [1977, S. V]

"Ohne eine sachgerechte, kritische Anwendung der zum Teil recht komplizierten mathematischen und statistischen Verfahren sind hinreichend abgesicherte Forschungsergebnisse im Rahmen der 'Quantitativen Geographie' nur noch selten zu erreichen. Die Suche nach neuen mathematischen Verfahren und Modellen wird daher weiterhin ein Kernpunkt der ‚Quantitativen Geographie' bleiben."

E. Giese [1980, S. 256]

"Never concerned that the answer may prove disappointing, with pleasure and confidence we turn over each new stone to find unimagined strangeness leading on to more wonderful questions and mysteries – certainly a grand adventure."

Richard P. Feynman [1988, S.243]

„Uncertainties in data lead to uncertainties in the result of analysis."

P. LONGLEY [2001, S.137]

„Zweiundvierzig!" kreischte Luunquuoal los. „Ist das alles, nach siebeneinhalb Millionen Jahren Denkarbeit?"

D. Adams, „Per Anhalter durch die Galaxis" [1985]

Inhalt

Abbildungsverzeichnis[1]

[1] Sollten in dieser Arbeit Abbildungen nicht durch eine Quellenangabe besonders gekennzeichnet sein, so wurden diese Abbildungen vom Verfasser der Arbeit eigenständig angefertigt.

Tabellenverzeichnis[2]

[2] Sollten in dieser Arbeit Tabellen nicht durch eine Quellenangabe zusätzlich gekennzeichnet sein, so wurden diese Tabellen vom Verfasser der Arbeit eigenständig angefertigt.

Abkürzungsverzeichnis

A-BL	Alte Bundesländer
ADV	Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland
AIBAU	Aachener Institut für Bauschadensforschung und angewandte Bauphysik
ALB	Automatisiertes Liegenschaftsbuch
ALK	Automatisierte Liegenschaftskarte
ALKIS®	Amtliches Liegenschaftskataster-Informationssystem
BA	Bundesagentur für Arbeit
BBR	Bundesamt für Bauwesen und Raumordnung
BGBl.	Bundesgesetzblatt
BKG	Bundesamt für Kartographie und Geodäsie
BL	Bundesländer
BRD	Bundesrepublik Deutschland
CA	Clusteranalyse
CART	Classification of Regression Trees
CLC	CORINE Land Cover
COLIDO	Computergestützte Liegenschaftsdokumentation
DA	Diskriminanzanalyse
DEÜV	Datenerfassungs- und -übermittlungsverordnung
DFG	Deutsche Forschungsgemeinschaft
DFD	Deutsches Fernerkundungsdatenzentrum
DGF	Digitale Flurkarte
DGM	Digitales Gelände-Modell
DLR	Deutsches Zentrum für Luft- und Raumfahrt
EAD	Einwohner- und Arbeitsplatzdichte
ESOM	Emergente Selbstorganisierende Merkmalskarten
EU	Europäische Union
FA	Faktorenanalyse
FEtN	Flächenerhebung nach tatsächlicher Nutzung
FPK	Fuzzy-Pattern-Klassifikation
GBWZ	Gebäude- und Wohnungszählung
GI-Sytem	Geographisches Informationssystem
GMM	Gaußsche-Mixtur-Modelle
GVHK	Gemeinschaft zur Verbreitung der Hauskoordinaten
HABITAT II	Konferenz der vereinten Nationen über menschliche Siedlungen
IBIS	Isoplan-Bau-Informations-System
ICA	Independent Component Analysis
IFIB	Institut für industrielle Bauproduktion
IÖR	Leibniz-Institut für ökologische Raumentwicklung e.V.
KDD	Knowledge Discovery in Databases
KNN	Künstliche Neuronale Netze
k-NN	k-Nächster Nachbar (k-nearest neighbour)
LK	Landkreis
LVA	Landesvermessungsamt
MDS	Multidimensionale Skalierung
MG	Megagramm
MZ	Mikrozensus
N-BL	Neue Bundesländer
OSKA	Objektschlüsselkataloge
OBAK	Objektabbildungskataloge
PDE	Pareto Density Estimation
QQ-Plot	Quantil/Quantil-Plot
RA	Regressionsanalyse
ROG	Raumordnungsgesetz
SK	Stadtkreis
SOM	Self-Organizing Maps
StaBu	Statistisches Bundesamt
StaLaBW	Statistisches Landesamt Baden-Württemberg
WSVO	Wärmeschutzverordnung

1. Einführung in die Problemstellung

1.1. Ausgangslage

Die gebaute Umwelt, bestehend aus Gebäuden und Infrastruktur, bildet als Kulturraum das größte physische, wirtschaftliche und soziokulturelle Kapital einer Gesellschaft. Im Hinblick auf eine nachhaltige Entwicklung kommt dem historisch gewachsenen deutschen bzw. europäischen Gebäudebestand und den dadurch geprägten Stadtstrukturen eine große Bedeutung zu, da diese in ihrer Qualität und historischen Dichte nicht reproduzierbar sind. Der natürlichen Umwelt als ursprünglichem ökologischem System der Natur bleibt in Europa nur noch ein geringer Raum zwischen den technischen, vom Menschen geschaffenen Systemen Stadt und Kulturlandschaft.[3] Bei der Generalversammlung der Vereinten Nationen (HABITAT II[4]) heißt es zur Förderung einer Ressourcen schonenden und umweltverträglichen Siedlungs- und Stadtentwicklung, dass neben der Ressourcennutzung in den Städten und den stofflichen Austauschprozessen der Städte mit ihrem Umland für eine nachhaltige Stadtentwicklung die räumlichen Strukturen von Interesse sind. Eine zentrale Herausforderung für eine Neugestaltung nach den Zielvorgaben einer nachhaltigen zukunftsverträglichen Entwicklung ist demzufolge der Bereich Bauen und Wohnen, in dem sich die Wechselwirkungen zwischen Umweltbeeinflussung und Lebensstilen, sozialen Strukturen und Bedürfnissen, Arbeits- und Konsumgewohnheiten sehr deutlich zeigen. Die Studien der ENQUETE-KOMMISSION[5] haben die Bedeutung des Gebäudebestandes zum ersten Mal in die gesellschaftliche Diskussion eingebracht. Obwohl der Gebäudebestand seit geraumer Zeit Gegenstand zahlreicher Untersuchungen gewesen ist, bestehen Daten- und Erkenntnislücken[6]. Der Umfang, die Struktur und die Dynamik der Veränderungen von Gebäude- und Stadtstrukturen sind recht unzureichend bekannt.[7] Es existieren wenige Grundlagen, die eine Quantifizierung des Rohstoffverbrauchs und der Umweltbelastungen ermöglichen.

Die Entwicklung von Erklärungs- und Messmodellen wird zukünftig eine grundlegende Interpretation von Gebäude- und Stadtstrukturen fördern.[8]

---

3 Vgl. HABER [1996, S. 78/79]

4 BMBau [1996, 45f.]: In Istanbul wurde 1996 die Deklaration und die Habitat-Agenda verabschiedet. Kern der Deklaration ist die gemeinsame Verantwortung für Städte und Siedlungen, keinen Raubbau an den natürlichen Ressourcen zu Lasten des ländlichen Raumes und dem restlichen Teil der Welt zu betreiben.

5 Vgl. ENQUETE-KOMMISSION: „Schutz des Menschen und der Umwelt“ [1999]

6 Vgl. WÜEST [1989], DT. BUNDESTAG [1996], HEBER / LEHMANN [1996], ARLT et al. [2001], ÖKO-INSTITUT [1998], ENQUETE-KOMMISSION [1999], HOLZKAMP [1999], ENGELBACH [2000], THUVANDER [2002], GRUHLER et al. [2002], SCHWAIGER [2002], BRADLEY / FERRARA [2004], BRADLEY et al. [2005]

7 Vgl. AIBAU (Hrsg.), HOFMAN, F. G. [2001, S. 13 ff.]: Der Gesamtbestand an Hochbauten mit Ausnahme einiger Basisdaten zum Wohnungsbau, den öffentlichen Bauten und Denkmälern ist nicht erfasst.

8 Vgl. BMVBW (Hrsg.) [2003, S. 8]: Forderung statistisch repräsentativer u. kontinuierlicher Beschreibungen.

1.1.1 Nachhaltige Entwicklung[9]

Die Nachhaltigkeitsidee bezieht sich nicht mehr allein auf die Nutzung natürlicher Ressourcen, sondern dehnt sich auf die Gesellschaft als Ganzes aus.[10] In den Berichten der ENQUETE-KOMMISSION[11] findet sich die folgende Begriffserklärung: „Das Leitbild einer nachhaltig zukunftsfähigen Entwicklung zielt darauf ab, die Natur als Produktivkraft und Lebensgrundlage einschließlich ihres kulturellen, ästhetischen und Erholungswertes zu erhalten und damit eine wichtige Voraussetzung für eine stabile wirtschaftliche und soziale Entwicklung zu sichern. Es verlangt nach längerfristiger Absicherung der ökonomischen und sozialen Entwicklungschancen sowie nach Verteilungsgerechtigkeit, sowohl was die heute lebenden Menschen als auch die zukünftigen Generationen angeht."

Der Begriff ‚Entwicklung' steht in diesem Zusammenhang für den Prozess der Zustandsänderung von Regionen, Nationen oder der globalen Gemeinschaft. Die Aussagen über die Qualität und Quantität einer Zustandsveränderung müssen sich auf ein kontextuelles und gesellschaftlich ausgehandeltes Zielsystem beziehen, in dem ‚Nachhaltigkeit' die Rolle des Oberziels zukommen kann.

Die ‚Nachhaltige Entwicklung' bildet eine Synthese wissenschaftlicher, politischer und gesellschaftlicher Zielvorstellungen. Bei der Zieldefinition sind drei Dimensionen[12] zu nennen: Eine ökologische, ökonomische und sozial-kulturelle, wobei diese sich in verschiedenen Handlungsfeldern konkretisieren. Damit der dauerhafte Erhalt der natürlichen Lebensgrundlagen gewährleistet wird, ist eine nachhaltige Entwicklung grundlegende Voraussetzung und beinhaltet nach COENEN:[13] „einen […] ständigen Such- und Lernprozess, in dem das Verständnis der Nachhaltigkeit, Prioritätensetzungen und Abwägungen sowie Maßnahmen einer dauernden Weiterentwicklung unterzogen werden." Um eine ökologisch tragfähige Entwicklung anzustreben, ist zukünftig eine Reduzierung der Ressourceninanspruchnahme notwendig.

---

9 Vgl. BROCKHAUS [1979], Stichwort: ‚Nachhaltige Nutzung' - Nachhaltige Forstwirtschaft wurde 1713 im Buch ‚Sylvicultura Oeconomica' des sächsischen Oberberghauptmanns von Carlowitz erwähnt. Das Konzept ‚Nachhaltigkeit' stammt ursprünglich aus der Forstwirtschaft, in der es bereits Anfang des 19. Jahrhunderts als Leitprinzip des damals eingeführten Waldbaus proklamiert wurde. Charakterisiert wird eine Form der Waldbewirtschaftung, die berücksichtigt, dass in einem Wald nicht mehr Holz geschlagen werden darf als in dieser Zeit wieder nachwachsen kann (engl. „sustainable yield, vgl. BÄCHTHOLD [1998]).

10 Vgl. BUND/MISEREOR [1996], INTERDEPARTEMENTALER AUSSCHUSS RIO [1996, 1997] und WORLD COMMISSION on Environment and Development („Brundtland Committee") [1987]: „Nachhaltige Entwicklung ist eine Entwicklung, welche die heutigen Bedürfnisse zu decken vermag, ohne für künftige Generationen die Möglichkeit zu schmälern, ihre eigenen Bedürfnisse zu decken."

11 Vgl. ENQUETE KOMMISSION des deutschen Bundestages: „Schutz des Menschen und der Umwelt" [1994, S.54]

12 Andere Autoren ergänzen diese drei Dimensionen nachhaltiger Entwicklung um weitere Dimensionen.

13 Vgl. COENEN/GRUNEWALD [2003, S.35]

Im Kontext der nachhaltigen Entwicklung sind drei Grundstrategien zu nennen, wobei keine für sich allein genommen hinreichend[14] sein kann: Effizienz,[15] Suffizienz[16] und Konsistenz.[17] Unterschiedliche Leitstrategien[18] (siehe Tabelle 1-1) verfolgen eine Umsetzung des Konzepts ‚Sustainable Building': Langzeitprodukt, Produktnutzungsdauerverlängerung, Nutzungsintensivierung. Diese beziehen sich auf die Zielsetzungen der Stoffwirtschaft und tragen unterschiedlich zur Steigerung einer relativen oder absoluten Ressourcenproduktivität bei:

Leitstrategie	**Maßnahmen**
Konzeption des Gebäudes als Langzeitgut: Dauerhaftes Bauen und Konstruieren	▪ Eco-Design, Ziel: Langlebigkeit und Flexibilität von Bauweisen und Bauprodukten ▪ Recyclinggerechtes Konstruieren, Ziel: Planung der Wiederverwendung / -wertung, Einsatz von Sekundär-Bauteilen und Baustoffen
Verlängerung der Gebäudenutzungsdauer: Bestandsorientiertes Stoffstrommanagement	▪ Instandhaltung, Ziel: Erhalt des Gebäudebestandes ▪ Facility Management, Ziel: Gebäudebewirtschaftung
Intensivierung der Gebäudenutzung: Bestandsorientiertes Stoffstrommanagement	▪ Demand-Side-Management, Ziel: Ausgleich zwischen Nutzungsanforderungen und Gebäudepotential
Schließen von Stoffkreisläufen: Optimierung des Bauabfallrecyclings	▪ Recyclinggerechter, abfallarmer Baustellenbetrieb, Ziel: Wiederverwertung von Baustellenabfällen ▪ Bauteilorientierter bzw. selektiver Rückbau, Ziel: Wiederverwendung von Bauteilen bzw. Bauabfällen

Tabelle 1-1: Leitstrategien und Maßnahmen zur Abfallvermeidung und -verminderung im Bauwesen[19]

[14] Vgl. BROSKA [2000]

[15] Vgl. WEIZÄCKER et al. [1995], BRINGEZU [2000] und HUBER [1995, S.131]; Ziel der Effizienzstrategie ist die noch konsequentere Anwendung betrieblicher Prinzipien der Wirtschaftlichkeit selbst als auch deren Adaption auf ökologische Zusammenhänge. Durch bessere Ausnutzung vorhandener Ressourcen mit Hilfe technischer und organisatorischer Prozessoptimierung wird der Ressourcenverbrauch relativ und im besten Fall sogar absolut minimiert, also verlangsamt (z.B. die Erhöhung der Lebensdauer bestimmter Verbrauchsgüter, Vielfachnutzung von Produkten usw.).

[16] Vgl. BUND/MISEREOR (Hrsg.) [1996]: Die Suffizienzstrategie verfolgt den Gedanken eines genügsamen Umgangs mit Produkten und Gütern der Konsumgesellschaft unter Berücksichtigung der Zufriedenheit der Gesellschaft (z.B. gemeinschaftliche Nutzung bestimmter Güter).

[17] Vgl. HUBER [1998, S. 27]: Mit Konsistenzstrategie wird die Neuentwicklung ökologiefähiger Produkte und Verfahren beschrieben (z.B. Substitution fossiler Energieträger durch andere, ökologisch weniger problematische Energieträger wie Wind- oder Solarenergie). Im Zusammenhang mit der Konsistenzstrategie wird von notwendigen „ökologisch-dauerhaft verträglichen Basisinnovationen" gesprochen.

[18] Vgl. STAHEL [1993, S.53]: L = Langzeitgut; P = Produktnutzungsdauer; N = Nutzungsintensivierung

[19] Vgl. HOLZKAMP [1999, S. 68]

Eine nachhaltige Regionalentwicklung kombiniert Aspekte der globalen Nachhaltigkeit mit den Merkmalen einer eigenständigen Regionalentwicklung, die durch eine Dezentralisierung ökonomischer und politischer Strukturen und durch das Übernehmen von Verantwortung für den eigenen Raum gekennzeichnet werden kann.[20] Zur Reduzierung von Transportvorgängen und zur Verringerung von Transportentfernungen können Maßnahmen wie die regionale Vernetzung von Betrieben, die Entwicklung regionaler Absatz- und Bezugsmöglichkeiten und die Schließung regionaler Stoffkreisläufe beitragen.[21]

Der Begriff der nachhaltigen Raumentwicklung umfasst in Deutschland die Umsetzung des Konzeptes der nachhaltigen Entwicklung in die räumliche Dimension. Seit 1997 ist eine nachhaltige Raumentwicklung Planungsleitsatz der Raumplanung,[22] eine nachhaltige städtebauliche Entwicklung Planungsleitsatz der Stadtplanung,[23] und auch das im Entwurf vorliegende Umweltgesetzbuch[24] ist einer nachhaltigen Entwicklung verpflichtet.

Die nachhaltige Raumentwicklung bringt die sozialen und wirtschaftlichen Ansprüche an den Raum mit seinen ökologischen Funktionen in Einklang und führt zu einer dauerhaften, großräumig ausgewogenen Ordnung im Sinne des Ziels gleichwertiger Lebensverhältnisse. Die vorhandene Raumnutzung und deren Entwicklung sollen den Bedürfnissen derzeitiger Generationen gerecht werden, ohne die Entfaltung künftiger Generationen zu beeinträchtigen. Dazu müssen die Gestaltungsmöglichkeiten der Raumnutzung sorgfältig untereinander abgewogen und möglichst langfristig offen gehalten werden. Die Raumentwicklungspolitik hat in diesem Zusammenhang eine den gesellschaftlichen Entwicklungserfordernissen dienende Funktion zu erfüllen.

Für die nachhaltige Siedlungsentwicklung wurde in Deutschland ein nationaler Aktionsplan[25] mit Bezug auf die HABITAT II erstellt, der Zielvorstellungen für ländliche und städtische Siedlungen in den drei Nachhaltigkeitsdimensionen entwirft.

---

20 Vgl. SPEHL [1998, S. 24]: „Sie ist aus Sichtweise der Theorie des Föderalismus eine Regionalpolitik von unten bzw. eigenständige Regionalpolitik, eine wichtige Wirkungs- und Handlungsebene."

21 WIEGANDT [1999, S. 96]: „Regionen der Zukunft" (BBR) erforscht die „Möglichkeiten der Lösung ökonomischer, sozialer und ökologischer Probleme mittels Kooperation auf stadtregionaler Ebene."

22 Raumordnungsgesetz (ROG) vom 18. August 1997 (BGBl. I S. 2081, 2102), geändert durch Artikel 2 des Gesetzes zur Anpassung des Baugesetzbuches an EU-Richtlinien (Europarechtsanpassungsgesetz Bau – EAGBau) vom 24. Juni 2004 (BGBl. I S. 1359, 1379)

23 § 1 Abs. 5 Baugesetzbuch (BauGB)

24 Langfristiges Ziel der bundesdeutschen Umweltpolitik ist es, das deutsche Umweltrecht in einem Umweltgesetzbuch (UGB) zusammenzuführen. Im Herbst 1997 hat die Unabhängige Sachverständigenkommission zum Umweltgesetzbuch beim Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit ihren Vorschlag für ein Umweltgesetzbuch vorgelegt.

25 BMBAU (Hrsg.) – Bundesministerium für Raumordnung, Bauwesen und Städtebau [1996]: Nationaler Aktionsplan zur nachhaltigen Siedlungsentwicklung. Deutsches Nationalkomitee HABITAT II. Bonn

Bei der nachhaltigen Stadtentwicklung[26] geht es zum einen um die Art und Weise, wie Ressourcen in der Stadt genutzt werden, zum anderen um die stofflichen Austauschprozesse der Stadt mit anderen Gebieten und darüber hinaus um die räumlichen Strukturen der Stadt. Über die kompakte Stadt als Leitbild für eine nachhaltige Stadtentwicklung gibt es heftige Diskussionen (compact city versus suburban quality of life).[27] Es basiert konzeptionell auf der kompakten Struktur der historischen Stadt und verfolgt planungsbezogen die folgenden vier Zielelemente: hohe Bebauungsdichte, hohe Durchmischung, gute Erschließung durch öffentlichen Personennahverkehr und Verbesserung öffentlicher Räume und Wohnumfelder.

Eine Stadtregion leicht verständlich zu charakterisieren und zu bewerten sowie seine Dynamik zu untersuchen ist eine interdisziplinäre Aufgabe angefangen vom Städtebau bis zur Geoinformatik.

Im Sektor Bauen und Wohnen wurden durch die ENQUETE-KOMMISSION[28] Leitbilder für eine nachhaltige Entwicklung formuliert, die für die Bereiche Stoffe, Energie, Kosten, Problemstoffe und Umweltbelastung gelten. Darüber hinaus wurden Zieldimensionen[29] definiert. Festzustellen ist, dass oftmals kurzfristig ökonomisch orientierte Interessen den Belangen des Allgemeinwohls im Wege stehen, so dass die Aufgabe für die ausführenden Behörden und die Planung darin enden, ihre jeweiligen Kompetenzbereiche zu konkretisieren.[30]

Tabelle 1-2 nennt Forderungen an eine nachhaltige Politik für den Baubereich.

Einige Forderungen an eine nachhaltige Politik für den Baubereich sind zu stellen, d.h.:
„dass die vorhandenen Baukonstruktionen möglichst lange auf einem hohen Niveau weitergenutzt werden,"
„dass der existierende Gebäudebestand effizient gepflegt und genutzt wird,"
„dass der Energiebedarf für die Produktion und Nutzung von Gebäuden weiter gesenkt wird,"
„dass möglichst wenig bisher unbebaute Flächen neu bebaut werden,"
„dass bei Baumaßnahmen ein hoher Anteil von bereits existierenden Baustoffen wieder verwendet wird,"
„dass möglichst wenig neu gebaut wird,"
„dass neue Baukonstruktionen sowohl dauerhaft, reparaturfähig, pflegefreundlich und energiesparend im Betrieb geplant werden,"
„dass ungiftige, umwelt- und gesundheitsverträgliche, trennbare und weiterverwendbare Baustoffe entwickelt und eingesetzt werden,"
„dass kulturelle Kapitalien in ihrer Bedeutung für ganzheitliche Werterhaltungsstrategien erkannt und berücksichtigt werden,"
„dass der arbeitsmarktpolitische / soziale Effekt eines Ersatzes von Ressourcen durch Arbeit im Rahmen einer Bestandspflege erkannt und genutzt wird."

Tabelle 1-2: Forderungen an eine nachhaltige Politik für den Baubereich[31]

[26] BBR (Hrsg) [2004]: Nachhaltige Stadtentwicklung
[27] BREHENY [1992, S. 142-156], DEIMER [1998, S. 2], ALBERS [2000, S. 23], BEATLEY [2000, S. 29-75]
[28] Vgl. ENQUETE-KOMMISSION, Hrsg., Hassler, U. / Kohler, N. / Paschen, H. [1999, S. 17]
[29] Vgl. ENQUETE [1998, S.127], Abbildung 15, Zieldimensionen für den Bereich „Bauen und Wohnen"
[30] Vgl. HÜBLER, KAETHER (Hrsg.) [1999]
[31] Vgl. ENQUETE-KOMMISSION, Hassler, U. / Kohler, N. / Paschen, H. [1999, S. 2 und 3]

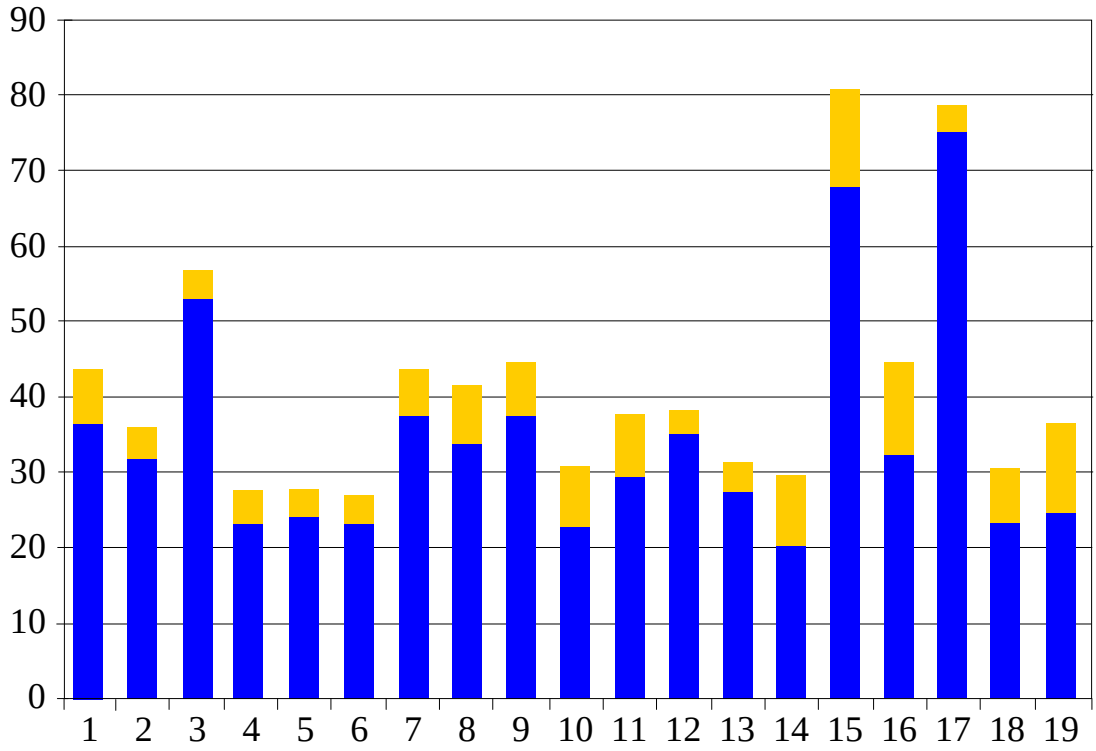
Tabelle 1-3 zeigt positive bzw. negative Trends im Bereich Bauen und Wohnen.

Indikator	**Entwicklung**	**Ziel**
Belastung der Wohnbevölkerung durch Lärm	1999 waren ca. 20 % der Bevölkerung (alte BL) durch Wertüberschreitung betroffen.	Die Lärmbelastung in der Wohnung sollte tagsüber einen Wert von 65 dB (A) und nachts einen Wert von 55 dB (A) nicht überschreiten.
Innenraumbelastung	Haushalte, deren Wohnungen durch Gerüche, Abgase und Staub beeinträchtigt sind. 1978: 22,9 % (alte BL), 1993: 32,3 % (Deutschland)	Einhaltung der Richtwerte für die Innenraumluft nach der Ad-hoc-Arbeitsgruppe für die Erarbeitung von Richtwerten für die Innenraumluft aus Mitgliedern der Kommission „Innenraumlufthygiene“ des Umweltbundesamtes und des Ausschusses für Umwelthygiene der Arbeitsgemeinschaft der Obersten Landesgesundheitsbehörden.
Wohnungslose	Über 1 Mio.	Anzahl soll sinken
Einkommensanteil, der für die Wohnung ausgegeben werden muss.	1960: 10 %, 1998: 25 %	Anstieg soll gestoppt werden
Flächenverbrauch für Siedlungs- und Verkehrszwecke	1981-1985: 112 ha/Tag, 1997-2001: 129 ha/Tag	2020: 30 ha/Tag (Bundesregierung)
Versiegelungsgrad	1960: 0,88 Mio. ha, 1993: 1,58 Mio. ha (alte BL)	Verlangsamung der Zunahme
Leerstandsquote	6,0 %	Reduktion
Relation Wohnungsbauleistungen im Bestand zu denen im Neubau	1,15:1	Anstieg
Relation der Neubauten von Mehrfamilien zu Ein- /Zweifamilienhäusern	1:11,6	Reduktion
Recyclingquote von Baureststoffen ohne Boden	1996: 69,8 %, 1998:71,6 %	2010: 80 % (EU)
Entnahme mineralischer Rohstoffe für die Bauwirtschaft	1996: 675 Mio. t, 2000:730 Mio. t	Reduktion auf 50 % gegenüber dem Wert von 1994 (Bundesregierung)
Durchschnittlicher Wärmeverbrauch der Wohngebäude	1998: 220 kWh/qm/a	2020: 110 kWh/qm/a (Anlehnung an Enquete-Kommission, Energieversorgung)
CO^2-Emissionen durch Wohnungsnutzung (Raumheizung und Warmwasserbereitung)	1990: 128 Mio. t, 1998:136 Mio. t	In Deutschland Reduktion der Gesamtemissionen zwischen 1990 und 2020 um 40 % (Bundesregierung)

Tabelle 1-3: Wesentliche Nachhaltigkeitsdefizite des Aktivitätsfeldes Wohnen und Bauen[32]

[32] Vgl. COENEN/GRUNWALD (Hrsg.) [2003, S.169]

Abbildung 1-1 charakterisiert das Aktivitätsfeld ‚Wohnen und Bauen' anhand ausgewählter ökonomischer, ökologischer und ressourcenbezogener Parameter.[33] Die Nachhaltigkeitsanalyse bezieht sich dabei auf den Ansatz der Aktivitätsfelder, die eine Beschreibung einer gesellschaftlichen Aktivität auf Basis einer bedürfnisorientierten Einteilung des gesamtgesellschaftlichen Konsummusters unter Einbeziehung der ökonomisch-technischen Zusammenhänge darstellt.



Ökonomische Parameter		Ökologische Parameter		Ressourcenbezogene Parameter	
1	Abschreibung	7	Primärenergieverbrauch	14	Wasserverwendung
2	Nettowertschöpfung	8	Treibhausgase	15	Verwertete Rohstoffe
3	Sonstige Einkommen	9	CO_2	16	Nicht verwertete Rohstoffe ohne Bodenaushub
4	Arbeitseinkommen	10	Versauerungsgase		
5	Arbeitnehmer	11	Staub	17	Abfall mit Bauschutt
6	Erwerbstätige	12	NMVOC	18	Abfall ohne Bauschutt
		13	CO	19	Besonders überwachungsbedürftiger Abfall

= Im Inland ausgelöste Effekte
= Im Ausland ausgelöste Effekte, bez. auf den jeweiligen inländischen Gesamtwert des Parameters

Abbildung 1-1: Parametrisierte Beschreibung des Aktivitätsfeldes Wohnen und Bauen 1993 (=100)[34]

[33] Vgl. COENEN/GRUNWALD (Hrsg.) [2003, S.154]: „Da sich die Angaben in der Literatur in der Regel auf einzelne Teilbereiche beschränken und nicht die gesamte Vorleistungskette beachten, sind die Ergebnisse nicht unmittelbar mit in den meisten anderen Veröffentlichungen zu findenden Daten vergleichbar."

[34] Vgl. COENEN/GRUNWALD (Hrsg.) [2003, S.156], Abbildung 4.8: Die Berechnungen inklusive sämtlicher Vorprodukte erfolgen auf Grundlage der Input-Output-Tabellen des Statistischen Bundesamtes [1997]. Die Ergebnisse sind aus den Angaben zur Methodik und Daten in KLANN und SCHULZ [2001] reproduzierbar.

1.1.2 Urbane Systeme[35]

Die Stadt, die als ein Baugefüge im Raum, jedoch im großen Maßstab definiert wird,[36] ist im historischen Kontext immer dynamisch gewesen, immer in Unordnung und im Aufbruch, so dass in Abhängigkeit von ökonomischen, sozialen oder anderen Faktoren Prozesse der Ausdehnung mit denen der Schrumpfung wechseln. Als Forschungsgegenstand bietet die Raum- und Stadtstruktur für verschiedene wissenschaftliche Disziplinen interessante Arbeitsfelder und der Erkenntnisgewinn erfolgt in jüngster Zeit verstärkt durch interdisziplinäre Ansätze.[37] Zunächst werden einige Begriffe definiert und ausgewählte relevante historische Forschungsansätze dargestellt, um schließlich die Anwendungsmöglichkeiten von Ähnlichkeitsuntersuchungen vorzustellen.

Die Begriffe Raum- und Stadtstruktur gehören zu den zentralen Begriffen in Raumforschung, Raumordnung und Raumplanung. Der Ausdruck Stadtstruktur lässt sich durch viele andere Begriffe wie z.B. städtische Siedlungsstruktur, städtische Bebauungsstruktur, städtisches Beziehungsgefüge oder städtische Gebäudestruktur ersetzen. Mit dem Bundesraumordnungsgesetz werden Zielaussagen für die Raumordnungspolitik und damit für die nachgeordnete regionale Wirtschaftspolitik gesetzlich festgelegt.[38] Es sieht vor, dass eine Entwicklung des Gesamtraumes einer ausgewogenen Siedlungs- und Freiraumstruktur unterliegt, die zugleich die Funktionsfähigkeit des Naturhaushaltes berücksichtigt. Weiterhin soll eine Zersiedlung der Landschaft vermieden werden, jedoch eine effektive Infrastruktur aufrecht erhalten bleiben. Ländliche Räume bedürfen struktureller Entwicklung und Erholungsgebiete sind gezielt aufzubauen. Für die Erfüllung dieser und weiterer daran anknüpfender Vorschriften ist eine Abgrenzung der administrativen Einheiten nach einer Vielzahl an Merkmalen notwendig. Maßnahmen, die zu beobachtende bzw. erwartende Diskrepanzen zwischen angestrebten Soll- und zu erwartenden Ist-Zuständen abbauen sollen, erfordern klare Zielvorstellungen, eine umfassende Analyse der Ausgangslage und der vergangenen Entwicklung, alternative Prognosen und Kenntnis der Wirksamkeit der zur Verfügung stehenden Mittel.

---

35 Vgl. BACCINI/OSWALD [1999]: „In einem urbanen System haben mindestens 90 % aller Einwohnerinnen in rund einer halben Stunde Reisezeit Zugang zu sämtlichen „urbanen Angeboten“ in den Bereichen wie Güterversorgung, Gesundheitspflege, Ausbildung, Arbeit, Kulturpflege und sind mit ihren Privathaushalten an die technische Infrastruktur eines großräumigen Versorgungs- und Entsorgungssystems essenzieller Massengüter (Wasser, Nahrungsmittel, Baumaterialien, Energieträger) angeschlossen.“

36 Vgl. LYNCH [2001, S. 26-27]

37 Die interdisziplinäre Arbeit ist in der heutigen Forschung von großer Bedeutung. Laut einer Prognose des ehemaligen DFG-Präsidenten WOLFGANG FRÜHWALD (1997) werden sich die Einzelwissenschaften systematisch annähern. Durch Verbindung der Fachdisziplinen ist es möglich, Richtungen und Quellen zu neuen Erkenntnissen zu finden.

38 Raumordnungsgesetz (ROG) vom 18. August 1997 (BGBl. I S. 2081, 2102), geändert durch Artikel 2 des Gesetzes zur Anpassung des Baugesetzbuches an EU-Richtlinien (Europarechtsanpassungsgesetz Bau – EAGBau) vom 24. Juni 2004 (BGBl. I S. 1359, 1379).

Für Stadtregionen oder Stadtlandschaften, die durch unvermindertes Ausufern der Stadt in das Umland beschrieben werden, ist der Begriff des räumlichen Ordnungsgefüges ein analoger Begriff, der in diesem Fall aber präziser das Phänomen beschreibt und als die Gesamtheit der räumlichen Konfiguration im Sinne von Einteilung, Gliederung, Gruppierung, Rangordnung und Staffelung von Flächen unterschiedlicher Nutzungsarten zu verstehen ist. Ein räumliches Ordnungsgefüge wird in verschiedene Dimensionen unterteilt (z.B. physisch - urban physics, funktional - urban biology oder auch kulturell). In diesem Zusammenhang ist auf so genannte Nominal- und Realdefinitionen zu verweisen, um die Begriffe zu bewerten und genauer zu beschreiben.[39] Ergänzend ist die Begriffsdefinition von THOMAS SIEVERTS [1997] anzufügen, der zur Vereinfachung die Form der verstädterten Landschaft bzw. der verlandschafteten Stadt durch die ‚Zwischenstadt' ergänzt und feststellt, dass der Übergang zwischen Stadt und Land fließend ist und nicht mehr an klaren Stadtkanten abzulesen ist.[40]

Historisch betrachtet sind seit etwa zwei Jahrhunderten die räumlichen Strukturen ein Gegenstand theoretischer Überlegungen und praktischer Forschungen. Im deutschsprachigen Raum wurden diese zu Beginn des 19. Jahrhunderts zum Inhalt reger autodidaktischer wissenschaftlicher Arbeit bei Johann Heinrich von THÜNEN (1783-1850).[41] Seine Arbeit führt modellanalytisch zu einer konzentrischen Ringstruktur um einen zentralen Ort in einem geschlossenen Wirtschaftsystem. Es begründet die landwirtschaftliche Standortlehre und wahrscheinlich die Raumforschung im engeren Sinne.

Die Theorie der Zentralen Orte von WALTER CHRISTALLER ist eine der bedeutendsten Theorien in der Anthropogeografie.[42] Die Arbeit vollendet nicht nur die von HANS BOBEK[43] über Innsbruck und von GEER[44] über Stockholm begründete Phase der funktionalen Siedlungsgeografie, sondern legt auch den Grundstein für eine eigene Forschungsrichtung: die Zentralitätsforschung. BLOTEVOGEL[45] fasst die in den letzten Jahren immer stärker werdende Ablehnung des Zentralen-Orte-Konzeptes zusammen und nennt Gründe dafür, dass eine behutsame Anpassung des Konzeptes an die heute existierenden Bedingungen notwendig ist. Er verweist auf die Möglichkeit einer empirischen funktionalen Landschaftsgliederung.

---

39 Vgl. THINH [2004 a, S. 4], JENKIS [1996, S. 18]

40 Vgl. SIEVERTS [1997], siehe auch MUMFORD [1938 / 1970]

41 THÜNEN [1826]

42 CHRISTALLER [1933]

43 BOBEK [1928]

44 GEER [1923]

45 BLOTEVOGEL [1996, S.12]: „Insofern interessierte weniger die Zentrale-Orte-Theorie selbst, sondern vor allem die empirische Erfassung und Darstellung der Zentralität von Siedlungen sowie die Abgrenzung ihrer Bereiche, um auf diese Weise zu einer funktionalen Landschaftsgliederung zu gelangen."

In den letzen Jahren zeigten einige Stadtforscher,[46] dass die räumliche Ausdehnung der Stadt zu ganzen Stadtlandschaften durchaus inneren Ordnungsprinzipien folgt und weiterhin die Formen der heutigen Stadtlandschaften den Gesetzmäßigkeiten der fraktalen Geometrie unterliegen können.

In den nächsten zwei Generationen wird die urbane Lebensform weltweit dominant.[47] Urbanität ist definiert als eine bestimmte Organisation des Politischen, als Demokratie, als Organisation des Ökonomischen und schließlich als eine bestimmte Art zu leben.[48] Die vergangenen fünf Jahrzehnte haben gezeigt, dass eine beschleunigte Verschiebung menschlicher Siedlungen vom Ruralen ins Urbane stattgefunden hat.[49] Die Entfernung der Orte oder Räume ist in Anbetracht des technologischen Fortschritts von sekundärer Bedeutung.[50] Einerseits ist die Geschwindigkeitszunahme der Verkehrssysteme dafür verantwortlich, und andererseits beinhaltet der virtuelle Daten-Highway[51] die Auflösung des geographischen Maßstabs. Zunehmende Kapitalkonzentration in den Zentren der wirtschaftlichen Entwicklung und gleichgerichtete Wanderungsbewegungen haben zu Wachstums- und Wohlstandsunterschieden zwischen den einzelnen Teilräumen in Deutschland geführt, die den gesellschaftspolitischen Zielvorstellungen – insbesondere dem Ziel gleichwertiger Lebensbedingungen für alle Teile der Bevölkerung – z.T. 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 wird und daher durch Maßnahmen der regionalen Wirtschaftspolitik den unerwünschten Entwicklungstendenzen entgegenwirkt und ein Ausgleich der interregionalen Unterschiede im jeweils erreichten Entwicklungsstand angestrebt werden muss.

---

46 BATTY&LONGLEY [1994], FRANKHAUSER [2000,2002] und HUMPERT et al. [1996, 2002]

47 Vgl. LICHTENSTEIGER [2006, S.1]: Einführung von PETER BACCINI

48 HÄUSSERMANN/SIEBEL [1999, S.19-21]

49 Vgl. BACCINI, KYTZIA, OSWALD [2002]

50 Vgl. THOMAS SIEVERTS [1997]: „Die heutigen Stadtbewohner wählen, unabhängig von der Dichte der Städte, ihre Sozialkontakte weniger nach räumlicher Nähe und Nachbarschaft, als vielmehr nach nichträumlich vermittelten Interessen und Neigungen."

51 Vgl. PRIGGE: „Wie urban ist das digitale Zeitalter?", In: Kulturregion Stuttgart e.V. (Hrsg.) [1999, S.54]

1.1.3 Herausforderungen der Informationsgesellschaft

Durch den schnellen Fortschritt in der Informationstechnologie und das immer schnellere Anwachsen der Datenmengen steigen die Anforderungen an Systeme, die Wissen in irgendeiner Form aus Daten extrahieren und darstellen.[52] Man schätzt, dass sich die weltweit vorhandene Informationsmenge alle 20 Monate verdoppelt,[53] jedoch die Bedeutung und die Erkenntnisse, die man aus diesen Daten zieht, steigen nicht entsprechend. KEIM[54] prognostiziert im Jahr 2002, „[…] dass in den nächsten 3 Jahren mehr Daten generiert werden als in der gesamten menschlichen Entwicklung zuvor […]“ und beruft sich auf eine Berechnung von Forschern der Universität Berkeley, wonach jedes Jahr ca.1 Exabyte (= 1 Million Terabyte) Daten generiert werden. Das stetige Wachstum der Datenbestände macht den Zugriff auf die gewünschten Informationen immer schwieriger, eine manuelle Analyse immer zeitaufwendiger, personalintensiver und dadurch kostspieliger und für einen Menschen nahezu unmöglich. Es werden daher immer Zeit sparendere und effektivere Systeme und Methoden zur Wissensgewinnung gesucht. Es ist festzustellen, dass die Daten ursprünglich meist für andere Zwecke als die Verwendung gesammelt und routinemäßig archiviert werden.

Die Bezeichnung Data Mining stammt ursprünglich aus dem Bereich der Statistik und kennzeichnet dort die selektive Methodenanwendung zur Bestätigung vorformulierter Hypothesen.[55] Noch heute beruhen zahlreiche Data-Mining-Methoden auf statistischen Verfahren.[56] Data Mining wird definiert als Erforschung und Analyse großer Datenmengen mit automatischen oder halbautomatischen Werkzeugen, um bedeutungsvolle Muster und Regeln aufzufinden.[57] Muster sind Ausdrücke, die eine Teilmenge dieser Daten beschreiben und das zu extrahierende oder bereits gewonnene Wissen repräsentieren.[58] Data Mining wird als eine Methodik zur Problemlösung verstanden, um logische oder mathematische, zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen zu entdecken.[59] Es wird als Teilschritt des KDD-Prozesses (Knowledge Discovery in Databases) angesehen, der aus der Anwendung von Datenanalysealgorithmen besteht und zu einer

---

52 Vgl. CLAUS [2003, S. 9-50]: „In der Praxis sind im Bereich der Modellbildungsmethoden für Entscheidungshilfesysteme verschiedene Verfahren bekannt: Markovketten, Bayes-Netze, Petri-Netze, Künstliche Neuronale Netze, Expertensysteme, Fuzzy-Logic und Fuzzy-Pattern Klassifikator-Netze. Jede dieser Methoden kann auf seine Weise Wissen abbilden und mit diesem Wissen für die Lösung einer gestellten speziell für den Ansatz angepassten Aufgabe unterstützend oder vollständig dienlich sein.

53 Vgl. QUECKBÖRNER [2004]

54 Vgl. KEIM [2002]

55 GROB / BENSBERG [1999]

56 FAYYAD et al. [1996]

57 BERRY / LINOFF [1997]

58 PETRAK, J. [1997]

59 DECKER / FOCARDI [1995], ALPAR [2005, S.38]

Auflistung von Mustern führt, die aus den Daten gewonnen wurden.[60] Abbildung 1-2 verdeutlicht den Einsatz des Data Mining zur Erkenntnisgewinnung sowie die Verbindung zwischen Aktion und Wissen in Abhängigkeit von der Datenveränderung.

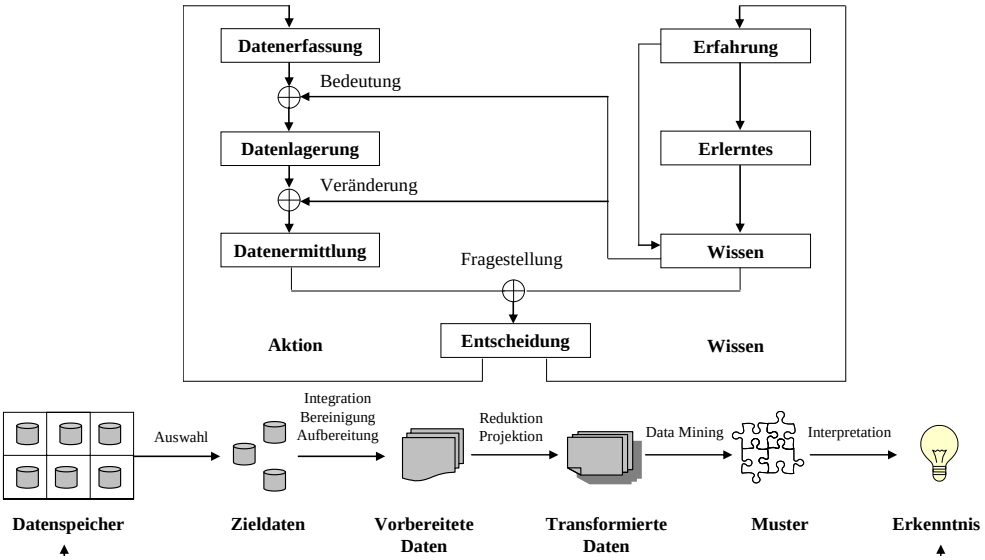


Abbildung 1-2: Aktion und Wissen, Daten, Data Mining, Erkenntnis[61]

Die Analyseaufgabe definiert die Art nach welcher die Muster gesucht werden. Die beschreibenden Analysen suchen nach Gruppen (Cluster), während die vorhersagenden verstärkt an Verbindungsmustern (Link), zeitlichen Mustern (Sequenz), Regeln und Abhängigkeiten sowie Formeln und Gesetzmäßigkeiten interessiert sind. Eine konkrete Analyseaufgabe ist die Klassifikation. Klassifizierungsverfahren stellen ein wichtiges multivariates statistisches Instrument zur Durchdringung komplexer Strukturen dar, die die in der Vielzahl der zu berücksichtigenden Variablen enthaltenen Informationen ordnen und verdichten. Es handelt sich um die Aufdeckung inhärenter Gruppenstrukturen unter der Annahme, „dass die betrachtete Objektmenge in mehrere kleine, prinzipiell gut unterscheidbare Gruppen zerfällt; indessen sind weder die Zuordnung der Objekte zu den einzelnen Gruppen noch die kennzeichnenden Charakteristika dieser Gruppe bekannt“.[62] Einzelne, homogene Objektklassen,[63] die BOCK[64] als „Objekt-Typen“ bezeichnet, lassen sich hinsichtlich gewählter Eigenschaften interpretieren – wie etwa bei Krankheitstypen – und können jederzeit gesondert behandelt werden. Zwischen den Gruppen lassen sich ggf. Zusammenhänge aufdecken. Statistiker und Taxonomen haben sich bemüht, objektive Kriterien für eine derartige Aufgliederung zu entwickeln.

60 Vgl. FAYYAD et al. [1996]
61 Eigene Bearbeitung unter Anlehnung an FAYYAD et al. [1996]
62 Vgl. BOCK [1974, S. 13]
63 Vgl. SCHULZE [1980, S. 52]
64 Vgl. BOCK [1974, S. 1]

1.1.4 Räumliche und zeitliche Ähnlichkeitsmuster

Die Raumtypisierung und Regionalisierung dient der gedanklichen Ordnung der Vielfalt der Realität und ermöglicht dadurch die Erfassung und Charakterisierung der räumlichen Struktur- und Interaktionsmuster.[65] Im Hinblick auf die ‚Klassifizierung von Städten'[66] ist festzustellen, dass diese geeignet ist, allgemeine Sätze über das räumliche Muster von Städten zu formulieren und die Beziehung zwischen Städten mit bestimmter Funktion und ihren Hinterländern darzustellen.[67] Im Bereich der Regional- und Stadtforschung gibt es eine Reihe von Versuchen, die Vielzahl der Städte auf wenige überschaubare Gruppen von Städten zu reduzieren. In der Anfangsphase fehlte statistisches Zahlenmaterial, um eine umfassende Untersuchung der Städte eines Landes vornehmen zu können. Mit der Einführung regelmäßiger Volkszählungen und anderer statistischer Erhebungen konnten diese Schwierigkeiten beseitigt werden. Die ‚Klassifizierung von Städten' ermöglicht es, statt einzelner Merkmale sämtliche relevante Merkmale simultan zu berücksichtigen, und mit zunehmender Entwicklung der Leistungsfähigkeit von Computern und Algorithmen werden die Klassifizierungsansätze komplexer. Abbildung 1-3 gibt eine chronologische Übersicht der thematisch relevanten Untersuchungsansätze.[68]

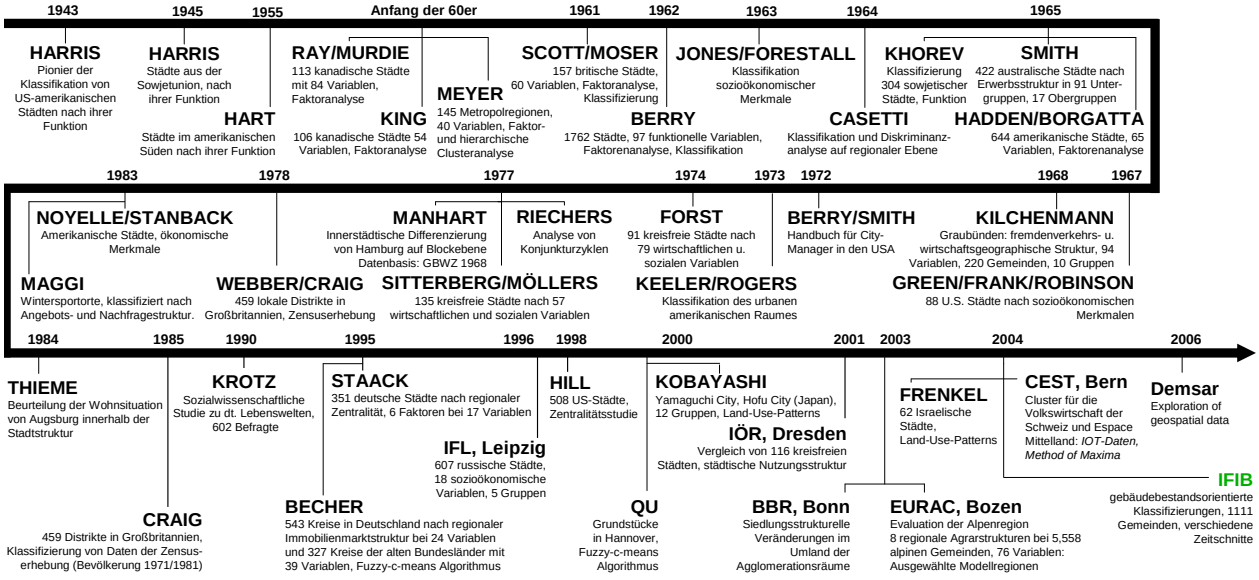


Abbildung 1-3: Chronologischer Überblick zur Klassifizierung von Städten

---

65 Vgl. FISCHER, M. [1982, S. 24], der die zwei Hauptphasen der Forschungsperspektive Regionalisierung zusammenfassend darstellt und ausführlich die phänomenologisch-ontologische Phase (z.B. HAHN [1892], HERBERTSON [1905], THORNWAITE [1933]) sowie die empirisch analytische Phase beschreibt (z.B. BARTELS [1968], RODOMAN [1967], GALE/ATKINSON [1979]).

66 Der Begriff „Klassifizierung von Städten" ist in den vergleichbaren Arbeiten verwendet und beschreibt zunächst im Allgemeinen die Arbeit mit Klassifizierungsverfahren auf dem Gebiet der Regional- und Stadtforschung. Der Begriff ist als abstrakter Oberbegriff anzusehen und ist unabhängig von den Untersuchungsobjekten, die sich auf unterschiedliche administrative Verwaltungseinheiten beziehen.

67 Vgl. LICHTENBERGER [1998, S. 34]

68 Es handelt sich um eine Auswahl, wobei die Abbildung keinen Anspruch auf Vollständigkeit tragen kann.

Das Entdecken von Ähnlichkeitsbeziehungen unter Voraussetzung bestimmter Rahmenbedingungen eignet sich, differenzierte Orientierungen für Städte mit weitgehend ähnlichen Problemlagen und Potentialen zu entwickeln. Die Positionierung einzelner Städte im Wettbewerb wird ermöglicht, so dass die Leistungen der eigenen Kommune auf Grundlage von Städteindikatoren bewertbar und mit anderen Kommunen vergleichbar werden (z.B. Vergleichsmöglichkeit für Städte auf dem Weg zu einer nachhaltigen Entwicklung).[69] Der Städtevergleich ist nach ARLT[70] ein zentrales methodisches Instrument zur Klärung der Frage nach einer nachhaltigen nutzungsstrukturellen Entwicklung, die ebenso zur Ausgewogenheit sozioökonomischer und ökologischer Leistungen in den Städten beiträgt. Es ist mit Hilfe von Flächenleistungen und Basisindikatoren möglich, auf empirische Weise Skalen und Bewertungsmaßstäbe zu entwickeln sowie eine stadttypologische Differenzierung zu definieren. Als methodische Instrumente dienen Städtevergleiche und stadttypologische Gliederungen dazu, auf Grundlage städtestatistischer Informationen Maßstäbe und Bewertungsskalen städtischer Phänomene aufzubauen.

In der Regionalforschung und Regionalpolitik kommt der ‚Klassifizierung von Städten' eine erhebliche Bedeutung zu, da die Objekte als regionalwirtschaftliche Einheiten betrachtet werden können und somit wichtig für Strukturentwicklungskonzepte sind. Für eine detaillierte Regionalpolitik sind das ökonomische und ökologische Leistungsvermögen ein wichtiges Kriterium, um Städte zu bewerten.[71]

Für die Unternehmenspolitik stellt die Klassifizierung eine wichtige Entscheidungshilfe dar, weil Städte wichtige Absatzgebiete für viele Konsumgüter darstellen und somit eine große Bedeutung für das Marketing haben. Eine Klassifizierung der städtischen Märkte führt beispielsweise dazu, dass unternehmensspezifische Testmärkte[72] gefunden werden können. Im letzen Jahrzehnt wächst in Deutschland das Interesse, die regionalen Immobilenmärkte mit Klassifizierungsverfahren strukturell zu bewerten, und es werden Gruppen gebildet, die als Zuordnungsbasis für zukünftige Kauffälle dienen können.[73]

---

69 Vgl. DAVID S. ARNOLD in: BERRY [1972, Chapter 12, S.361 ff.]: „For the urban administrator classifications of cities are one part of an urban information framework within administrative analysis, research, budgeting, programming, and planning can be operational. […] Classification increasingly will be used as a tool of urban planning, partly because of substantial changes that have taken place in the scope of urban planning in recent years. […] Classifications are fluid rather than final and will change as methodologies are improved through future research, as better quality data become available, and as urban characteristics, both properties and their relationships, change over time."

70 Vgl. ARLT et al. [2001, S. 33]

71 Vgl. ARLT et al. [2001, S. 66]

72 Vgl. GREEN, FRANK und ROBINSON [1967]

73 Vgl. BECHER [1995] und QU [2000]

Die Klassifizierungen nach bestandsorientierten[74] Merkmalsträgern verbessern die Möglichkeiten, grundlegende Informationen über den Gebäudebestand dokumentieren und analysieren zu können, wobei Struktur- und Veränderungsprozesse des Gebäudebestandes dabei nicht unwesentlich sind.

In Deutschland beziehen sich die Untersuchungen räumlich bzw. administrativ in der Regel auf die Kreisebene, die sich aus Stadt- und Landkreisen zusammensetzt. Sehr häufig wird die Objektmenge der Klassifizierung speziell aus Stadtkreisen bzw. kreisfreien[75] Städten gebildet. In Einzelfällen wird eine gewählte Gemeinde mit Hilfe von Klassifizierungsverfahren auf Ebene der Wohnblöcke oder Grundstücke untersucht und strukturell auch mit Hilfe selbst erhobener Statistiken analysiert. Bei einer umfassenderen Klassifizierung auf Gemeindeebene ist festzustellen, dass es sich um eine sehr hohe Betrachtungsstufe handelt, die zusätzliche Herausforderungen beinhaltet.[76] Auf den hohen Zeitaufwand bei der Datenerhebung ist besonders hinzuweisen. Nach aktuellem Kenntnisstand ist ein Klassifizierungsmodell nicht existent, welches sowohl die Ebene der Kreise als auch der Gemeinden berücksichtigt bzw. inhaltlich Bezüge zwischen beiden Ebenen zu definierten Fragestellungen herstellt.

Zeitlich betrachtet, wählen bisherige Arbeiten als Datenreferenzpunkt ein festes Jahr bzw. einen eindeutigen Zeitschnitt, um die Objektmenge in Teilgesamtheiten aufzugliedern. In den meisten Fällen handelt es sich um abgeschlossene Untersuchungsmodelle, die nicht die Möglichkeit vorsehen, eine Datenbasis für darauf aufbauende Untersuchungen zu erzeugen. Eine nachträgliche Zuordnung von Objekten zu bestehenden Klassifizierungen ist in der Regel nicht vorgesehen und der Einsatz von so genannten Klassifikatornetzen ist in den bisherigen Arbeiten nicht diskutiert. Die dynamische ‚Klassifizierung von Städten' ermöglicht die Abbildung von Objektveränderungen innerhalb berechneter Gruppenstrukturen aufgrund von Änderungen in der Variablenstruktur.

Es sind in der Zukunft weitere Untersuchungsmöglichkeiten denkbar, da methodisch daran geforscht wird, die zeitliche Prognose in die Klassifikationsmodelle mit einzubeziehen und Modelle des zeitlichen Verhaltens zu erzeugen.[77]

---

74 Vgl. FERRARA, C. [2004]

75 Die kreisfreie Stadt (in Baden-Württemberg als Stadtkreis bezeichnet) ist eine kommunale Gebietskörperschaft, die nach dem Kommunalrecht Deutschlands ihre Aufgaben in eigener Zuständigkeit erledigt. In der Regel handelt es sich um Großstädte, also Städte mit mehr als 100.000 Einwohnern oder größere Mittelstädte. Die kleinste kreisfreie Stadt in Deutschland ist Zweibrücken in Rheinland-Pfalz mit 38.000 Einwohnern (2004), der größte Stadtkreis die bayerische Landeshauptstadt München mit 1,3 Millionen Einwohnern. Berlin und Hamburg sind zwar größer, aber nicht kreisfreie Städte, sondern Stadtstaaten.

76 Vgl. STAACK [1995], SIEDENTOP et al. [2003] und BEHNISCH, MARTIN [2004]

77 Vgl. PÄßLER, M. [1999, S. 26]

1.2 Zielsetzung

Die Ziele und Grundsätze der Raumordnung in Deutschland sind nicht als absolut und statisch anzusehen, sondern bedürfen stetig einer Überprüfung und ggf. Neuausrichtung, insbesondere zur Festlegung von Prioritäten.[78] Die Bewertung räumlicher Strukturen und damit verbundenener Entwicklungsmuster ist auf unterschiedliche Weise möglich, wobei ein anerkanntes Zielsystem nicht existiert. Auch wenn die Erarbeitung von Leitbildern für eine räumliche Entwicklung als wichtige Aufgabe der Raumordnung verstanden wird, gibt es z.B keinen politischen Konsens über die damit verbundenen Messgrößen und zulässigen Ausprägungen.

Der Begriff des ‚Urban Data Mining'[79] bezeichnet einen Forschungsansatz, der durch die Erarbeitung einer im wesentlich methodisch orientierten Vorgehensweise charakterisiert wird und sich eignet, Hypothesen zu formulieren und zu überprüfen. Neben Methoden, die eine kritische Bestandsaufnahme und Auseinandersetzung mit vorhandenen räumlichen Eigenschaften und Entwicklungstendenzen ermöglichen, werden vor allen Dingen methodische Vorgehensweisen gesucht, die sich dazu eignen, bereits vorhandene Informationen oder Erkenntnisse auf weitere Untersuchungsgebiete zu übertragen. Die Aufgabe besteht somit darin, für Forschungsfragen im urbanen Kontext geeignete Methoden auszuwählen und diese mit Hilfe eines im Wesentlichen empirisch-analytischen[80] Untersuchungsansatzes am deutschen Gemeindesystem darzustellen. Dokumentiert werden einerseits Methoden, die Ähnlichkeiten zwischen Untersuchungsobjekten entdecken und andererseits Methoden, die geeignet sind, daraus Erkenntnisse abzuleiten und zu bewerten. Auf Grundlage dieses methodischen Diskurses ist zukünftig ggf. der Aufbau eines umfassenden Regelwerkes möglich, welches sich zur Bewertung von räumlichen Strukturen eignen könnte. Mit Blick auf die Regionalisierung[81] als traditionelles Anliegen der Geografie kann die Gliederung von räumlichen Untersuchungseinheiten zur räumlichen Theoriebildung herangezogen werden, aber auch zur Verwerfung bereits existierender Theorien einen entscheidenden Beitrag leisten.

---

78 Siehe Abschnitt 1.1.1 - Nachhaltige Entwicklung

79 Ein wesentliches Kennzeichen von empirisch-analytischen Theorien im Allgemeinen ist die empirische Testbarkeit. Der Begriff des ‚Urban Data Mining' wurde vom Autor selbst definiert und soll eine Überprüfungsmöglichkeit von gewählten Hypothesen über die Raum- und Stadtstruktur und den damit verbundenen georeferenzierten Objekten beschreiben. Dabei wird grundsätzlich auf bereits allgemein anerkannte Methoden des Data Mining zurückgegriffen und zusätzlich wird der Einsatz von üblichen GI-Werkzeugen zur Verarbeitung von Geoinformationen berücksichtigt.

80 Vgl. NARR [1971, S. 41], der drei Kategorien der Theoriebildung unterscheidet: (1) ontologisch-normative Theorie, (2) dialektisch-historische Theorie, (3) empirisch-analytische Theorie.

81 Vgl. FISCHER, M [1982, S. 22]: „Unter einer räumlichen (empirisch-analytischen) Theorie kann man ein informatives (d.h. empirisch gehaltvolles) Aussagesystem verstehen, dessen Explanandum räumlich indizierte Begriffe enthält, wobei die räumliche Indizierung des Explanandums zu einer solchen der Anfangs-, Randbedingungen und/oder der (quasi) nomologischen Hypothesen des Systems führen kann."

Im Jahre 1972 wird von ARNOLD[82] darauf hingewiesen: „The justifications for classification of cities can be elegant, rational, and logical (as they should be!), but they are also of practical necessity. They provide the framework for better understanding of urban phenomena that are susceptible to quantitative measurement and the checkpoints for comparison, compilation and evaluation of information. Mayor city managers and other chief administrators must take a greater personal interest in and responsibility for information; if they do not, it is inevitable that other governments will step in to do the job."

Nach LICHTENBERGER[83] beruhen Stadtklassifikationen auf statistischen Schwellenwerten, die keineswegs dauernde Gültigkeit beanspruchen können und mit der Veränderung der städtischen Gesellschaft und des arbeitsteiligen Prozesses einer Korrektur bedürfen.

Die Klassifikation ist mit Blick auf die in den vergangenen Jahren durchgeführten Ansätze sicherlich ein sehr wichtiges Element im ‚Urban Data Mining'. Durch induktive[84] Verallgemeinerungen über die Untersuchungsobjekte kann ein gemeinsamer Begriff (Semantik) gefunden werden, der die Benennung der gruppierten Phänomene ermöglicht, die sich in bestimmten Aspekten ähnlich sind bzw. nur unwesentlich unterscheiden. Das ‚Urban Data Mining' soll aber darüber hinaus eine nachträgliche Zuordnung von Objekten in bereits berechnete Klassensysteme vorsehen. Für gewählte Zeitabschnitte werden Informationen über bisher nicht zugeordnete Objekte schnell abschätzbar, da ein Vergleich mit den Eigenschaften der bekannten Klassenstruktur erfolgt und charakteristische Merkmale erkennbar werden. In Zukunft lassen sich bei verbesserter Datenlage auch Hypothesen über die gebaute Umwelt[85] formulieren und überprüfen.

Die als besonders relevant erachteten Methoden sind in dieser Arbeit sowohl theoretisch zu beschreiben als auch in einem empirischen Teil an ausgewählten Untersuchungsaufgaben anzuwenden. Im Vordergrund der Betrachtung stehen die Beschäftigungsschwerpunkte, die Verflechtungen zwischen Kernbereich und Umland und insbesondere die Schätzung des Gebäudebestandes. Dieser Forschungsansatz spannt einen Bogen von der Deskription regionaler Strukturen und Kennzahlen über das Problem der Beschaffung verwendbarer Daten und über die Theorie ihrer mehrdimensionalen Verwendung bis hin zur Anwendung des Data Mining und des Knowledge Discovery.

---

82 Vgl. DAVID S. ARNOLD in: BERRY [1972, Chapter 12, S.377]

83 Vgl. LICHTENBERGER [1998, S. 33]

84 Vgl. NARR [1971], der die empirisch analytische Theorie in die deduktiv-empirische Theorie und die empirisch generalisierende induktive Theorie unterteilt.

85 Der Begriff ‚gebaute Umwelt' dient als zusammenfassende Beschreibung des Gebäudebestandes und der Infrastruktur.

Die Forschungsfragen für die gewählten Untersuchungsaufgaben lauten wie folgt:

- Gibt es einen methodischen Untersuchungsansatz, um Beschäftigungsschwerpunkte von Gemeinden zu ermitteln bzw. den Grad der Diversität zu identifizieren?
- Wie kann eine Klassenbildung auf Basis von ausgewählten raumstrukturellen Messgrößen aufgebaut werden, die zur Erfassung und Strukturierung des Agglomerations- und Verdichtungsprozesses geeignet ist?
- Welche Methoden sind geeignet, um die Bipolarität von Schrumpfungs- oder Wachstumstendenzen abzubilden?
- Gesucht wird nach einer Zuordnungsvorschrift, welche sich eignet, eine Gemeinde in bereits bestehenden Klassenstrukturen mit Hilfe weiterer Variablen einzuordnen (Klassifikator)?
- Ist es möglich, räumliche Information zwischen Untersuchungsobjekten zu übertragen oder sogar eine gemeindescharfe Prognose zum Gesamtgebäudebestand zu stellen?

1.3 Aufbau der Arbeit

Die Arbeit gliedert sich in einen theoretischen Teil zur Methodik des ‚Urban Data Mining' (Kapitel 2) und einen allgemeinen Datenteil, der den Status quo der Datenerhebung und das Prinzip der Datenregularisierung definiert (Kapitel 3 und 4). Der empirisch-analytische Teil (Kapitel 5 bis 8) schildert die Anwendungsmöglichkeiten von Verfahren des ‚Urban Data Mining'.

Im Kapitel 2 wird das methodische Instrumentarium für das ‚Urban Data Mining'vorgestellt, welches dann später im empirisch-analytischen Teil zu großen Teilen nochmals in seiner Anwendung gezeigt wird.

Das Kapitel 3 dient dazu, die Möglichkeiten der Datenbeschaffung auf Gemeinde- und Kreisebene weitgehend transparent darzustellen und die Qualität von möglichen Datenquellen zu diskutieren.

Das Kapitel 4 schildert das in dieser Arbeit einheitlich verwendete Arbeitsprinzip der Datenaufbereitung (Datenregularisierung).

In den Kapiteln 5 bis 8 werden die Ergebnisse des ‚Urban Data Mining' vorgestellt. Zu Beginn des jeweiligen Kapitels wird die jeweilige Untersuchungsaufgabe definiert, für die dann auf Grundlage des Datenmaterials spezifische methodische Untersuchungsansätze vorgestellt werden. Die Ergebnisdarstellung endet jeweils mit einer Diskussion der methodischen Vorgehensweise und und schließt mit einem kurzen Fazit zu den gewonnenen Erkenntnissen.

Das Kapitel 9 enthält ein Resümee und gibt einen Überblick zu Bestandteilen dieser Arbeit.

Im Kapitel 10 erfolgt ein Ausblick auf die zukünftig zusätzlich erweiterbaren Möglichkeiten der Anwendung des ‚Urban Data Mining'. Es wird auf andere Stadt- und Raumstrukturtypenansätze verwiesen, um weitere Integrationsmöglichkeiten anzudeuten.

Im Nebenteil A (siehe CD-ROM) sind Kartendarstellungen gezeigt, die im Wesentlichen vor Beginn der Auswertung der Daten zu Kontrollzwecken aus der aufgebauten Datenbank erstellt wurden. Darüber hinaus sind Karten des Hauptteils hinterlegt, die sich auf einzelne Teilergebnisse beziehen.

Mit Hilfe des Nebenteils B (siehe CD-ROM) ist es möglich, die Ergebnisberechnung des empirischen Bearbeitungsteils in ergänzender Form zum Kapitel 5 nachzuvollziehen.

2 Methodik des ‚Urban Data Mining' (Theoretischer Teil)

Die methodische Vorgehensweise ist in das Gebiet der experimentellen System-[86] und Strukturanalyse[87] einzuordnen. Als Ziel wird die Erstellung von Modellen[88] oder die Einordnung eines betrachteten Phänomens verfolgt auf der Basis von Beobachtungen bzw. Messergebnissen. Die Klassifikation ist eines der wichtigsten Ziele des Data Mining und in diesem Zusammenhang werden die erstellbaren Modelle Klassifikatoren genannt, die eine Zuordnung unbekannter Objekte zu der richtigen Klasse ermöglichen.[89] Data Mining ist als ein zyklischer Prozess zu verstehen, so dass in jedem Schritt gewonnene Erkenntnisse immer wieder validiert und als Eingangsstufe eines nachfolgenden Schrittes verwendet werden (Abbildung 2-1). Die für diese Arbeit relevanten Verfahrensschritte des ‚Urban Data Mining' werden im Folgenden beschrieben, wobei die Gliederungsstruktur dieser Arbeit eine zukünftige methodische Erweiterung nicht ausschließt.[90] Über die Arbeitsschritte ist ein Protokoll zu führen, so dass dieses am Ende der Bearbeitung für jeden zugänglich ist und auf diese Weise die Bearbeitung nachvollziehbar ist.

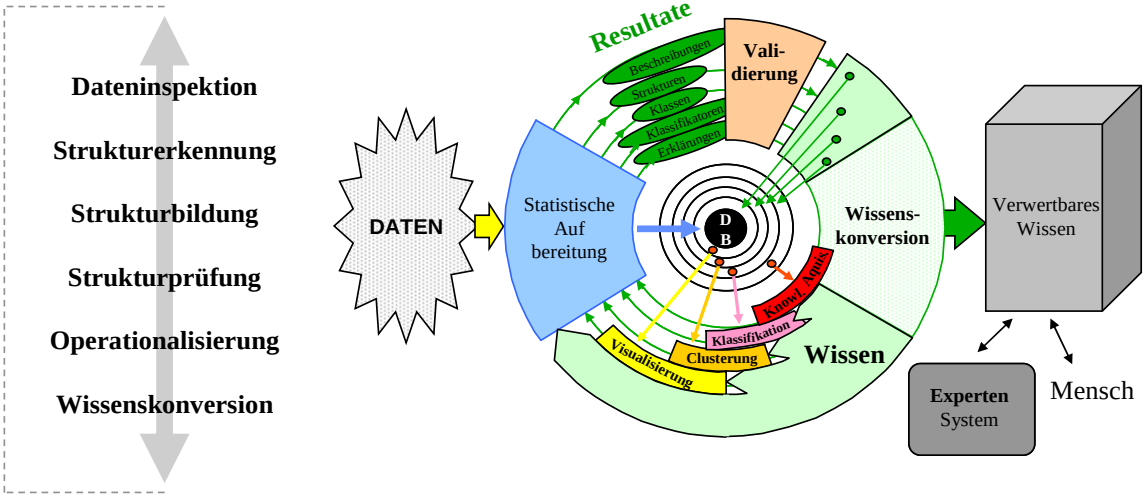


Abbildung 2-1: Zyklischer Prozess des Data Mining[91]

---

86 Vgl. BOSSEL [1987, S. 12]: „Ziel der Systemanalyse ist es […], ein Modell des realen Systems zu entwickeln, dessen Verhalten dem Verhalten des realen Systems sehr nahe kommt. Es ist klar, dass dieses Modell im Allgemeinen nicht so komplex wie die Realität selbst sein kann."

87 Vgl. BOSSEL [1987, S. 10]: „Ein System besteht aus einem oder mehreren strukturell verbundenen Elementen, deren Zustände von anderen Elementen (oder sich selbst) abhängen und die die Zustände anderer Elemente (oder sich selbst) beeinflussen. […] Ein System hat einen Zweck (bzw. ist es möglich, ihm einen Zweck zuzuschreiben). […] Ein System hat eine Systemgrenze, die es von seiner Systemumwelt trennt."

88 Vgl. SESTER [1995, S. 34]: „Modellierung kennzeichnet den Prozess, der implizit gegebenes Wissen in eine explizite Form überführt.", vgl. SOMMER et al. [1993]: Im Prozess der Modellierung ergibt sich oft erst die Art der zu erstellenden Modelle, d.h. die verwendeten Merkmale; vgl. REIMER [1991]: Unterschiedliche Repräsentationsstrukturen der Realität; Vgl. BOSSEL [1987, S. 12]: Systemmodelle

89 Vgl. GHOLAMREZA, N. [1998 S.8], BOCKLISCH [1987, S. 18]

90 Methodische Erweiterungen sind u.a. mit BAHRENBERG [2003] und BACKHAUS [2006] möglich. Eine Hauptkomponentenanalyse ist beispielsweise zusätzlich im Rahmen der Dateninspektion einsetzbar.

91 Quelle: ULTSCH [2006 c] und eigene Ergänzungen

2.1 Dateninspektion

2.1.1 Grundeigenschaften

2.1.1.1 Einzelbetrachtung

Die Dateninspektion beginnt mit der Durchsicht der Objektdaten jeder einzelnen Variablen, indem man sich einen Überblick von Anzahl, Art, Wertebereichen und Verteilungen verschafft. Es sind Vorkehrungen bez. der Datensicherheit zu treffen, damit Daten nicht unbeabsichtigt verändert werden oder sogar verloren gehen.

Gegeben sei $O = \{O_1, O_2, ..., O_n\}$ mit $n \in \mathrm{N}$ eine Menge von n Objekten, die mit Verfahren des Data Mining (z.B. der Clusteranalyse) untersucht werden sollen.[92] Bei Objekten, die mehrere Variablen $X_1, ..., X_m$, $m \in N$ aufweisen, lassen sich die Beobachtungswerte in Form einer Datenmatrix

$$X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{pmatrix} = \left(x_{ij}\right)_{\substack{1 \le i \le n \\ 1 \le j \le m}} \in \mathrm{R}^{(n,m)} \qquad (1)$$

zusammenfassen, wobei x_{ij} die Ausprägungen der Variable j beim Objekt i sein soll.

Der Zeilenvektor $x_i = (x_{i1} ... x_{im}) = \left(x_{ij}\right)_{1 \le j \le m} \in R^m$ wird durch das Objekt i beschrieben und Informationen über die Variable j sind in der j-ten Spalte von X enthalten.

Die Untersuchung der Grundeigenschaften der Datenmenge umfasst erste statistische Beschreibungen der Daten sowie eine Prüfung auf Datenvollständigkeit. Es ist möglich, eine Messgenauigkeit bzw. Rechengenauigkeit abzuleiten. Die Fehlstellenbereinigung ist durch Elimination von Daten oder Ergänzungsverfahren durchführbar.[93] Der Begriff Skala stellt die Messlatte dar, auf der die Ausprägungen von Variableneigenschaften abgetragen werden. Tabelle 2-1 gibt einen Überblick der verschiedenen Skalen und Messniveaus.

Skala		**Variable**	**Rechenweise**
Nicht metrisch	nominal	Klassifizierung qualitativer Eigenschaftsausprägungen	Bildung von Häufigkeiten
	ordinal	Rangwert mit Ordinalzahlen	Ermittlung des Median
metrisch	intervall	Skala mit gleichgroßen Abschnitten ohne natürlichen Nullpunkt	Addition, Subtraktion
	Ratio	Skala mit gleichgroßen Abschnitten und natürlichem Nullpunkt	Addition, Subtraktion, Division, Multiplikation

Tabelle 2-1: Skalen und Messniveaus[94]

92 Vgl. BOCK, H.H. [1974, S.25 ff.] oder BOCKLISCH [1987, S. 59 ff.]

93 Vgl. SCHAFER, J.L. [1997], LITTLE / RUBIN [1987]. und ULTSCH [2006 c] z.B. plausibelster Wert, Nearest Neighbour, Neuronales Netz, 5 Nearest Neighbour als Imputationsverfahren (Fehlstellenbehandlung)

94 Eigene Bearbeitung nach SCHUCHARD-FICHER et al. [1985, S. 5]

Ein erstes Bild von den Daten gewinnt man aufgrund der Beschreibung jeder einzelnen Variablen durch so genannte Lage- und Streuungsmaße.[95] Ein Lagemaß (z.B. Mittelwert, Minimum, Maximum, Median, Modus) charakterisiert die Variable bez. der generellen Lage der Variablen auf der Skala. Mit einem Streuungsmaß (z.B. Varianz und Standardabweichung, Spannweite, Quartilsabstand) wird bewertet, ob die Datenausprägungen über den Wertebereich verteilt bzw. konzentriert sind. Inwieweit ein gewähltes Lage- und Streuungsmaß überhaupt geeignet ist, um spezifische Eigenschaften abzubilden (Eingipfelige, U-förmige, Mehrgipfelige Verteilung)[96] sollte geprüft werden, da es sich um einen zusammenfassenden Eindruck von Wertebereichen einer Variablen handelt.[97] Als zusätzliche Maßzahlen zur Beurteilung eingipfliger Häufigkeitsverteilungen sind die Schiefe und der Exzess (Kurtosis, Wölbung) zu nennen.[98] Da die Verteilung der Variablen (z.B. Gleichverteilung, Normalverteilung, LogNormal Verteilung, Chi-Quadrat Verteilung, Multimodale Verteilungen) üblicherweise nicht vorab bekannt ist, besteht die Aufgabe darin, eine Hypothese über eine empirisch beobachtete Variable zu gewinnen. Hierfür geeignet ist die Visualisierung von Sachverhalten, die neben den genannten Maßzahlen eine Variablenbeschreibung ergänzen (z.B. Histogramme,[99] Box-Plots,[100] Quantil/Quantil- bzw. QQ-Plots,[101] Pareto Density Estimation, siehe Abschnitt 2.1.1.2).[102] Die QQ-Plots vergleichen zwei Verteilungen bildhaft miteinander, in dem die Quantile der beiden Verteilungen in einem Koordinatensystem gegeneinander aufgetragen werden. Wenn die auf diese Weise entstandenen Punkte annähernd eine Gerade bilden, ist davon auszugehen, dass die beiden Verteilungen gleich sind. Es ist zu beachten, dass bei multimodalen Verteilungen der QQ-Plot in der Regel mehrfach gekrümmt ist. In dieser Arbeit sind die Quantile der Untersuchungsvariablen auf der Y-Achse aufgetragen und das Ablesen des Wertebereichs ist hier möglich.

---

95 Vgl. HARTUNG [2005, S. 31 ff.] und HARTUNG [2005, S. 40 ff.]

96 Vgl. ULTSCH [2006 c]: „Verteilungen sind Modelle für die Wahrscheinlichkeitsdichte der Daten. Dies bedeutet, dass zu jeder Ausprägung ein Wert angegeben werden kann, welcher besagt, wie wahrscheinlich es ist an dieser Stelle einen Datenpunkt zu finden."

97 Im Gegensatz zu so genannten eingipfligen oder unimodalen Verteilungen sind bei mehrgipfligen und insbesondere bei U-förmigen Häufigkeitsverteilungen, die Lagemaße oft nicht charakteristisch bzw. sogar irreführend für die Häufigkeitsverteilung (siehe HARTUNG [2000, S.37ff.]). Bei schiefen eingipfligen Verteilungen ist der Mittelwert oder die Varianz wenig aussagekräftig, und es empfiehlt sich der Median.

98 Vgl. HARTUNG [2005, S. 47 ff.]

99 Vgl. HARTUNG [2005, S.22]: Der Wertebereich einer Variablen wird in Intervalle (Bins) eingeteilt. Die Anzahl der Daten pro Bin bzw. die Häufigkeit des Auftretens in den Bins wird durch Werte ermittelt, die als aneinander stoßende Rechtecke dargestellt werden, deren Flächeninhalt proportional zur Anzahl der Daten in den jeweiligen Bins ist. Vgl. SCOTT/KEATING [1999, pp. 16-22], für die Wahl optimaler Bin-Weiten.

100 Vgl. HARTUNG [2005, S. 835 ff.], Boxplot, Kastenzeichnungen oder auch Whiskers-Plot

101 Es werden zum Vergleich bekannte Verteilungen (z.B. Standardnormalverteilung) mit einbezogen.

102 Vgl. ULTSCH [2001] und [2003 a]: Es handelt sich um die empirische Schätzung der Dichte von Daten anhand der informationsoptimalen Menge – Pareto Density Estimation (PDE).

ULTSCH[103] verweist darauf, dass die meisten statistischen Verfahren die behandelten Variablen als normal verteilt bzw. einen ähnlichen Verteilungsverlauf voraussetzen, so dass die Variablenvorbehandlung eine besonders wichtige Rolle einnimmt. Sind Abweichungen von Standardnormalverteilungen festzustellen, ist darüber zu entscheiden, mit welchen Umformungen (Transformationen) die Daten in eine bekannte Verteilung transformiert werden können.[104] Die Zielsetzung von Datentransformationen besteht somit darin, dass die transformierten Daten Voraussetzungen für bestimmte Methoden erfüllen.[105]

In der explorativen Datenanalyse werden nichtlineare Transformationen $y = x^p$ auf die Daten angewendet, wobei eine Auflistung in der so genannten ‚ladder of power' (Tabelle 2-2) erfolgt, die die Größe (power) des Exponenten p beschreibt. Eine nichtlineare Transformation verfolgt den Gedanken, eine Verteilung so zu transformieren, dass die transformierte Verteilung einer theoretischen entspricht. Wird beispielsweise mit den zuvor gezeigten Methoden festgestellt, dass eine Datenreihe eine rechtsschiefe oder linksschiefe Verteilung besitzt, so kann diese durch diesen Transformationsansatz symmetrisiert werden.

p	**…**	**-2**	**-1**	**0**	**½**	**1**	**2**	**3**	**…**
Transformation	…	x^{-2}	$x^{-1} = \frac{1}{x}$	$\ln(x)$	$x^{\frac{1}{2}} = \sqrt{x}$	x	x^2	x^3	…

Tabelle 2-2: Exponentenleiter (ladder of power)[106]

Eine Überprüfung der Gültigkeit der aus den Daten gezogenen Schlussfolgerungen und Erkenntnisse ist in Form von statistischen Tests[107] mit zu berücksichtigen, da insbesondere die darauf folgenden Verfahrensschritte des Data Mining (z.B. Strukturerkennung, Strukturbildung) von diesen Entscheidungen abhängen. Eine Untersuchung möglicher Ausreißer[108] ist einzubeziehen, wobei Schritte zur Behandlung von Ausreißern wie das Weglassen des Datensatzes oder eine Begrenzung des Wertebereichs darzustellen sind. Die Plausibilität auch aus der Datenquelle selbst ist zu hinterfragen, da z.B. bei flächenhaften oder volumenhaften Verteilungen die Transformationen wie Wurzel oder dritte Wurzel wahrscheinlich oder Wachstumsvorgänge mit Hilfe einer Logarithmierung modellierbar sind.

---

103 Vgl. ULTSCH [2006 c], siehe zusätzlich auch bei HARTUNG [2005, S. 832 ff.] oder ERB [1990, S. 57 ff.]

104 Vgl. ULTSCH [2006 c]: „Zeigt der QQ-Plot eindeutig einen konvexen oder konkaven Bogen, so deutet dies auf eine nicht normale, ‚schiefe' Verteilung hin."

105 Vgl. Hartung [2005, S. 833]: „Bei der explorativen Datenanalyse dienen Datentransformationen (re-expressions) dazu, das Datenmaterial übersichtlicher zu gestalten. Treten zum Beispiel sehr viele kleine Werte x_i und einige wenige sehr große Werte x_i auf, so ist eine übersichtliche Darstellung auch bei Klassenbildung kaum zu erreichen, wenn man gleiche Klassenbreiten wählen möchte."

106 Eigene Bearbeitung nach HARTUNG [2005, S. 833]

107 Vgl. HARTUNG [2005, S.133 ff.], siehe z.B. Anpassungstest oder KOLMOGOROFF-SMIRNOV-Test

108 Vgl. Hartung [2005, S. 343 ff.], siehe DAVID-HEARTLEY-PEARSON-Test, DIXON-Test, GRUBBS-Test und mit Hilfe von Box-Plots kann ein erster Hinweis auf Ausreißer gefunden werden.

2.1.1.2 Pareto Density Estimation

Die Darstellung der Wahrscheinlichkeitsdichte bietet in Ergänzung zu der Betrachtung von Daten durch ein Histogramm eine weitere Möglichkeit, die Verteilung von Beobachtungswerten zu charakterisieren. Die Wahrscheinlichkeitsdichte[109] ermöglicht die Beschreibung von Wahrscheinlichkeitsverteilungen und wird als Dichtefunktion $f(x)$ bezeichnet. Die Wahrscheinlichkeit P wird definiert als Integral mit den Integrationsgrenzen a und b:

$$\int_a^b f(x)dx = P(a \le X \le b) \tag{2}$$

In dieser Arbeit wird die so genannte Pareto-Dichte-Schätzung[110] (Pareto Density Estimation, PDE) eingesetzt. Es ist ein Verfahren, welches eine stetige Schätzung einer unbekannten Verteilung von Beobachtungswerten ermöglicht. Für eindimensionale Daten kann die Schätzung der Wahrscheinlichkeitsdichte anhand von PDE wie folgt berechnet werden:

$$PPDE(x) = \frac{NN(x, r_p)}{Fläche} \qquad \text{wobei die Fläche definiert ist als } \int_{-\infty}^{\infty} NN(x, r_p)dx \tag{3}$$

Definiert ist $NN(x, r)$ als Nachbarschaftszahl. Die Nachbarschaftszahl stellt in einem Datenraum eine Menge von Datenpunkten dar, die kugelförmig mit einem Radius r um einen Punkt x angeordnet sind. Um die optimale Datensatzgröße zu ermitteln, die den größten Informationsgehalt trägt, wird der Radius Pareto-Radius r_p bestimmt. Dieser Radius r_p wird in Abhängigkeit des Medians von $NN(x, r)$ bestimmt und soll 20,13 % aller Datenpunkte in einem Datensatz umfassen. Die Fläche wird näherungsweise ermittelt durch Dichteschätzer, Trapezoid: $(x_i, NN(x_i, r_p))$. Die Formel setzt voraus, dass das Integral über der PPDE-Darstellung (x) den Wert 1 annimmt, um eine gültige Funktion der Wahrscheinlichkeitsdichte zu erhalten. Trägt man $PPDE(x)$ gegen x auf, so handelt es sich hierbei um die PDE-Darstellung, die eine genauere Sicht auf die Verteilung ermöglicht.

---

[109] Vgl. SCOTT [1992]

[110] Vgl. ULTSCH [2001] und [2003 a]: Optimierung des Informationsgehaltes
Sei S eine Teilmenge von n Punkten und $|S| = s$ die Anzahl der Elemente in S. Dann ist $p = s/n$ die relative Größe dieser Menge. Wenn ein Punkt x mit gleicher Wahrscheinlichkeit betrachtet wird, dann entspricht p der Wahrscheinlichkeit $p = P(x \text{ in } s)$. Entsprechend der Informationstheorie berechnet sich die Entropie oder Teilinformation mit Hilfe von p. Mit einer Skalierung von 0 bis 1 berechnet sich die Information eines Datensatzes: $I(S) = -e\, p \ln(p)$. Um eine optimale Datensatzgröße zu finden, wird das nicht erreichte Potential $\mathrm{URP(S)}$ eines Datensatzes als Euklidischer Abstand vom Idealpunkt definiert. Ausgehend vom Idealpunkt gilt: $URP(S) = \sqrt{p^2 + (1 + e \cdot p \cdot \ln(p))^2}$. Die Ergebnisse für $\mathrm{URP(S)}$ werden auf die optimale Datensatzgröße von $p_u = 20{,}13\%$ reduziert. Diese Datensatzgröße liefert 88 % der maximal möglichen Information.

Abbildung 2-2 gibt ein Beispiel für die Anwendung einer PDE, um sich einen Überblick über die Verteilung und ihre Zusammensetzung (Struktur) zu verschaffen. Auf der x-Achse ist jeweils die Ausprägung der Daten aufgetragen. Dargestellt sind 4 transformierte Variablen (Log(Data)): ‚LogBevoelkerung', ‚LogBeschaeftigte',LogBevoelkerungUndBeschaeftigte' und LogGebaeudeUndFreiflaeche' (siehe ergänzend Abschnitt 4). Da es sich um transformierte Daten handelt, kann durch Rückrechnung der nichttransformierte Wert beurteilt werden. Auf der Y-Achse ist ein Wahrscheinlichkeitswert ablesbar, der eine Aussage über die Menge der Untersuchungsobjekte erlaubt, die eine Ausprägung an der Stelle X aufweisen.

Bei dieser Abbildung handelt es sich um Untersuchungen mit einer sogenannten PDE-Mischung, d.h. die Verteilung wird zusätzlich nach gewählten Klassen gesondert unterteilt. In diesem Beispiel beziehen sich die Klassen auf die im urbanen Kontext gültigen Zentrale-Orte-Kategorien in Deutschland. Es handelt sich um sechs Kategorien, die zusätzlich eine Aussage über die Zusammensetzung der Verteilung erlauben. Die Abbildungen zeigen für alle vier Variablen, dass die Verteilung sich nach diesen gegebenen sechs Klassen (‚Oberzentrum', ‚Mittelzentrum' usw.) zusätzlich strukturieren lässt bzw. dass jede Zentrale-Orte-Kategorie eigene Wertebereiche und damit charakteristische Ausprägungen besitzt. Die Variable ‚LogBevoelkerung' zeigt beispielsweise, dass die meisten Oberzentren wesentlich größere Einwohnerzahlen besitzen als die Kleinzentren. Die Kleinzentren zeigen ein Maximum der PDE bei einem x-Wert von ungefähr 7. Dies entspricht ca. 1000 Einwohnern.

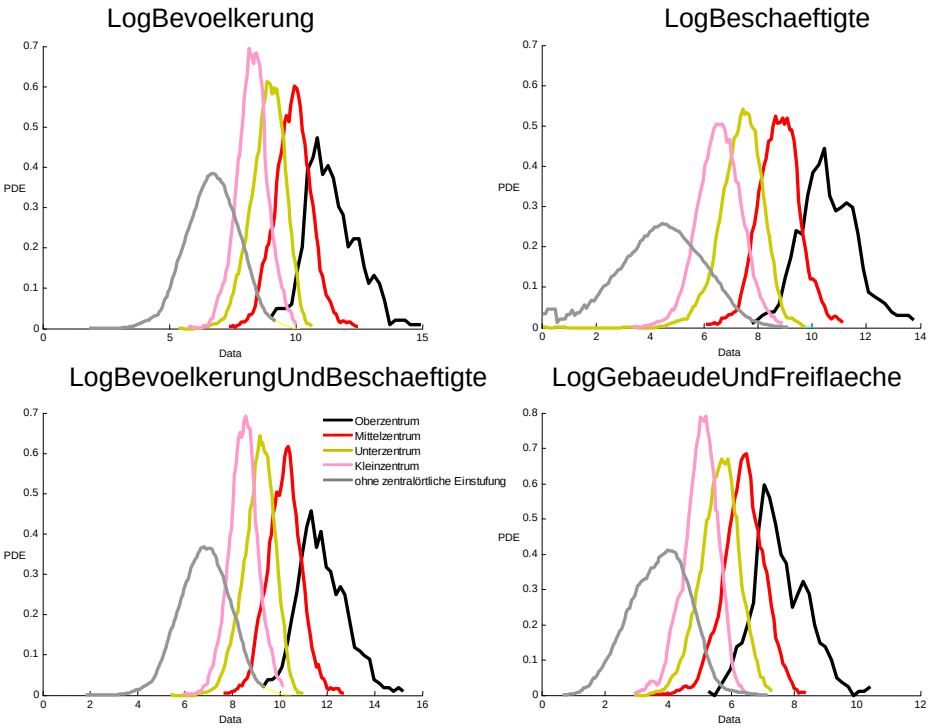


Abbildung 2-2: Beispiel für eine Variablenuntersuchung mit PDE-Mischungen

2.1.1.3 Messung der Konzentration

Die Messung der Konzentration gibt eine Vorstellung über die Gleich- bzw. Ungleichverteilung von Merkmalsausprägungen. Zur graphischen Veranschaulichung wird die so genannte Lorenzkurve[111] eingesetzt. Gemessen wird die Konzentration bei einem Merkmal mit nicht negativen Ausprägungen und gefragt wird, welcher Anteil an der Merkmalssumme $\sum_{j=1}^{n} x_j$ .bzw. $\sum_{i=1}^{k} a_i n_i$ auf einen vorgegebenen Anteil $\alpha\,(0 \le \alpha \le 1)$ der Merkmalsträger mit den kleinsten Merkmalsausprägungen entfällt. Die empirische Lorenzkurve ist an den Stellen $u_i = \frac{i}{n}, 0 \le i \le n$ wie folgt gegeben:

$$L_n\left(\tfrac{i}{n}\right) = \begin{cases} 0 & ,i = 0 \\ \dfrac{\sum_{j=1}^{i} x(j)}{\sum_{j=1}^{n} x(j)} = \dfrac{\frac{1}{n}\sum_{j=1}^{i} x(j)}{\bar{x}} & ,i = 1,...,n\,. \end{cases} \tag{4}$$

Linear interpoliert wird zwischen den Punkten $u_i = i/n \quad i = 0,...,n$. Handelt es sich um gruppierte Daten, so definiert man entsprechend für

$$u_0 = 0, u_i = \sum_{j=1}^{i} r_j,\ 1 \le i \le k \text{ und } a_1 < a_2 < ... < a_k$$

$$L_n(u_i) = \begin{cases} 0 & ,i = 0 \\ \dfrac{\frac{1}{n}\sum_{j}^{i} a_j n_j}{\bar{x}} & ,i = 1,...,k\,. \end{cases} \tag{5}$$

und mit a_i werden die Klassenmitten bezeichnet. Zwischen den Punkten $u_i, L_n(u_i)$ wird linear interpoliert. Der empirischen Lorenzkurve lassen sich folgende Eigenschaften zuweisen:

a) $L_n(0) = 0, L_n(1) = 1$ und L_n ist konvex im Intervall $[0,1]$;

b) $L_n(x) \le x$ und für $x \in (0,1)$ gilt $L_n(x) = x$, wenn $x_1 = x_2 = ... = x_n$ (keine Konzentration);

c) Für gruppierte Daten liegt die Lorenzkurve stets oberhalb (höchstens auf) der Lorenzkurve für die ungruppierten Daten.

Auf der Grundlage von Lorenzkurven lässt sich insbesondere eine optimale Situation finden, die als konsistent mit der empirisch beobachteten PARETO-80/20 Regel angesehen werden kann.[112] In Bezug auf die Merkmalsträger und die Merkmale lässt sich auf diese Weise ein Wert bestimmen, bei dem 20 % der Merkmalsträger bereits 80 % der Information enthalten.

---

[111] Eingeführt wurde die Lorenz-Kurve im Jahr 1905 von Max Otto Lorenz zur grafischen Darstellung von statistischen Verteilungen und der Veranschaulichung des Ausmaßes an Konzentration bzw. Ungleichheit.

[112] Vgl. ULTSCH [2001]: Mit Hilfe des unrealisierten Potentials und des entropischen Nutzens kann die Frage geklärt werden, warum sich viele Projekte in Teilprojekte mit ca. 20 % Aufwand und ca. 80 % Nutzen aufteilen lassen. (Eine Begründung der Pareto-80/20 Regel – Optimierung des Informationsgehaltes).

2.1.1.4 Korrelationsanalyse

Ist die Beschreibung einzelner Variablen erfolgt, beginnt nachfolgend die Untersuchung der Datensätze auf Zusammenhänge bzw. Abhängigkeiten zwischen zwei oder mehreren Variablen. Es wird überprüft, ob redundante Information in den Datensätzen existiert und Hinweise auf die Struktur des Datensatzes bzw. des Grundproblems zu erkennen sind. Hierzu werden einerseits visuelle Methoden wie Streu-Diagramme[113] (Scatter-Plots) und andererseits statistische Maßzahlen wie Korrelationsmaße[114] eingesetzt, die den Grad der Abhängigkeit verschiedener Variablen im Bereich $[-1;+1]$ messen. Die Streudiagramme eignen sich dazu, Zusammenhänge abzulesen, wie z.B. Gruppenbildung, Abhängigkeiten untereinander und vermutliche Artefakte[115] in den Daten. Von besonderer Bedeutung ist der lineare Zusammenhang zwischen zwei Variablen X und Y, d.h. es gilt:

$$Y = a \cdot X + b \qquad \text{mit zwei reellen Konstanten } a \text{ und } b\,. \tag{6}$$

Der Korrelationskoeffizient von Pearson wird als Maßzahl wie folgt definiert:[116]

$$pc(X,Y) = \frac{1}{\sqrt{\sum_{i=1}^{d}(x_i-\bar{x})^2 \sum_{i=1}^{d}(y_i-\bar{y})^2}} \cdot \sum_{i=1}^{d}(x_i-\bar{x})(y_i-\bar{y}) \tag{7}$$

Man spricht von positiver Korrelation, wenn die Werte der Variablen Y mit den Werten der Variablen X wachsen oder beide fallen. Eine negative Korrelation ist dadurch gekennzeichnet, dass die Werte von Y fallen und von X wachsen oder umgekehrt. Der Korrelationskoeffizient liegt nahe bei 0, wenn kein linearer Zusammenhang erkennbar ist. Um die Möglichkeit der Entdeckung nichtlinearer Korrelationen zu verfolgen, sind die transformierten und die originalen Variablen zu untersuchen. Falls keine Normalverteilung der Variablen vorauszusetzen ist, wird auf Rangkorrelationskoeffizienten[117] zurückgegriffen.

---

[113] Vgl. FAHRMEIER et al. [2004, S. 128], Üblicherweise werden in einem x-y Koordinatensystem zwei Variablen gegeneinander aufgetragen, indem jeder Variablen eine Achse zugewiesen wird. Die Skalierung der Achsen erfolgt so, dass Minimal- und Maximalwerte der Variablen ein Quadrat bilden.

[114] Vgl. HARTUNG [2005, S. 72 ff.] und FAHRMEIER et al. [2004, S. 134 ff.], Pearsonscher Korrelationskoeffizient, Spearmans Rangkorrelationskoeffizient, Kendalls Rangkorrelationskoeffizient.

[115] In der Wissenschaft wird der Begriff dazu verwendet, beobachtete Phänomene zu charakterisieren, die auf Fehler im Aufbau eines Experiments zurückzuführen sind. Diese Beobachtungen sind dann wissenschaftlich wertlos, weil sie nichts über den eigentlichen Untersuchungsgegenstand aussagen.

[116] Vgl. ULTSCH [2006 c]: Es wird die Gültigkeit des Mittelwertes und der Varianz vorausgesetzt, und in der Regel erfordert dies eine Normalverteilung der Variablen.

[117] Vgl. HARTUNG [2005, S. 79-83]: SPEARMANS Rangkorrelationskoeffizient und KENDALLS Rangkorrelationskoeffizient. Es handelt sich um nicht parametrische statistische Verfahren, welche auf der Vergabe von Rängen $R(x_i)$ beruhen, wobei der kleinsten Beobachtung einer Variablenausprägung der Wert 1, der zweitkleinsten der Wert 2 zugewiesen wird, bis hin zur größten Beobachtung, die den Rang n erhält (n = Anzahl der Beobachtungen). Nach ULTSCH [2006 c] sollte im Allgemeinen bei metrischen Daten nicht auf einen Rangkorrelationskoeffizienten zurückgegriffen werden, da mit der Rangvergabe auch ein beachtlicher Informationsverlust verbunden ist. In dieser Arbeit werden primär metrische Daten verwendet.

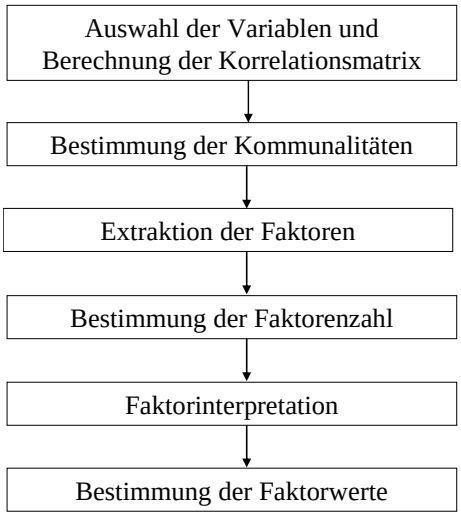
Nach VOGEL[118] sind Korrelationen innerhalb der Merkmalsstruktur unvermeidbar, doch kann unterschiedlich reagiert werden. Durch Eliminierung ausgewählter ähnlicher Variablen ist in einigen Fällen der Verlust wertvoller Informationen vorhanden.[119] Mit Hilfe einer Faktorenanalyse (siehe Abschnitt 2.1.1.5) oder einer Hauptkomponentenanalyse vor der Klassifikation kann eine vorhandene, untereinander korrelierende Merkmalsstruktur durch Bildung fiktiver unkorrelierender Merkmale ersetzt werden. Die an eine Klassifikation sich anschließende Klassendiagnose ist jedoch durch Interpretationsschwierigkeiten gekennzeichnet, da hierzu oftmals wieder die nicht fiktiven Merkmale verwendet werden müssen.[120] Nach VOGEL[121] ist der Einsatz von Korrelationskoeffizienten als Ähnlichkeitsmaß ebenso problematisch, um auf vorhandene korrelierende Daten zu reagieren. BECHER[122] äußert sich kritisch zur Aufdeckung redundanter Informationsträger durch Korrelationsberechnungen: „Zusammenhänge zwischen den Variablen anhand von Korrelationen zu erkennen, ist [...] methodisch unbefriedigend, weil hierbei offensichtlich versucht wird, mangelnde theoretische Vorüberlegungen bei der Zusammenstellung der Variablen durch die Anwendung formaler Methoden zu kompensieren.“ Die Vorgehensweise von BECHER berücksichtigt ein Gewichtungsschema, dem eine strukturierte Datenanalyse vorausging. BECHER führt zur Begründung des gewählten Vorgehens zusätzlich an, dass neben der Festlegung von Gewichtungsfaktoren weitere subjektive Manipulationsmöglichkeiten bestehen, z.B. die Wahl eines Clusteranalyseverfahrens und der damit verbundenen Parameter. FISCHER[123] sagt: „Hat man eine gewisse Anzahl relevanter und sinnvoller Attribute ausgewählt, deren Entdeckung und Formulierung sicherlich wissenschaftliche Kreativität erfordert, so muss man sich entscheiden, ob und gegebenenfalls wie man die einzelnen Attribute gewichtet (externes Gewichtungsproblem).“ SITTERBERG[124] stellt fest: „In der Regel wird es logisch schlüssige Gewichtungsmethoden nicht geben. Für eine Gleichgewichtung der Merkmale spricht unter diesen Umständen, dass eine Entscheidung über die Gewichtung in jedem Fall nötig ist. Die Zuordnung gleicher Gewichte zu den Merkmalen kann dann das ‚geringste Übel’ bedeuten.“ VOGEL[125] diskutiert die Gleichgewichtung der Variablen und analysiert das von vielen ‚Taxonomen’ vielfach unantastbare Prinzip der Gleichgewichtung.

---

118 Vgl. VOGEL [1975, S.52, S. 54 und S. 59 ff]
119 Vgl. GOWER [1969]
120 Vgl. VOGEL [1975, S. 67]
121 Vgl. VOGEL [1975, S. 92]
122 Vgl. BECHER [1995, S. 56]
123 Vgl. FISCHER, M. [1982, S.54]
124 Vgl. SITTERBERG [1977, S. 24]
125 Vgl. VOGEL [1975, S. 57 und 68 ff.]

2.1.1.5 Faktorenanalyse

Die Faktorenanalyse (FA)[126] eignet sich dazu, das Korrelationsmuster innerhalb einer größeren Anzahl von Variablen (messbare Beobachtungsgrößen) durch Bildung von Faktoren (nicht messbare Einflussgrößen) zu erklären. Es werden dabei diejenigen Variablen zu einem Faktor zusammengefasst, die untereinander starke Korrelationen aufweisen. Die Faktorenanalyse stellt die Frage, welches die einfachste Struktur ist, die die vorliegenden Daten genügend genau reproduziert und erklärt. SCHUCHARD-FICHER[127] sagt: „Je größer die Zahl der notwendigen Erklärungsvariablen wird, um so weniger ist gesichert, dass diese auch alle tatsächlich unabhängig voneinander zur Erklärung des Sachverhaltes notwendig sind." Die Teilschritte zur Durchführung einer Faktorenanalyse sind in Abbildung 2-3 dargestellt und werden ergänzt durch das Fundamentaltheorem der Faktorenanalyse.



Fundamentaltheorem der Faktorenanalyse (THURSTONE)

„Jeder Beobachtungswert einer Ausgangsvariablen x_j oder der standardisierten Variablen z_j lässt sich als eine Linearkombination mehrerer (hypothetischer) Faktoren beschreiben." (Vgl. BACKHAUS et al., 2006)

Mathematisch formuliert:

$$x_{kj} = a_{j1} \cdot p_{k1} + a_{j2} \cdot p_{k2} + ... + a_{jQ} \cdot p_{kQ}$$

bzw. für standardisierte *x*-Werte:

$$z_{kj} = a_{j1} \cdot p_{k1} + a_{j2} \cdot p_{k2} + ... + a_{jQ} \cdot p_{kQ} = \sum_{q=1}^{Q} a_{jq} \cdot p_{kq}$$

in Matrixschreibweise überführt:

$$Z = P \cdot A'$$

Die Korrelationsmatrix R wird bei standardisierten Daten aus der Datenmatrix Z ermittelt:

$$R = \frac{1}{K-1} \cdot Z' \cdot Z \quad \Rightarrow \quad R = \frac{1}{K-1} \cdot (P \cdot A')' \cdot (P \cdot A')$$

Nach den Regeln der Matrixmultiplikation gilt:

$$R_h = \frac{1}{K-1} \cdot A \cdot P' \cdot P \cdot A' = A \cdot \underbrace{\frac{1}{K-1} \cdot P \cdot P'}_{=C} \cdot A'$$

Der Faktor C bezeichnet die Korrelationsmatrix der Faktoren, wobei C einer Einheitsmatrix entspricht: $\Rightarrow R_h = A \cdot A'$

Die Korrelationsmatrix lässt sich durch die Faktorladungen (Matrix A) und die Korrelationen zwischen den Faktoren (Matrix C) reproduzieren.

Abbildung 2-3: Ablaufschema und Fundamentaltheorem der Faktorenanalyse[128]

Nach BACKHAUS[129] ist das Fundamentaltheorem durch eine nicht erklärte Komponente U (Restvarianz) zu ergänzen, die potentielle Messfehler beschreibt, so dass folgendes gilt:

$$R_h = A \cdot A' + U \tag{8}$$

Den Ausgangspunkt für die Faktorenanalyse bildet wie gezeigt die Datenmatrix bzw. die standardisierte Datenmatrix Z, aus der eine Korrelationsmatrix $R = A \cdot A'$ ermittelt wird. Auf der Hauptdiagonalen der reduzierten Korrelationsmatrix $R_h = A \cdot C \cdot A'$ stehen die reduzierten Varianzen der Matrix R. Die reduzierten Varianzen sind in der Regel kleiner als Eins, da die gemeinsamen Faktoren nicht die Gesamtvarianz erklären (Kommunalitätenproblem).

---

126 Die FA wurde in den für diese Arbeit exemplarischen Disziplinen von z.B. BERRY, B.J.L. [1972], BERRY, B.J.L.; KASARDA, J.O. [1977]; MÖLLERS [1977] und DEITERS [1978] eingesetzt.

127 Vgl. SCHUCHARD-FICHER et al. [1985, S. 215 und S. 221 ff.]

128 Eigene Bearbeitung nach BACKHAUS et al. [2006, S. 331] und eigene Erweiterungen.

129 Vgl. BACKHAUS et al. [2006, S.290]

Der Begriff der Kommunalität[130] einer Variablen gibt an, wie hoch der dazugehörige Varianzanteil ist, der durch k gemeinsame Faktoren nach folgender Gleichung gelöst wird:

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + ... + a_{jk}^2 \quad (9)$$

Die Kommunalität wird definiert als die Summe der Quadrate der Ladungen der gemeinsamen Faktoren einer Variablen. Der unerklärte Teil der Varianz einer Variablen j berechnet sich aus $1 - h_j^2$. Die Bestimmung der Kommunalitäten ist ein Problem, da die Entscheidung über ihre Größe vor der Faktorenextraktion zu erfolgen hat und nur geschätzt werden kann. Die Extraktion der Faktoren basiert auf der Korrelationsmatrix R_h, wobei verschiedene Verfahren[131] eingesetzt werden können. Die Matrix A enthält die gesuchten Faktoren und wird Faktorenmuster genannt. Die darin enthaltenen Faktorladungen sind eine Maßgröße für den Zusammenhang zwischen Variablen und Faktor (Korrelationskoeffizient zwischen Faktor und Variablen).

Zur Bestimmung der Faktorenzahl gibt es keine eindeutigen Vorgaben, so dass der Anwender eine subjektive Entscheidung zu treffen hat, die mit Hilfe des KAISER-Kriteriums[132] oder des SCREE-Tests[133] unterstützt werden kann. Die aufgrund eines mathematischen Vorgangs berechneten Faktoren werden mit Hilfe der Variablen mit hohen Ladungen auf einen Faktor interpretiert. Die Interpretation der Faktorladungsstrukturen erfordert eine große Sachkenntnis des Anwenders bez. des konkreten Untersuchungsobjektes und bei zunehmender Variablenanzahl erschwert sich die eindeutige Zuweisung von Variablen. Durch Rotation[134] wird in einigen Fällen die Interpretation erleichtert.

Zuletzt erfolgt die Bestimmung der Faktorenwerte zur vertiefenden Faktoreninterpretation.[135] Die Untersuchungsobjekte lassen sich im Faktorenraum darstellen. Ein negativer Faktorenwert bedeutet, dass das Objekt in Bezug auf den Faktor im Vergleich zu den anderen Objekten unterdurchschnittlich bewertet wird. Ein Wert um 0 bedeutet, dass das Objekt durchschnittlich ist, und ein positiver Wert bedeutet eine überdurchschnittliche Bewertung.

---

130 Vgl. ÜBERLA [1971, S. 155]

131 Vgl. JANSSEN LAATZ [2005, S.507]: Hauptachsen-Faktorenanalyse, Hauptkomponenten-Analyse u.a.

132 Vgl. BAHRENBERG [2003, S.224]: Dem KAISER-KRITERIUM nach, ist die Zahl der Faktoren gleich der Zahl der Faktoren mit einem Eigenwert > 1. Die Eigenwerte entsprechen der Summe der quadrierten Faktorenladungen eines Faktors über alle Variablen und sind eine Beurteilungsgröße für die erklärte Varianz der Variablen des jeweiligen Faktors.

133 BACKHAUS et al. [2006, S.315]: Beim SCREE-TEST werden die Eigenwerte in einem Koordinatensystem nach abnehmender Wertefolge angeordnet und durch die Bestimmung einer eindeutigen Knickstelle die Faktorenanzahl gefunden.

134 Vgl. KAISER [1958], ÜBERLA [1971, S. 167 ff.]: Rotationsmöglichkeit, rechtwinklige Varimax-Rotation

135 Vgl. BACKHAUS et al. [2006, S.302,303], Mathematische Herleitung zur Bestimmung der Faktorwerte

2.1.2 Vergleichbarkeit

Um mehrdimensionale Objekte miteinander vergleichen zu können, ist es notwendig, ein quantifizierbares Maß zu verwenden, welches die Prinzipien zur Bestimmung der Gleichheit, Ähnlichkeit bzw. Verschiedenheit berücksichtigt. Die Vergleichbarkeit von Datensätzen erfordert zuvor die Klärung der Behandlung von Ausreißern und die Fehlstellenbereinigung. Die Auswahl der letztendlich zu berücksichtigenden Variablen ist ein fundamentaler Schritt des Data Mining und unter Beachtung der Variablenkorrelation wird über redundante Informationen entschieden. Das Wissen über unterschiedliche Variablenverteilungen und Variablenskalierungen ist mit einzubeziehen. Berücksichtigt man an dieser Stelle die Grundeigenschaften der Daten nicht, so hat dies einen entscheidenden Einfluss auf die nachfolgenden Verfahren. In den meisten Fällen liegen die Variablen nicht in gleicher Dimension vor, so dass z.B. bei den quadrierten Differenzen ein Problem auftreten würde. Deshalb sind die Variablen geeignet zu skalieren, z.B. durch Normierung, Standardisierung oder Lineare Transformation. DEIMER[136] sagt: „Im allgemeinen ist die Standardisierung immer dann sinnvoll, wenn zur Abstandsmessung metrischer Variablen die euklidische Metrik herangezogen wird, um eine isomorphe Abbildung des Variablenraumes in den euklidischen Raum zu ermöglichen.“ Tabelle 2-3 zeigt die Möglichkeiten der Skalierung.

Skalierungsmöglichkeiten für Variablen	
Normierung, d.h. die Ausprägungen der Variablen werden in das Intervall [0,1] transformiert.	
$x_{ik}^t = \frac{x_{ik} - x_{k\min}}{x_{k\max} - x_{k\min}}$	mit: x_{ik}^t = skalierte Ausprägung der Variable k bei Objekt i, $x_{k\min}$ = minimale Ausprägung der Variable k, $x_{k\max}$ = maximale Ausprägung der Variable k.
Standardisierung, d.h. die Variablen werden so transformiert, dass die entstehenden Ausprägungen den Mittelwert Null und die Standardabweichung Eins haben (z-Transformation).	
$x_{ik}^t = \frac{x_{ik} - \bar{x}_k}{s_k}$	mit: x_{ik} = Ausprägung der Variablen k bei Objekt i, n = Anzahl der Objekte $\bar{x}_k = \bar{x}_k = \frac{1}{n}\sum x_{ik}$ (Mittelwert), $s_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2}$ (Standardabweichung)
Lineare Transformation: Abbildung des gesamten Wertebereichs $[x_{k\min}, x_{k\max}]$ auf den Bereich $[x_{ku}, x_{ko}]$.	
$x_{ik}^t = \frac{(x_{ik} - offset)}{quotient}$	mit: x_{ko} = obere Ausprägung der k-ten Variable und x_{ku} = untere Ausprägung der k-ten Variable, $quotient = \frac{x_{k\max} - x_{k\min}}{x_{ko} - x_{ku}}$ und $offset = x_{k\max} - quotient \times x_{ko} = x_{k\min} - quotient \times x_{ku}$

Tabelle 2-3: Möglichkeiten für die Skalierung von Variablen

[136] DEIMER [1987, S.53], VOGEL [1977, S. 105-129], EVERITT [1980] und MILLIGAN [1979, S.343-346], KILCHEMANN [1970, S. 5]

Für die Algorithmen der Clusteranalyse ist die Wahl des Ähnlichkeits- bzw. Distanzmaßes von entscheidender Bedeutung, da über Ähnlichkeit[137] bzw. Unähnlichkeit von Objekten entschieden wird. Bei nichtmetrischen Skalen werden Koeffizienten gebildet, die bei der Überprüfung der Variablen nach Ähnlichkeit herangezogen werden. Man geht von einem paarweisen Vergleich aus und berechnet die erforderlichen Koeffizienten. Binäre Variablen können bei nur zwei Variablenausprägungen verwendet werden. Bei mehr als zwei Ausprägungen ist eine Transformation in binäre Variablen voranzustellen. Die Verwendung von Distanzmaßen bzw. Ähnlichkeitskoeffizienten ist abhängig vom Skalenniveau, so dass bei unterschiedlichem Skalenniveau spezielle Verfahren zur Konvertierung anzuwenden sind.

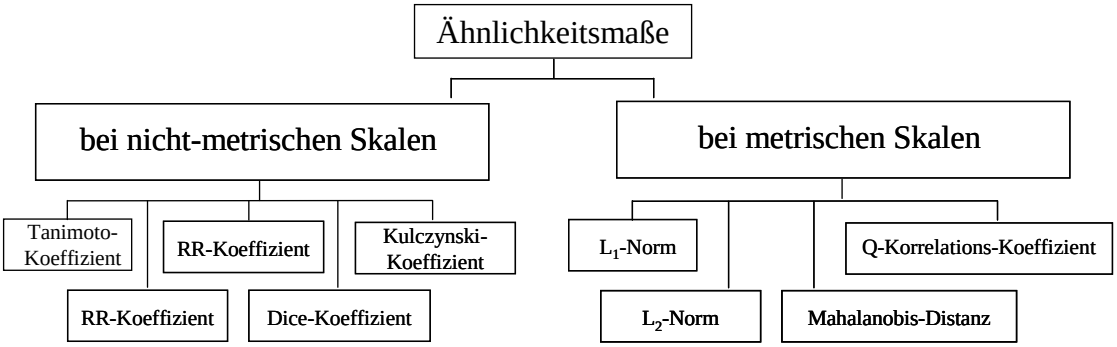


Abbildung 2-4: Ähnlichkeitsmaße im Überblick[138]

Tabelle 2-4 zeigt beispielhaft Ähnlichkeitsfunktionen für binäre Variablen.[139]

Ähnlichkeitskoeffizienten		
Tanimoto-Koeffizient	**Russel u. Rao (RR)-Koeffizient**	**M-Koeffizient (per cent matching M)**
$\frac{a}{a+b+c}$	$\frac{a}{m}$	$\frac{a+e}{m}$
mit: $a =$ Anzahl übereinstimmender Variablen, $b+c =$ Anzahl Variablen, nur bei einem Objekt vorhanden, $e =$ Anzahl der Variablen, die bei beiden Objekten nicht vorhanden ist, $m = a+b+c+e$		

Tabelle 2-4: Beispiel für Ähnlichkeitsmaße bei nicht-metrischer Skala

Tabelle 2-5 enthält beispielhaft Distanzfunktionen für metrische Variablen.[140]

Distanzmaße
Klasse der Minkowski-Metriken: $d_{k,l} = \left(\sum_j \left| \omega_{jk} - \omega_{jl} \right|^{\lambda} \right)^{\frac{1}{\lambda}} \qquad \lambda > 0$
L_1-Norm $(\lambda = 1)$: $d_{k,l} = \sum_j \left| \omega_{jk} - \omega_{jl} \right|$ und L_2-Norm $(\lambda = 2)$: $d_{k,l} = \sqrt{\sum_j \left(\omega_{jk} - \omega_{jl} \right)^2}$
mit: $d_{k,l} =$ Distanz zwischen Objekten k und l, $\omega_{jk} =$ Objektmesswert auf der Variablen $j (j = 1, 2, \ldots, J)$

Tabelle 2-5: Beispiele für Distanzmaße bei metrischer Skala

137 Vgl. EVERITT [1980, S. 19], BOCK [1974, S. 44 ff. und S. 77] und GORDON [1981, S. 13]
138 Vgl. SCHUCHARD-FICHER et al. [1985, S. 109]
139 Vgl. DEIMER [1987, S.39] und STEINHAUSEN/LANGER [1977, S. 51-67] zu Ähnlichkeitskoeffizienten
140 Vgl. DEIMER [1987, S.47] und BERGS [1980, S.79] zu Distanzfunktionen

Nach Beendigung der Variablenauswahl sollte ein validiertes Abstandsmaß zur Messung der Ähnlichkeit oder Unähnlichkeit von hochdimensionalen Daten vorliegen. Es ist zu bedenken, dass durch die Verwendung von hoch korrelierten Variablen eine Gewichtung[141] einzelner Variablen eines Datensatzes erfolgt. Statt der Datenmatrix X wird eine Distanzmatrix:

$$D = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots \\ d_{n1} & \dots & d_{nn} \end{pmatrix} = \left(d_{ij}\right)_{1 \le i,j \le n} \in \mathrm{R}^{(n,n)} \tag{10}$$

oder eine Ähnlichkeitsmatrix

$$S = \begin{pmatrix} s_{11} & \dots & s_{1n} \\ \dots & \dots & \dots \\ s_{n1} & \dots & s_{nn} \end{pmatrix} = \left(s_{ij}\right)_{1 \le i,j \le n} \in \mathrm{R}^{(n,n)} \tag{11}$$

verwendet.

$d(i,j)$ ist ein Maß für die Unähnlichkeit der Objekte i und j, welche bei geometrischer Interpretation die Distanz ausdrückt und nicht unbedingt metrische Eigenschaften haben muss.[142] Es wird aber vorausgesetzt, dass folgendes gilt:

$$\begin{aligned} &d(i,j) \ge 0 && \forall \mathrm{i,j} \in \mathrm{N} \\ &d(i,i) = 0 && \forall \mathrm{i} \in \mathrm{N} \\ &d(i,j) = d(j,i) && \forall \mathrm{i,j} \in \mathrm{N} \end{aligned} \tag{12), (13), (14}$$

Eine Distanzmatrix ist deshalb symmetrisch und enthält auf der Hauptdiagonalen nur Nullen.

Ein Maß für die Ähnlichkeit $s(i,j)$, welche sich durch die Nähe der Objekte ausdrückt, erfüllt die folgenden Bedingungen:

$$\begin{aligned} &0 \le s(i,j) \le 1 && \forall \mathrm{i,j} \in \mathrm{N} \\ &s(i,i) = 1 && \forall \mathrm{i} \in \mathrm{N} \\ &s(i,j) = s(j,i) && \forall \mathrm{i,j} \in \mathrm{N} \end{aligned} \tag{15), (16), (17}$$

Nach dem Festlegen eines Distanzmaßes ist es nötig, die Vergleichbarkeit der Datensätze zu untersuchen und besonders ähnliche als auch unähnliche Daten zu bewerten.

---

141 Vgl. VOGEL [1975, S. 67 ff.], Ausführungen zur externen und internen Gewichtung.

142 Eine Metrik d ist eine Abbildung, die je zwei Punkten i und j eines Vektorraums $d(i,j)$ eine nicht negative reelle Zahl zuordnet, wobei die folgenden drei Bedingungen eingehalten werden:

$$\begin{aligned} &d(i,j) = 0 \leftrightarrow i = j \\ &d(i,j) = d(j,i) && \text{(Symmetrie)} \\ &d(i,j) \le d(i,k) + d(k,j) && \text{(Dreiecksungleichung)} \end{aligned}$$

2.2 Strukturerkennung

2.2.1 Multidimensionale Skalierung

Die Multidimensionale Skalierung (MDS) dient als multivariates Verfahren und nichtlineare Projektionsmethode dazu, die Struktur der Objekte als Ähnlichkeitsbeziehung in einem n-dimensionalen Raum abzubilden. Bei der Interpretation der Ergebnisse ist zu beachten, dass es sich um ein iteratives Optimierungsverfahren handelt.

2.2.1.1 Grundbegriffe

Je nach vorhandenem Datenmaterial spricht man bei intervall- oder verhältnisskalierten Daten von metrischer MDS und bei ordinalen Daten von nicht metrischer MDS. Nicht-metrisch bezieht sich nur auf die Eingangsdaten, während die Ergebnisse immer metrisch sind. Die verwendeten Daten dieser Arbeit werden alle metrisch skaliert sein. Das Verfahren der MDS eignet sich zur Darstellung von Daten zu einem einzigen Zeitpunkt oder zu aufeinander folgenden Zeitpunkten, um auch Entwicklungen und Tendenzen auszuwerten. Voraussetzung hierfür ist jedoch eine einheitlich durchgängige Variablenstruktur, die auf gleichen statistischen Erfassungsmethoden beruht. Die MDS wird auch zur Reduktion der Variablenanzahl eingesetzt.

Allgemein gliedert sich das Verfahren in die Phase der Ähnlichkeitsmessung, die Phase der Distanzmodellwahl, die Phase der grafischen Darstellung (Konfiguration) und die Phase der Interpretation und Validierung. Damit Distanzen zwischen allen Paaren der Objekte berechnet werden können, muss eine bestimmte Abbildungsvorschrift bzw. ein Distanzmodell angegeben werden. Verwendet wird das euklidische Distanzmaß, welches aufgrund der vielseitigen Eigenschaften bevorzugt im Rahmen der MDS angewendet wird.[143] Der Vorteil der Euklidischen Metrik liegt darin, dass ermittelte Konfigurationen nachträglich veränderbar sind. Verschiebungen des Koordinatenursprungs sind erlaubt, wie auch eine rechtwinklige Drehung der Koordinatenachsen um den Ursprung (Rotation). Die Distanzen bleiben dabei unverändert. Alle Distanzen können bei euklidischer Metrik proportional vergrößert oder verkleinert werden. Die Gesamtheit der Positionen der Objekte im Wahrnehmungsraum in ihrer relativen Lage zueinander wird Konfiguration genannt. Die Angabe der Dimension ist ein wichtiger Bestandteil des Wahrnehmungsraumes und in dieser Arbeit wird die 2-dimensionale Darstellung gewählt. Die Ergebnisse einer MDS sind schwierig zu interpretieren, da zwischen den Dimensionen des Wahrnehmungsraumes und erhobenen Objekteigenschaften kein direkter Bezug besteht.

[143] Vgl. SCHUCHARD-FICHER et al. [1985, S. 275], GREEN et al. [1989].

2.2.1.2 Visualisierung

Um die Konfiguration zu bestimmen, wird iterativ vorgegangen, d.h. man startet mit einer Ausgangskonfiguration und versucht, über einen Anpassungsalgorithmus diese schrittweise zu verbessern (Abbildung 2-5).

Ziel der metrischen MDS ist es, ausgehend von einer beliebigen $n \times n$-Distanzmatrix $(d(i,j))$ zu n Objekten eine Dimension p und n Vektoren $x_{1,\ldots,n} \in \mathbb{R}^p$ zu finden, so dass der Abstand $d(i,j)$ zwischen den Objekten i und j mit dem euklidischen Abstand $\|x_i - x_j\|$ der Vektoren $x_i, x_j \in \mathbb{R}^p$ übereinstimmt:

$$d(i,j) = \|x_i - x_j\|, \qquad i,j = 1,\ldots,n.$$

Indem das Objekt i durch den Vektor x_i repräsentiert ist, kann die Abstandsmatrix $d(i,j)$ visualisiert werden.

Abbildung 2-5: Ziel der metrischen Multidimensionalen Skalierung[144]

Als Maß für die Güte einer Konfiguration und damit als Zielkriterium für deren Optimierung wird das so genannte Stress-Maß[145] eingesetzt. Das Stress-Maß dient der Überprüfung, inwieweit die erwähnte Monotoniebedingung erfüllt ist bzw. stellt ein Gütemaß dar. Das Stress-Maß kann Werte zwischen 0 und 1 annehmen, und je höher es liegt, desto schlechter ist die Anpassung der Distanzwerte an die Ähnlichkeit (badness to fit). Ist die monotone Anpassung exakt erfüllt, würde das Stress-Maß den Wert 0 annehmen.

$$STRESS = \sqrt{\frac{\sum_k \sum_l \left(d_{kl} - \hat{d}_{kl}\right)^2}{Faktor}} \quad \text{mit} \quad \begin{matrix} d_{kl}: \text{ Distanz zwischen Objekten } k \text{ und } l \\ \hat{d}_{kl}: \text{ Disparität für Objekte } k \text{ und } l \end{matrix} \tag{18}$$

Besonders gebräuchlich sind die Stress-Formeln 1 und 2 von KRUSKAL,[146] wobei $\bar{d}$ den Mittelwert der Distanzen umfasst.

$$STRESS\,1 = \sqrt{\frac{\sum_k \sum_l \left(d_{kl} - \hat{d}_{kl}\right)^2}{\sum_k \sum_l d_{kl}^{\,2}}} \quad \text{und} \quad STRESS\,2 = \sqrt{\frac{\sum_k \sum_l \left(d_{kl} - \hat{d}_{kl}\right)^2}{\sum_k \sum_l \left(d_{kl} - \bar{d}\right)^2}} \tag{19), (20}$$

Zur Beurteilung der Anpassungsgüte schlägt KRUSKAL[147] Erfahrungswerte vor, die fünf Klassen bilden und Anhaltswerte für das Stressmaß 1 bzw. durch Semikolon getrennt für das Stressmaß 2 liefern (gering: 0,2; 0,4, ausreichend: 0,1; 0,2, gut: 0,05; 0,1, ausgezeichnet: 0,025; 0,05, perfekt: 0; 0).

144 Eigene Bearbeitung nach FALK et al. [1995]
145 Vgl. BACKHAUS et al. [2006, S. 639]
146 Vgl. KRUSKAL [1964 a, S. 1 ff. und 1964 b, S. 115 ff.]
147 Vgl. KRUSKAL / CARMONE [1973]

2.2.2 Emergente Selbst-Organisierende Merkmalskarten

Im Jahr 1982 wurden die selbstorganisierenden[148] Karten (Self-Organizing Maps, SOM) oder auch Kohonen-Karten von T. Kohonen,[149] motiviert durch die Selbstorganisation rezeptiver Felder im menschlichen Gehirn in Form künstlich Neuronaler Netze, entwickelt. Es handelt sich um ein unüberwachtes Lernverfahren der Neuroinformatik, dessen Ziel eine topologische Darstellung und Clusterbildung ist, ohne vorher bekannte Ergebniswerte zu kennen. Die SOM sind geeignet, inhärente Strukturen eines meist hochdimensionalen Eingaberaumes sich selbst ordnend auf eine zweidimensionale Gitterstruktur zu projizieren. Übliche SOM sind durch eine geringe Anzahl Neuronen charakterisiert, wobei diese im Ergebnis einer k-Means-Clusterung entsprechen ($k =$ Anzahl Knoten der Karte). Verwendet man dagegen eine sehr große Anzahl Neuronen, so ist es möglich, Strukturen in der Merkmalskarte durch Emergenz abzubilden. Diese sogenannte Emergente[150] SOM realisiert die Visualisierung von Datenstrukturen in Form einer dreidimensionalen Landschaftsdarstellung.[151] Dabei entsprechen Berge einer großen Entfernung der Daten und ähnliche Daten ordnen sich in Tälern an.

Es findet zunächst eine Auseinandersetzung mit der SOM im Allgemeinen statt, bevor im Speziellen die emergenten selbstorganisierenden Merkmalskarten erläutert werden.

2.2.2.1 Eigenschaften der Selbstorganisierenden Merkmalskarten

Eine SOM setzt sich aus nur zwei Schichten zusammen: der Input-Schicht, die so viele Neuronen enthält wie die untersuchten Objekte Variablen besitzen, sowie der Output-Schicht – der Neuronen-Karte. Es handelt sich um die Abbildung eines p-dimensionalen Eingangsraumes E mit $x = [x_0,...,x_{p-1}]$, $x \in E^p$ auf einen q-dimensionalen Ausgangsraum A. Die Neuronen der Inputschicht $i_1...i_n$ werden mit einer Gewichtung $w_1...w_n$ mit allen Neuronen der Outputschicht $o_1...o_n$ verbunden. Die Ausgabe ist die Position des Neurons auf der Karte, dessen Prototypenvektor (model vector) $m_c = [m_0,...,m_{p-1}]$ der Eingabe x am ähnlichsten ist.

---

148 Selbstorganisation ist die Fähigkeit eines Systems sich ohne gerichtete Vorgaben von außen zu ordnen. Im Bereich der Datenverarbeitung bedeutet dies, dass Daten nicht durch Vorgaben organisiert werden, sondern sich durch geeignete Computertechnik selbstständig organisieren. Datensätze sollen sich selbstständig zu zusammengehörigen Gruppen zusammenfinden. Weiterhin sollten solche Gruppen eine die wesentlichen Variablen der Gruppe charakterisierende Beschreibung entwickeln.
Vgl. SERUGENDO et al. [2004, S. 1 ff.]

149 Vgl. KOHONEN [1982], [1995], [2001]: Die biologische Abbildung von Datenklassen wird künstlich nachgebildet. Das menschliche Gehirn besteht z.B. aus ca. 10^{11} Neuronen, die untereinander verknüpft sind und gemeinsam alle geistigen Funktionen steuern. Nervenzellen lassen sich räumlich in Bereiche unterteilen.

150 Emergenz bedeutet das Auftauchen einer übergeordneten Struktur, die sich aus der Kooperation vieler elementarer Prozesse ergibt (z.B. ist das Auftauchen einer ‚La Ola Welle' in einem Fußballstadion, mit der Selbstorganisation vieler Menschen zu erklären).Vgl. STEPHAN, A. [1997, S. 244 ff.], [1999, S. 232 ff.]

151 Vgl. ULTSCH [1999]: In emergenten Systemen entstehen Distanz- und Dichtestrukturen in Datensätzen die als U-Matrix, P-Matrix oder U*-Matrix betrachtet werden können. Emergente Strukturen beschreiben das System der elementaren Prozesse auf einem neuen, übergeordneten Niveau.

Die Gewichtsmatrix $M = \left[{m_0}^T, ..., {m_{k-1}}^T \right]$ enthält k_{som} Prototypenvektoren mit:

$$k_{SOM} = \prod_{0}^{q-1} n_i \qquad ; n_i = \text{Anzahl Neuronen in Richtung } i \tag{21}$$

entsprechend der Gesamtzahl der Neuronen in der Ausgabeschicht. Es folgt daraus für die Gesamtzahl der Gewichte $k_{W_{SOM}}$ einer SOM:

$$k_{W_{SOM}} = k_{SOM} \cdot \dim(X) \tag{22}$$

Als Gewinner-Neuron bezeichnet man das Neuron, welches sich durch einen Vergleich aller Prototypenvektoren m_i mit der Eingabe x ergibt. Es handelt sich um einen Wettbewerb der Neuronen um die stärkste Aktivierung für eine identische Eingabe. Wird ein Ähnlichkeitsmaß $\Lambda(u,v)$ als Funktion der Ähnlichkeit zweier Vektoren definiert, so gilt:

$$c = \arg\max_i \Lambda(x, m_i) \tag{23}$$

Bei den SOM werden als Ähnlichkeitsmaße in der Regel Minkowski-Metriken, sowie das Skalarprodukt eingesetzt:

$$\Lambda = \begin{cases} \sum_{i=0}^{p-1} |x_i - m_i| & \text{, L1- Norm} \\ \sqrt{\sum_{i=0}^{p-1} (x_i - m_i)^2} & \text{, L2 – Norm} \\ \sum_{i=0}^{p-1} (x_i m_i) = \langle x, m \rangle & \text{, Skalarprodukt} \end{cases} \tag{24}$$

Die maximale Ähnlichkeit zweier Vektoren in der L1- bzw. L2-Norm wird durch das Minimum des Abstandes gefunden:

$$\begin{aligned} \overset{L2}{c} &= \arg\min_i \|x - m_i\|_{L2} \\ \overset{L1}{c} &= \arg\min_i \|x - m_i\|_{L1} \end{aligned} \qquad \text{mit } i = 0,...,k-1 \tag{25}$$

Bei Anwendung des Skalarproduktes ist dessen Maximum die maximale Ähnlichkeit zweier Vektoren.

$$\overset{Skal}{c} = \arg\max_i \langle x, m_i \rangle \qquad \text{mit } i = 0,...,k-1 \tag{26}$$

Die Entfernung der Neuronen der Outputschicht wird als Nachbarschaft mit entsprechendem Radius definiert. Das Ziel des Verfahrens ist die Veränderung der Gewichtsvektoren auf der Neuronenkarte in der Form, dass sich in der Topologie der Karte die Struktur der Objekte widerspiegelt und ähnliche Objekte über eine enge Nachbarschaft verfügen.

Die Ähnlichkeit der Gewichtsvektoren beieinander liegender Neuronen wird erreicht, indem die Gewichtsvektoren der Siegerneuronen und ihrer benachbarten Neuronen in Richtung des Eingabevektors im Maße einer anfänglich größeren, aber mit jedem Lernschritt abnehmenden Lernrate und Nachbarschaft verändert werden.

Die gewünschte, z.B. zweidimensionale, Repräsentation der Objektstruktur wird erreicht durch eine immer feinere Anpassung aufgrund der geringeren Lernrate und der kleineren Nachbarschaft. Zu derselben Klasse gehören alle Objekte mit demselben Gewinnerneuron. Je nach gewünschter Klassenanzahl können außerdem Nachbarschaften zu Klassen zusammengefasst werden.

Die Tabelle 2-6 enthält die Abbildungseigenschaften der SOM.

Eigenschaften	**Beschreibung**
Approximation des Eingaberaumes	Die Gewichtsmatrix der SOM approximiert den Eingaberaum E, so dass eine vollständige Abdeckung besteht.
Abschnittsweiser Zusammenhang der Ausgaben	Infolge einer lokalen Nachbarschaft der Ausgabeneuronen beeinflussen sich die Neuronen gegenseitig. Die Änderung der synaptischen Gewichte geschieht lokal und ist durch Nachbarschaftsbeziehung gekennzeichnet.
Topologische Ordnung	Der Ausgaberaum A ist so strukturiert, dass im Sinne eines Distanzmaßes wie z.B. der Euklidischen Distanz, ähnliche Eingaben auf benachbarte Koordinaten in A abgebildet werden. Cluster im Eingaberaum lassen sich in der SOM auffinden.
$SOM : E \rightarrow A$: A kann nicht linear gegenüber E verzerrt sein	Die Verzerrung ergibt sich aus der statistischen Häufigkeitsverteilung der Eingaben. Teile des Eingaberaumes, die eine hohe Auftretenswahrscheinlichkeit besitzen, belegen einen größeren Teil der SOM-Ausgabeschicht als Eingaben, die nur selten vorkommen. Die lokale Auflösung der Karte ist der statistischen Häufigkeit der Eingaben angepasst.
Die am besten diskriminierenden Variablen bestimmen die Ausgabe	Ähnlich einer Hauptkomponentenanalyse stellt sich die SOM anhand der Variablen ein, welche die größte Varianz aufweisen. Die SOM lässt sich als nichtlineare Form einer Hauptkomponentenanalyse betrachten, bei der die Anzahl der verwendeten Hauptkomponenten der Dimension q der Karte entspricht.
Dimensionsreduktion	Für die Abbildung $SOM : E \rightarrow A$ gilt im allgemeinen $q \leq p$, mit $p = \dim(E)$ und $q = \dim(A)$

Tabelle 2-6: Eigenschaften der selbstorganisierenden Merkmalskarten[152]

Die topologische Nachbarschaftsfunktion muss der Forderung genügen, dass der Maximalwert der Nachbarschaftsfunktion mit der Zeit abnimmt und der Wert der Nachbarschaftsfunktion für große Abstände vom Gewinnerneuron gegen 0 geht, so dass gilt:

$$\lim_{n\to\infty} h_{c,l} = 0 \text{ und } \lim_{d_{c,l}\to\infty} h_{c,l} = 0 \qquad (27)$$

[152] Eigene Bearbeitung nach ELSEN [2000, S.48/49]

Dies bedeutet nicht, dass eine monoton fallende Funktion vorliegen muss, und in Frage kommen sowohl stetig differenzierbare (Gleichung 26 oder 27) als auch nicht stetig differenzierbare Funktionen (Gleichung 28 oder 29).

$$h_{c,l} = \left(1 - \frac{d_{c,l}^2}{\sigma(n)^2}\right) \frac{1}{\sqrt{2\pi}\sigma(n)^3} e^{-\frac{d_{c,l}^2}{2\sigma(n)^2}} \quad \text{oder} \quad h_{c,l} = \frac{1}{\sqrt{2\pi}\,\sigma(n)} e^{-\frac{d_{c,l}^2}{2\sigma^2(n)}} \qquad (28) \text{ und } (29)$$

$$h_{c,l} = \begin{cases} -\frac{a_0}{b_0}\left|d_{c,l}\right| + a_0 ; \left|d_{c,l}\right| > b_0 \\ 0 \quad ; \text{ sonst} \end{cases} \quad \text{oder} \quad h_{c,l} = \begin{cases} a_0 ; \left|d_{c,l}\right| > b_0 \\ 0 \; ; \text{ sonst} \end{cases} \qquad (30) \text{ und } (31)$$

Das Training einer SOM wird in drei Phasen untergliedert, die in Tabelle 2-7 dargestellt sind.

Phase	**Beschreibung**
Wettbewerb	Jedes Neuron c strebt nach der stärksten Aktivierung und das Neuron, dessen Gewichtsvektor die größte Ähnlichkeit zur Eingabe besitzt, stellt den Gewinner: $c = \arg\max_i \Lambda(x, m_i)$
Laterale Interaktion	Ein Gewinnerneuron bildet das Zentrum einer lokalen Nachbarschaft, innerhalb derer eine laterale Interaktion stattfindet. Die Nachbarschaftsfunktion ist radialsymmetrisch und beschränkt. Der Grad der synaptischen Gewichte der Neuronen während der Adaption wird durch den Wert der Nachbarschaftsfunktion $h_{c,l}(d_{c,l})$, (Gleichungen 28, 29) bestimmt und hängt vom Abstand d des zu modifizierenden Neurons l vom Gewinnerneuron c ab. Die Distanz der beiden Neuronen im R^q wird durch den Abstand $d_{c,l}$ gebildet: $d_{c,l}^2 = \left\|\rho_c - \rho_l\right\|^2$
Adaption	Die synaptischen Gewichte werden bez. der Ähnlichkeit der Eingabe verändert. Der Prozess der Modifikation wird bestimmt durch eine Lernrate $\eta(t)$ und die Nachbarschaftsfunktion. Eine Funktion g, die nur von der Aktivierung des Neurons abhängt, berücksichtigt die Änderung des synaptischen Gewichtes mit: $\Delta m_l = \eta y_l x - g(y_l) m_l$ Der erste Term der rechten Seite entspricht dem so genannten HEBBschen Lernen[153] und der zweite Term dient dem teilweisen Vergessen des ursprünglichen Gewichtes. Durch Einsetzen von ηy_l für $g(y_l)$ und ersetzen von y_l durch $h_{c,l}$ gilt: $\Delta m_l = \eta h_{c,l}(x - m_l)$ Die zeitdiskrete Form der Gleichung zur Anpassung der synaptischen Gewichte ist von Bedeutung, da das Training iterativ erfolgt und die zu lernenden Muster nacheinander an der SOM anliegen: $m_l(n+1) = m_l(n) + \eta(n) h_{c,l}(n)(\mathrm{x} - \mathrm{m}_\mathrm{l}(n))$

Tabelle 2-7: Trainingsphasen einer SOM und deren Bedeutung[154]

Für die zeitliche Modifikation des Parameters σ gilt: $\sigma(n+1) - \sigma(n) < 0$. Dieser bestimmt die Breite der Nachbarschaftsfunktion. Wird die Breite der Nachbarschaftsfunktion am Beginn des Trainings zu klein gewählt, kann sich die SOM nicht dem Eingaberaum anpassen, und es entstehen so genannte Topologische Defekte.[155]

---

153 Der Neuropsychologe HEBB [1949] postuliert das nach ihm benannte Lernverfahren für die Änderung der synaptischen Verbindung zwischen zwei Neuronen, vgl. BRAUSE [1995, S. 79], GRAUEL [1992, S. 35]

154 Eigene Bearbeitung nach ELSEN [2000, S.52 ff.] und GRAUEL [1992, S. 161 ff.]

155 ELSEN [2000, S.63] und BORGELT et al. [2003, S. 109]

Abbildung 2-6 stellt das Problem anhand des Trainings zweidimensionaler Koordinaten als Eingabe für die SOM dar, wobei sich die Eingabe aus Koordinatenpaaren (x, y) im Intervall $[0,1[$ zusammensetzt und die durch einen Zufallsgenerator erzeugt werden.

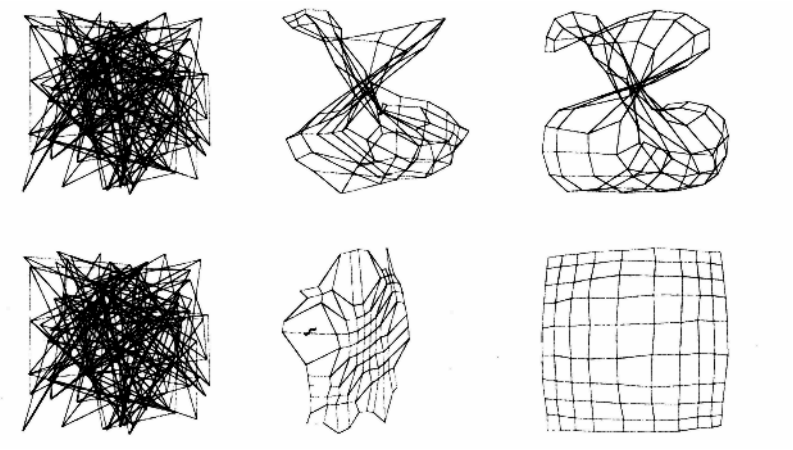


Abbildung 2-6: Entstehen eines Topologischen Defektes durch zu geringe Breite bzw. zu schnelle Abnahme der Breite der Nachbarschaftsfunktion (oben) und korrekte Entfaltung der Karte (unten), dargestellt nach 0, 100, 1000 Trainingsschritten[156]

Für die Visualisierung wird in der Regel eine zweidimensionale Gitterstruktur verwendet. Die Neuronen können in einem Quad-Gitter oder Hex-Gitter angeordnet sein. Durch Erweiterung zu einer toroidalen Topologie wird die Problematik der so genannten Randeffekte vermieden, die ihren Ursprung in der kleineren Nachbarschaft haben, d.h. der geringeren Zahl benachbarter Neuronen an den Rändern im Vergleich zur Nachbarschaft der Neuronen im Zentrum. Es erfolgt dadurch keine Stauchung an den Rändern der Karte und keine Streckung der Abstände der Cluster im Zentrum der Karte (siehe Tabelle 2-8).

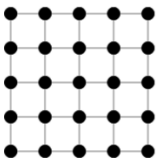
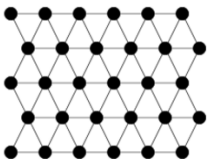
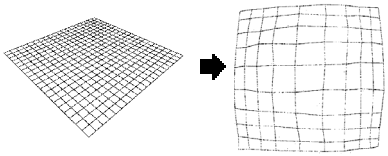
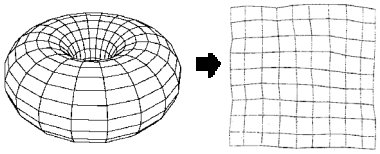
Quad-Gitter	**Hex-Gitter**
	
Planare Topologie mit Rand	**Toroidale Topologie**
$d_{c,l}^2 = \lVert \rho_c - \rho_l \rVert^2$ 	$d_{c,l}^2 = \sum_{k=0}^{q-1} \min\left(\left(n_k - \lvert \rho_c - \rho_l \rvert \right)^2, \left(\lvert \rho_c - \rho_l \rvert \right)^2 \right)$ 

Tabelle 2-8: Gitterstrukturen und Nachbarschaftsfunktionen der SOM (Randeffekt)[157]

156 Vgl. ELSEN [2000, S.64]
157 Eigene Bearbeitung mit ULTSCH [2006 a] als Quelle der Abbildungen.

Die Eigenschaften der SOM lassen sich wie folgt darstellen:

- komplex,
- multiagent,
- dynamisch,
- nicht-deterministisch,
- verzweigend,
- irreversibel und
- nichtlinear.

Die SOM beschreibt ein komplexes System, da diese sich, wie zuvor gezeigt, aus multiplen miteinander kommunizierenden Elementen zusammensetzt, den Neuronen und ihren Nachbarn.

Bei der SOM können die Neuronen im Sinne eines Multi-Agenten-Systems als die Agenten angesehen und die Modifikation der Gewichtung eines Neurons innerhalb seiner Nachbarschaft als Zusammenwirken bezeichnet werden. Das Ziel eines derartigen Multi-Agenten-Systems besteht in der Adaption der Struktur der Eingangsdaten.

Die SOM kann als dynamisches System betrachtet werden, wobei unterschiedliche Zustände von den Gewichtungen der Neuronen beeinflusst werden.

Eine SOM trägt nicht-deterministische Eigenschaften, da eine Zufälligkeit von zukünftigen Systemzuständen besteht. In den SOMs gibt es üblicherweise zwei Quellen von Zufälligkeiten. Einerseits liegt diese in der Wahl der Anfangskonfiguration (Initialisierung) der Gewichtungen begründet und andererseits in der Auswahl des nächsten zu lernenden Eingangsvektors.

Da das Auffinden des Gewinnerneurons (Best Matching Unit) von all den möglicherweise geringfügigen Veränderungen der Gewichtungsvektoren abhängt, wird es mit Blick auf die iterativen und von der Vorgeschichte abhängigen Lerneigenschaften als verzweigendes System bezeichnet. Dabei bestimmt die Reihenfolge der Abarbeitung der Eingangsvektoren die Identifizierung des Gewinnerneurons und den zukünftigen Berechnungsweg.

Bei der SOM handelt es sich um einen unumkehrbaren (irreversiblen) Lernprozess, der durch den Ein- bzw. Ausschluss von Neuronen von einer Nachbarschaft beschrieben wird und darüber hinaus die Größe der Nachbarschaft mit steigender Iterationszahl langsam sinkt.

Gerade der Ein- oder Ausschluss von Neuronen von einer Nachbarschaft führt zu einer nichtlinearen Eigenschaft der SOM.

2.2.2.2 Erkennung von Strukturen durch Emergente SOM

Die Darstellung der Positionen von Datensätzen im Kartenraum ist zur Erkennung von Strukturen vielfach ungenügend, da die hochdimensionalen Distanzen verzerrt dargestellt sind. Die Entfernung von jeweils benachbarten Neuronen auf dem Gitter kann im Datenraum sehr unterschiedlich sein. Methoden wie die U-Matrix, P-Matrix oder U*-Matrix fördern die Strukturerkennung in den Daten. Die U-Matrix[158] stellt die lokalen Distanzbeziehungen im hochdimensionalen Datenraum als dreidimensionale Landschaft mit Bergen und Tälern dar. Die P-Matrix[159] bildet die Dichte der Daten über dem Gitter ab. Unter Verwendung der Pareto-Dichte-Schätzung als eine informationstheoretisch optimale Dichteschätzung wird die Dichte errechnet. Die Kombination von Distanz und Dichte berücksichtigt die U*-Matrix,[160] wobei in dichten Bereichen die Distanz zwischen Daten weniger ins Gewicht fällt und in dünn besiedelten Bereichen des hochdimensionalen Datenraumes die Distanzstrukturen der U-Matrix betont werden. Tabelle 2-9 enthält beispielhaft randlose gekachelte Abbildungen und eine redundanzfreie Inselansicht.[161]

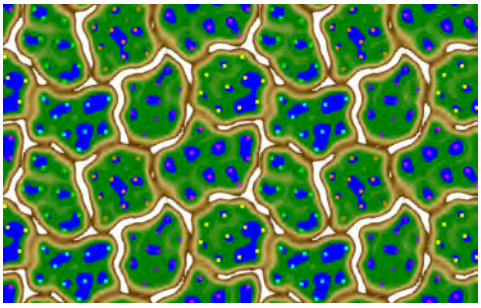
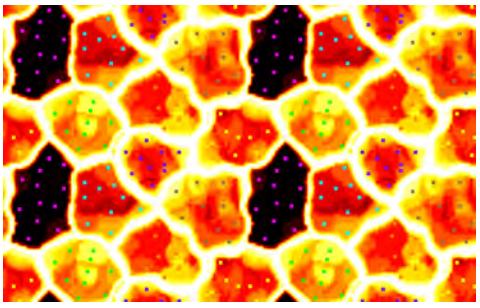
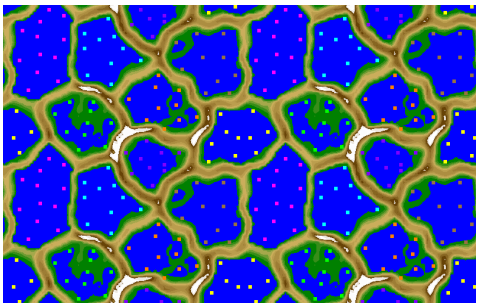
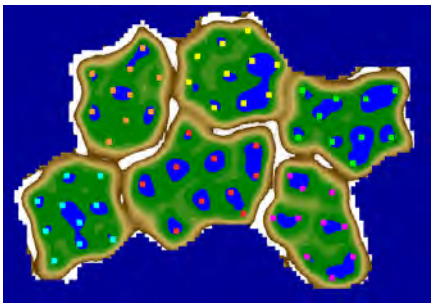
U-Matrix (Distanzbeziehung = Berge und Täler)	**P-Matrix (hohe Dichte = dunklere Darstellung)**
	
U*-Matrix (Kombination aus U- und P-Matrix)	**Inselansicht (redundanzfreie Darstellung)**
	

Tabelle 2-9: Visualisierung von Daten mit U-Matrix, P-Matrix, U*-Matrix und Inselansicht[162]

158 Vgl. ULTSCH [1993]
159 Vgl. ULTSCH [2003 b]
160 Vgl. ULTSCH [2003 c]
161 Vgl. ULTSCH [2006 b]: Im Rahmen dieser Arbeit werden die Emergenten Selbstorganisierten Merkmalskarten mit einer Software der Philipps-Universität Marburg des Fachbereichs Mathematik und Informatik berechnet, die kostenlos jedem Anwender zur Verfügung steht und durch viele Projektbearbeitungen hinsichtlich der Stabilität geprüft ist.
162 Eigene Bearbeitung mit ULTSCH [2006 a] als Quelle der Abbildungen, die sich auf Beispieldatensätze des Softwareurhebers beziehen.

Die Anzeige der U-Höhenwerte (U-height) an der Spitze der Neuronen eines Kartenraums wird U-Matrix[163] genannt. Ein U-Höhenwert $uh(N)$ eines Neurons N ist der durchschnittliche Datenabstand des Gewichtungsvektors von N zu den Gewichtungsvektoren, die mit den Neuronen in ihrer Nachbarschaft verknüpft sind. Ein Schritt (step) in einer U-Matrix ist eine Bewegung von einem Neuron A zu einem unmittelbar benachbarten Neuron A'. Ein Pfad ist eine verbundene Reihenfolge dieser Schritte. Ein Schritt von einem Neuron A mit U-Höhenwert uh_A zu einem unmittelbar benachbarten Neuron B mit U-Höhenwert uh_B wird ansteigend genannt, wenn folgendes gilt: $uh_B > uh_A$. Neuron A fällt zu Neuron B, wenn es einen Pfad p von A nach B gibt und jeder Schritt in p nicht ansteigend ist und folgendes gilt: $uh_B < uh_A$.

Ein Auffangbecken oder Tal (catchment basin) ist eine Teilmenge S von Neuronen einer SOM insofern, dass alle Neuronen in S auf dasselbe lokale Minimum fallen. Wenn derartige lokale Minima unmittelbare Nachbarn sind, sind ihre Auffangbecken vermischt. Das auffälligste aller Neuronen (attractor) innerhalb eines Auffangbeckens ist ein einzigartiges Neuron, das aus den Minima des Auffangbeckens ausgewählt wurde. Wenn es mehr als einen Kandidaten für diese Rolle gibt, kann eines dieser Neuronen entsprechend von Datenverteilungskriterien, wie z.B. der lokalen Dichte im Datenraum, ausgewählt werden.

Die Grenzlinien zwischen Auffangbecken sind die sogenannten Wasserscheiden (watersheds). Es gibt effiziente Algorithmen, die sich für eine U-Matrix eignen, z.B. bei LUC / SOILLE,[164] um die Auffangbecken und dergleichen zu berechnen. Die Ordnung der Wasserscheiden $WO(U)$ der U-Matrix U ist die Zahl der unterscheidbaren Auffangbecken (=Zahl der unterschiedlichen attractors) in U. Eine U-Matrix wird nicht-trivial genannt, wenn die dazugehörige Ordnung der entsprechenden Wasserscheiden $WO(U)>1$ und $WO(U)$ wesentlich kleiner ist als die Zahl der Eingangsdaten und die Zahl der Neuronen der SOM.

Eine SOM ist lokal geordnet, wenn sie eine nicht triviale U-Matrix U entstehen lässt und U mit der Clusterstruktur der Daten konform ist: z.B. gehören alle Neuronen innerhalb eines Auffangbeckens zum selben Cluster. In diesem Fall repräsentiert das Auffangbecken ein Cluster von spezifischen Daten. Die Wasserscheiden in U repräsentieren (lokale) Clustergrenzen. Eine U-Matrix ist clusterkonform, wenn jedes Cluster in den Daten entweder von einem einzelnen oder einem Satz von direkt benachbarten Auffangbecken repräsentiert wird.

---

[163] Vgl. ULTSCH [2003 b, S. 225-230]

[164] Vgl. SOILLE, P. / LUC, V. [1991] und siehe Abschnitt 2.3.4 (Dichtebasierte Clusteranalyseverfahren)

2.3 Strukturbildung

2.3.1 Wissenschaft des Systematisierens

2.3.1.1 Grundbegriffe

Die Taxonomie ist die methodisch ausgerichtete Wissenschaft von den Prinzipien (Taxonomische Grundprinzipien), Grundlagen (Taxonomisches Grundproblem), Verfahren (Clusteranalyse, Distanzgruppierung), Ergebnissen (Raumtypisierung, Regionalisierung) und Verallgemeinerungen des Systematisierens (Systematik). Die Taxonomie (Abbildung 2-7) wurde durch die Biologie maßgeblich geprägt. Andere Fachbereiche verwenden den Begriff der Taxonomie allgemein für ein Klassifikationssystem, eine Systematik oder den Vorgang des Klassifizierens. Für die „Räumliche Taxonomie“ ist von Bedeutung, dass die betrachteten Individuen räumliche Objekte darstellen. Die so genannte räumliche Kontingenz als Teil des taxonomischen Grundproblems bezieht sich auf die räumliche Lage der zu untersuchenden Objekte. Je nachdem ob die räumliche Lage berücksichtigt wird oder nicht, ist die Zerlegung der Menge der räumlichen Individuen in Taxa (Cluster)[165] eine Regionalisierung bzw. eine Raumtypisierung. Eine Gruppierung der Menge von räumlichen Basiseinheiten[166] ohne Berücksichtigung der räumlichen Kontingenz ist eine Raumtypisierung. Es handelt sich dabei um eine Zerlegung von räumlichen Objekten z.B. in Klein-, Hafen- und Industriestädte. Die Regionalisierung umfasst eine Gruppierung räumlicher Basiseinheiten unter Berücksichtigung räumlicher Zusammenhänge, so dass z.B. Arbeitsmarktregionen beschrieben werden.

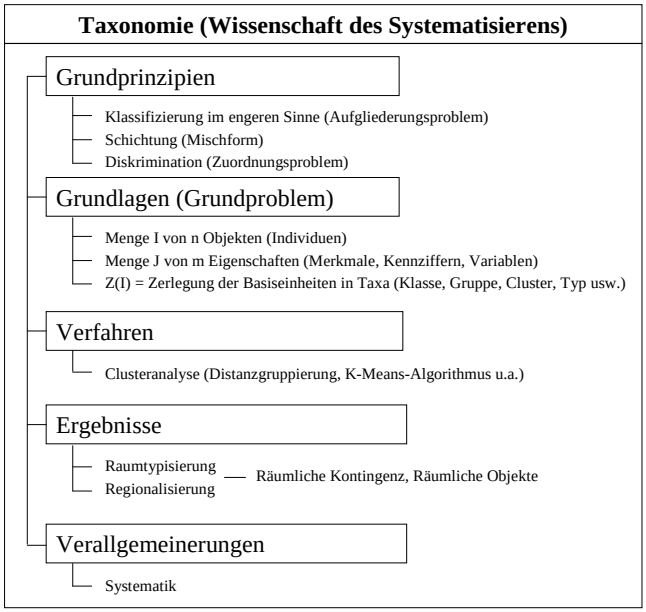


Abbildung 2-7: Wissenschaft des Systematisierens[167]

[165] Cluster = engl. Traube, Haufen.

[166] Die Basiseinheit wird als kleinste Einheit definiert, die sich der Bearbeiter eines ganz bestimmten Problems als hierarchisch-systematisch unteilbar vorstellt, siehe Taxonomie: „operational taxonomic unit“ (OTU).

[167] Eigene Bearbeitung, siehe zur Vertiefung: MARGRAF [2005]

Seit Aristoteles wird mit den Begriffen Gattung und Art klassifiziert und als Grundlage für eine Klassifizierung ist der Begriff des Merkmals bereits existent. Der Klassifizierungsbegriff ist grundsätzlich vom griechischen Wort κλαω (klao) abgeleitet, mit dem Sinn: ich zerteile, ich zerbreche. Der Begriff ging ins Lateinische ein zur Bezeichnung einer bestimmten Abteilung von Bürgern mit dem Terminus „classis".[168] PESCHEL[169] sagt: „Die Klassifizierung ist eine Hauptform des Abstraktionsprozesses und befördert den Übergang zu einer höheren Stufe der Erkenntnis, zur theoretischen Erkenntnis." WISHART[170] sagt: „the act of sorting things into categories is one of the most primitive and common pursuits of man". Die Arbeit des schwedischen Naturwissenschaftlers CARL VON LINNÉ[171] umfasst die erste systematische Klassifikation der Lebewesen im Jahre 1758 und dokumentiert den Versuch zur Lösung eines Klassifizierungsproblems durch empirische und intuitive Vorgehensweise. Die Einsatzfelder der Klassifikation sind vielfältig und liegen besonders in den nichttechnischen Disziplinen, die stark vergleichend arbeiten wie z.B. Biologie, Medizin, Ökonomie (z.B. JUGLAR: Konjunkturzyklus, 1862), Marktforschung, Verlagswesen, Soziologie, Psychologie, Kriminalstatistik, Agrarwissenschaften, Meteorologie, Astronomie (z.B. Herzsprung-Russel-Diagramm, 1913), Anthropologie, Paläontologie, Geologie und Chemie (z.B. Periodensystem, 1869). In den Ingenieurwissenschaften und der Informationsverarbeitung wird die Klassifizierung beispielsweise zur Signalanalyse (EKG, EEG), Bildverarbeitung (z.B. Luft-, Röntgen-, Mikroskopbilder), Situationsanalyse (z.B. bei Robotern), Diagnose (z.B. Maschinen, Anlagen) oder der Identifikation (z.B. Prozesszustände) verwendet. Die Numerische Taxonomie[172] (Biologie), Taxonometrie (Psychologie), Automatische Klassifikation (Mathematik) beschreibt den Teil der Taxonomie, der die rechentechnische Bestimmung von Clustern mit Hilfe von Algorithmen ermöglicht. Heute findet der Begriff Clusteranalyse die größte Verbreitung zur Beschreibung eines Prozesses der Numerischen Taxonomie, der als ein Instrument der mehrdimensionalen Datenreduktion und Informationskonzentration dient.[173] Nach PIRKTL[174] ist: „die uneinheitliche Terminologie im Bereich der Verfahren und Algorithmen zur multivariaten Datenanalyse ... größtenteils auf die Vielzahl unterschiedlichster Schulen und Forschungseinrichtungen zurückzuführen."

---

168 Vgl. SERVIUS TULLUS, 6. römischer Kaiser, 578-534 v.u.Z., 5 Klassensystem nach Alter und Vermögen.
169 Vgl. PESCHEL [1991]: Klassifizierung geowissenschaftlicher Informationen.
170 WISHART, D. [1978]: CLUSTAN – user manual, S. 1
171 Vgl. LINNÉ, Carl [1770]
172 Vgl. AMBROSI, K [1980, S.1]: „mathematische-statistische Verfahren, die die Gewinnung und Analyse von Ähnlichkeitsbeziehungen einer endlichen Menge O von Objekten, also $O = \{1, 2, ..., n\}$ als Zielsetzung haben."
173 Vgl. weitere Begriffe bei BERGS [1981, S. 3]
174 Vgl. PIRKTL, L. [1983, S. 33]

2.3.1.2 Klassifizierungsproblem

Mit Hilfe mathematisch-statistischer, d.h. rechentechnisch umsetzbarer Verfahren wird das gestellte Klassifizierungsproblem gelöst. Es lassen sich drei Aufgabenstellungen nennen, die sehr häufig als Klassifizierungsprobleme bezeichnet werden:[175]

I. Das Aufgliederungsproblem (Klassifizierung im engeren Sinne)[176]

Es liegen keine Informationen über Gruppen in der Gesamtheit vor. Die Aufgabe besteht darin, die Gesamtheit oder Stichproben aus der Gesamtheit in eine zunächst unbekannte 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 einzige Informationsquelle.

II. Das 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.

III. Das Zuordnungsproblem (Diskriminanzproblem)

Es liegen Informationen über Anzahl 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.

Im Wesentlichen entspricht das Problem der Klassifizierung dem unter I. dargestellten Aufgliederungsproblem. Indem das Aufgliederungsproblem von den Statistikern einer methodischen Untersuchung unterzogen wurde, begann sich erst mit Beginn der 70er-Jahre die Clusteranalyse durch die Möglichkeit des vermehrten Computereinsatzes als eine eigenständige Analyseform zu etablieren. Ihre Entwicklung verdankt die Clusteranalyse dem Wunsch, den Klassifikationsprozess systematisch und quantitativ erfassen zu wollen und durch Berücksichtigung numerischer Kriterien die Güte von Gruppierungen „objektiv“ zu vergleichen. Es handelt sich um Verfahren, die sich auf Objekte stützen und als Technik geeignet sind, diese Objektmenge, von der meist zunächst keine Gruppenstruktur bekannt ist,

---

175 Vgl. SCHÄFFER [1969, S. 3 ff.], VOGEL [1975, S. 3-5], DEICHSEL / TRAMPISCH [1985, S. VII]
176 Vgl. LEUSCHNER, D. [1974]

in homogene Teilmengen zu zerlegen, d.h. im Ergebnis eine Konfiguration von Clustern zu bilden. Ein Cluster ist eine Menge von homogenen Daten, während eine Klasse aus mehreren Clustern bestehen kann. Eine Gruppierung g ist eine Menge von nicht leeren Teilmengen $A_1,...,A_k, k \in N$, der Indexmenge I, d.h.

$$g = \{A_1,...,A_k\} \text{ mit } A_I \subseteq I, 1 \le i \le k \tag{32}$$

Man bezeichnet eine Gruppierung disjunkt, d.h. jedes Objekt gehört zu höchstens einer Klasse, wenn gilt:

$$A_i \cap A_j = 0 \text{ für } i \ne j, 1 \le i, j \le k \tag{33}$$

Eine Gruppierung heißt exhaustiv, d.h. jedes Objekt wird mindestens einer Klasse zugeordnet, wenn folgendes gilt:

$$\sum_{i=1}^{k} A_i = I \tag{34}$$

Eine Partition α der Länge k charakterisiert eine disjunkte, exhaustive Gruppierung mit k Clustern. Die Partition $\alpha = \{A_1,...,A_k\}, k \in N$ ist feiner als $\beta = \{B_1,...,B_l\}, l \in N, l < k$, wenn jedes Cluster von α vollständig in einem Cluster von β enthalten ist, d.h. wenn gilt: $\forall A_i \in \alpha \, \exists B_j \in \beta : A_i \subseteq B_j$.

Die Clusteranalyse wird in die Gruppe der Verfahren der automatischen Strukturbildung eingeordnet. Die Strukturbildung ist ein Prozess, ein Vorgang, dem eine Strategie zugrunde liegt, die im Wesentlichen durch ein spezielles Clusterverfahren gekennzeichnet ist. Mit Hilfe des Clusterverfahrens erfolgt eine Durchmusterung und Auswertung der Ähnlichkeits- bzw. Distanzmatrizen. CORMACK[177] fordert die externe Isolierung (=Heterogenität zwischen den Clustern) und die interne Kohäsion (=Homogenität innerhalb der Cluster) für die Klassifizierung. Beide Forderungen werden durch so genannte natürliche Cluster erfüllt.[178] BOCKLISCH[179] nennt Kriterien, die an einen Clusteralgorithmus gestellt werden:

- Es sollten sehr unterschiedliche Klassentypen entdeckt werden können.
- Das Datenmaterial ist in möglichst viele inhaltlich interpretierbare Klassen zu strukturieren, die in sich sehr kompakt aber untereinander gut getrennt vorliegen.
- Klassen sollten durch einen möglichst hohen Prozentsatz der Objekte gestützt werden.

177 Vgl. CORMACK [1971, S. 321-367]
178 Vgl EVERITT [1980, S. 60], ALDENDERFER/BLASHFIELD [1984, S. 33-34], VOGEL [1975, S.16-17], SCHULZE [1980, S. 52]
179 Vgl. BOCKLISCH [1987, S.64]

Die Verfahren der Clusteranalyse[180] sind außerordentlich zahlreich und können nach verschiedenen Kriterien systematisiert werden. Abbildung 2-8 zeigt sowohl scharfe als auch unscharfe Clusteranalyse-Verfahren. Die deterministischen Clusteranalyseverfahren unterscheiden zwischen hierarchischen und partitionierenden Verfahren. In der Literatur[181] wird weiterhin von Verfahren der unvollständigen Clusteranalyse gesprochen.[182] Zusätzlich zu den Algorithmen der Clusteranalyse ist es möglich, Mischungsmodelle (Abschnitt 2.3.4) zur Klassifizierung von Objektdaten einzusetzen.

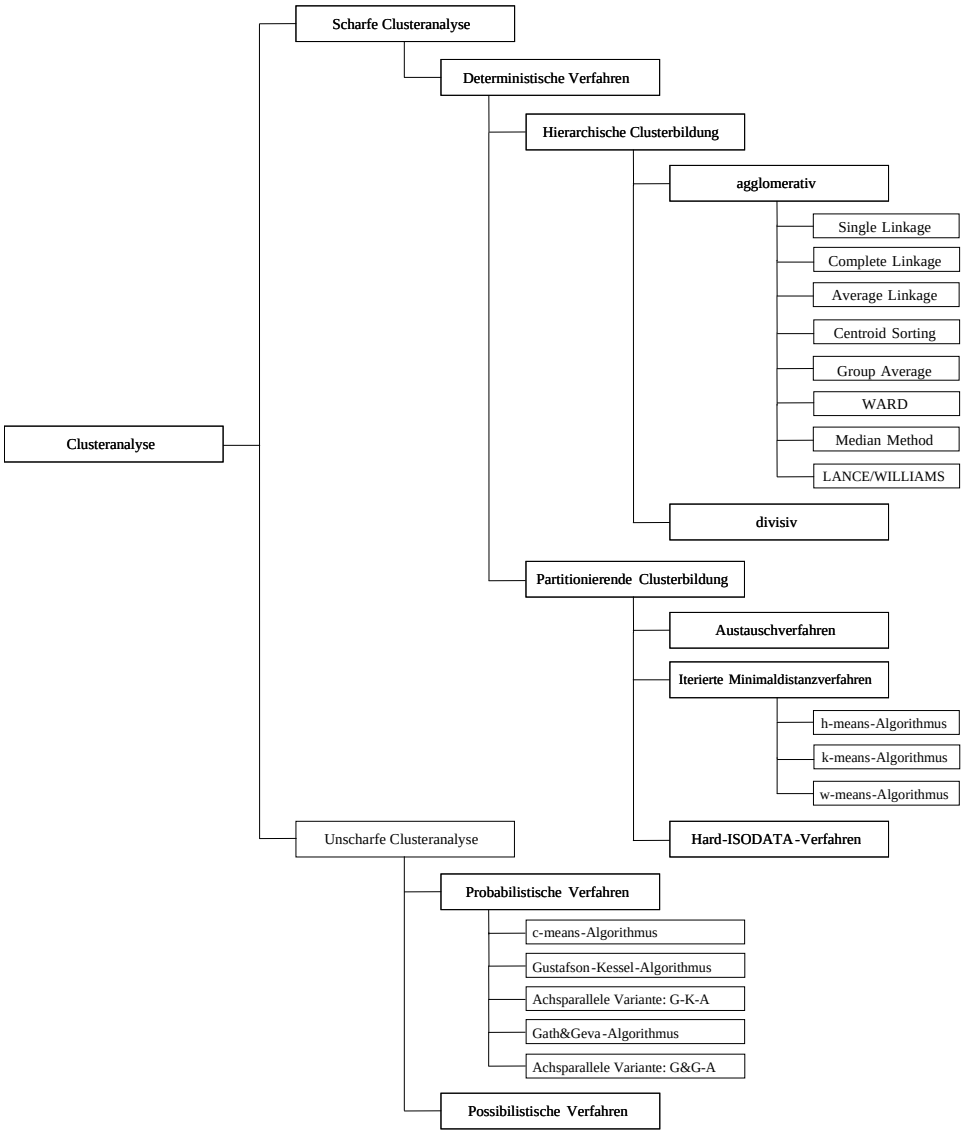


Abbildung 2-8: Verfahren der Clusteranalyse im Überblick

180 Vgl. SOKAL [1963], ANDERBERG [1973], HARTIGAN [1975], BOCK [1974], SPÄTH [1975, 1977]

181 Vgl. HÖPPNER et al. [1997, S. 8]: Es handelt sich um geometrische Methoden, Repräsentations- oder Projektionstechniken. Mehrdimensionale Daten werden einer Dimensionsreduktion unterzogen, um sie zwei- oder dreidimensional graphisch darzustellen. Die Clusterbildung erfolgt manuell durch augenscheinliche Betrachtung der Daten.

182 Da die so genannten unvollständigen Verfahren keine Zuordnung der Klassifikationsobjekte zu Clustern vornehmen, werden in dieser Arbeit derartige Verfahren der Strukturerkennung (Abschnitt 2.2) zugewiesen.

Der allgemeine Ablauf eines Verfahrens der Clusteranalyse von der Datenerhebung bis zur Interpretation der in Cluster klassifizierten Objekte ist in Abbildung 2-9 wiedergegeben.

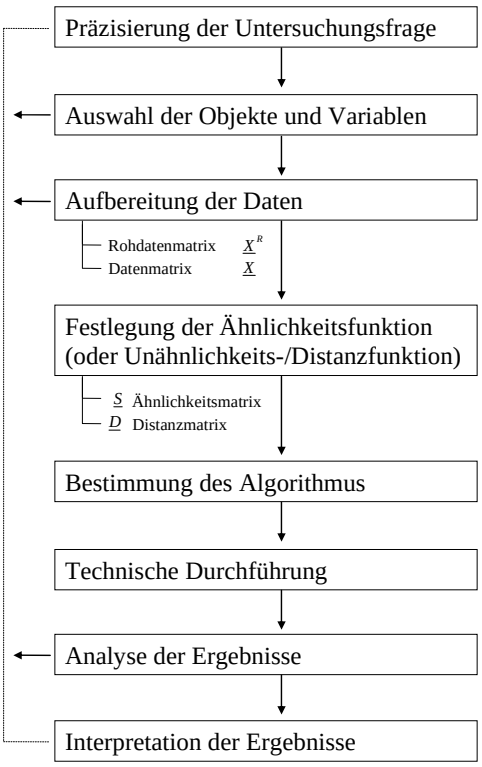


Abbildung 2-9: Ablaufschema zur Lösung des Klassifizierungsproblems[183]

Innerhalb des Ablaufplans besteht die Möglichkeit der Entscheidung zur Modifikation und der Wiederholung von Verfahrensschritten.

2.3.2 Deterministische Clusteranalyseverfahren

Die deterministischen Clusteranalyseverfahren berechnen Cluster, so dass die Klassifikationsobjekte mit einem Grad der Zuordnung von 0 oder 1 einem oder mehreren Clustern zugehören.

2.3.2.1 Hierarchische Clusterbildung

Hierarchische Klassifikationsverfahren[184] erzeugen eine Folge von Gruppen, die jeweils hinsichtlich eines bestimmten Gütekriteriums homogen sind. Die hierarchischen Verfahren werden nach agglomerativen und divisiven Algorithmen unterschieden. Die agglomerativen Algorithmen gehen von der feinsten Startpartition aus (= Gesamtzahl der Untersuchungsobjekte), während bei den divisiven Algorithmen zu Anfang alle Untersuchungsobjekte in einer Gruppe vorliegen. Demzufolge wird bei der agglomerativen Verfahrensart ein Gruppen-

183 Eigene Bearbeitung nach DEIMER [1987, S.23] und BOCKLISCH [1987, S.54]

184 Als Begründer der hierarchischen Verfahren sind zu nennen: RAO [1952], SNEATH, P.H.A. [1957, S.201-226], SOKAL UND MICHENER [1958, S.1409-1438] und WARD [1963, S. 236-244].

zusammenfassungsprozess ausgeführt, der nach einer Bottom-Up-Strategie einen der Fusionierungsschritte von Objekten vollzieht. Die divisive Verfahrensart ist durch die Gruppenaufteilung charakterisiert und führt den zerlegenden Prozess bis zu einelementigen Klassen fort (Top-Down-Strategie). Für die Ergebnisdarstellung hierarchischer Clusterverfahren sind Entscheidungsbaumdiagramme (Dendrogramme) üblich, bei denen die Objekte je nach Zugehörigkeit entsprechend ihrer Distanzen zusammengefasst werden. Abbildung 2-10 zeigt das Ablaufschema für einen agglomerativen Algorithmus.

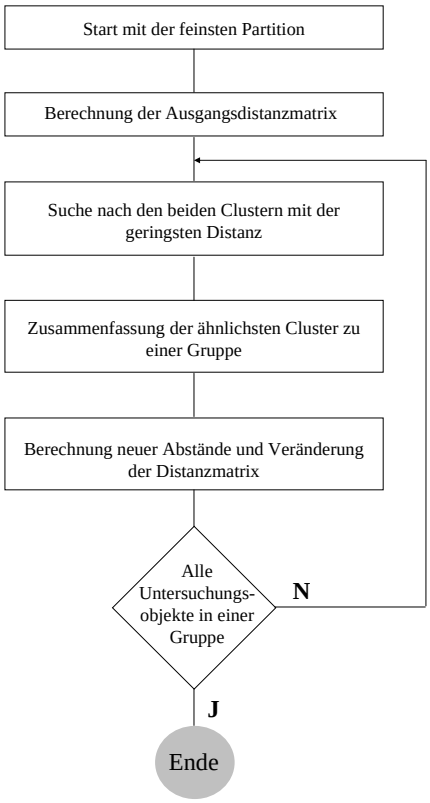


Abbildung 2-10: Ablaufschema der hierarchischen Verfahren mit agglomerativem Algorithmus[185]

Schritt 1: Jedes Klassifikationsobjekt bildet zu Beginn ein selbständiges Cluster. Setze daher die Clusterzahl c gleich der Klassifikationsobjektzahl n.

Schritt 2: Suche das Clusterpaar $(\{p\},\{q\})$ mit der größten Ähnlichkeit bzw. der geringsten Unähnlichkeit, verschmelze das Clusterpaar zu einem neuen Cluster $\{p,q\}$ und reduziere die Clusterzahl c um 1 $(c = c-1)$.

Schritt 3: Prüfe ob c gleich 1 ist. Ist dies der Fall, beende den Algorithmus, da alle Klassifikationsobjekte einem einzigen Cluster angehören. Andernfalls fahre mit Schritt 4 fort.

Schritt 4: Berechne die Ähnlichkeit bzw. Unähnlichkeit des neu gebildeten Clusters $\{p,q\}$ zu den verbleibenden Clustern i.

Schritt 5: Gehe zu Schritt 2.

185 Eigene Bearbeitung nach SCHUCHARD-FICHER et al. [1985, S. 129]

Innerhalb der hierarchisch-agglomerativen Verfahren wird zwar zwischen mehreren Verfahren unterschieden, doch gleichen sie sich in der Methodik und unterscheiden sich lediglich in der Berechnung der so bezeichneten Zwischenklassenverschiedenheit.[186]

Das Single-Linkage-Verfahren definiert die Distanz $\tilde{d}_{ij}$ zwischen zwei Clustern A_i und A_j durch die Distanz der beiden nächstgelegenen Elemente O_P aus A_i und O_q aus A_j, d.h.

$$\tilde{d}_{ij} = \min\left\{d_{pq} \middle| p \in A_i, q \in A_j\right\} \tag{35}$$

Beim Complete Linkage Verfahren wird die Unterschiedlichkeit zweier Cluster am Abstand der beiden am weitesten voneinander entfernten Elemente gemessen, d.h.

$$\tilde{d}_{ij} = \max\left\{d_{pq} \middle| p \in A_i, q \in A_j\right\} \tag{36}$$

Average-Linkage verwendet alle Distanzen zwischen Objekten aus A_i und A_j zur Berechnung des Abstandes von A_i und A_j und stellt einen Kompromiss zwischen Single-Linkage und Complete-Linkage dar. Es werden diejenigen Objekte stärker gewichtet, die zuletzt zu A_i und A_j hinzugekommen sind unter Berücksichtigung der Anzahl der Fusionsschritte, die jedes Objekt bereits durchlaufen hat. Der Clusterabstand wird berechnet, wobei die Anzahl der Elemente $n_i = |A_i|$, $1 \le i \le k$ von Cluster i ist. Man bezeichnet die Anzahl der Fusionsschritte, die die Objekte $O_p \in A_i$ bzw. $O_q \in A_j$ hinter sich haben, mit c_p bzw. c_q, d.h.

$$\tilde{d}_{ij} = \frac{1}{n_i n_j} \sum_{p \in A_i} \sum_{q \in A_j} \frac{1}{2^{c_p + c_q}} d_{pq} \tag{37}$$

Das Weighted-Average-Linkage-Verfahren berechnet stattdessen die Zwischengruppendistanzen ohne Gewichtungsfaktoren, d.h.

$$\tilde{d}_{ij} = \frac{1}{n_i n_j} \sum_{p \in A_i} \sum_{q \in A_j} d_{pq} \tag{38}$$

Die Verfahren (33) bis (36) arbeiten jeweils mit einer Distanzmatrix $D = \left(d_{ij}\right)$, die beliebig definierte Distanzen enthalten kann. Die folgenden 3 Verfahren erwarten als wesentliche Voraussetzung, dass die vorgegebene Matrix D die quadrierte Euklidische Distanz $d_{ij} = \sum_{l=1}^{m} \left(x_{il} - x_{jl}\right)^2$ enthält, andernfalls werden die Ergebnisse u.U. unbrauchbar und die einzelnen Fusionierungsschritte unverständlich.

[186] Vgl. BERGS [1981, S. 27 ff.]

Das Median-Verfahren misst die Unterschiedlichkeit zweier Cluster A_i und A_j an der Distanz zwischen zwei Repräsentanten x_i^* und x_j^* mit $x_i^*,\ x_j^* \in R^m$, d.h.

$$\tilde{d}_{ij} = d\left(x_i^*, x_j^*\right) \tag{39}$$

Ein Cluster mit nur einem Element wird von diesem Element selbst repräsentiert und bei der Fusionierung zweier Cluster i und j wird der Repräsentant x^* des neuen Clusters aus x_i^* und x_j^* berechnet, d.h.

$$x^* = \tfrac{1}{2}\left(x_i^* + x_j^*\right) \tag{40}$$

Das Centroid-Verfahren verwendet einen Gruppenmittelwert (Centroid) $\bar{x}_j$, also

$$\bar{x}_j = \tfrac{1}{n_j} \sum_{i \in A_j} x_i \tag{41}$$

als Vertreter des Clusters zur Berechnung der Abstände, und die unterschiedliche Mächtigkeit der Cluster wird anders als beim Median-Verfahren berücksichtigt, d.h.

$$\tilde{d}_{ij} = d\left(\bar{x}_i, \bar{x}_j\right) \tag{42}$$

Beim WARD-Verfahren wird das so genannte Varianzkriterium als Maß für die Heterogenität der Cluster minimiert, d.h.

$$K_1(\alpha, X) = \sum_{j=1}^{k} \sum_{\mathrm{i} \in \mathrm{A}_j} \left\| x_i - \bar{x}_j \right\|^2 \tag{43}$$

Es gilt

$$\left\| x_i \right\| = \sqrt{\sum_{j=1}^{m} x_{ij}^2} \tag{44}$$

die euklidische Norm des R^m und folglich ist

$$\left\| x_i - x_j \right\| = \sqrt{\sum_{p=1}^{m} \left(x_{ip} - x_{jp}\right)^2} \tag{45}$$

die euklidische Distanz von x_i und x_j. Es ergibt sich der durch die Fusion von A_i und A_j bewirkte Zuwachs des Varianzkriteriums als Distanz, d.h.

$$\tilde{d}_{ij} = \frac{n_i n_j}{n_i + n_j} \left\| \bar{x}_i - \bar{x}_j \right\|^2 \tag{46}$$

Mit Hilfe einer Formel von LANCE/WILLIAMS[187] lassen sich für die Verfahren 1 bis 7 durch Nutzung der Rekursionsformel weitere Gruppierungsalgorithmen konstruieren.

[187] Vgl. LANCE, G.N., WILLIAMS, W.T. [1966, S. 373-380]

Zwei agglomerative Algorithmen werden in Tabelle 2-10 genauer betrachtet.

	SINGLE-LINKAGE[188]	**WARD**[189]
Kennzeichen	Auf jeder Agglomerationsstufe werden die beiden Klassen vereint, in denen sich die zueinander am nächsten liegenden Nachbarobjekte befinden. Für die Fusion genügt eine einzige Distanz.	Diejenigen beiden Gruppen werden vereint, die fusioniert bei der Innenklassenverschiedenheit den geringsten Zuwachs aufweisen. Das Maß der Heterogenität bildet die so genannte Fehlerquadratsumme.
Verschiedenheitskoeffizient	$\tilde{d}_{ij} = \min\left\{d_{pq} \middle\| p \in A_i, q \in A_j\right\}$	$V_g = \sum_{k=1}^{K_g}\sum_{j=1}^{J}\left(x_{kjg} - \bar{x}_{jg}\right)^2$, wobei x_{kjg} ... Wert der Variablen j bei Objekt k (für alle Objekte $k = 1,...,K_g$ in Gruppe g). x_{jg} ... Mittelwert der Werte der Variablen j in Gruppe g.
Eigenschaften	Nach SITTERBERG[190] führt das Verfahren oft zu lang gestreckten Clustern, so dass der Interpretationsgrad schwierig ist. Einzelne Objekte werden tendenziell mit bereits existierenden Klassen zusammengefasst (Ketteneffekt). Im Ergebnis stehen oft sehr große Gruppen praktisch nur Einzelobjekten gegenüber. FORST[191] nennt einen „chaining effect", d.h. zwei recht unterschiedliche Klassen werden fusioniert, nur weil ein einziges Objekt genau zwischen ihnen liegt. Die Gruppentrennung ist nicht möglich.	Nach KADAS[192] liefert WARD aus rechentechnischer Erfahrung in vielen Anwendungsfällen das beste Ergebnis. Laut VOGEL[193] zählt das Verfahren zu den leistungsfähigsten und wirtschaftlichsten Klassifikationsverfahren". Auch WISHART[194] sagt: „possibly the best of the hierarchy options". MÖLLERS/SITTERBERG[195] verweisen darauf, dass eine starke Tendenz zur Bildung gleich großer und kompakter Gruppen vorliegt. Nach FORST[196] werden kleinere Gruppen oder Ausreißer den größeren Gruppen angegliedert.
Ergebnis	Das Verfahren ist geeignet zum Test auf multivariate Ausreißer.[197] Das Verfahren ist in der Regel nicht geeignet, eine sinnvolle Gruppierung der Untersuchungsgesamtheit vorzunehmen.	Aufgrund des Dendrogramms allein lässt sich nach SITTERBERG[198] nicht entscheiden, wie viele Gruppen gebildet werden sollen. Erst mit der Darstellung im Merkmalsraum ist eine zweckmäßigere Aufteilung möglich.

Tabelle 2-10: Das Verfahren WARD und Single-Linkage

In der Literatur ist kein Konsens zum besten agglomerativen Verfahren zu finden. Unter der Voraussetzung metrischer Daten und homogener Clusterstruktur bevorzugen einige Autoren[199] das WARD-Verfahren, während andere Autoren[200] Complete-Linkage vorziehen.

188 Vgl. Das Verfahren geht zurück auf FLOREK et al. [1951], SNEATH [1957] und MCQUITTY [1957], LANCE/WILLIAMS [1966] nennt es Nearest-Neighbour- und JOHNSON [1967] die Minimum-Methode.
189 Vgl. WARD, J.H. [1963, S.237]
190 Vgl. SITTERBERG [1977, S. 51, 52]
191 Vgl. FORST [1974, S. 31]
192 Vgl. KADAS [1981, S. 10]
193 Vgl. VOGEL [1975, S. 332]
194 Vgl. WISHART [1978, S. 33]
195 Vgl. MÖLLERS, H. [1977, S. 139]
196 Vgl. FORST [1974, S. 32]
197 Vgl. F.J. ROHLF [1975, S. 92-101]
198 Vgl. SITTERBERG [1977, S. 61)
199 Vgl. VOGEL [1975], MOJENA [1977], KUIPER/FISHER [1975] und BLASHFIELD [1976]
200 Vgl. CUNNINGHAM/OGILVIE [1972]

2.3.2.2 Iterative partitionierende Clusterbildung

Die iterativen partitionierenden Verfahren[201] sind dadurch gekennzeichnet, dass jedes Element von Cluster zu Cluster beliebig verschoben werden kann, während bei den hierarchischen Verfahren ein einmal konstruiertes Cluster nicht aufgelöst werden kann. Von Vorteil ist eine größere Variabilität, als Nachteil wirkt sich die Voraussetzung kugeliger Klassen[202] aus. BECHER[203] stellt fest: „Die partitionierenden Verfahren verbessern eine bereits bestehende Partition im Sinne der Zielfunktion durch schrittweise Umgruppierung einzelner Objekte." Die Verbesserung der Gruppenbildung erfolgt mit Hilfe eines Beurteilungskriteriums. Als beendet gilt der iterative Gruppierungsprozess, wenn alle Objekte bezüglich ihrer Verlagerung untersucht wurden und sich keine Verbesserung des Kriteriums erreichen lässt. Man spricht dann von einem zumindest lokal erreichten Optimum.[204] In der Praxis ist eine Kombination von hierarchischen und partitionierenden Verfahren üblich. Der Ablauf der Algorithmen wird zusammenfassend in Abbildung 2-11 dargestellt: 1. Determination der Clusterzahl, 2. Vorgabe der Anfangspartition, 3. Wahl des Distanzmaßes und des Gütekriteriums, 4. Bestimmung der Zuordnungsregel für die einzelnen Objekte bezüglich der Cluster, 5. Stoppregel für die Beendigung des Iterationsprozesses.

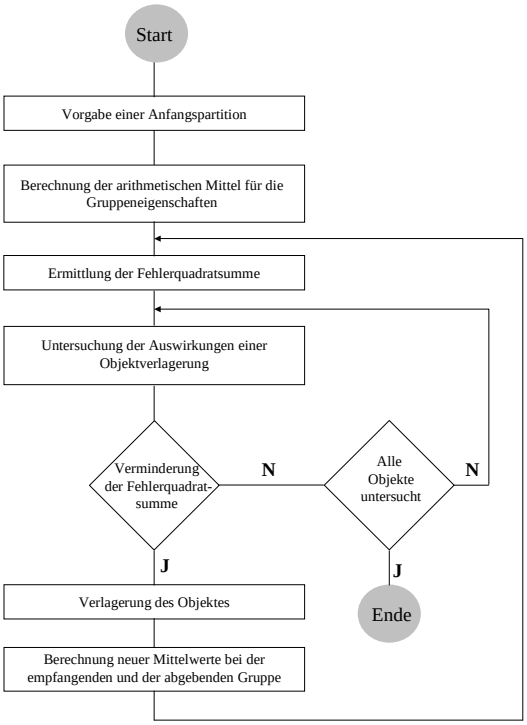


Abbildung 2-11: Ablauf des Austauschverfahrens von RUBIN[205]

201 Als Begründer der iterativen Verfahren sind zu nennen: RUBIN, J. [1967, S. 103-144], FRIEDMAN, H.P. und RUBIN, J. [1967, S. 1159-1178] sowie BALL und HALL [1965] – Hard-ISODATA-Verfahren.
202 Vgl. BOCKLISCH [1987, S.69]
203 Vgl. BECHER [1995, S. 34]
204 SITTERBERG [1977, S. 66]
205 Eigene Bearbeitung nach SCHUCHARD-FICHER et al.[1985, S. 142]

2.3.3 Probabilistische und Possibilistische Clusteranalyseverfahren

In einigen Anwendungsfällen ist festzustellen, dass es sich um eine unnötig scharfe Restriktion handeln kann, jedes Objekt einem einzigen Cluster zuordnen zu wollen und bei konventionellen iterativen Gruppierungsmethoden und -algorithmen wurde bereits erkannt, dass sich Objekte ggf. nicht eindeutig einem der Cluster zuordnen lassen. Im Jahr 1965 veröffentlicht LOFTI ZADEH[206] das Konzept der unscharfen Mengen und schafft damit einen Ansatz zum Umgang mit Vagheit und legt den Grundstein für die unscharfen Techniken in der Mustererkennung, zu der die Clusteranalyse im weitesten Sinne gehört.[207] Im Rahmen der der Fuzzy-Theorie[208] wird dem Umgang mit gewissen Unsicherheiten, z.B. durch Messfehler oder fehlende Informationen Rechnung getragen. Die entstehenden Beschreibungen bezeichnet man als unscharfe Objekte. Neben der Berücksichtigung der Unsicherheiten einer Beobachtung in Form unscharfer Objekte können weiteren Unsicherheiten (Lage der Objekte zueinander, wechselnde Umgebungsbedingungen usw.) im Rahmen eines (globalen) unscharfen Modells Rechnung getragen werden.

2.3.3.1 Partitionsbedingung

Die probabilistischen Clusteranalyseverfahren berechnen für ein Klassifikationsobjekt i einen Zuordnungsgrad μ_{ic} (Möglichkeitsgrad), der die Möglichkeit angibt, mit der dieses einem Cluster c angehört. Die Interpretation klassischer Partitionen fällt zwar zunächst leichter, jedoch besteht oftmals die Gefahr der Fehlinterpretation. Die Zuordnung der Objekte zu den Clustern erfolgt im Vertrauen auf die Exaktheit der zugrunde liegenden mathematischen Verfahren, erfordert aber ein zusätzliches kritisches Verständnis. DEIMER[209] sagt: „Der Übergang von einem Cluster zu einem anderen [ist] in der Regel mehr gradueller als abrupter Natur.“ Mit Hilfe der deterministischen Cluster-Algorithmen wird die Zuordnung von Objekten zu einem bestimmten Cluster aufgrund sehr geringer Differenzen in den Distanzen des Objektes zu verschiedenen Clustern festgelegt.[210] Die gefundenen Partitionen sind nach BECHER[211] wenig robust, da selbst kleine Änderungen in den Variablen der Objekte die Clusterzuordnung beeinflussen können. Die deterministischen Clusteranalysemethoden

---

206 Vgl. eine Ergänzung zur Theorie der unscharfen Mengen findet sich im Nebenteil B.

207 Unscharfe Mengen als Basis für Clusteralgorithmen wurden zuerst von BELLMANN, KALABA UND ZADEH [1966] angewendet. WEE [1967] nutzt die Mechanismen der linearen Entscheidungstheorie zur Begründung einer Zugehörigkeitsfunktion. FLAKE und TURNER [1968] sowie GITMAN und LEVINE [1970] diskutieren Clustermethoden, die durch Schwellenwerte unscharfe Teilmengen identifizieren. RUSPINI [1969,1970] entwarf die erste objektive Funktionsmethode der unscharfen Clusteranalyse.

208 Vgl. ZADEH [1965], im Nebenteil B ist eine Ergänzung zur Theorie der unscharfen Mengen zu finden.

209 Vgl. DEIMER [1986, S. 115]: „Unscharfe Clusteranalysemethoden“

210 Siehe Abschnitt 2.3.2

211 Vgl. BECHER [1995, S. 35]

erzielen lediglich dann einigermaßen robuste Resultate, wenn bereits eine natürliche Clusterstruktur gegeben ist, d.h. theoretische Überlegungen eine bestimmte Gruppierung der Objekte nahe legen. Ein unscharfer Clusteralgorithmus ist geeignet, um schlecht strukturierte Objektmengen zu klassifizieren bzw. wird dieser eingesetzt, wenn sich nicht alle Objekte eindeutig genau einem Cluster zuordnen lassen (Brückenobjekte).[212] Die Zugehörigkeit der Objekte zu einzelnen Clustern wird mit Hilfe der Distanz zu den jeweiligen Clustern bestimmt. Je geringer die Distanz eines Objekts zu einem Cluster ist, desto eher besteht die Möglichkeit, dass dieses Objekt diesem Cluster zugeordnet werden kann und somit geringere Zugehörigkeitsgrade zu den anderen Clustern besitzt. Die Distanz wird wie bei den klassischen Clusteranalyseverfahren auf unterschiedliche Art definiert. In der probabilistischen Clusteranalyse gelten einige Restriktionen (Tabelle 2-11). Ein Zugehörigkeitsgrad μ_{ic}=0 bedeutet, dass es unmöglich ist, das Objekt i dem Cluster c zuzuordnen und μ_{ic}=1 bedeutet, dass in keiner Weise eingeschränkt wird, dass das Objekt i dem Cluster c zuzuordnen ist. Es ist zu beachten, dass eine Interpretation der Zugehörigkeitsgrade μ_{ic} zu ungenauen bzw. fehlerbehafteten Aussagen führen kann, wenn trotzdem nicht weitere Informationen, z.B. über die Lage der Cluster, mit berücksichtigt werden.

Bedingungen für eine unscharfe Partition der probabilistischen Clusteranalyse	
$\bigwedge_{\substack{i=1...n \\ c=1...p}} 0 \le \mu_{ic} \le 1,\ \bigwedge_{i=1...n} \sum_{c=1}^{p} \mu_{ic} = 1,\ \bigwedge_{c=1...p} \sum_{i=1}^{n} \mu_{ic} > 0$	**mit** n = Anzahl der Objekte, p = Anzahl der Cluster, μ_{ic} = Zugehörigkeitsgrad des Objekts i zum Cluster c (Normierung der Zugehörigkeiten je Objekt).

Tabelle 2-11: Partitionsbedingungen bei der probabilistischen Clusteranalyse[213]

Eine Alternative zu probabilistischen Zugehörigkeitsgraden stellt die Verwendung possibilistischer Zugehörigkeitsgrade dar.[214] Auf die Restriktion $\bigwedge_{i=1...n} \sum_{c=1}^{p} \mu_{ic} = 1$ wird verzichtet und zusätzlich eine leicht veränderte Zielfunktion verwendet (siehe 2.3.3.3). Im Gegensatz zu der probabilistischen Clusteranalyse wird bei der possibilistischen Clusteranalyse der Zugehörigkeitsgrad für ein Klassifikationsobjekt i zu einem Cluster c nur aus der Relation des Abstands zu dem betreffenden Cluster bestimmt. Der Abstand zu den anderen Clustern wird nicht berücksichtigt. Im Sinne der Possibilitätstheorie[215] repräsentieren die Zugehörigkeitsgrade die Möglichkeit, dass ein Objekt zu dem entsprechenden Cluster gehört.

---

212 Vgl. DEIMER [1987, S. 116] und BEZDEK, J.C. [1974, S. 58]
213 Eigene Bearbeitung unter Einbeziehung von QU [2000, S. 34]
214 KRISHNAPURAM / KELLER [1993, S. 98-110]
215 DUBOIS / PRADE [1988]

Tabelle 2-12 enthält Beispiele, die die Interpretationsschwierigkeiten bei der Objektzuordnung bei deterministischer als auch unscharf beschriebener Objektzuordnung berücksichtigen. Bei der probabilistischen Clustereinteilung gibt der Zugehörigkeitsgrad eher die relative Zuordnung eines Objekts zu einem Cluster an, während bei der possibilistischen Clustereinteilung aufgrund der direkten Abstandsrelation der Zugehörigkeitsgrad ein Maß darstellt, wie typisch ein Objekt für ein Cluster ist.

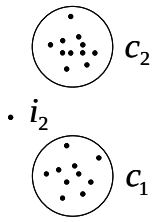
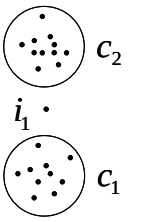
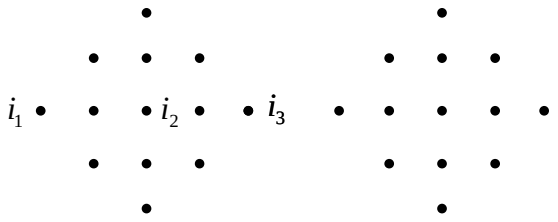
Brückenobjekt	Brückenobjekt, Stördaten	Separierte Struktur (2 Cluster)
		

Tabelle 2-12: Interpretationsschwierigkeit bei der Objektzuordnung

Die Zugehörigkeitsgrade der beiden Daten i_1 und i_2 (mittlere Abbildung, Tabelle 2-12) sind bei einem probabilistischen Zugehörigkeitsgrad zu den beiden Clustern c_1 und c_2 jeweils 0,5. Es findet keine Unterscheidung statt, dass das Objekt i_1 eher beiden Clustern angehört, während das Objekt i_2 eher als Störddatum bezeichnet werden kann und keinem der beiden Cluster angehören sollte.[216]

Die possibilistische Beschreibung ist besser geeignet, Störddaten und Ausreißer während der Klassifizierung zu behandeln und auch nichtdisjunkte Klassen lassen sich aufgrund der stärkeren Orientierung an der Form der Daten besser interpretieren. Bei einer probabilistischen Beschreibung würden die Objekte i_1 und i_3 (rechte Abbildung, Tabelle 2-12) einen unterschiedlichen Zugehörigkeitsgrad erhalten, obwohl beide den gleichen Abstand zum Clusterzentrum des linken Clusters haben, welches sich beim Objekt i_2 befindet. Die Robustheit von Fuzzy-Clusterverfahren lässt sich durch Einsatz von possibilistischen Clusterverfahren erhöhen.[217] Im Gegensatz zu probabilistischen Fuzzy-Clusterverfahren, die einen Datensatz partitionierend aufteilen, können bei possibilistischen Fuzzy-Clusterverfahren auch Cluster identisch sein. Der Grund hierfür besteht darin, dass bei possibilistischen Verfahren nicht berücksichtigt wird, ob und inwieweit Daten schon Clustern zugeordnet wurden.[218]

[216] Vgl. HÖPPNER et al. [1997, S. 18/19]: Normierung und Zugehörigkeitsgrade
[217] Vgl. Benutzerdokumentation der MIT GmbH [1998]
[218] Vgl. TIMM [2002, S. 45 ff.]

2.3.3.2 Konfigurationsparameter

Im Wesentlichen ist das Ergebnis der Fuzzy-Clusteranalyse von einigen Konfigurationsparametern beeinflusst: Distanzmaß,[219] Zielfunktion, Gewichtungsexponenten, Anzahl der Cluster, Startpartition, Eingabereihenfolge, Genauigkeitsschwelle.

Nach BECHER[220] führt ein unscharfer Klassifikationsalgorithmus „...im allgemeinen nicht zum globalen Optimum, sondern zu einem von der Startpartition und der Eingabereihenfolge abhängigen lokalen Minimum der Zielfunktion." Durch mehrfache Anwendung des Verfahrens mit unterschiedlichen Ausgangspartitionen und Eingabereihenfolgen lässt sich nach BECHER dieser Einfluss verringern. Um die Ausgangspartition zu definieren, bieten sich verschiedene Möglichkeiten an.[221] Zum einen ist die Verwendung aktueller bzw. bereits bestehender Klassenzentren möglich, zum anderen ist eine benutzerdefinierte Zuordnung denkbar und weiterhin ist eine zufällige Belegung der Ausgangsmatrix erlaubt, die durch softwarespezifische Trainingswerkzeuge optimiert wird.[222]

Mit Hilfe des Gewichtungsexponenten r wird die Unschärfe der resultierenden Partition festgelegt. Für $r = 1$ erhält man eine scharfe Partition und mit wachsendem Gewichtungsexponenten wird die Partition unschärfer, d.h. der Grenzübergang $r \to \infty$ spiegelt die maximale Unschärfe wieder und die Zugehörigkeitsgrade der Objekte sind zu allen Clustern gleich. BOCK[223] schlägt Exponent zwischen 2,0 und 3,0 vor. Bei einigen Klassifikationen[224] ist es möglich, dass der Schärfegrad weiter heruntergesetzt werden muss, um gut interpretierbare Klassenstrukturen zu erhalten.

Die Anzahl der Cluster wird maßgeblich mit Hilfe von Gütekriterien festgelegt, falls keine natürliche Clusterstruktur vorliegt, die auf eine bestimmte Clusterzahl hindeutet. Es ist davon auszugehen, dass eine Clusterstruktur umso besser identifiziert wird, je stärker sich die Punkte um ein Clusterzentrum konzentrieren. Es gilt dann die Bedingung, dass der Partitionskoeffizient (pk) und der Proportionsexponent (pex) möglichst groß und die Werte der Partitionsentropie (pe) möglichst klein sind.

---

219 In DataEngine® wird als Distanzmaß der unscharfen Clusteranalyse die Euklidische Distanz vorgegeben. DataEngine® stand kostenlos zur Verfügung und ermöglicht eine unscharfe Clusteranalyse.

220 Vgl. BECHER [1995, S. 39]

221 Initialisierungsmöglichkeiten im Programm DataEngine, vgl. ZIMMERMANN [1995a, S. 150]

222 Vgl. BOCK [1979, S. 156] und BECHER [1995, S. 46/51]: „Führen unterschiedliche Startpartitionen immer zur gleichen unscharfen Partition, ist die berechnete Lösung also stabil, so kann dies als Indiz für eine gute Clusterstruktur der Objektmenge und die Wahl der richtigen Clusteranzahl gewertet werden."

223 Vgl. BOCK [1979, S. 144]

224 Diese Vorgehensweise wird bei BECHER [1995, S. 133] beschrieben und angewendet. Die Ursache des von ihm festgestellten schlechten Konvergenzverhaltens vermutet BECHER in der relativ großen Zahl nur schwach korrelierter Variablen in Verbindung mit einer fehlenden natürlichen Clusterstruktur.

Genaue Informationen über die Anzahl der Cluster gewinnt man durch Vergleich von verschiedenen Partitionen und die sich daran anschließende grafische Darstellung der Gütemaße. Mit Hilfe der grafischen Darstellungen lassen sich Hinweise auf eine geeignete Klassenzahl ermitteln.[225] Es werden die Gütemaßwerte über die Klassenanzahl aufgetragen.[226] Nach ZIMMERMANN[227] befindet sich mit Blick auf die Klassifikationsentropie eine geeignete Klassenzahl c an der Stelle, für die beim Übergang c-1 nach c der Wert der Entropie unter dem Trend der Kurve liegt. Eine weitere Möglichkeit besteht nach ZIMMERMANN[228] darin, den Partitionskoeffizienten und die Klassifikationsentropie in Abhängigkeit von der Klassenzahl hinsichtlich des monotonen Verhaltens zu untersuchen. Eine sinnvolle Klassenzahl c* liegt an der Stelle, für die beim Übergang von der Klassenzahl c*-1 zur Klassenzahl c* der Wert für die Entropie unter den steigenden Trend fällt bzw. der Wert des Partitionskoeffizienten über den fallenden Trend steigt.

Als weitere Parameter sind die Abbruchkriterien der unscharfen Clusteranalyse zu nennen. Die Genauigkeitsschwelle $\varepsilon_{\min}$ entscheidet über das Fortsetzen des Algorithmus. Solange sich wenigstens einer der Zugehörigkeitsgrade während einer Iteration um mehr als den Betrag verändert, wird der Algorithmus weitergeführt. Wird die Grenze nicht mehr überschritten, endet der Klassifikationsprozess und die Endpartition steht fest. Als Abbruchgenauigkeit dient in einigen Arbeiten[229] ein Wert von $\varepsilon_{\min} = 0{,}001$ bis $\varepsilon_{\min} = 0{,}00001$.

Das zweite Abbruchkriterium bildet der vom Anwender festzulegende Wert über die Höchstzahl der Iterationen. Dieser Parameter dient nur dazu, aussichtslose Rechenprozesse frühzeitig abbrechen zu können. Dies ist für den Fall notwendig, dass eine Objektmenge mit fehlender Clusterstruktur zur Untersuchung vorliegt und eine lange Rechenzeit erforderlich würde, ohne Ergebnisse zu erhalten. Generell ist die Anzahl der Iterationen abhängig von der Startpartition. Je schärfer die Startpartition ist, desto weniger Iterationen werden benötigt, um die Endpartition zu berechnen.[230]

---

225 Vgl. DEIMER [1986, S. 159]: „Bei der Interpretation der Validitätskriterien muss beachtet werden, dass sich aus ihnen keine allgemeingültigen Aussagen ableiten lassen. Sie bestimmen nur eine so genannte ‚heuristische Validität', so dass bei der Strukturanalyse einer unscharfen Partition auf keines der drei Kriterien verzichtet werden kann, ohne einen Informationsverlust hinzunehmen"

226 Vgl. BECHER [1995, S. 135]: „Eine günstige Clusterzahl ist dadurch gekennzeichnet, dass der Partitionskoeffizient gegenüber der nächstniederen Clusterzahl schwächer abfällt als bei einer Glättung des Kurvenverlaufs zu erwarten wäre. Gleichzeitig sollte der Anstieg der Klassifikationsentropie unterdurchschnittlich und der Anstieg des Proportionsexponenten überdurchschnittlich sein."

227 Vgl. ZIMMERMANN [1995a, S. 129]

228 Vgl. ZIMMERMANN [1995a, S.154]

229 Vgl. BECHER [S.134], QU [2000, S. 59], ZIMMERMANN [1995a, S.151]

230 Vgl. BOCK [1979, S. 156]

2.3.3.3 Klassenbildung

Die Fuzzy-Clusteranalyse zählt zu den zielfunktionsbasierten Klassifikationsverfahren, d.h. das Klassifikationsproblem wird durch eine Zielfunktion beschrieben, die unter Berücksichtigung von Restriktionen zu optimieren ist.[231] Ein probabilistisches Fuzzy-Clusterverfahren löst das Optimierungsproblem im Allgemeinen wie folgt:

$$\text{Minimiere } J(X,U,\beta)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\cdot d^2\left(\vec{\beta}_i,\vec{\chi}_j\right) \text{ unter Berücksichtigung der Restriktionen}$$

$$\sum_{j}^{n}=1\ u_{ij}>0 \text{ für alle } i\in\{1,...,c\} \text{ und } \sum_{i}^{c}=1\ u_{ij}=1 \text{ für alle } j\in\{1,...,n\} \qquad (47)$$

Dabei ist $U=\left\{u_{ij}\middle| i\in\{1,2,...,c\}, j\in\{1,2,...,n\}\right\}$ die Menge der Zugehörigkeitsgrade der Daten zu den Klassen, $\beta=\{\beta_1,\beta_2,...\beta_c\}$ die Menge der Klassen und $d(\beta_i,\chi_j)$ der Abstand zwischen der Klasse i und dem Objekt j. Der Exponent $m\in(1,\infty)$ bestimmt, wie stark Daten, die einer Klasse nur mit einer geringen Zugehörigkeit zugeordnet wurden, diese Klassen beeinflussen. Ein Datensatz wird klassifiziert, indem ausgehend von einer Anfangszuordnung abwechselnd die Klassen und Zugehörigkeitsgrade der Daten zu Klassen berechnet werden, wobei der alternierende Berechnungsprozess von Klassen und Zugehörigkeitsgraden solange wiederholt wird, bis sich die Zugehörigkeitsgrade nicht mehr wesentlich ändern:

$$u_{i,j}=\begin{cases}\dfrac{1}{\sum_{k=1}^{c}\left(\dfrac{d^2\left(\vec{x}_j,\vec{\beta}_i\right)}{d^2\left(\vec{x}_j,\vec{\beta}_k\right)}\right)^{\frac{1}{m-1}}}, & \text{falls } I_j=0\\ 0, & \text{falls } I_j\neq 0 \text{ und } i\notin I_j\\ x, x\in[0,1] \text{ , so daß } \sum_{i\in I_j}u_{i,j}=1 \text{ gilt,} & \text{falls } I_j\neq 0 \text{ und } i\in I_j\end{cases} \qquad (48)$$

Die possibilistische Fuzzy-Clusteranalyse verwendet eine modifizierte Zielfunktion, die mathematisch nachfolgend beschrieben wird:

$$\text{Minimiere } J(X,U,\beta)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\cdot d^2\left(\vec{\beta}_i,\vec{\chi}_j\right)+\sum_{i=1}^{c}\eta_i\sum_{j=1}^{n}\left(1-u_{i,j}\right)^m$$

$$\text{unter Berücksichtigung der Restriktion } \sum_{j}^{n}u_{ij}>0 \text{ für alle } i\in\{1,...,c\}. \qquad (49)$$

Der Parameter $\eta_i\in IR_{>0}$ gibt den Abstand an, bei dem der Zugehörigkeitsgrad zu jedem Cluster $\beta_i=\frac{1}{2}$ betragen soll. Weiterhin werden die Zugehörigkeitsgrade $u_{i,j}$ berechnet durch:

$$u_{i,j}=\frac{1}{1+\left(\dfrac{d^2\left(\vec{x}_j,\beta_i\right)}{\eta_i}\right)^{\frac{1}{m-1}}} \qquad (50)$$

[231] Vgl. TIMM [2002, S. 14 ff.], HÖPPNER et al. [1999] und BORGELT et al. [2003, S. 170 ff.]

Der Fuzzy-C-Means-Algorithmus[232] ist einer der bekanntesten Algorithmen, der in verschiedenen Mustererkennungssystemen[233] Einsatz gefunden hat. Der Ablauf des Fuzzy-C-Means-Algorithmus ist ausführlich wie folgt definiert:[234]

1. Initialisiere die Zugehörigkeiten μ_{ik} des k-ten Objektes x_k zur i-ten Klasse für alle $i = 1,...,c$ und $k = 1,...,K$, wobei gelten muss:
   $$\sum_{i=1}^{c} \mu_{ik} = 1 \quad \forall k = 1,...,,K \text{ und } \mu_{ik} \in [0,1] \qquad \begin{matrix} \forall i = 1,...,c \\ \forall k = 1,...,k \end{matrix}$$
2. Berechne aus den aktuellen Zugehörigkeiten μ_{ik} die Klassenschwerpunkte v_i der Klassen: $$v_i = \frac{\sum_{k=1}^{K} (\mu_{ik})^m \cdot x_k}{\sum_{k=1}^{K} (\mu_{ik})^m}, \quad \forall_i = 1,...,c$$
3. Berechne aus den aktuellen Zugehörigkeiten μ_{ik}^{neu} die Klassenschwerpunkte v_i:
   $$\mu_{ik}^{neu} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|v_i - x_k\|}{\|v_j - x_k\|} \right)^{\frac{2}{m-1}}}, \qquad \begin{matrix} \forall i = 1,...,c \\ \forall k = 1,...,K \end{matrix}$$
4. Falls $\left\| \mu^{neu} - \mu \right\| > \in$, setze $\mu = \mu^{neu}$ und fahre fort mit Schritt 2.

Für die Berechnung der Vektorabstände in Schritt 3 ist ein geeignetes Abstandsmaß zu wählen. Das Verfahren ist beendet, falls der Abstand zweier aufeinander folgender Zugehörigkeitsmatrizen μ kleiner ist als eine anzugebende Konvergenzschwelle ε. Für diese Abstandsberechnung muss eine geeignete Matrixnorm gewählt werden. Das Verfahren terminiert über den Vergleich zweier aufeinander folgender Klassenschwerpunktsmatrizen, als Matrixnorm wird die Summe der komponentenweisen Abstände verwendet. Der Fuzzy-C-Means liefert neben der Lage der Klassenschwerpunkte mit Hilfe der Formel aus Schritt 3 die Zugehörigkeitswerte der einzelnen Objekte zu den verschiedenen Klassen. Mittels dieser Formel können bei einer Klassifikation neuer Objekte die Zugehörigkeiten der Objekte zu den durch die Klassenschwerpunkte vorgegebenen Klassen berechnet werden.[235]

---

232 Vgl. BEZDEK, J.C. [1981] und BEZDEK, J.C. [1993]

233 Das Programm DataEngine® ist ein Software-Produkt für die effektive Datenanalyse und zum Data Mining (Vgl. ROIGER, R und GEATZ, M. [2003]). Es wurde eine zeitbegrenzte Studentenversion der MIT GmbH, Aachen - Version 4.01 verwendet. Des Weiteren ist der Algorithmus in einer Toolbox von Matlab enthalten.

234 Vgl. QU [2000, S. 37] und BECHER [1995, S.39]

235 In DataEngine® wird eine leicht modifizierte Version des Algorithmus verwendet, die bezüglich des Laufzeit- und Speicherplatzbedarfes Vorteile gegenüber dem oben beschriebenen Verfahren bietet. Der modifizierte Algorithmus ist äquivalent zu dem oben beschriebenen Originalalgorithmus. Anstelle der Zugehörigkeitsmatrix wird die Matrix der Klassenschwerpunkte *v* initialisiert. Aus diesem Grund sind die Schritte 2 und 3 des obigen Algorithmus vertauscht. Als Abstandsmaß dient in diesem Programm der euklidische Abstand, siehe MIT GmbH [2000]

Tabelle 2-13 gibt einen Überblick ausgewählter unscharfer Clusterverfahren.[236] Ob eine Partition brauchbar ist, hängt grundsätzlich von deren Interpretierbarkeit ab.

Algorithmus	**Klassenform und -größe**	**Beschreibung**
Fuzzy-C-Means[237]	Zugeschnitten auf die Unterteilung eines Datensatzes in kreis- bzw. kugel- oder hyperkugelförmige (sphärische) Klassen gleicher Größe.[238] Der Algorithmus ist für eine feste Anzahl von Clustern vorgesehen, wobei diese nicht selbstständig bestimmt wird.	Verzicht auf die Berechnung der Form und der Größe der einzelnen Klassen. Gefahr einer Fehlklassifikation besteht für Datensätze mit schlecht separierten Klassen, die nicht kreisförmig sind. Jedes Cluster wird durch seinen Mittelpunkt dargestellt. Die Repräsentation des Clusters wird auch Prototyp genannt, da sie oft als Stellvertreter aller zugeordneten Daten angesehen wird. Der Algorithmus zeichnet sich durch eine einfache Berechnungsvorschrift aus (kurze Rechenzeit).
Gustafson-Kessel-Algorithmus (G-K-A)[239]	Der Algorithmus verwendet zusätzlich zum Clusterzentrum so genannte Geometrieparameter, die die Form und Ausdehnung der Datenvektoren der dem Cluster zugeordneten Datenpunkte beschreiben. Variable Clusterform, d.h. ellipsoide Klassen können auch erkannt werden. Die Erkennung unterschiedlicher Clustergrößen setzt Vorwissen über die Cluster voraus (Festlegung einer Konstante).	Bestimmung der Form erfolgt individuell für jede Klasse, und es besteht höhere Flexibilität (Berechnung einer Kovarianzmatrix für jedes Cluster). Bei höherdimensionalen Datensätzen existiert eine längere Rechenzeit. Es empfiehlt sich zur Verringerung der Iterationsschritte die Initialisierung mit den Prototypen eines vorangegangenen Fuzzy-C-Means-Durchlaufes.
Gath&Geva-Algorithmus (G&G-A)[240]	Es existiert eine hohe Flexibilität durch individuell bestimmte Form und Größe jeder Klasse. Erkennung auch von ellipsoiden Klassen verschiedener Größe und Dichte.	Die Zugehörigkeiten zeigen oftmals eine sehr genaue Trennung der Datenstruktur (Berechnung einer Kovarianzmatrix für jedes Cluster). Mit zunehmender Komplexität wird der Algorithmus anfälliger in einem lokalen Minimum zu konvergieren. Für die richtige Einteilung müssen die Prototypen bereits in der Nähe der endgültigen Prototypen initialisiert werden, wobei diese Entscheidungen nicht ohne Vorwissen zu treffen ist. Aufgrund der Grundvoraussetzungen ist eine possibilistische Vorgehensweise mit dem Algorithmus nicht durchführbar.
Achsparallele Varianten des G-K-A und des G&G-A[241]	Aufteilung eines Datensatzes in achsparallele Klassen. Form und Größe der Klassen werden individuell bestimmt.	Es wird auf das Invertieren von Matrizen bei der Berechnung verzichtet, so dass eine geringere Rechenkomplexität vorliegt. Die Flexibilität bezüglich der Lage der Klassen ist reduziert. Besser geeignet für die Erzeugung von Fuzzy-Regeln.

Tabelle 2-13: Algorithmen der Fuzzy-Clusteranalyse im Überblick

236 Vgl. MIT (Hrsg.) [1998, S. 5-9], BORGELT / TIMM [2006], HÖPPNER et al. [1997], HÄNDEL [2003]

237 In früheren Veröffentlichungen wurde der Fuzzy-C-Means-Algorithmus auch als Fuzzy-Isodata bezeichnet, vgl. DUBOIS / PRADE [1980, S. 325], HÖPPNER et al. [1997, S. 35]

238 HÖPPNER et al. [1997, S. 37] verweist darauf, dass bei der Bildverarbeitung langgestreckte Klassen bei Verwendung der Euklidischen Distanz durch den Algorithmus Fuzzy c-Means ggf. nicht erkannt werden konnten. Liegen solche Klassenformen vor, empfiehlt sich die Verwendung verallgemeinerter Distanzmaße oder anderer Cluster-Algorithmen. Alternative Verfahren, denen so genannte Hyperebenen als Clusterrepräsentanten bzw. adaptive Distanzen zugrunde liegen, sind in BOCK [1979, S. 158 ff.] kurz beschrieben. Der Rechenaufwand dieser Algorithmen liegt wesentlich höher.

239 Vgl. GUSTAFSON [1979, S. 761-766] und TIMM [2002, S. 19 ff.]

240 Vgl. GATH [1989, S. 773-781]

241 Vgl. KLAWONN / KRUSE [1995] und KLAWONN / KRUSE [1997]

2.3.3.4 Gütekriterien

Die Anzahl der Cluster ist bei den Verfahren der Fuzzy-Clusteranalyse zu Beginn der Klassifizierung vorzugeben. Da oft die Anzahl der Cluster nicht bekannt ist, erhält die Bewertung der Ergebnisse einer Fuzzy-Clusteranalyse große Bedeutung. Man unterscheidet zwischen globalen Gütekriterien, die eine Klassifikation als Ganzes bewerten und lokalen Gütekriterien, bei denen jedes Cluster separat bewertet wird. Zur Bewertung einer Clustereinteilung sollten mehrere Gütekriterien betrachtet werden. Mit dem Zugehörigkeitsgrad $\mu_{ic} \in [0,1]$, Objektanzahl n und Clusteranzahl P werden hier drei Gütekriterien definiert:[242]

- Partitionskoeffizient (partition coefficient) : $pk = \sum_{i=1}^{n}\sum_{c=1}^{p}\frac{\mu_{ic}^{\ 2}}{n}$ (51)
- Partitionsentropie (partition entropy) $pe = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{p}\mu_{ic}\ln\mu_{ic}$ (52)
- Proportionsexponent (proportion exponent) $pex = -\ln\left[\prod_{i=1}^{n}\sum_{j=1}^{\mu_i^{-1}}(-1)^{j+1}\binom{P}{j}(1-j\mu_i)^{(p-1)}\right]$ (53)

  mit $\mu_i = \max\{\mu_{ic}\}$ und
  μ_i^{-1} = die größte ganze Zahl $\leq \frac{1}{\mu_i}$ $\quad i = 1,...,n$ und $c = 1,...,P$

Ein Clusterergebnis gilt umso schärfer, je größer sich der Wert pk bzw. kleiner pe ausbildet. Für $pk = 1$ bzw. $pe = 0$ liegen scharfe Clusterstrukturen vor. Die unschärfste aller Partitionen liegt dann vor, wenn alle Objekte gleichmäßig auf alle Cluster verteilt sind. Dies ist der Fall bei einem Wert von $pk = \frac{1}{P}$ oder $pe = \ln P$. Die Wertebereiche für den Partitionskoeffizienten (pk) und die Partitionsentropie (pe) sind von der Clusterzahl P abhängig. Bei einer großen Clusterzahl werden demzufolge die Wertebereiche größer, so dass sich diese Gütekriterien nicht für den Vergleich zweier Partitionen mit unterschiedlichen Clusterzahlen eignen. Der Proportionsexponent (pex) ist entwickelt worden, um eine Unabhängigkeit von der Clusterzahl P zu gewährleisten. Für vollständig unscharfe Partitionen liefert der Proportionsexponent pex den Wert 0 und wächst mit zunehmender Schärfe der Partition über alle Grenzen. Tabelle 2-14 enthält einige Beurteilungsmöglichkeiten.

Gütemaß	**min**	**max**	**scharfe Trennung**
Partitionskoeffizient (pk)	1/c	1	1
Klassifikationsentropie (pe)	0	ln c	0
Proportionsexponent (pex)	0	∞	maximal

Tabelle 2-14: Gütemaße zur Beurteilung des Klassifikationsergebnisses

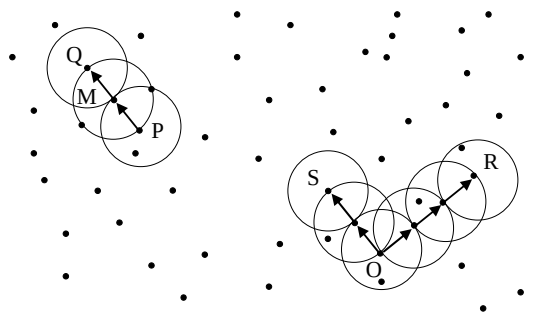
[242] Vgl. DEIMER [1986, S. 148-160], ZIMMERMANN [1993, S. 77], ZIMMERMANN [1995a, S. 43-44] und BECHER [S.40-42], weitere Gütekriterien sind beschrieben bei TIMM [2002, S. 32-36]

2.3.4 Dichtebasierte Clusteranalyseverfahren

2.3.4.1 Grundbegriffe

Die Idee der dichtebasierten Clusteranalyse besteht darin, Regionen im Merkmalsraum zu bestimmen, die eine hohe Anzahl von Objekten (also eine hohe Dichte) aufweisen. Weiterhin sollen diese Regionen durch Bereiche mit einer kleinen Anzahl von Objekten (geringe Dichte) abgegrenzt sein. Es gelten die folgenden Begriffe in Bezug auf dichtebasierte Cluster, die sich auf eine Datenmenge D beziehen (siehe Abbildung 2-12):

- Für jedes Objekt eines Clusters überschreitet die lokale Punktdichte einen gesetzten Grenzwert. Die Menge von Objekten, welche ein Cluster charakterisieren, ist räumlich zusammenhängend. Es gilt für die ε-Umgebung des Punktes p die Definition: $N_\varepsilon(p) = \{q \in D | d(p,q) \leq \varepsilon\}$, wobei die Metrik d die Form der ε-Umgebung festlegt.
- Es ist zwischen Kern- und Randobjekten eines Clusters zu unterscheiden, dabei unterliegt ein Kernobjekt $q \in D$ der Bedingung $|N_\varepsilon(q)| \geq MinPts$ (Minimale Anzahl von Objekten)
- Ein Punkt p ist direkt Dichte erreichbar von einem Punkt q bezüglich ε und $MinPts$, wenn $p \in N_\varepsilon(q)$ und q ein Kernobjekt ist.
- Ein Punkt p ist Dichte erreichbar von einem Punkt q bezüglich ε und $MinPts$, wenn es einen Weg von Punkten $p_1,...,p_n \wedge p_1 = q \wedge p_n = p$ gibt, so dass p_{i+1} direkt Dichte erreichbar von p_i ist.
- Ein Punkt p ist verbunden mit Punkt q bezüglich ε und $MinPts$, wenn Punkt o existiert, so dass sowohl p als auch q von o bez. ε und $MinPts$ erreichbar sind.



M,P,O,R sind Kernobjekte, S und Q nicht.

Q ist direkt Dichte erreichbar von M.
M ist direkt Dichte erreichbar von P.
P ist direkt Dichte erreichbar von M.

Q ist Dichte erreichbar von P.
R und S sind Dichte erreichbar von O.
O ist Dichte erreichbar von R.

O, R, S sind alle Dichte verbunden.

Abbildung 2-12: Begriffe der Dichte-basierten Clusterverfahren

Das Auffinden von dichtebasierten Clustern erfolgt durch Überprüfung aller Objekte hinsichtlich ihrer Eigenschaft als Kernobjekt und der Beurteilung der Dichte-Erreichbarkeit.

2.3.4.2 U*C-Algorithmus

Der Algorithmus U*C[243] wird in seinen Teilschritten dargestellt und danach ausführlich in seinem Ablauf erläutert. Gegeben sei gemäß der Abbildung 2-13 eine U-Matrix, P-Matrix und U*-Matrix sowie die noch zu erläuternde Immersion $I = \{\ \}$;

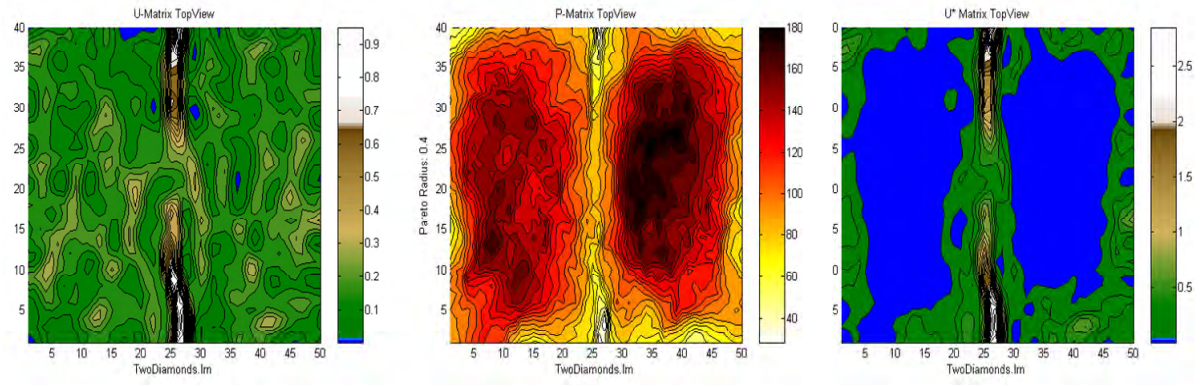


Abbildung 2-13: U-Matrix (Distanzstruktur), P-Matrix (Dichtestruktur), U*-Matrix (Clusterstruktur)

Immersion: Für alle Neuronen n einer Emergenten SOM gilt:

1. Folge vom Neuron n einem Gradientenabstieg auf der U-Matrix bis ein Minimum im Neuron u erreicht ist.
2. Folge vom Neuron u einem Gradientenanstieg auf der P-Matrix bis ein Maximum im Neuron p erreicht wird.
3. $I = I \cup \{p\};\ \text{Immersion}(n) = p$

Clusterzuordnung:

1) Berechne die Wasserscheiden (Watersheds) für die U*-Matrix
2) Partitioniere I unter Verwendung der Wasserscheiden innerhalb der Cluster $C_1,...,C_c$
3) Ordne den Datenpunkt x zu einem Cluster C_j zu, wenn $\text{Immersion}(bm(x)) \in C_j$.

Eine topologiegetreue[244] Emergente SOM projiziert ein Cluster auf eine kohärente Fläche einer Karte (Cluster-Fläche). Punkte innerhalb der Cluster werden ins Innere der Clusterfläche abgebildet. Datenpunkte am Rand des Clusters werden an den Rand der Clusterfläche projiziert. Betrachte einen Datenpunkt x am Rand des Clusters C mit $n_i = bm(x)$. Der Gewichtungsvektor seiner Nachbarn $N(i)$ ist entweder innerhalb des Clusters, in einem unterschiedlichen Cluster oder interpoliert zwischen Cluster. Wenn man annimmt, dass der Abstand zwischen Clustern lokal größer wird als die lokalen Abstände innerhalb der Cluster ist, dann wird der U-Höhenwert in $N(i)$ groß in den Richtungen, die von dem Cluster C wegzeigen. Das bedeutet, dass ein Gradientenabstieg auf der U-Matrix von den Clustergrenzen wegführen wird.

---

243 Vgl. ULTSCH [2005, pp.240-246], ULTSCH [2005, pp. 75-82]
244 Vgl. BORGELT et al. [2003, S. 98]

Eine Bewegung von einem Neuron n_i zu einem anderen Neuron n_j mit dem Ergebnis, dass sich *wj* mehr innerhalb des Clusters C befindet als *wi*, heißt immersiv. Für Datenpunkte, die sich deutlich im Innern von C befinden, wird ein Gradientenabstieg auf der U-Matrix notwendigerweise immersiv sein.

Die Höhenwerte der P-Matrix folgen der Dichtestruktur eines Clusters. Unter der Annahme, dass die Kernbereiche eines Clusters die Regionen mit der großen Dichte sind, ist ein Gradientenabstieg auf der P-Matrix stets immersiv. Cluster können auch über die Dichte allein anstatt mit einem Distanzmaß definiert werden. An der Grenze eines Clusters ist die Dichtemessung dennoch kritisch. Am Clusterrand sollte die lokale Punktdichte signifikant abnehmen. In den meisten Fällen werden die Clustergrenzen entweder durch niedrige Punktdichtewerte oder durch freien Raum zwischen den Clustern (große Abstände) definiert. Für empirische Schätzungen der Punktdichte wird ein Gradientenaufstieg auf der P-Matrix deshalb nicht immersiv für Punkte im Randbereich sein.

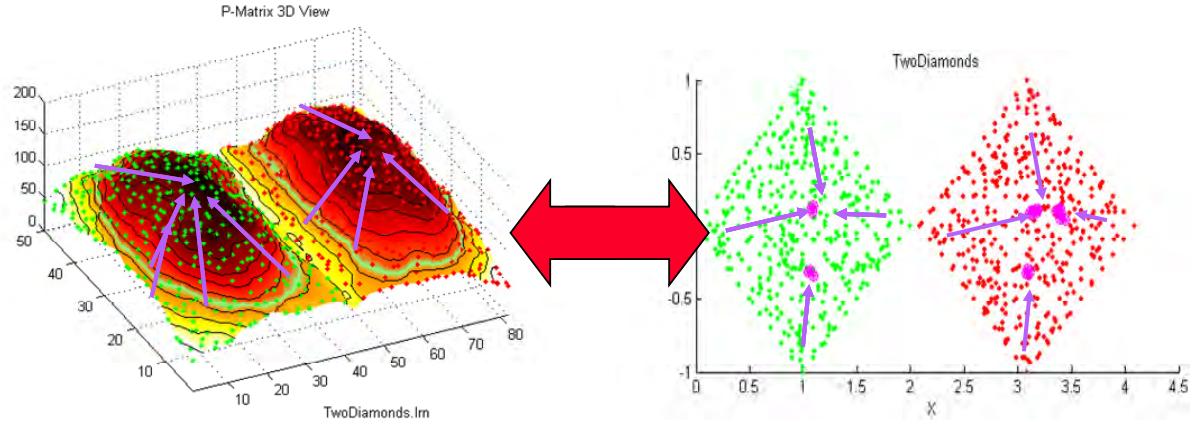


Abbildung 2-14:Gradientenaufstieg auf der P-Matrix[245]

Eine Bewegung, die auf der SOM-Karte erst einen Gradientenabstieg auf der U-Matrix und dann einen Gradientenaufstieg auf der P-Matrix folgt, heißt Immersion. Bezeichne I die Endpunkte der Immersion, die von jedem Neuron auf der Karte beginnen. Wenn die Dichte innerhalb der Cluster konstant ist, wird die Immersion für keinen Startpunkt innerhalb der Cluster zu einem einzelnen Punkt eines Clusters konvergieren. Die U*-Matrix wird dann dazu verwendet, um die Punkte in I zu bestimmen, die zu dem gleichen Cluster gehören. Die Wasserscheiden (Watersheds) der U*-Matrix. werden mit einem Algorithmus[246] berechnet. Die Punkte, die durch eine Watershed getrennt sind, werden unterschiedlichen Clustern zugeordnet. Punkte innerhalb desselben Beckens werden einem einzigen Cluster zugeordnet.

245 Vgl. ULTSCH [2003 c]
246 Vgl. SOILLE, P. /LUC, V. [1991]: An Efficient Algorithm Based on Immersion Simulations.

2.3.5 Datengenerierung durch Mischungsmodelle

2.3.5.1 Gaußsche Mixtur-Modelle

Gaußsche Mixtur-Modelle (Gaussian Mixture Models, GMM)[247] sind Kombinationen von Gaußverteilungen[248] bzw. additiv zusammengesetzt aus mehreren einzelnen Gaußschen Normalverteilungen.[249] GMMs sind sehr eng mit dem Bayes'schen Klassifizierer verwandt und gelten als ein probabilistisches Modell für multivariate Wahrscheinlichkeitsdichten.[250] In der Anwendung wird mit Hilfe von GMMs die zugrunde liegende klassenspezifische Wahrscheinlichkeitsdichte berechnet, auf Basis derer ein Likelihood-Ratio-Klassifizierer dann ein vorliegendes Muster einer Kategorie zuweist.

Die Verteilungsdichte eines Merkmalsvektors x lässt sich meistens nur ungenau mit Hilfe einer einzelnen Normal- bzw. Gaußverteilung beschreiben. Eine genauere Approximation ist mit einer Gaußschen Mischverteilung möglich. Für einen n-dimensionalen Merkmalsvektor x ist die Gaußsche Mischverteilung (Gaussian mixture density) definiert als eine gewichtete Summe von K Gaußdichten:

$$p(X|\lambda) = \sum_{k=1}^{K} p_k \cdot b_k(X) \tag{54}$$

für die Gewichtungen p_k gilt: $p_k \geq 0$ und $\sum_{k=1}^{K} p_k = 1$ for $k \in \{1,...,K\}$.

Die Gaußsche Wahrscheinlichkeitsdichtefunktion (probability density function, PDF) wird durch folgende Gleichung beschrieben:

$$b_k(\mathrm{X}) = \frac{1}{\sqrt{2\pi}^d \sqrt{|\sum|}} e^{-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)} \tag{55}$$

wobei x ein Spaltenvektor mit d Elementen ist, μ der Mittelwertvektor, $\sum$ die d-mal-d-Kovarianzmatrix, $|\sum|$ ihre Determinante und $\sum^{-1}$ ihre inverse Matrix. Des Weiteren ist $(x-\mu)^T$ die Transponente von $(x-\mu)$.

Durch Veränderung der Anzahl der Gaußverteilungen K, den Gewichtungen p_k und den Parametern μ sowie der Kovarianzmatrix $\sum$, können Gaußmixturen dazu verwendet werden, um komplexe Wahrscheinlichkeitsdichtefunktionen (PDF) abzubilden.

---

247 GMM wurden in den für diese Arbeit exemplarischen Disziplinen von z.B. THINH; BEHNISCH und ULTSCH [2006] eingesetzt.
248 Vgl. LAURITZEN [1996]
249 Vgl. G. MCLACHLAN UND D. PEEL [2000]
250 Vgl. REYNOLDS et al. [2000]

2.3.5.2 Expectation-Maximization-Algorithmus

Mit dem Expectation-Maximization (EM) - Algorithmus wird das Ziel verfolgt, die Parameter eines GMMs so zu schätzen, dass das GMM die Verteilung der Trainingsvektoren am besten darstellt.[251] Die grundsätzliche Idee des EM-Algorithmus (Abbildung 2-15) besteht darin, iterativ durch eine abwechselnde Klassifikation (Expectation, E-Schritt) und eine anschließende Anpassung der Modellparameter $\lambda = \{p_k, \mu_k, \sigma_k^2\}$ (Maximization, M-Schritt), die Wahrscheinlichkeit für das Auftreten eines stochastischen Prozesses X bei gegebenem Modell zu maximieren.[252] Es ist wichtig, dass mit guten Initialisierungsparametern begonnen wird, da der Algorithmus ein lokales und nicht ein globales Optimum findet. Es müssen zur Maximierung die Modellparameter λ angepasst werden, weil der stochastische Prozess X durch die Trainingsdaten x der Klasse gegeben ist. Als Voraussetzung für das Auffinden des Maximums gilt, dass nach jedem Induktionsschritt und der Berechnung des neuen Modells $\bar{\lambda}$ gilt:

$$p\left(X\middle|\bar{\lambda}\right) \geq p\left(X|\lambda\right) \qquad \text{(Monotoner Anstieg der Wahrscheinlichkeit)} \qquad (56)$$

Das Verfahren wird bis zum Erreichen eines Konvergenzschwellwertes fortgeführt.

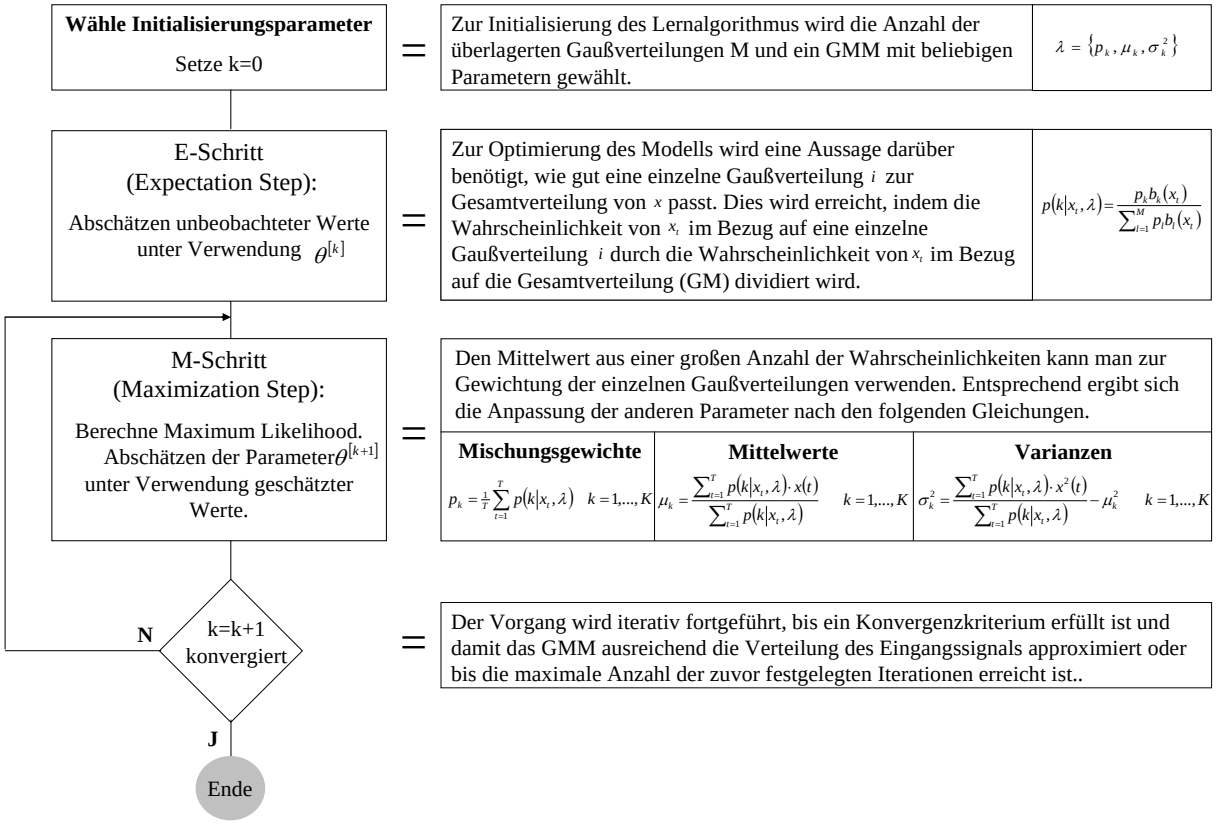


Abbildung 2-15: Ablauf des EM-Algorithmus (Approximation einer Gaußschen Mischverteilung)[253]

[251] Vgl. HAND et al. [2001, S. 281], MCLACHLAN / BASFORD [1989], STÖLZEL [1999, S. 61]
[252] Vgl. BILMES [1997], DEMPSTER et al. [1977], MCLACHLAN AND KRISHNAN [1997]
[253] Eigene Bearbeitung nach MOON, T. [1996] und eigene Ergänzung mit RHODENBURG [2003, S.16 ff.]

2.3.5.3 Bayes'sche Entscheidungsgrenze

Um die Vorgehensweise für die Klassenbildung anhand der Bayes'schen Entscheidungsgrenze zu erläutern, werden die Gaußschen Mixtur-Modelle an einem exemplarischen Beispiel für eine Variable beschrieben. Abbildung 2-16 zeigt eine Gauß-Mixtur (gestrichelte Linie), die mit Hilfe des EM-Algorithmus gefunden wurde und aus zwei einzelnen Gaußverteilungen (breite Linie) zusammengesetzt wird. Da die gefundene Lösung besonders von den Initialisierungsparametern abhängt, enthält die Optimierungsprozedur eine Berechnung für zwei bis fünf Gaußverteilungen. Als zusätzliches Gütekriterium dient die Pareto Dichte-Schätzung (Pareto Density Estimation), welche einen geeigneten Schätzer für Wahrscheinlichkeitsdichten von Mischungen von gaußverteilten Daten darstellt[254]. Die Bayes'sche Entscheidungsgrenze wird durch den Schnittpunkt der Kurvenverläufe der einzelnen Gaußverteilungen (Modus) gebildet und dient dazu, die Objekte aufgrund der Variablenausprägungen in zwei Gruppen einzuteilen (vertikale Markierung).

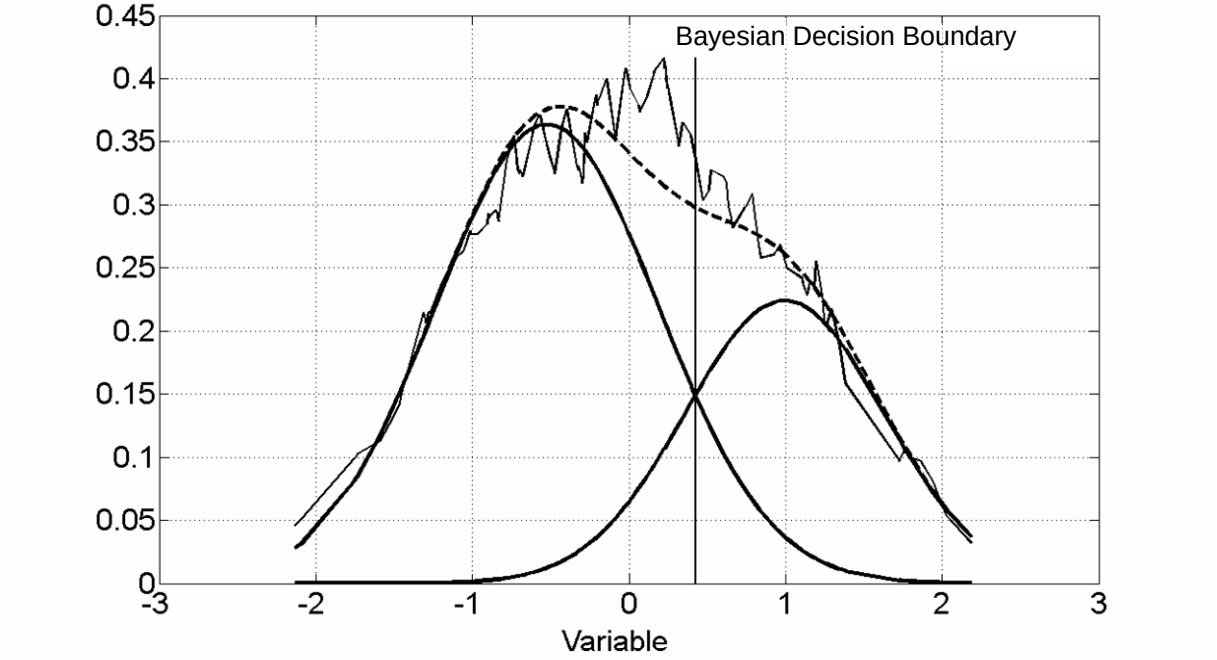


Abbildung 2-16: PDE (feine Linie), EM-Gauß-Mixtur (gestrichelte Linie), Gaußverteilung (starke Linie)

Für mehrere Variablen werden jeweils getrennt voneinander weitere Gauß-Mixtur-Modelle berechnet und interpretiert, so dass eine endgültige Variablenauswahl erfolgen kann. Der Klassenbildung liegen die jeweils gefundenen Bayes'schen Entscheidungsgrenzen zu Grunde, die den gewählten Gauß-Mixtur-Modellen entnommen werden. Anhand gewählter Codierungsregeln werden die Klassen unter Berücksichtigung der eingesetzten Variablen und entsprechenden Ausprägungen systematisch erzeugt.

[254] SCOTT, D.W. [1992], ULTSCH [2003 a]

2.4 Strukturprüfung

2.4.1 Kriterien zur Validierung und Charakterisierung

Die kontextabhängige Interpretation der Klassifikationsergebnisse bildet den Abschluss einer Clusterzuordnung. Es ist darauf hinzuweisen, dass die recht ungenau beschriebene Zielsetzung der Clusteranalyse, Cluster möglichst ähnlicher Objekte zu bilden, unterschiedliche Interpretationen zulässt. Denn es ist nicht zu unterscheiden zwischen richtigen oder falschen Gruppierungen, sondern vielmehr im Sinne der jeweiligen Anwendung nach brauchbaren bzw. unbrauchbaren Lösungen.

Die statistischen Eigenschaften der Variablenwerte sind zur Beurteilung der Clusterstruktur von großer Bedeutung. Zur Untersuchung dieser reduzierten Objektmenge eines Clusters eignen sich beispielsweise Minimum, Maximum, Mittelwert, Varianz, Standardabweichung, Schiefe und Exzess. Die Analyse des Klassifikationsergebnisses sollte die Homogenität der gebildeten Cluster, die relative Lage der Clusterrepräsentanten, den Einfluss bestimmter Variablen und Objekte und die Bedeutung der Ausgangspartition mit berücksichtigen. Ergebnisse einer Clusterzuordnung lassen sich anhand von einigen zusätzlich berechneten Werten genauer beschreiben. Im Folgenden erläutert seien die F-Werte und die t-Werte. Zur Beurteilung der Homogenität eines gefundenen Clusters dient der F-Wert, der für jede Klassifikationsvariable in einem Cluster wie folgt zu berechnen ist:

$$F = \frac{V(J,C)}{V(J)} \tag{57}$$

mit: $V(J,C)$: Varianz der Variable J in Cluster C
$V(J)$: Varianz der Variable J in der Erhebungsgesamtheit

Zur inhaltlichen Interpretation bzw. der Charakterisierung der jeweiligen Cluster wird der t-Wert verwendet, der sich für jede Variable in einem Cluster ermitteln lässt. Die t-Werte stellen normierte Werte dar:

$$t = \frac{\overline{X}(J,C) - \overline{X}(J)}{S(J)} \tag{58}$$

mit: $\overline{X}(J,C)$: Mittelwert der Variable J über die Objekte in Cluster C
$\overline{X}(J)$: Gesamtmittelwert der Variable J in der Erhebungsgesamtheit
$S(J)$: Standardabweichung der Variable J in der Erhebungsgesamtheit

Je kleiner ein F-Wert ist, desto geringer ist die Streuung dieser Variablen in einem Cluster im Vergleich zur Erhebungsgesamtheit. Der F-Wert sollte den Wert 1 nicht überschreiten, da in diesem Fall die entsprechende Variable in dem zu Grunde liegenden Cluster eine größere

Streuung aufweist als in der Erhebungsgesamtheit. Die F-Werte werden für alle Variablen in den mit Hilfe der Clusterverfahren gefundenen Clustern berechnet. Ein Cluster ist als vollkommen homogen anzusehen, wenn alle F-Werte kleiner als 1 sind. Negative t-Werte deuten darauf, dass eine Variable in einem betrachteten Cluster im Vergleich zur Erhebungsgesamtheit unterrepräsentiert ist. Positive t-Werte zeigen, dass eine Variable in einem betreffenden Cluster im Vergleich zur Erhebungsgesamtheit überrepräsentiert ist. Anhand der t-Werte lassen sich somit Aussagen über die Variablenausprägungen in einem gefundenen Cluster treffen und Vergleiche bilden.

Bei den Ergebnissen der unscharfen Clusteranalyse sind folgende Kriterien zu berücksichtigen: Kardinalität, typischstes Objekt,[255] Klassenzentren und Heterogenität der Cluster. Die Kardinalität definiert die Summe der Zugehörigkeitsgrade aller Objekte eines Clusters und bestimmt die Größe eines unscharfen Clusters. Die Summe der Kardinalitäten aller Cluster einer unscharfen Partition ist unter der unten genannten Bedingung gleich der Gesamtzahl der Objekte, so dass sich die Kardinalitäten auch in Prozent angeben lassen.

$$K_c = \sum_{i=1}^{n} \mu_{ic} \qquad c = 1,...,P \text{ und der Bedingung } \bigwedge_{i\ =\ 1...n} \sum_{c=1}^{p} \mu_{ic} = 1 \tag{59}$$

mit K_c = Kardinalität der unscharfen Cluster c, μ_{ic} = Zugehörigkeitsgrad des Objektes i zum Cluster c, n = Anzahl der Objekte, P = Anzahl der Cluster.

Das typischste Objekt eines Clusters wird definiert als das Objekt, welches die geringste Distanz zum Zentrum hat. Im Allgemeinen handelt es sich um das Objekt mit dem höchsten Zugehörigkeitsgrad. Die Lage des Clusterzentrums lässt sich durch das typischste Objekt besonders gut veranschaulichen.

Durch Berechnung der Klassenzentren (vgl. Zielfunktion des Fuzzy-c-means-Algorithmus) wird es möglich, die einzelnen Cluster bezüglich eines fiktiven Objektes zu interpretieren. Diese Objekte stimmen nur im günstigsten Fall mit einem der Untersuchungsobjekte überein. Es handelt sich um einen imaginären besten Vertreter der jeweiligen Klasse. Die Cluster lassen sich hinsichtlich ihrer Heterogenität untersuchen. Ausschlaggebend ist dabei die Größe des höchsten Zugehörigkeitsgrades und die Anzahl der Objekte, deren Zugehörigkeitsgrad zur besten Repräsentation der Clustereigenschaften aufgrund der Zielfunktion des Algorithmus größer als 0,5 ist.[256]

---

255 Vgl. BOCK [1974, S. 100 ff.]

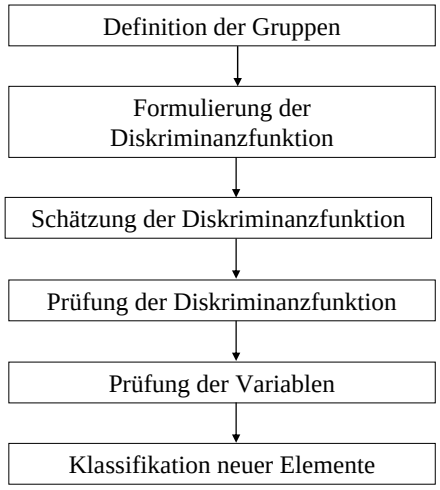
256 Vgl. QU [2000, S. 61,62]: „Die Objekte mit Zugehörigkeitsgraden größer als 0,5 können einen Cluster am besten erklären, weil die Zugehörigkeitsgrade zu den anderen Clustern wegen der Nebenbedingung [...] des Algorithmus kleiner als 0,5 sein müssen."

2.4.2 Diskriminanzanalyse

Die Diskriminanzanalyse (DA)[257] wird als multivariates Verfahren dazu verwendet, um Gruppenunterschiede hinsichtlich einer vorgegebenen Variablenanzahl zu analysieren. Es lassen sich folgende Aufgabenstellungen[258] lösen und Fragen[259] dadurch beantworten:

- Unterscheiden sich die Gruppen bezogen auf die Variablen signifikant voneinander?
- Welche Variablen sind besonders zur Gruppenunterscheidung (bestmögliche Gruppentrennung) geeignet bzw. ungeeignet? Wie lassen sich Klassifikationen verbessern?
- Welcher der bereits entwickelten Gruppen lassen sich weitere bisher nicht klassifizierte Objekte aufgrund ihrer Variablenstruktur zuordnen?
- Analyse der Unterschiede zwischen Objekten. Test aufgestellter Hypothesen.

Die wesentlichen Teilschritte zur Durchführung einer Diskriminanzanalyse sind in Abbildung 2-17 aufgeführt und werden ergänzt durch die allgemeine Form der Diskriminanzfunktion.



Formulierung der Diskriminanzfunktion

Die Diskriminanzfunktion Y lässt sich als Linearkombination der I Variablen X_i in folgender Form darstellen:

Y = Diskriminanzvariable

$Y = v_0 + v_1 X_1 + v_2 X_2 + \ldots + v_I X_I$ mit X_i = Merkmalsvariable $i\,(i = 1,2,\ldots,I)$

v_i = Diskriminanzkoeffizient der Merkmalsvariablen X_i

Die Parameter v_0 und $v_i (i = 1,\ldots,I)$ sind auf der Basis von Daten für die Merkmalsvariablen zu schätzen.

Für jedes Element $i (i = 1,\ldots,I_g)$ einer Gruppe $g (g = 1,\ldots,G)$ mit den Merkmalswerten $X_{jgi} (j = 1,\ldots,J)$ liefert die Diskriminanzfunktion einen Diskriminanzwert Y_{gi} .

Formulierung des Diskriminanzkriteriums

Das Diskriminanzkriterium misst die Unterschiedlichkeit der Gruppen:

$$r = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung in den Gruppen}} = \frac{SS_b}{SS_w} = \frac{(\text{Sum of squares between})}{(\text{Sum of squares within})}$$

$$= \frac{\sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2}{\sum_{g=1}^{G}\sum_{j=1}^{n_g} (y_{gj} - \bar{y}_g)^2}$$ mit

$\bar{y}_g$ = Mittlerer Diskriminanzwert in Gruppe g

$\bar{y}$ = Gesamtmittel der Diskriminanzwerte aller N Elemente

n_g = Fallzahl in Gruppe g

y_{gj} = Diskriminanzwert von Element j in Gruppe g

$$SS = SS_b + SS_w = \sum_{g=1}^{G}\sum_{i=1}^{I_g} (Y_{gi} - \bar{Y})^2 \qquad \gamma = \text{Max}\{r\} \qquad \text{EA}_k = \frac{\gamma_k}{\gamma_1 + \gamma_2 + \ldots + \gamma_K}$$

Die Diskriminanzwerte selbst und damit auch deren Streuung sind abhängig von den zu bestimmenden Koeffizienten v_i der Diskriminanzfunktion. Wähle die Koeffizienten $v_i (i = 1,\ldots,I)$ so, daß das Diskriminanzkriterium r maximal wird. Zu jeder Diskriminanzfunktion gehört ein Eigenwert (Maximal-Wert) Die Wichtigkeit einer Diskriminanzfunktion wird durch den Eigenwertanteil EA_k ausgedrückt.

Abbildung 2-17: Ablaufschema und Ergänzungen zur Formulierung der Diskriminanzfunktion[260]

Den Ausgangspunkt für die Diskriminanzanalyse bildet wie gezeigt die Strukturbildung durch Festlegung von Gruppen. Vorraussetzung für die Anwendung der Diskriminanzanalyse sind metrisch skalierte Daten für die Variablen der Objekte und die zuvor festgelegte Gruppenzugehörigkeit in Form einer nominal skalierten Variablen. Die Formulierung und Schätzung der Diskriminanzfunktion ermöglicht eine optimale Trennung zwischen den Gruppen und eine Prüfung der diskriminatorischen Bedeutung der Variablen.

---

257 DA wurde in den für diese Arbeit exemplarischen Disziplinen zuerst von KING [1967, 1969, 1970] und CASETTI [1964 a, b] eingesetzt. In Deutschland seien genannt: KILCHENMANN [1968], ERB [1990], STEINER [1975] und ARLT et al. [2005, S. 106 ff.] sowie THINH [2004 a, S. 95 ff.]

258 Vgl. BAHRENBERG [2003, S. 318] und ERB [1990, S.9/10]

259 Vgl. SCHUCHARD-FICHER et al. [1985, S. 153 ff.]

260 Eigene Bearbeitung nach BACKHAUS et al. [2006, S. 159] und eigene Erweiterungen.

Um eine Diskriminanzfunktion zu formulieren, ist es notwendig Variablen auszuwählen, die zunächst infolge theoretischer oder sachlogischer Überlegung zwischen den Gruppen differieren und zur Unterscheidung der Gruppen bzw. Erklärung der Gruppenunterschiede geeignet sein könnten. Die Schätzung der Diskriminanzfunktion bzw. der unbekannten Koeffizienten v_i ermöglicht dann die Überprüfung der diskriminatorischen Eignung der gewählten Variablen. Die Diskriminanzwerte Y und deren Streuung sind abhängig von den zu bestimmenden Koeffizienten v_i der Diskriminanzfunktion. Eine Verschiebung der Diskriminanzwerte erfolgt durch den Wert v_0. Die Wahl der anderen Koeffizienten $v_i(i = 1,..., I)$ unterliegt einem Optimierungsproblem dahingehend, dass das Diskriminanzkriterium maximal wird. Zu jeder Diskriminanzfunktion gehört ein Maximalwert, der bei mehrfachen Diskriminanzfunktionen Eigenwert lautet. Von den nach und nach bestimmten Diskriminanzfunktionen erfolgt die Beurteilung der Wichtigkeit mit dem Maßkriterium des Eigenwertanteils (EA_k), wobei die diskriminatorische Bedeutung mit der Größe der Eigenwerte abnimmt. Die Eigenwertanteile bilden in der Summe 100 %. Um das Ergebnis der Diskriminanzanalyse besser interpretieren zu können, wird eine Normierung vorgenommen, so dass die Innergruppen-Varianz von allen Diskriminanzwerten Eins ergibt:

$$s^2 = \frac{SS_W}{I - G} = \frac{\sum_{g=1}^{G} \sum_{j=1}^{n_g} \left(y_{gj} - \bar{y}_g \right)^2}{I - G} \tag{60}$$

Im weiteren Verlauf wird eine Prüfung der Diskriminanzfunktion durchgeführt. Hier werden verschiedene Gütekriterien verwendet, wobei das so genannte inverse Gütemaß WILKS' LAMBDA von BACKHAUS[261] als gebräuchlichstes bezeichnet wird:

$$\Lambda = \frac{1}{1 + \gamma} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}} \tag{61}$$

Je kleiner der Wert von Λ, desto verschiedener sind die Gruppen voneinander. Durch Transformation lässt sich dieses Kriterium in eine probabilistische Variable umwandeln und Wahrscheinlichkeitsaussagen über die Unterschiedlichkeit von Gruppen treffen. Eine weitere Möglichkeit die Diskriminanzfunktion zu prüfen bietet sich an, indem die durch Diskriminanz erzeugte Klassifizierung der Untersuchungsobjekte mit deren tatsächlicher Gruppenzugehörigkeit verglichen wird.[262]

---

[261] Vgl. BACKHAUS et al. [2006, S. 182], weitere Kriterien sind u.a. der kanonische Korrelationskoeffizient und der Eigenwert der Diskriminanz sowie der Multivariate Wilks'Lambsda.

[262] Vgl. BACKHAUS et al. [2006, S. 179 ff.]

Die Trennkraft der einzelnen Variablen wird mit Hilfe des standardisierten Diskriminanzkoeffizienten b_i* beurteilt:

$$b_i^* = s_i \cdot b_i \tag{62}$$

mit b_i = normierter Diskriminanzkoeffizient der Merkmalsvariablen X_i

s_i = Standardabweichung der Variablen X_i

Zusätzlich lassen sich nach BAHRENBERG[263] Korrelationskoeffizienten von Variablen und Diskriminanzwerten berechnen, die als Strukturkoeffizienten bezeichnet werden und analog zur Faktorladung einer Faktorenanalyse die Diskriminanzladung ausdrücken.

Da im Falle von mehrfachen Diskriminanzfunktionen für jede Variable mehrere Diskriminanzkoeffizienten vorliegen, werden mittlere Diskriminanzkoeffizienten berechnet:

$$\bar{b}_i^* = \sum_{k=1}^{K} \left| b_{ik}^* \right| \cdot EA_k \tag{63}$$

mit b_{ik}^* = Standardisierter Diskriminanzkoeffizient der Variablen X_i bezüglich der Diskriminanzfunktion Y_k

EA_k = Eigenwertanteil der Diskriminanzfunktion Y_k

Zur Interpretation der Ergebnisse einer vorgeschalteten Clusteranalyse ist zwischen aktiven und passiven Variablen zu unterscheiden, die mit einer Diskriminanzanalyse beurteilt werden. Man unterscheidet zwischen der Eignungsprüfung von verwendeten Variablen im Rahmen der Clusteranalyse (aktiv) und den zusätzlich relevanten Variablen, die man zur Erklärung einer Gruppierung mit Hilfe der Diskriminanzanalyse bewertet (passiv).

Die Klassifizierung von Objekten mit einer unbekannten Gruppenzugehörigkeit zu vorgegebenen Gruppen wird durch die Diskriminanzanalyse unterstützt. Zu unterscheiden sind das Distanzkonzept, das Wahrscheinlichkeitskonzept und die Klassifizierungsfunktionen.[264] Das Distanzkonzept wird näher betrachtet. Ein Objekt wird demnach derjenigen Gruppe zugeordnet, zu deren Gruppenmittelpunkt es den geringsten Abstand besitzt. Unter der Voraussetzung, dass K Diskriminanzfunktionen extrahiert werden, spannen diese einen K-dimensionalen orthogonalen Diskriminanzraum auf. Es lassen sich die quadrierten Distanzen zwischen dem Objekt und dem Gruppenmittelpunkt berechnen:

$$d_{jg}^2 = \sum_{k=1}^{K} \left(y_{kj} - \bar{y}_{kg} \right)^2 \qquad (g = 1,2,...,G) \tag{64}$$

mit y_{kj} = Diskriminanzwert von Objekt j bezüglich der Diskriminanzfunktion k

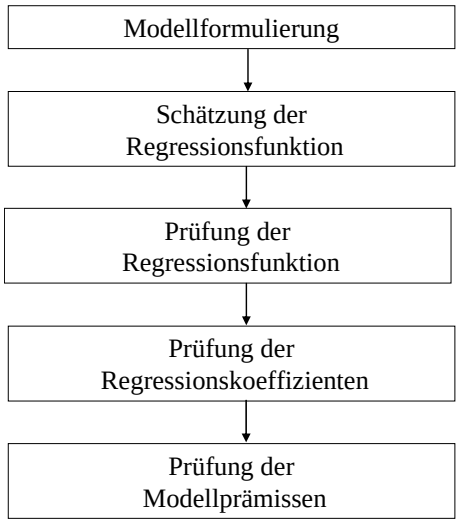
$\bar{y}_{kg}$ = Mittlerer Diskriminanzwert in Gruppe g bezüglich der Diskriminanzfunktion k

263 Vgl. BAHRENBERG et al. [2003, S. 337]

264 Vgl. BACKHAUS [2006, S. 188 ff.], vgl. ERB [1990, S. 51 ff.]

2.4.3 Regressionsanalyse

Mit Hilfe der Regressionsanalyse (RA)[265] wird das Ziel verfolgt, ein funktionales Modell zwischen einer abhängigen (Regressand, erklärte Variable [target variable]) und einer oder mehreren unabhängigen Variablen (Regressoren, erklärende Variablen [explanatory variable]) zu finden. Damit wird nicht nur die Stärke des Zusammenhangs von gleichberechtigten Variablen beschrieben, wie es bei der Korrelationsanalyse der Fall ist, sondern es interessiert die Art und Weise, in der die abhängige Variable von unabhängigen Variablen beeinflusst wird. Es ist möglich, Zusammenhänge in quantitativer Form zu beschreiben, zu erklären und schließlich Werte der abhängigen Variablen zu schätzen bzw. zu prognostizieren. Das Skalenniveau der Variablen sollte metrisch sein. Der Anwendungsbereich der Regressionsanalyse umfasst die Ursachenanalyse, Wirkungsprognose und Zeitreihenanalyse.[266] Durch die Regressionsanalyse wird lediglich eine Hypothese von möglichen Ursache-Wirkungsbeziehungen aufgestellt, so dass zusätzlich eine Plausibilitätsprüfung erfolgen muss.[267] Die wesentlichen Teilschritte zur Durchführung einer Regressionsanalyse sind in Abbildung 2-18 aufgeführt und werden ergänzt durch die Darstellung des Regressionsansatzes.



Formulierung des Regressionsansatzes

Das Modell der multiplen Regression wird für eine zu untersuchende Variable Y (Regressand) dargestellt, die durch zahlreiche Größen (Regressoren) beeinflußt wird, d.h. $Y = f(X_1, X_2, ..., X_m, ..., X_M)$:

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m \quad \text{mit}$$

$\hat{Y}$ = Schätzung der abhängigen Variable Y
X_m = unabhängige Variablen $(m = 1,2,...,M)$
β_m = Regressionskoeffizienten $(m = 1,2,...,M)$
α = Regressionskonstante

Der Parameter β_m heißt partieller Regressionskoeffizient und gibt den spezifischen Einfluß der Variablen X_m wieder. Dadurch drückt sich aus, um wieviele Einheiten Y zunimmt, wenn X_m um eine Einheit größer wird und alle anderen unabhängigen Variablen konstant bleiben.
Die Parameter α und $\beta_m (m = 1,2,...,M)$ müssen im Regelfall aus einer Stichprobe mit dem Umfang n geschätzt werden. Die unbekannten Parameter werden durch Minimierung der Summe der Abweichungsquadrate (Methode der kleinsten Quadrate=KQ-Schätzung) bestimmt.

Zielfunktion der multiplen Regressionsfunktion

Nicht erfasste Einflußgrößen der empirischen Y -Werte führen zu Abweichungen von der Regressionsgeraden. Die Variable ε repräsentiert diese Abweichungen und deren Werte ε_k heißen Residuen.

$$\sum_{k=1}^{K} \varepsilon_k^2 = \sum_{k=1}^{K} [y_k - \hat{y}_k]^2 \rightarrow \min$$

$\hat{y}_k = \alpha + \beta_1 x_{2k} + \beta_2 x_{2k} + ... + \beta_m x_{mk} + ... + \beta_M x_{Mk}$
ε_k = Werte der Residualgröße $(k = 1,2,...,K)$ bzw. des Abweichungsparameters
mit y_k = Werte der abhängigen Variablen $(k = 1,2,..., K)$, $\hat{y}_k$ = ermittelter Schätzwerte von Y für x_k
x_{jm} = Werte der unabhängigen Variablen $(m = 1,2,..., M)$
K = Zahl der Beobachtungen, M = Zahl der unabhängigen Variablen

Abbildung 2-18: Ablaufschema und Ergänzungen zur Formulierung des Regressionsansatzes[268]

---

265 Der Regressionsbegriff geht zurück auf den englischen Wissenschaftler SIR FRANCIS GALTON (1822-1911), der die Abhängigkeit der Körpergröße von Söhnen in Abhängigkeit von der Körpergröße ihrer Väter untersuchte, vgl. BULMER [2003]. Die RA wurde in den für diese Arbeit exemplarischen Disziplinen von z.B. GÜSSEFELDT [1988] und ARLT et al. [2001, S.112 ff.], [2005, S. 99 ff.] eingesetzt.

266 Vgl. BACKHAUS [2006, S. 49]: „Wie stark ist der Einfluss der unabhängigen Variablen auf die abhängige Variable?“, „Wie verändert sich die abhängige Variable bei einer Änderung der unabhängigen Variablen?“, „Wie verändert sich die abhängige Variable im Zeitablauf und somit ceteris paribus auch in der Zukunft?“

267 Vgl. BACKHAUS [2006, S. 47, 57], verwiesen sei auf außerstatistisches Wissen, wie theoretische oder sachlogische Überlegung, Experimente und auf Residualgrößen oder so genannte Zufallsfehler.

268 Eigene Bearbeitung nach BACKHAUS et al. [2006, S. 52] und eigene Erweiterungen nach BAHRENBERG et al. [2003, S. 31 ff.]. Historisch verwiesen sei an dieser Stelle auf GAUSS [1887].

Die methodisch sachgerechte Formulierung eines Regressionsmodells verlangt zu Beginn vom Bearbeiter einige Vorüberlegungen, die primär fachdisziplinäre und weniger methodenanalytische Gesichtspunkte einbeziehen. Der Schwerpunkt liegt auf der Abbildung der Ursache-Wirkungs-Beziehungen, um die Abhängigkeit einer Untersuchungsvariablen von anderen Variablen zu charakterisieren.

Für die Schätzung der Regressionsfunktion ist die Anzahl der unabhängigen Variablen von Bedeutung, da entweder bei monokausaler Beziehung ein einfacher oder bei mehreren unabhängigen Variablen ein multipler Regressionsansatz verfolgt wird. Es wird die Regressionsfunktion aufgestellt (Abbildung 2-18), die zweidimensional eine Regressionsgerade, dreidimensional eine Regressionsebene bzw. mit mehr als drei unabhängigen Variablen eine Hyperebene beschreibt. Die Bedeutung einer Änderung von unabhängigen Variablen und die damit verbundenen Auswirkungen auf die abhängige Variable werden durch die Regressionskoeffizienten ausgedrückt. Um eine Vergleichbarkeit von Regressionskoeffizienten herzustellen und somit Aussagen über die Wichtigkeit der Variablen treffen zu können, ist bei unterschiedlichen Messeinheiten zusätzlich eine Standardisierung der Koeffizienten durchzuführen. Man spricht von partiellen Regressionskoeffizienten:[269]

$$\hat{Z}_o = \alpha_s + \beta_{s1} Z_1 + \beta_{s2} Z_2 + ... + \beta_{sm} Z_m \qquad (65)$$

mit Z_i = standardisierte Variable der Ausgangsvariablen X_i

b_{si} = partieller Regressionskoeffizient der Variablen Z_i

Bei der Regressionsanalyse ist zwischen der Möglichkeit der blockweisen (=Einschlussverfahren) oder der schrittweisen Einbeziehung von unabhängigen Variablen zu unterscheiden, und darüber hinaus kann durch ein nachträglich festgelegtes Ausschlusskriterium ein späterer Verzicht auf Variablen ermöglicht werden.[270]

Die Güte der geschätzten Regressionsfunktion wird in zwei Schritten geprüft, d.h. es erfolgt zunächst eine globale Prüfung der Regressionsfunktion[271] und anschließend eine Prüfung der Regressionskoeffizienten.[272] Einerseits wird die Regressionsfunktion als Ganzes geprüft, um zu untersuchen, inwieweit die abhängige Variable Y durch das Regressionsmodell erklärt wird. Andererseits wird der Beitrag geprüft, den jede einzelne Variable zur Erklärung leistet. Eine Variable ohne nennenswerten Erklärungsbeitrag, kann aus der Regressionsfunktion entfernt werden.

---

269 Vgl. BAHRENBERG [2003, S.33]

270 Vgl. BÜHL/ZÖFEL [2002] und JANSSEN/LAATZ [2005, S.436]

271 Vgl. BACKHAUS et al. [2006, S.64 ff.]: Geeignete Maße sind das Bestimmtheitsmaß, die F-Statistik und der Standardfehler.

272 Vgl. BACKHAUS et al. [2006, S.73 ff.]: Geeignete Maße sind der t-Wert und der Beta-Wert.

Die Prüfung der Modellprämissen bzw. die Prüfung auf Verletzung von Modellbedingungen wird mit Hilfe der grafischen Analyse[273] und statistischer Testverfahren[274] durchgeführt. Bei Einhaltung der von BACKHAUS[275] genannten Prämissen führt die Methode der kleinsten Quadrate zu linearen Schätzfunktionen für Regressionsparameter, die alle wünschenswerten Eigenschaften von Schätzern besitzen (= Best Linear Unbiased Estimators), d.h. unverzerrt (erwartungstreu) und effizient.[276] Die Modellannahme der Normalverteilung der Störgrößen ist vorteilhaft zur Durchführung von Signifikanztests. Die wichtigsten Prämissen eines linearen Regressionsmodells[277] werden mit den Konsequenzen eines Verstoßes gegen Modellbedingungen in Tabelle 2-15 dargestellt.

Prämisse	Verletzung der Modellbedingung	Konsequenz
Linearität in den Parametern	Nichtlinearität	Verzerrung der Schätzwerte
Vollständigkeit des Modells (Berücksichtigung aller relevanten Variablen)	Unvollständigkeit	Verzerrung der Schätzwerte
Homoskedastizität[278] der Störgrößen	Heteroskedastizität[279]	Ineffizienz
Unabhängigkeit der Störgrößen	Autokorrelation	Ineffizienz
Keine lineare Abhängigkeit zwischen den unabhängigen Variablen	Multikollinearität[280]	Verminderte Präzision der Schätzwerte
Normalverteilung der Störgrößen	Nicht normalverteilt	Ungültigkeit der Signifikanztests (F-Test und t-Test), wenn die Zahl der Beobachtungen K klein ist.

Tabelle 2-15: Konsequenzen des Verstoßes gegen Modellbedingungen eines linearen Regressionsmodells[281]

Die Überprüfung der aufgestellten Regressionsgleichung an der Realität bildet den Abschluss.

[273] Vgl. Empfehlungen von STADEL [2006, S. 117 ff.]: „a) Nicht-Linearitäten: Streudiagramme der (unstandardisierten) Residuen gegen angepasste Werte (Tukey-Anscombe plot) und gegen die (ursprünglichen) erklärenden Variablen, Wechselwirkungen: Pseudo-dreidimensionales Diagramm der (unstandardisierten) Residuen gegen je zwei erklärende Variablen; b) Gleiche Streuungen: Streudiagramme der (standardisierten) absoluten Residuen gegen angepasste Werte und gegen die (ursprünglichen) erklärenden Variablen (meist nicht speziell dargestellt, mit den Streudiagrammen unter (a) mitbetrachtet); c) Normalverteilung: QQ-plot (oder Histogramm) der (standardisierten) Residuen; d) Unabhängigkeit: (Unstandardisierte) Residuen gegen die Zeit oder gegen den Ort auftragen."

[274] Beispielsweise DURBIN / WATSON-Test, siehe HARTUNG [2005, S. 740] oder GOLDFELD / QUANDT-Test, siehe ECKEY et al. [2005, S.79 ff.] und GREENE, W.H. [2003]

[275] BACKHAUS et al. [2005, S. 79]

[276] Vgl. Ergänzende Literatur zum damit ausgedrückten GAUSS-MARKOV-THEOREM findet sich bei BLEYMÜLLER et al. [2004, S.150] und KMENTA [1997, S. 162].

[277] Vgl. BACKHAUS et al. [2006, S. 79 ff.]

[278] Homoskedastizität ist dann vorhanden, wenn die Residuen (Abweichung des Schätzwertes vom Beobachtungswert) eine konstante Varianz aufweisen, vgl. SACHS [2000, S. 551].

[279] Heteroskedastizität liegt dann vor, wenn die Streuung der Residuen in einer Reihe von Werten der prognostizierten abhängigen Variablen nicht konstant ist, vgl. JANSSEN / LAATZ [2005, S. 410].

[280] Multikollinearität besteht dann, wenn zwischen den erklärenden Variablen (Regressoren) eine lineare Abhängigkeit existiert, d.h. eine Korrelation der Variablen vorliegt, vgl. BAHRENBERG [2003, S. 40 ff.] und WETHERILL, G. [1986].

[281] Eigene Bearbeitung nach BACKHAUS [2005, S.93]

2.4.4 Räumliche Autokorrelation

Allgemein versteht man unter räumlicher Abhängigkeit oder räumlicher Autokorrelation die Existenz einer funktionalen Beziehung zwischen dem, was an einem Ort passiert und den Ereignissen anderswo. Die Problematik wird durch das so genannte ‚first law of geography' von WALTER TOBLER[282] präzisiert: "Everything is related to everything, but near things are more related than distant things." Das Phänomen der räumlichen Autokorrelation ist vor allem durch die gleichnamige Arbeit von CLIFF und ORD[283] bekannt geworden.

Bei der Modellierung räumlicher Abhängigkeit unterscheidet ANSELIN[284] zwei verschiedene Ansätze. Die erste Vorgehensweise versucht auf theoretischer Basis die räumliche Abhängigkeit a priori zu berücksichtigen, indem ein entsprechendes formales Modell definiert wird. Bei der zweiten Vorgehensweise werden die Daten als Ausgangspunkt genommen und über verschiedene Maßzahlen wird versucht, ein Bild über die räumliche Abhängigkeit zu bekommen.

Der Index Moran's I[285] misst den Grad der räumlichen Assoziation innerhalb des gesamten Datensatzes $\{x_i : i = 1,...,n\}$. Die Wahl einer räumlichen Gewichtsmatrix wird vorausgesetzt $\{w_{ij} : i = 1,...,n \quad ; \quad j = 1,...,n\}$, und die Elemente w_{ij} charakterisieren die räumliche Nachbarschaft zwischen den Raumeinheiten i und j:

$$I = \frac{n}{S_o} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{w_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{i=1}^{n}(x_i - \mu)^2} \tag{66}$$

mit μ = Mittelwert der Variablen x

w_{ij} = Elemente einer räumlichen Gewichtsmatrix (einfachster Fall: binär)

$$S_o = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$$

Für den Fall einer zeilenstandardisierten Gewichtsmatrix gilt: $S_o = 1$, da jede Zeile sich zu 1 aufsummiert. Es handelt sich um den Normalfall in der Praxis.[286] Moran's I berechnet sich dann aus dem Verhältnis eines räumlichen Kreuzprodukts (Zähler) zu einer Varianz (Nenner).

---

282 Vgl. TOBLER [1979, S. 379-386]

283 Vgl. CLIFF / ORD [1973]

284 Vgl. ANSELIN [1988]

285 Vgl. MORAN [1948, S. 243-251], die Formel ist ähnlich der des Korrelationskoeffizienten nach Pearson. Im Zähler steht die Kovarianz, über die die Werte aller Zahlenpaare verglichen werden. Im Nenner steht ein Schätzwert für die Varianz. Während beim Pearson-Korrelationskoeffizienten die Produkte der Standardabweichungen der beiden Variablen im Nenner stehen, ist es bei Moran's I nur die z-Variable.

286 Vgl. FISCHER, M. M. [2003], Institut für Wirtschaftsgeographie, Regionalentwicklung und Umweltwirtschaft, Abteilung für Wirtschaftgeographie & Geoinformatik, Wirtschaftsuniversität Wien.

Das Maß für die Autokorrelation entspricht der folgenden Gleichung:

$$\text{Moran's } I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{i=1}^{n}(x_i - \mu)^2} \tag{67}$$

Der Autokorrelationskoeffizient nimmt Zahlenwerte zwischen -1 und +1 an. Unter positiver räumlicher Autokorrelation versteht man das Phänomen, dass hohe Werte der gegenständlichen Variablen räumlich geclustert auftreten. Die negative räumliche Autokorrelation berücksichtigt, dass hohe Werte in der Nachbarschaft von niedrigen Werten auftreten. Der Erwartungswert beträgt Null, folglich berücksichtigt die Nullhypothese, dass keine räumliche autokorrelative Beziehung bei der untersuchten Variablen vorliegt. Die räumliche Interpolation der Variablen ist in einem solchen Fall nicht möglich.

In Matrixschreibweise wird das Maß für die Autokorrelation folgendermaßen ausgedrückt:

$$\text{Moran's } I = \frac{(x-\mu)^T w(x-\mu)}{(x_i-\mu)^T(x-\mu)} \tag{68}$$

wobei $(x-\mu)$ ein n – dimensionaler Vektor ist,
w eine (n,n) Matrix ist,
$(x_i-\mu)^T$ der $(x\text{-}\mu)$ zugehörige n - dimensionale Spaltenvektor ist und
T die Transposition eines Vektors bzw. einer Matrix kennzeichnet.

Zu beachten ist bei einer großen Anzahl von zu untersuchenden Raumeinheiten, dass Moran's I als globales Autokorrelationsmaß den Grad der räumlichen Assoziationen im gesamten Datensatz angibt und ggf. zur Identifizierung von lokal vorhandenen Autokorrelationen eine Beschreibung von Teilräumen besser geeignet ist.[287]

Im Fall des Vorhandenseins von Trends, d.h. z.B. die Datenwerte fallen in eine oder mehrere Richtungen ab, reagieren die räumlichen Autokorrelationskoeffizienten empfindlich und führen ggf. zu Fehleinschätzungen, die jedoch mit geeigneten Verfahren vermeidbar sind.[288] Zusätzlich sei auf das ‚Spatial Sampling' verwiesen, welches Einfluss auf die Ergebnisse der räumlichen Autokorrelation nimmt.[289] Es handelt sich um eine Reduzierung von Untersuchungseinheiten in einem Datenmodell (‚Sampling'), wobei zu berücksichtigen ist, dass eine unterschiedliche Verteilung von Stichprobenpunkten ggf. auch andere räumliche Zusammenhänge erzeugt.

[287] Vgl. ROBINSON [1998], WONG / LEE [2005, S. 260 ff. und S. 349 ff.]
[288] Vgl. NIPPER/STREIT [1977, S. 241-263], z.B. Trendoberflächenanalyse
[289] Vgl. LONGLEY [2001, S. 103 ff.] und LO / YEUNG [2002, S. 118 ff.]

2.5 Operationalisierung

Die Operationalisierung bildet die Grundlage für die nachfolgende Wissenskonversion und ermöglicht insbesondere die Wissensgewinnung aus bestehenden Daten unter Einbeziehung der bereits entdeckten Strukturen. Um eine Objektstruktur bzw. entwickelte Klassifikation operational zu gestalten, ist es notwendig Algorithmen zu suchen, die Objekte auf eine Klassenstruktur aufteilen können. Mit Hilfe von konstruierten Klassifikatoren werden Klassifikationen nachvollziehbar beschrieben und ein Verständnis für die Klassen geschaffen. Es werden Zuordnungsvorschriften gesucht, die die gewonnenen Klassifikationen charakterisieren und auch eine nachträgliche Zuordnung von nicht klassifizierten Daten realisieren. Im Kontext des Data Mining wird zwischen subsymbolischen und symbolischen Klassifikatoren unterschieden. Ein symbolischer Klassifikator stellt die Anforderung einer nahezu natürlichsprachigen Beschreibungsform an die einzusetzenden Algorithmen, während ein subsymbolischer Klassifikator die Aufgabe ohne ein genaues Verständnis der Klassen erledigt, d.h. eine regelbasierte Klassenbeschreibung wird nicht berücksichtigt.

2.5.1 Subsymbolische Klassifikatoren

2.5.1.1 K-Nearest Neighbour (KNN-) Klassifikation

Der k-Nearest Neighbour Algorithmus dient der nichtparametrischen Klassifikation und wurde erstmals Anfang der fünfziger Jahre eingeführt.[290] Eine Klassenzuordnung erfolgt unter Berücksichtigung seiner k nächsten Nachbarn. Zu einer gegebenen Beobachtung $x \in IR^d$ wird die Lernfolge nach aufsteigenden Werten von $\|x - X_i\|$ in einer gewählten Norm auf IR^d angeordnet. Dadurch gilt die folgende Gleichung:

$$\left\|x - X_{R_{1:n}}\right\| \le \left\|x - X_{R_{2:n}}\right\| \le \ldots \le \left\|x - X_{R_{n:n}}\right\| \tag{69}$$

wobei $(R_{1:n},\ldots,R_{n:n})$ eine zufällige Permutation des Tupels $(1,\ldots,n)$ ist. Verschiedene Punkte der Lernfolge können denselben Abstand zu x aufweisen. Ein k-NN Klassifikator $\delta_{k,n}$ entscheidet sich für die Klasse, die unter k-nächsten Nachbarn von x am häufigsten auftritt:

$$\begin{aligned} &\delta_{k,n}(x,((X_1,Y_1),\ldots,(X_n,Y_n))) = l \\ &\Rightarrow \left|\left\{i \in \{1,..,k\} \middle| Y_{R_{i:n}} = l\right\}\right| = \max_{j \in \{1,\ldots,m\}} \left|\left\{i \in \{1,\ldots,k\} \middle| Y_{R_{i:n}} = j\right\}\right| \end{aligned} \tag{70}$$

Falls es mehrere Musterklassen gibt, für die dieses Maximum zutrifft, so können Zufallsexperimente eingesetzt werden.

---

290 Vgl. FIX / HODGES [1951] und [1952]

Das Verfahren beinhaltet die Klassifikation mit einem Referenzdatensatz, der sowohl die Eingangs- als auch die Zielvariablen enthält. Dieser Datensatz dient als Grundlage für den Vergleich mit unbekannten Daten. In Bezug auf die Anzahl der k nächsten Nachbarn besteht für ein zu klein gewähltes k die Gefahr, dass Rauschen in den Trainingsdaten die Klassifikationsergebnisse verschlechtert. Für k =1 ergibt sich ein Voronoi-Diagramm. Falls k zu groß gewählt wird, besteht die Gefahr, Punkte mit großem Abstand zu x in die Klassifikationsentscheidung mit einzubeziehen. Besonders groß ist diese Gefahr, wenn die Trainingsdaten nicht gleichverteilt vorliegen oder nur wenige Beispiele vorhanden sind. Eine gewichtete Abstandsfunktion kann dazu dienen, den näheren Punkten ein höheres Gewicht zuzuweisen als den weiter entfernten.

Bei einer größeren Merkmalsanzahl besteht zusätzlich das Problem, dass das Abstandsmaß (z.B. Euklidische Distanz) immer über alle Punkte berechnet wird, so dass eventuell weniger relevante Attribute zu Verzerrungen führen können. Als Lösung für dieses Problem könnte auch die Gewichtung der Einzelmerkmale angesehen werden, um auf diese Weise eine Streckung der Achsen zu erzielen.

2.5.1.2 Fuzzy-Pattern-Klassifikation

Das Verfahren der Fuzzy-Pattern-Klassifikation (FPK) dient der unscharfen Beschreibung und dem Erkennen von Situationen, Zuständen, Verhältnissen durch Vektoren in einem Merkmalsraum und wird den Mustererkennungs- und -beschreibungsverfahren zugeordnet.[291] Während Clusteralgorithmen Strukturen in Datensätzen nach definierten formalen Kriterien (Distanzen im Merkmalsraum) aufdecken, geht die Fuzzy-Pattern-Klassifikation von strukturierten Datensätzen aus, und verallgemeinert die Struktur der Objekte, indem unscharfe Klassenbeschreibungen mittels Potentialfunktionen erzeugt werden.[292]

Die Strukturierung der Daten wird von einem Lehrer / Experten nach inhaltlichen (semantischen) Kriterien vorgenommen, wobei Objekte Klassen zugewiesen werden. Dazu ist es erforderlich, dass ein Satz von Merkmalen bestimmt wird, anhand dessen die Klassenzuordnung durch das System erfolgen soll. Bei der Bildung des Klassifikationsmodells handelt es sich um eine deterministische Berechnung der Klassifikatoren, d.h. nicht um eine multivariate Optimierung. Daraus resultieren extrem kurze Rechenzeiten in der Entwicklungsphase der Klassifikatoren. Die Klassifikatoren sind transparent d.h. der Beitrag, den jedes einzelne Merkmal zur Klassentrennung leistet, kann verfolgt werden.

---

[291] Vgl. BOCKLISCH [1987], PÄßLER [1998] und MANN [1983]
[292] Vgl. BOCKLISCH [1987, S. 114 ff.]

Es ist möglich, Expertenwissen bei der Klassenbildung mit einzubeziehen und die Klassifikationsstruktur ergibt sich automatisch. Es wird eine plattformunabhängige leicht portierbare Applikation sichergestellt, da die Klassifikatoren als Datensätze vorliegen.

Die FPK steht in enger Beziehung zu den Künstlichen Neuronalen Netzen (KNN),[293] die aus vorgegebenen Lerndatensätzen in iterativer Optimierung ein Übertragungsverhalten antrainieren, das in der Lage ist, unbekannte Daten zu klassifizieren und Muster zu erkennen. Probleme der Anwendung von KNN liegen jedoch darin, dass es keine gesicherten Methoden für die Wahl einer optimalen Netztopologie gibt, a priori vorhandenes Expertenwissen in das Modell nicht eingebracht werden kann und die Optimierung für das Lernen erhebliche Rechenzeit in Anspruch nehmen kann.

Grundbegriffe

Für die unscharfe Beschreibung einer elementaren oder globalen Information wird eine spezielle Zugehörigkeitsfunktion $\mu(x)$ erforderlich. Es haben sich parametrische Zugehörigkeitsfunktionen $\mu\left(x, \vec{p}\right)$ mit gut interpretierbaren Parametern durchgesetzt, die einen Kompromiss zwischen Flexibilität und mathematischer sowie rechentechnischer Handhabbarkeit ermöglichen. Eine gute Anpassungsfähigkeit an vorliegende Datenkonfigurationen und -strukturen ist zu erfüllen. Wichtige Vertreter parametrischer Zugehörigkeitsfunktionen sind Dreiecks-, Trapez-, Zadeh-, Potential-, Negations- und Exponentialfunktionen.[294]

Im Allgemeinen lassen sich unscharfe Mengen grafisch repräsentieren und werden durch die Aufzählung ihrer Elemente dargestellt: $A = \{(x_1, \mu(x_1)),...,(x_n, \mu(x_n))\}$. Eine unscharfe Menge A über $X = R^1$ wird beschrieben gemäß Abbildung 2-19:

$$\mu^A\left(x, \vec{p}^{\,A}\right) = \begin{cases} \dfrac{a}{1+\left(\frac{1}{b^r}-1\right)\left(\frac{x-u_0}{c^r}\right)^{d^r}} & \text{für } x \geq u \\ \dfrac{a}{1+\left(\frac{1}{b^l}-1\right)\left(\frac{u_0-x}{c^l}\right)^{d^l}} & \text{für } x \leq u \end{cases}$$

mit $\vec{p}^{\,A} = \left(\vec{p}_s, \vec{p}_u\right)$

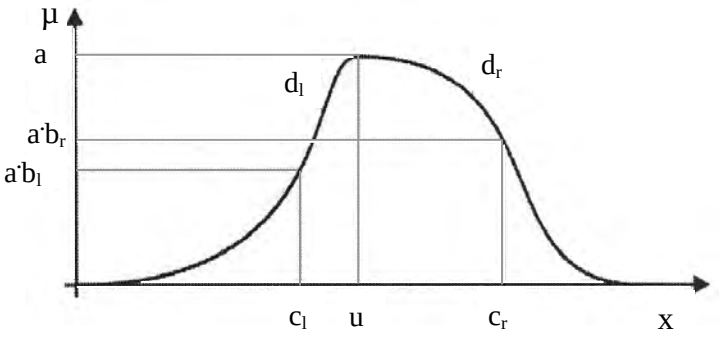


Abbildung 2-19: Darstellung der Aizermannschen Potentialfunktion und ihrer Parameter[295]

293 Vgl. BORGELT et al. [2003, S. 1 ff.] und ZIMMERMANN [1995b, S. 26 ff.]: künstliche neuronale Netze
294 Vgl. BOCKLISCH [1987, S. 75 ff.], PÄßLER [1998] und SCHEUNERT [2001, 24 ff.]
295 HEMPEL [2005, S. 11]

Der Vektor $\vec{p}_s = \left(\vec{u}_0\right)$ fasst dabei die Parameter der scharfen Lageinformation einer unscharfen Menge zusammen. Für die unscharfe Beschreibung, wie Ausdehnung und Form der unscharfen Menge, enthält $\vec{p}_u = \left(a, b^l, c^l, c^r, d^l, d^r\right)$ die modellierenden Parameter.

Die Lage einer unscharfen Menge wird durch den Parameter u_0 beschrieben. Dieser wird als Repräsentant bezeichnet und gibt den Ort der höchsten Zugehörigkeit an. Alle anderen Parameter beziehen sich auf den Repräsentanten.

Die Parameter c^l, c^r beschreiben das als scharf betrachtete Wirkungsgebiet einer unscharfen Menge. Man unterscheidet zwischen einer links- und rechtsseitigen Ausprägung. Die Parameter b^l, b^r legen die links- bzw. rechtsseitige Zugehörigkeit des Randes ($u_0 + c^r$ bzw. $u_0 - c^l$) einer unscharfen Menge fest.

Die Parameter d^l, d^r bestimmen den stetig fallenden Verlauf der Zugehörigkeitsfunktion ausgehend vom Repräsentanten. Der Parameter a gilt speziell für Klassenzugehörigkeitsfunktionen und gibt die Bedeutung einer Klasse relativ zu anderen an.

Für die Klassenzugehörigkeitsfunktion kann a eine Information über die Glaubwürdigkeit der Stützung der Klassen durch die Lernobjekte tragen.[296]

Das Parametrische Funktionenkonzept der Aizermanschen Potentialfunktion besitzt die folgenden Kennzeichen:

- Elementare und globale Ereignisse lassen sich mit Zugehörigkeitsfunktionen $\left(\mu_e(x) \text{bzw.} \mu_g(x)\right)$ beschreiben.[297]
- Beschreibungen erfolgen parametrisch und die Funktion ist unimodal und im Allgemeinen nicht symmetrisch.[298]
- Die Parameter des Funktionenkonzeptes sind physikalisch sehr anschaulich, gesondert interpretierbar und tragen jeweils voneinander unabhängige Informationen.
- Der Beschreibungsaufwand wird reduziert durch die Projektion über festgelegte Achsen der im n-dimensionalen Raum erklärten Zugehörigkeitsfunktionen auf n eindimensionale Funktionen.

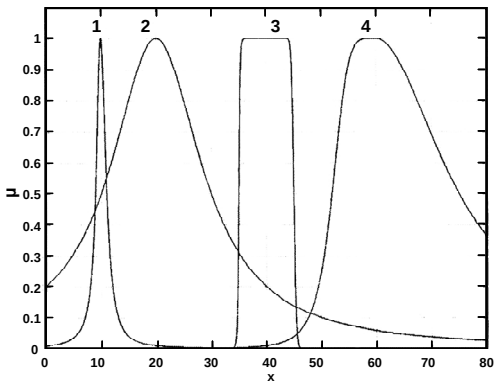
---

296 Vgl. BOCKLISCH [1987, S. 111], TAT INGENIEURBÜRO BURMEISTER [1997, S. 51]

297 Bei Kombination mit der hierarchischen Klassifikation entstehen Vorteile, da hier die Betrachtungsschärfe beispielsweise über den Schwellwert gesteuert werden kann, vgl. BOCKLISCH [1987, S. 110].

298 BOCKLISCH [1987, S. 110] verweist auf die Vorteile bei der Beschreibung von Klassen, die aus der Aggregation vieler elementarer Ereignisse (Beobachtungen, Messungen) hervorgehen.

In Abbildung 2-20 werden einige qualitative Möglichkeiten der Beschreibung unscharfer Mengen mittels Aizermannscher Potentialfunktion an ausgewählten Beispielen zusammenfassend dargestellt.



Parametersatz Kurve 1:
$\vec{p}_s(u_0 = 10),\ \vec{p}_u(a = 1,\ b^{lr} = 0.5,\ c^{lr} = 2,\ d^{lr} = 2)$

Parametersatz Kurve 2:
$\vec{p}_s(u_0 = 20),\ \vec{p}_u(a = 1,\ b^{lr} = 0.5,\ c^{lr} = 10,\ d^{lr} = 2)$

Parametersatz Kurve 3:
$\vec{p}_s(u_0 = 40), \vec{p}_u(a = 1, b^{lr} = 0.5, c^{lr} = 5, d^{lr} = 30)$

Parametersatz Kurve 4:
$\vec{p}_s(u_0 = 60),\ \vec{p}_u(a = 1,\ b^{l} = 0.5,\ b^{r} = 0.5,\ c^{l} = 8,\ c^{r} = 15,\ d^{l} = 5,\ d^{r} = 2)$

Abbildung 2-20: Beispiele von Aizermanschen Potentialfunktionen[299]

Aufgrund der ausdehnungsgebenden und formbestimmenden Parameter ist die Variation der Zugehörigkeitsfunktion von Peaks (Kurve 1), Glockenformen (Kurve 2) bis hin zu scharfen Kurven (Kurve 3) möglich. Die Kurve 4 zeigt die Flexibilität des Potentialfunktionenkonzeptes anhand der seitenspezifischen Kombination verschiedener Ausdehnungs- und Formparameter.

Als Vorteil des Konzeptes der Aizermannschen Potentialfunktion ist die einfache Deutung durch gut interpretierbare Parameter anzusehen. Des Weiteren sind die Modellierung mengeninterner Verteilungen und die Modellierung eines Zugehörigkeitscharakters hervorzuheben. Die seitenspezifischen Parameter ermöglichen eine hohe Flexibilität und Problemanpassung.

Elementare Unschärfe

Im Rahmen der Fuzzy-Pattern-Klassifikation werden die Elemente des Klassifikationsprozesses ausschließlich in Form von Potentialfunktionen beschrieben. Dies bedeutet, dass Objekte von einer scharfen (punktuellen) in eine unscharfe Beschreibung übergehen. Die Objekte werden dabei mittels gewöhnlicher mehrdimensionaler Potentialfunktionen über ihrem Grundbereich dem so genannten Merkmalsraum abgebildet. Durch die unscharfe Beschreibung erhalten die Objekte eine Ausdehnung (Wirkungsbereich), die als elementare Unschärfe bezeichnet wird. Die elementare Unschärfe wird durch die spezielle Indizierung c_e^{lr} des Ausdehnungsparameters gekennzeichnet.

[299] Vgl. HEMPEL [2005, S. 12]

Unter dem Blickwinkel der Fuzzy-Theorie werden unscharfe Objekte durch die Verknüpfung von unscharf beschriebenen Realisierungen der Merkmale eines Objektes gebildet. Die Basisfunktion der Objekte wird durch die Realisierung der Merkmale gebildet. Ein Beispiel der elementaren Unschärfe für eindimensionale Objekte zeigt Abbildung 2-21.

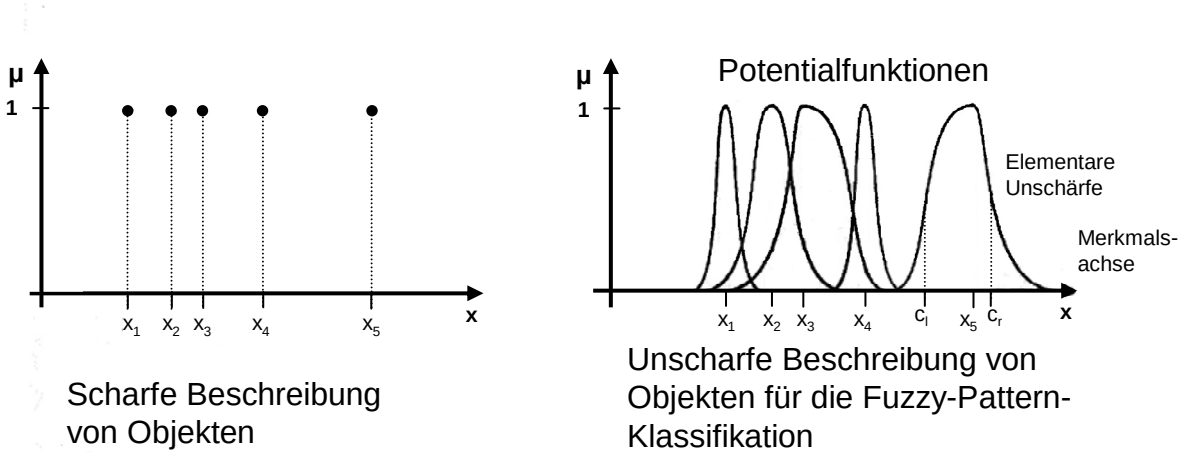


Abbildung 2-21: Elementare Unschärfe bei eindimensionalen Objekten[300]

Die Darstellung unscharfer Objekte mit den zugehörigen realisierten Merkmalen wird anhand eines zweidimensionalen Beispiels für den Merkmalsraum $X_1 \times X_2$ in Abbildung 2-22 gezeigt.

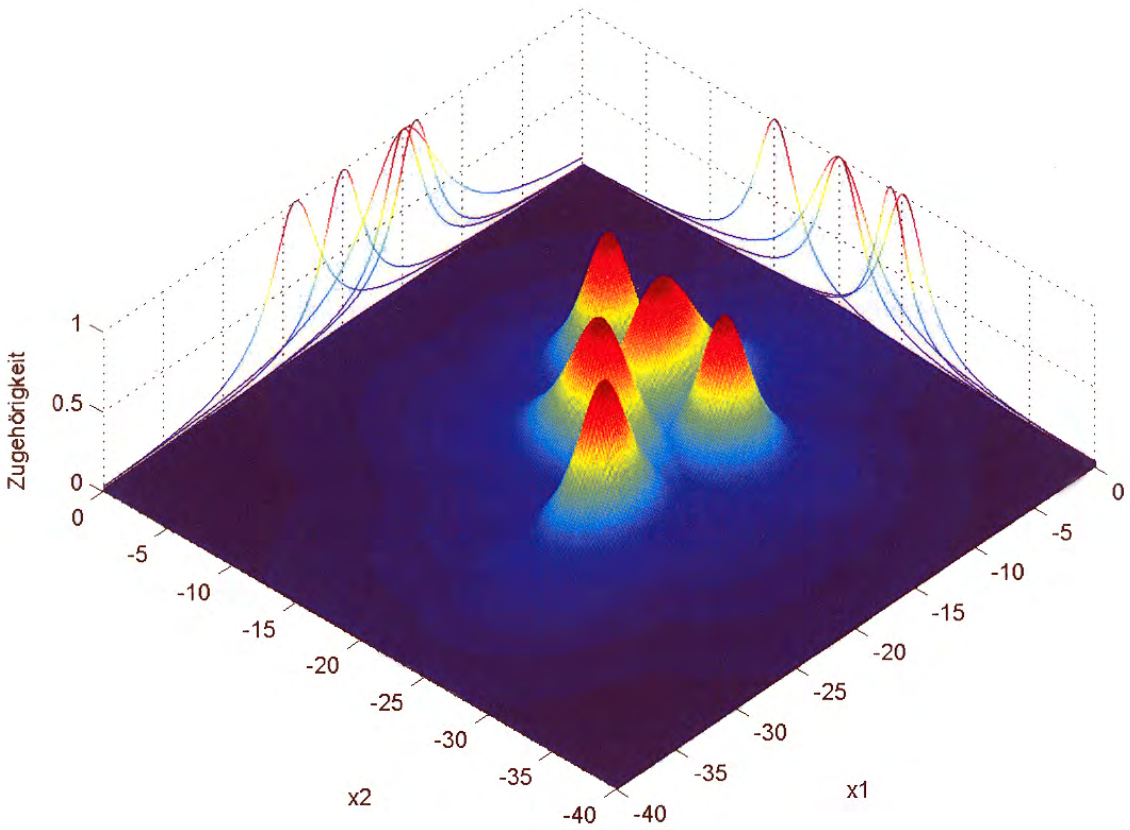


Abbildung 2-22: Unscharfe Objekte mit zugehörigen Merkmalsrealisierungen[301]

[300] Vgl. HEMPEL [2005, S. 35]
[301] Vgl. HEMPEL [2005, S. 36]

Aufbau eines Fuzzy-Pattern-Klassifikators

Man unterscheidet bei der FPK eine Lern- und eine Arbeitsphase, die den Aufbau und die Nutzung eines Klassifikators unabhängig von der Klassifikationsmethode ermöglicht. In der Lernphase werden merkmalsweise Zugehörigkeitsfunktionen der Klassen aggregiert, d.h. es werden die Parameter der klassenbeschreibenden Zugehörigkeitsfunktionen aus elementaren Informationen (Objekten) ermittelt. Der Bildungsprozess einer Klassenstruktur aus Objektdaten und deren Beschreibung wird Lernprozess genannt und die Zeitspanne, während dieser Prozess läuft, heißt Lernphase. Das Ergebnis der Lernphase sind Klassifikatoren, wobei anzustreben ist, dass die Güte des Klassifikators während der Lernphase zielgerichtet verbessert wird. Die Anwendung erfolgt in der Arbeitsphase mit dem Ziel, unbekannte Objekte mit Hilfe von Algorithmen oder auf deren Basis arbeitenden Programmen oder Datenstrukturen hinsichtlich einer Klassenzugehörigkeit zu identifizieren. Das Basismaterial (Trainingsobjekte) für die Klassifikatoren bilden die Lernobjekte. Ein Klassifikator kann im ‚offline'-Betrieb als Beratungssystem dienen und unter ‚on-line'-Bedingungen als Steuerungs- oder Diagnosesystem für einen Prozess (Abbildung 2-23). Man unterscheidet stationäre und instationäre Klassifikatoren, je nach Adaptionsfähigkeit zum Selbstlernen: $K = K(t)$.

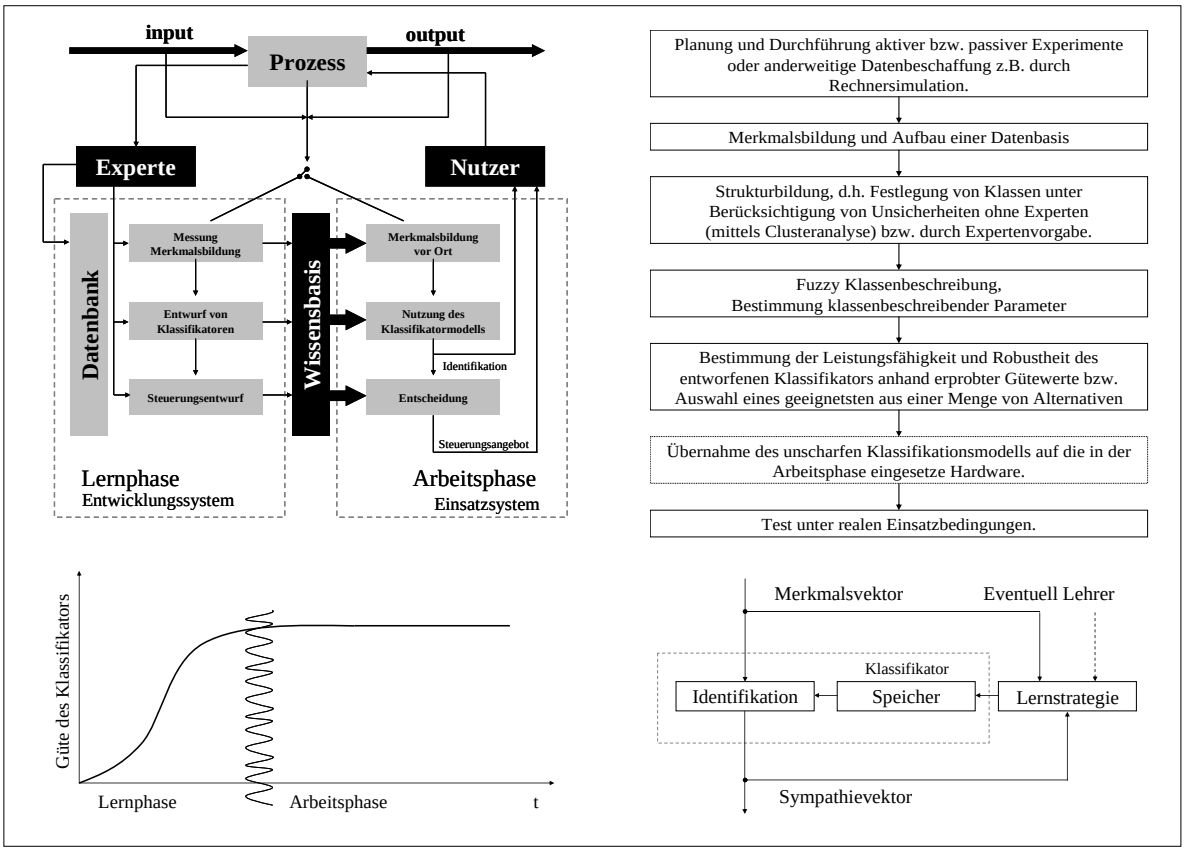


Abbildung 2-23: Entwicklungsschritte beim Aufbau eines Klassifikators[302]

[302] Eigene Bearbeitung aus TAT Ingenieurbüro BURMEISTER [1997, S.27,69], BOCKLISCH [1987, S.32,41]

Die unscharfe Modellierung eines Phänomens besteht in der Regel aus mehreren verschiedenen Klassen mit jeweils individuellen Bedeutungen und die Gesamtheit bildet den Klassifikator. Die Identifikation von Objekten ist dadurch charakterisiert, dass für den Vorgang der Klassifikation der Übergang von einer scharfen Merkmalsbewertung zu einer unscharfen Interpretation erfolgt und Objekte zu vorgegebenen Klassen zugeordnet werden. Wird ein unbekanntes Objekt im scharfen Sinne einer Ja-Nein-Entscheidung eindeutig einer Klasse zugeordnet oder nicht, so dient als Ausdrucksmittel der Zuordnungsvektor μ, der mit seinen Werten μ_K eine Zuordnung zu einer der K-Klassen anzeigt.[303] Es handelt sich um einen Klassifikator mit disjunkten Klassen (unscharfer Klassifikator), wenn während der Identifikation ein unscharfes Objekt dem Klassifikator präsentiert und bezüglich jeder unscharfen Klasse bewertet wird, so dass ein Sympathiewert zur jeweiligen Klasse vorliegt:

$$\mu = (\mu_1, \mu_2, ..., \mu_k, ..., \mu_K) \text{ mit } \mu_K = \begin{cases} 0 \text{ falls } X_i \notin K_k \\ 1 \text{ falls } X_i \in K_k \end{cases} \text{ und } \sum_{k=1}^{K} \mu_k = 1 \text{ bzw. } \sum_{k=1}^{K} \mu_k \geq 1 \qquad (71)$$

Man spricht von einer statischen Identifikation, wenn ein Objekt einem Klassifikator nur einmal vorgeführt und anhand der Merkmalswerte des Objektes eine Zuordnung vollzogen wird. Für die Identifikation dynamischer Vorgänge bieten sich zwei Möglichkeiten, d.h. entweder wird ein dynamisches Objekt in ein statisches überführt, indem die Abtastwerte des zeitlichen Vorganges als Merkmale betrachtet werden oder ein dynamischer Vorgang wird abgetastet und jeder Abtastwert wird als so genanntes zeitmarkiertes Objekt aufgefasst. Im erst genannten Fall entstehen leicht extrem hochdimensionale Merkmalsräume, und im zweiten Fall ergibt die Identifikation eine Trajektorie im Merkmalsraum (siehe Abbildung 2-24), die in ihrer Gesamtheit eine dynamische Prozessbeschreibung ermöglicht, und Trendbeschreibungen sowie Vorhersageprobleme schließen sich an.[304]

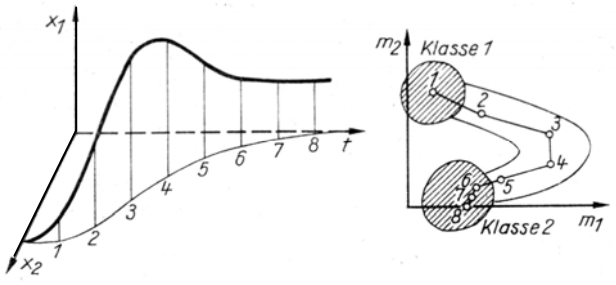


Abbildung 2-24: Übergangsvorgang im Zeitverlauf (links) und Trajektorie im Merkmalsraum (rechts)[305]

303 Vgl. BOCKLISCH [1987, S.33]

304 Vgl. BOCKLISCH [1987, S.34 ff.], PÄßLER [1998], [1999], es besteht noch großer Forschungsbedarf beim Einsatz von Fuzzy-Pattern Modellen, die sich für die Zeitreihenanalyse und -prognose eignen. Zukünftig ist deren Einsatz im Gebiet der Stadtklassifikation denkbar und bei ausreichend vorhandenem Datenmaterial geeignet, weitere Analyseergebnisse zu erarbeiten.

305 Quelle: BOCKLISCH [1987, S.36]

Bewertung eines Fuzzy-Pattern-Klassifikators

Die Leistungsfähigkeit und Robustheit von entworfenen Fuzzy-Pattern-Klassifikatoren ist anhand von Gütewerten zu beurteilen, und es lassen sich Aussagen treffen, inwieweit der untersuchte Klassifikator das zu modellierende System abbildet und wie stabil diese Abbildung bei Störungen ist. Kenngrößen[306] sind z.B. die Klassenzahl, der Prozentsatz der richtig in die Klassen eingeordneten Lernobjekte, Kompaktheitsmaße, Entropiewerte, Reklassifikationsaussagen oder zusätzlich subjektive Einschätzungen eines Experten.

Die Güte eines Klassifikators ist daran abzulesen, wenn ein Objekt zur semantisch richtigen Klasse zugeordnet wird und sich dies durch einen hohen Sympathiewert zu dieser Klasse ausdrückt (Hauptsympathiewert $\hat{\mu}$) sowie zu den anderen Klassen signifikant niedrigere Werte (Nebensympathiewerte) vorliegen. Ein Qualitätsmaß bezieht sich auf den Prozentsatz der auf diese Weise entstandenen richtigen Zuordnung und ist gültig für die Testobjekte, so dass für eine entsprechend globale Einschätzung genügend repräsentative Testobjekte vorhanden sein müssen.

Die Einschätzung der relativen Hauptsympathiewerte dient als Kenngröße für die Klassifikatorbewertung. Es wird der höchste Sympathiewert zum theoretisch maximal möglichen ins Verhältnis gesetzt und der Wert sollte möglichst nahe 1 liegen:

$$g_{1i} = \frac{\hat{\mu}_i^*}{\mu_{\max,i}^*} \tag{72}$$

Der Sympathiewert zur richtigen Klasse drückt μ_i^* aus und sollte hoch sein, der höchste muss er aber nicht sein. Die Maße

$$g_2 = \frac{\sum_{i=1}^{I} \mu_i^*}{\sum_{i=1}^{I} \hat{\mu}_i} \text{ und } g_3 = \frac{1}{I} \sum_{i=1}^{I} g_{1i} \tag{73}$$

drücken über alle Testobjekte die Erfüllung der Forderung aus, dass μ_i^* für jedes Objekt möglichst hoch sein soll, g_2 und g_3 sollten nahe 1 liegen. Für einen Wert 0 ist der Klassifikator unbrauchbar. Der Wert von g_2 und g_3 ist die Reklassifikationsrate, wenn nur einer der Sympathiewerte den Wert 1 annimmt und die anderen 0 sind und als Testobjekte die Lernobjekte benutzt werden. Die Zahl trifft eine Aussage über die richtig klassifizierten Lernobjekte bezogen auf die Gesamtzahl.

[306] Vgl. BOCKLISCH [1987, S. 42-46], TAT INGENIEURBÜRO BURMEISTER [1997, S. 70-73]

Es gelten weitere Bewertungsgrößen unter der Voraussetzung, dass g_2 und g_3 nahe 1 liegen. Der Quotient aus dem maximalen Nebensympathiewert und der Hauptsympathie sollte möglichst klein sein:

$$g_{4i} = \max_{\substack{k=1,\ldots,K \\ k \neq \hat{k}}} \frac{\mu_{k,i}}{\hat{\mu}_i} \tag{74}$$

Wenn der Mittelwert von allen Nebensympathien gebildet wird, ergibt sich die auf die Hauptsympathie bezogene mittlere Nebensympathie zu:

$$g_{5i} = \sum_{k=1}^{K} \frac{\mu_{k,i} - \hat{\mu}_i}{(K-1) \cdot \hat{\mu}_i} \tag{75}$$

Der Verhältniswert soll möglichst klein sein, um eine gute Trennung zwischen Haupt- und Nebensympathie zu erreichen.

Die Entropie h kann verwendet werden, um mit Hilfe eines formalen Kriteriums den Informationsgehalt der verwendeten Merkmale zu bewerten. Es gilt die folgende Definition für die Entropiebetrachtung unter Berücksichtigung der Normierung $\sum_{k=1}^{K} \mu_{k,i} = 1$:

$$h_i = -\frac{1}{ldK} \sum_{k=1}^{K} \mu_{k,i} \cdot ld\mu_{k,i} \tag{76}$$

Bei einer klaren Zuordnung liegt ein geringer Entropiewert vor, d.h. es ist eine hohe Relevanz des Merkmalsatzes für die Identifikation einer Klassenstruktur vorhanden.

Der Überdeckungsgrad kann als globales Kriterium für die Klassifikatorgüte verwendet werden und gibt Aufschluss darüber, inwieweit die Nebensympathiewerte die Hauptsympathiewerte überdecken. Wird bei der Fuzzy-Identifikation einer Klasse ein hoher Zugehörigkeitswert zugeteilt und den übrigen geringere, so ist ein geringer Überdeckungsgrad vorhanden. Dies ist nur dann der Fall, wenn Merkmale für die Klassifikation verwendet wurden, die den Sachverhalt besonders deutlich widerspiegeln. Definiert wird der Überdeckungsgrad als:

$$g_5' = \frac{1}{\sum_{i=1}^{I} \hat{\mu}_i} \sum_{i=1}^{I} \hat{\mu}_i \cdot g_{5i} = \sum_{i=1}^{I} \frac{\left(\sum_{i=1}^{I} \mu_{k,i} - \hat{\mu}_i \right)}{(K-1) \cdot \sum_{i=1}^{I} \hat{\mu}_i} \tag{77}$$

2.5.2 Symbolische Klassifikatoren durch Regelextraktion bzw. Entscheidungsbäume

2.5.2.1 Algorithmus: sig* (Signifikanz der merkmalsbasierten Klassenbeschreibung)

Der entwickelte Algorithmus des maschinellen Lernens sig*[307] generiert Regeln aus bereits klassifizierten Daten. Auf diese Weise wird die Entdeckung von neuem bislang unbekanntem Wissen in den Daten (Knowledge Discovery) unterstützt.

Den ersten Schritt bildet die Ermittlung der Variablen auf Grundlage einer Signifikanzmatrix, die für die jeweilige Klassenbeschreibung signifikant sind. Es können für jede gefundene Klasse andere und unterschiedlich viele Variablen sein. Dadurch werden für jede Klasse individuell charakterisierende Regeln geformt. Sollten sich zwei Klassen nicht eindeutig durch derartige Regelbildung unterscheiden lassen, so werden darüber hinaus so genannte Differenzierungsregeln entwickelt, um dennoch eine Klassenabgrenzung zu ermöglichen. Es handelt sich um eine differential-diagnostische Vorgehensweise, die in den verschiedensten Fachdisziplinen eingesetzt werden kann. Unter Berücksichtigung der Signifikanz werden für Klassen, die bereits durch wenige Merkmale ausreichend charakterisiert sind, auch einfache Regeln gefunden. Die Qualität der erzeugten Regeln bzw. der Klassenbildung wird mit zusätzlichen Maßzahlen wie z.B. Sensitivität oder Spezifität beurteilbar.

Der sig*-Algorithmus bildet in dieser Arbeit ein wichtiges Element, um die gefundenen Ähnlichkeitsmuster mit Hilfe von Regeln zu beschreiben. Die Regeln dienen zum Aufbau der Klassifikatoren. Im Folgenden wird eine Vorgehensweise gezeigt, um einen Klassifikator in die Lage zu versetzen auf nachträglich vorgelegte Objekte richtig zu reagieren, d.h. die Objekte der richtigen Klasse nachträglich zuzuordnen. Die Beurteilung der Klassifikatorleistung bezieht sich sowohl auf die Qualität der Klassenzuweisung der bereits untersuchten Objekte als auch auf das Zuordnungsergebnis neuer Fälle. In der Regel sind drei Arten von Datensätzen zu unterscheiden, die ULTSCH[308] wie folgt definiert:

1. „**Lerndatensatz**: Das ist der Datensatz, mit dem Klassifikatoren konstruiert werden."
2. „**Testdatensatz**: Das ist der Datensatz, mit dem die Generalisierungsfähigkeit des konstruierten Klassifikators gemessen und der Klassifikator weiterhin optimiert wird."
3. „**Validierungsdatensatz**: Mit diesem Datensatz wird, ohne dass eine weitergehende Optimierung des Klassifikators stattfindet, gemessen, inwieweit sich die mit den ersten beiden Datensätzen konstruierten Klassifikatoren generalisieren können."

---

[307] Der Algorithmus sig* wurde von ULTSCH [1991] entwickelt und steht für die Signifikanz der merkmalsbasierten Klassenbeschreibung, die sich ursprünglich auf SOM-Karten und die Ergebnisse der U*-Matrix (siehe Abschnitt 0) beziehen. Der Algorithmus wurde für medizinische Anwendungen in der Habilitationsschrift erfolgreich erprobt und später in anderen Fachdisziplinen ebenso angewendet.

[308] ULTSCH [2006 c] – Vorlesungsunterlagen

Abbildung 2-25 zeigt in schematischer Form zum Allgemeinverständnis die Arbeitsweise des entwickelten sig*-Algorithmus.

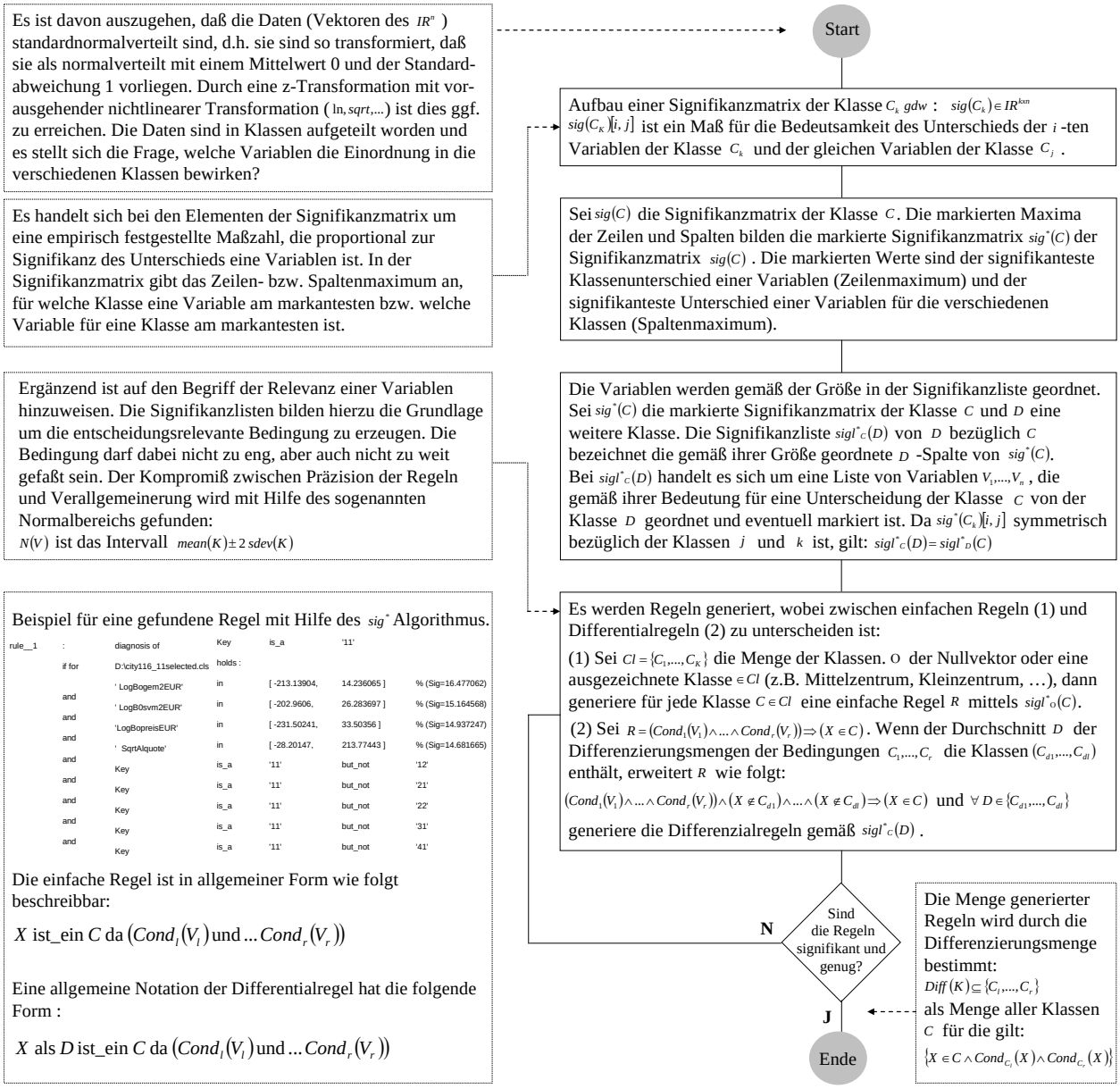


Abbildung 2-25: Ablauf des Algorithmus sig* (Generierung von Entscheidungsregeln)[309]

Es ist darauf hinzuweisen, dass die mit dem sig*-Algorithmus generierten Regeln oftmals über eine sehr hohe Sensitivität verfügen, um vorgelegte bisher nicht zugeordnete Fälle nachträglich korrekt einer Klasse zuzuordnen. Der sig*-Algorithmus ist nach Meinung der Entwickler besonders in Fachbereichen einsetzbar, bei denen es auf die Gewinnung bzw. die Kommunizierung von neuen Erkenntnissen ankommt. Nach aktuellem Wissensstand des Verfassers dieser Arbeit wurde der Algorithmus sig* bisher nicht im Bereich der Klassifikation von Städten eingesetzt.

309 Eigene Bearbeitung unter Berücksichtigung von ULTSCH [1991, S. 169 ff] als Urheber des Algorithmus.

2.5.2.2 Algorithmus: CART (Classification and Regression Trees)

Der Algorithmus Classification and Regression Trees (CART)[310] stellt ein maschinelles Lernverfahren dar und liefert eine Ausgabe in Form von Entscheidungsbäumen, deren Struktur von einem menschlichen Beobachter interpretierbar ist. Es gibt mehrere Arten einen Entscheidungsbaum wachsen zu lassen. Als zentrales Element von CART ist die Fähigkeit anzusehen, eine binäre Trennung aufgrund eines Optimierungsprozesses vorzunehmen. CART teilt jeden Eltern-Knoten in genau zwei Kind-Knoten auf, indem an jedem Entscheidungspunkt eine Frage mit Ja/Nein Antwort gestellt wird. Durch CART werden Fragen gesucht, welche die Knoten in relativ homogene Kind-Knoten teilen. Während der Baum wächst, werden die Knoten immer homogener, und wichtige Segmente werden identifiziert. Klassifikations- und Regressionsbäume unterscheiden sich von der Baumstruktur nicht. Während die Zielvariable bei Klassifikationsbäumen kategoriell ist, sind es bei Regressionsbäumen stetige Größen. Bei Klassifikationsbäumen kommt es darauf an, in einem Knoten möglichst nur noch Werte einer Gruppe der Zielvariablen zu versammeln, und bei Regressionsbäumen wird das Ziel verfolgt, die Abweichung zwischen den Werten im Knoten und dem geschätzten Wert des Knotens zu minimieren. Das übliche Maß ist die Summe der quadratischen Abweichungen. Die Hauptschwäche von Entscheidungsbäumen liegt darin, dass nur orthogonale Trennungen erzeugt werden (Treppenbildung). Die Bäume sind nur sehr begrenzt stabil gegenüber Änderungen der Eingabewerte. Zu jedem Baum ist in einem so genannten ‚Treemap' eine 2-dimensionale flächige Darstellung möglich. ‚Treemaps' sind wesentlich kompakter als Bäume, jedoch ist die Baumstruktur nur schwer erkennbar, da die Gliederung durch unterschiedlich gefärbte Flächen symbolisiert wird. Die Treemaps eignen sich, weitere Variablen darzustellen. Abbildung 2-26 zeigt beide Darstellungsformen.

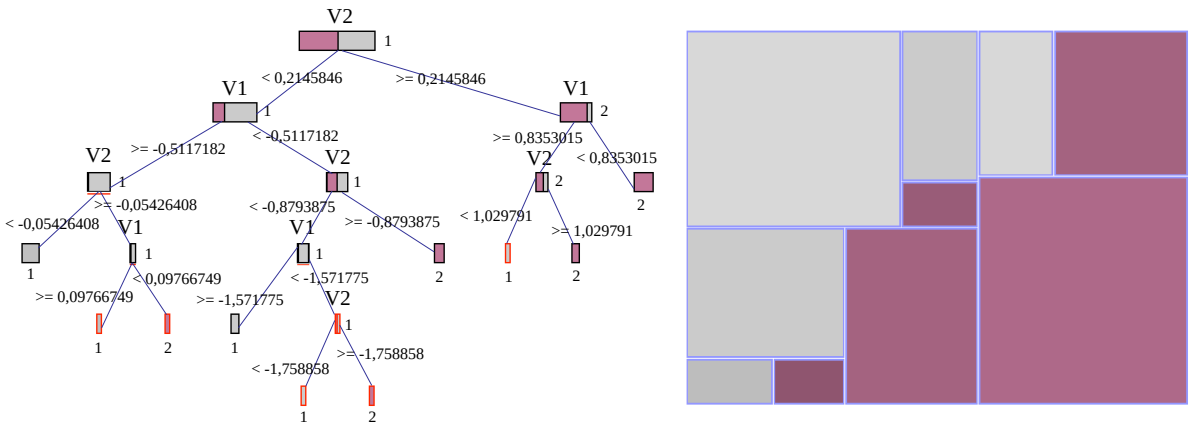


Abbildung 2-26: Beispiele für einen Entscheidungsbaum und einen so genannten ‚Treemap'[311]

---

310 Vgl. BREIMAN et al. [1984] publizieren diesen Algorithmus im Jahr 1984. Es gilt als Grundlagenwerk und schuf das Gebiet der hochentwickelten, mathematisch und theoretisch fundierten Entscheidungsbäume.

311 THEUSS [2004] – Vorlesungsunterlagen zu Multivariaten Statistischen Verfahren, Universität Augsburg

Die Kriterien, um eine binäre Trennung am Knoten bei Klassifikations- oder Regressionsbäumen vorzunehmen, sind in der Tabelle 2-16 kurz dargestellt.

Kriterium zur Trennung bei Klassifikationsbäumen
Gegeben sei die kategorielle Zielvariable mit Werten von $1, 2, ..., K$. Für eine Region R_m mit N_m Beobachtungen ist $\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$ der Anteil Beobachtungen von k Klassen im Knoten m. Die Zuordnung des Knotens zur Klasse geschieht über $k(m) = \arg\max_k \hat{p}_{mk}$. Als übliche Maße der Unreinheit $Q_m(T)$ der Knoten m eines Baums T seien genannt: ▪ Missklassifikation: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$ ▪ Gini Index: $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$ ▪ Entropy bzw. Devianz: $\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$ Im Fall von nur 2 Klassen (p sei der Anteil der 2. Klasse) vereinfachen sich die Maße in folgender Weise: ▪ Missklassifikation: $1 - \max(p, 1-p)$ ▪ Gini Index: $2p(1-p)$ ▪ Entropy: $-p \log p - (1-p)\log(1-p)$
Kriterium zur Trennung bei Regressionsbäumen
Für die Zielvariable y, eine potentielle Variable j zur Trennung (Split) aus $1, ..., p$ Variablen, und einem potentiellen Punkt s zur Trennung (Split), definieren sich zwei Halbebenen im IR^p in der folgenden Weise: $R_1(j,s) = \{X \mid X_j \leq s\}$ und $R_2(j,s) = \{X \mid X_j > s\}$ Gesucht wird daraufhin die Variable j zur Trennung (Split) und der Punkt der Trennung (Split) s, die $\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$ erfüllen. Für beliebige j und s wird die innere Minimierung durch $\hat{c}_1 = ave(y_i \mid x_i \in R_1(j,s))$ und $\hat{c}_2 = ave(y_i \mid x_i \in R_2(j,s))$ gelöst. In jedem Schritt wird also das optimale Paar (j,s) gesucht.

Tabelle 2-16: Kriterien zur Trennung bei Klassifikations- oder Regressionsbäumen[312]

Es existieren für kategorielle Daten bei k Ausprägungen einer Variable $2^{k-1} - 1$ verschiedene Trennungen (Splits). Das Problem vereinfacht sich für eine binäre Zielvariable, da die Klassen nach dem Anteil des positiven (oder negativen) Ausgangs sortiert werden können. Die optimale Trennung (Split) kann mittels Gini-Index berechnet werden. Für quantitative Zielvariablen funktioniert dies über die Sortierung nach dem Mittelwert der Gruppen.

[312] Vgl. THEUSS [2004] – Vorlesungsunterlagen zu Multivariaten Statistischen Verfahren, Universität Augsburg

2.5.2.3 Algorithmus: ID3 (Interactive Dichtomizer 3)

Der Interactive Dichtomizer 3 (ID3-) Algorithmus[313] geht von klassifizierten Beispieldaten aus, die in Form von Attribut-Wert-Listen gegeben sind. Jedes Beispiel besteht aus einer Liste von Variablen. Die Aufgabe besteht darin, eine Minimalkombination von Variablenausprägungen zu suchen, die ausreicht, um die Klassifizierung aufzubauen. Es genügen hierzu oft schon einige wenige Variablen aus einer großen Variablenliste. Das Ergebnis ist ein Entscheidungsbaum, der anschließend dazu genutzt werden kann, um neue, bisher unbekannte Beispiele zu klassifizieren. Begrifflich ist zwischen der extensionalen und intensionalen Beschreibung zu unterscheiden. Der erste Fall gibt in Form der Beispiele nur Informationen über den gegebenen Datensatz frei, während die daraus abgeleitete intensionale Beschreibung idealerweise danach auf alle beliebigen Fälle anwendbar ist. Als Knoten eines Entscheidungsbaumes werden die Konzeptattribute (Variablen) bezeichnet, die Kanten enthalten die Attributswerte (Variablenausprägung) und die Blätter charakterisieren die dazugehörigen Klassen. Es kommt den Knoten, die nahe an der Wurzel des Entscheidungsbaumes liegen, eine wichtigere (diskriminierendere) Bedeutung zu als den niedriger liegenden Knoten. Der Ablauf des ID3-Algorithmus ist in Abbildung 2-27 schematisch dargestellt:

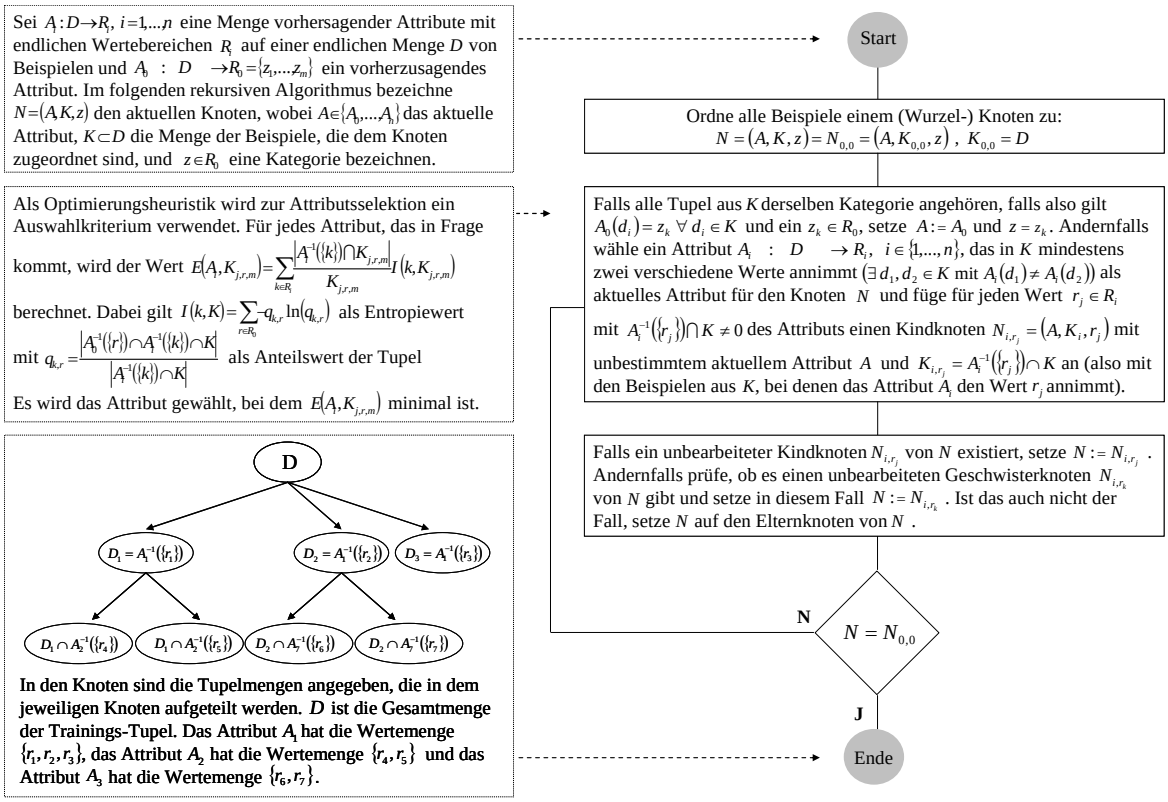


Abbildung 2-27: Ablauf des ID3-Algorithmus (Generierung eines Entscheidungsbaumes)[314]

313 Vgl. Der australische Forscher J. ROSS QUINLAN [1986, pp. 81-106] publizierte diesen Algorithmus. Weitere Folgeentwicklungen sind der Algorithmus C4.5 (QUINLAN [1993]) und C5.0 (QUINLAN [1997]).

314 Eigene Bearbeitung nach FERBER [1993] und eigene Ergänzungen.

Zu Beginn des Aufbaus eines Entscheidungsbaumes wird eine Variable mit Hilfe eines Auswahlkriteriums für den ersten, also den obersten Knoten bestimmt und danach wird der Baum von dort ausgehend sukzessiv Knoten für Knoten nach unten weiter aufgebaut (Top-Down-Ansatz). Von Bedeutung für die Auswahlentscheidung ist die Variable, welche an der jeweils gewählten Stelle den meisten Informationsgewinn ermöglicht. Für jeden der disjunkten, diskreten Werte, den die Variable annehmen kann, wird hinter dem neu gebildeten Knoten ein Ast entwickelt, der jeweils einem dieser Werte zugeordnet ist. An jedem Ende dieser Äste werden neue Knoten gebildet, und es ist einer der drei folgenden Fälle denkbar:

1. Der Ast ist als abgeschlossen zu betrachten, wenn alle Beispiele, die diesem Pfad bisher zugeordnet wurden, die gleiche Klassifikation besitzen, so dass diese in den Knoten abgelegt werden kann.
2. Falls die Beispiele nicht die gleiche Klassifikation besitzen, wird nach weiteren Variablen geschaut, die auf diesen Ast noch nicht geprüft wurden. Ist dieses nicht der Fall, so wird dem Knoten die Klassifikation zugewiesen, die den meisten der diesem Pfad entsprechenden Instanzen entspricht.
3. Es sind noch freie Variablen vorhanden, so dass der Algorithmus rekursiv selbst wieder von der Stelle an beginnt, an der dieser nach dem Prinzip des größtmöglichen Informationsgewinns eine Variable auswählt.

Tabelle 2-17 enthält einige Vor- und Nachteile zur Anwendung des ID3-Algorithmus.

Vorteile	Nachteile
Der Entscheidungsbaum gibt einen guten Einblick in die innere Struktur der Daten. Zudem wird eine gewisse Hierarchisierung durchgeführt, bei der ähnliche Objekte in den gleichen Teilbäumen angesiedelt sind. Ein Zusammenfassen von Blättern im Baum entspricht damit einer Generalisierung.	Der Algorithmus ID3 reagiert sehr empfindlich auf fehlerhafte bzw. widersprüchliche Daten. Kommt ein neu bewertetes Beispiel hinzu oder ändert sich ein Attributswert, ist eine erneute Generierung des Entscheidungsbaumes notwendig.
Das Ergebnis ist effizient, indem im Entscheidungsbaum nur diejenigen Objektmerkmale genutzt werden, die für eine eindeutige Klassifizierung nötig sind. Alle anderen müssen nicht berechnet werden.	Es können nur Attribute mit diskreten Attributwerten behandelt werden. Stetige Attribute müssen diskretisiert werden. Dies kann zu Einbußen in der Güte des Entscheidungsbaumes führen.
Das zugrunde liegende Bewertungsmaß der Entropie ist anschaulich und nachvollziehbar. Es lässt sich für sehr viele Fragestellungen einsetzen.[315]	Der Algorithmus ID3 ist ein „monotetic classifier“; an jedem Knoten wird immer nur ein Attribut betrachtet. Attributkombinationen sind nicht erlaubt.
Leichte Lesbarkeit des ermittelten Ergebnisses solange der Baum nicht zu verzweigt ist.	In Extremfällen kann es sein, dass für jedes Beispiel eine separate Verzweigung im Baum geschaffen werden muss.

Tabelle 2-17: Der Algorithmus ID3 und seine Vor- und Nachteile[316]

315 Vgl. FÖRSTNER [1991], VOSSELMAN [1992], LECLERC [1988]

316 Eigene Bearbeitung, siehe zur Ergänzung: WALKER / MOORE [1988], KODRATOFF [1994], SESTER [1995, S. 61 ff.], ELIAS [2006, S. 46 ff.]

2.5.3 Klassifikatornetze

Die Grundidee für den Einsatz eines Klassifikatornetzes besteht darin, die Vielzahl der an einem Gesamtidentifikationsprozess beteiligten Klassifikatoren zusammenzuführen und eignet sich, komplexe und komplizierte Systeme durch Teilidentifikationsprozesse zu ergründen. Ein Klassifikatornetz umfasst ein System von Klassifikator- und Auswertungsmodulen, die in einer bestimmten Hierarchie (Rangfolge) parallel oder sequentiell verschaltet sind und durch die Struktur von einem Experten oder einer Expertengruppe aus der Erfahrung in einen Kontext zueinander gestellt werden.[317] Bei sequentiellen Netzen wird genau ein Klassifikator aktiv und in Abhängigkeit des Identifikationsergebnisses folgt der Abbruch oder die Anwendung eines weiteren der nächsten Ebene. Bei parallelen Netzen erfolgt als Voraussetzung vor der Anwendung eines Klassifikators der aktuellen Ebene die Abarbeitung mehrerer Klassifikatoren der vorherigen Ebenen bzw. schließt sich an die Anwendung eines Klassifikators in Abhängigkeit von dessen Identifikationsergebnis der Abbruch oder die Anwendung mehrerer Klassifikatoren der nächsten Ebene an. Das sequentielle Netz kann als Sonderfall des Parallelen aufgefasst werden, falls die Klassifikatoren einer Ebene die gleiche Klassenstruktur aufweisen und ein Aggregationsprozess dieser Klassifikatoren erfolgt. Ein Klassifikatormodul (Knoten) in einem Klassifikatornetz besteht aus dem jeweiligen Klassifikator K und einer so genannten Steuerungs- und Bewertungsstrategie S, die z.B. entweder lehrergesteuert oder automatisch eine Klassifikatorvariation oder Gütebewertung der aktuellen Leistungsfähigkeit eines Klassifikators vornimmt (Adaptionsfähigkeit). Jeder Klassifikator kann im Prinzip auf Basismerkmale oder auf Ergebnisse vorheriger Klassifikatoren zugreifen. Man unterscheidet zwischen Primär- und Sekundärmerkmalen, da in einigen Fällen durch einen Klassifikator erst die Merkmalsbeschaffung ausgelöst werden kann, d.h. diese wurden vorher noch nicht benötigt (Abbildung 2-28).

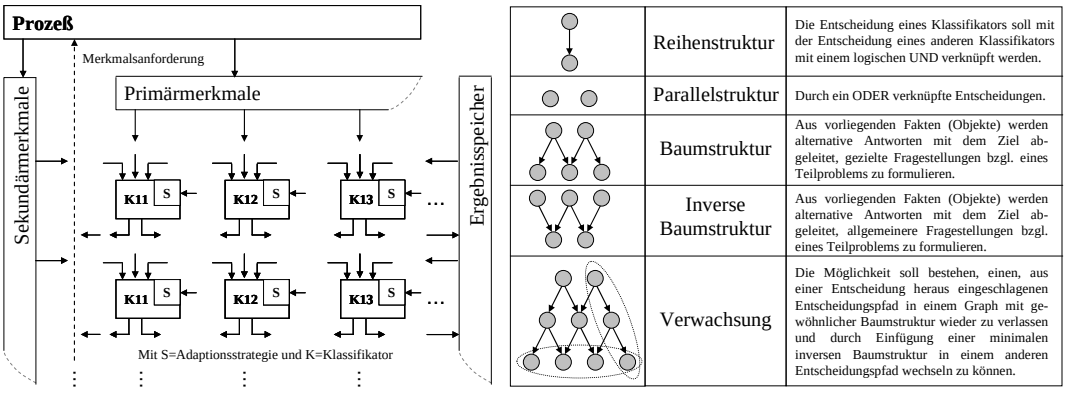


Abbildung 2-28: Klassifikatornetz gegliedert in Hierarchieebenen und Strukturformen[318]

[317] Vgl. GEISSLER, K. [1998], CLAUS [2003]

[318] Eigene Bearbeitung nach BOCKLISCH [1987, S.47] und eigene Ergänzung mit CLAUS [2003, S. 63-65]

Die Ketten von Klassifikatoren in einem Klassifikatornetz werden in Pfade zusammengefasst und die sich ergebenden errechneten Zugehörigkeitswerte dienen der Ergebnisvalidierung. Die Struktur von Klassifikatornetzen ist ein Abbild für die Chronologie bzw. Abfolge von Teilklassifikationsproblemen (Top-Down- oder Bottom-Up-Strategien). Für zeitlich nicht nebenläufige bzw. problembezogene, zusammenhängende Teilklassifikationsprobleme eignet sich die Nutzung von Reihenstrukturen. Für nebenläufige bzw. nicht problembezogene Teilklassifikationsprobleme bietet sich die Verwendung von Parallelstrukturen an. Tabelle 2-18 nennt Vorteile, die sich aus einer Vernetzung von Klassifikatoren ergeben.

Argumente für die Verwendung von Klassifikatornetzen
▪ Hochdimensionale Klassifikatoren werden auf niederdimensionale Klassifikatoren aufgeteilt.
▪ Es lassen sich unterschiedliche Methoden für den Entwurf von Einzelklassifikatoren anwenden.
▪ Der Anwender hat eine höhere Übersicht und damit verbesserte Interpretier- oder Nachvollziehbarkeit.
▪ Die Fortschritte des Identifikationsprozesses zwischen den einzelnen Klassifikatorebenen lassen sich nachverfolgen.
▪ Es sind Zwischenergebnisse verfügbar, und es besteht daher die Möglichkeit, Identifikationsprozesse auch vorzeitig abzubrechen.
▪ Einzelne Klassifikatoren lassen sich an die Zeit oder an bestimmte Zeitpunkte verknüpfen.
▪ Einzelne Klassifikatoren lassen sich an die Lage oder an bestimmte räumliche Ordnungsgefüge verknüpfen.
▪ Alternative Entscheidungsstränge können zur gleichen Zeit abgearbeitet werden.

Tabelle 2-18: Vorteile für den Einsatz von Klassifikatornetzen

Zum Entwurf eines Klassifikatornetzes kann der nachfolgende Ablauf verwendet werden:[319]

1. Das vorliegende Gesamtklassifikationsproblem soll in mehrere Teilklassifikations-probleme zerlegt werden.
2. Die Teilklassifikationsprobleme sind zu strukturieren, d.h. die Reihenfolgen bzw. Abhängigkeiten der Teilklassifikationsprobleme sind in einer Grundstruktur zu entwerfen. Es sind Fragestellungen für die einzelnen Teilklassifikationsprobleme zu formulieren, um Antworten (Entscheidungen) ggf. daraus ableiten zu können und weitere Teilklassifikationsprobleme einzubeziehen.
3. Eine entworfene Grundstruktur der Teilklassifikationsprobleme wird in einem Entscheidungsbaum abgebildet unter Einsatz von z.B. Reihenstrukturen, Parallel-strukturen, gewöhnlichen Baumstrukturen und inversen Baumstrukturen.
4. Die einzelnen Klassifikatoren werden für die einzelnen Teilklassifikationsprobleme erzeugt. Voraussetzung ist die Bereitstellung einer ausreichend großen Anzahl von Objekten, die einen Klassenberechnungsprozess durchlaufen. Der Klassifikator ist zu bewerten und zu testen (Güte – Überlappungsgrad, Re-identifikationsgrad).
5. Das Klassifikatornetz wird, wie in 3 entwickelt, aufgelöst.
6. Ein abschließender Test des Klassifikatornetzes wird durchgeführt.

Nach Entwicklung des Klassifikatornetzes beginnt die Arbeitsphase, d.h. die Klassifikation eines angebotenen Objektes startet ausgehend von einem Bezugsklassifikator i.

[319] Vgl. CLAUS [2003, S. 74]

2.6 Wissenskonversion

Der geschilderte zyklische Prozess des Data mining erfordert stets einen kritischen Umgang mit Daten, Informationen und Wissen. Im Hinblick auf die Erklärung des Begriffes der Wissenskonversion als wesentlicher Teilschritt des Data Mining, findet im Folgenden eine skizzenhafte Ausseinandersetzung mit grundsätzlichen Begriffen der Wissensthematik statt.

STREICH[320] definiert: „Wissen ist die intellektuelle Vernetzung von Informations-‚atomen' 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." POLANYI[321] unterscheidet zwischen implizitem und explizitem Wissen. Das explizite Wissen bezeichnet das Wissen des Verstandes und lässt sich in formaler Sprache ausdrücken bzw. weitergeben. Das implizite Wissen (engl.: implicit knowledge; tacit knowledge) umfasst dagegen persönliches, kontextspezifisches und analoges Wissen der Erfahrung. Dieses kann man nicht formalisieren und eindeutig kommunizieren. Beispielsweise ist nach GUNDRY[322] das implizite Wissen als schwer beschreibbare Fertigkeit (Know-How) oder als ein für selbstverständlich erachtetes mentales Modell anzusehen. Mit diesem Kontext setzt sich BAUMARD[323] mit den klassischen griechischen Begriffen metis (konjekturale Intelligenz), episteme (abstrakte Generalisierung), techné (Fähigkeit), phronesis (praktisches und soziales Wissen) auseinander. HEDLUND und NONAKA[324] fügen unter dem Aspekt der Dimension weitere Kategorisierungen des Wissens hinzu, wie z.B. individuell oder kollektiv und Untergliederungen wie Individuum, Gruppe oder Organisation. Es verstärkt sich die Einsicht[325] in die Notwendigkeit, die vorhandene Ressource ‚Information' anhand von so genannten ‚selektiven', ‚interpretatorischen' und ‚wertenden' Prozessen zu veredeln, die zu dem führen, was in einem umfassenden Sinne als Wissen bezeichnet wird.

WILKE[326] setzt bei der so genannten Wissensarbeit voraus,: „[…] dass das relevante Wissen (1) kontinuierlich revidiert, (2) permanent als verbesserungsfähig angesehen, (3) prinzipiell nicht als Wahrheit, sondern als Ressource betrachtet wird und (4) untrennbar mit Nichtwissen gekoppelt ist, so dass mit Wissensarbeit spezifische Risiken verbunden sind."

---

320 Vgl. STREICH [2005, S. 17 ff.]
321 Vgl. POLANYI, MICHAEL [1966]
322 Vgl. GUNDRY [2006], unterscheidet als Direktor des britischen Knowledge Ability Ltd. das implizite Wissen als „know-why" und im Gegensatz dazu das explizite Wissen als „know-what".
323 Vgl. BAUMARD [1999]
324 Vgl. HEDLUND / NONAKA [1993]
325 Vgl. MÜLLER, A. v. [1997, S. 471]
326 Vgl. WILKE [1998, S.161-177]

Abbildung 2-29 zeigt das unterschiedliche Aktionspotential von verschiedenen Kategorien des Wissens, wobei die Ebenen ‚Understanding' und ‚Wisdom' über einen hohen Gehalt an implizitemWissen verfügen. Ergänzend ist die Entstehung von Wissen aufgeführt.

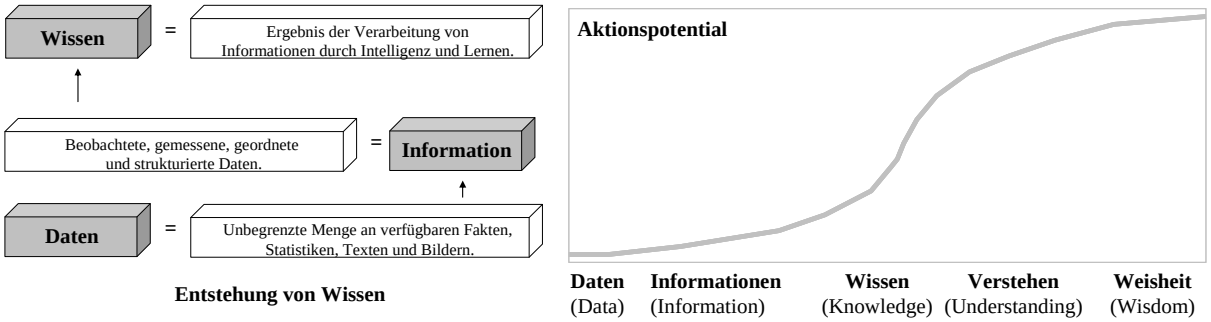


Abbildung 2-29: Wachsendes Aktionspotential von Daten zu Informationen und Ebenen des Wissens.[327]

KNOBLAUCH[328] verweist auf einen gesellschaftlichen Zusammenhang im Umgang mit Wissen: „Die Wissensgesellschaft und ihre Verwandten, die Informationsgesellschaft, die Wissenschaftsgesellschaft oder die Netzwerkgesellschaft, sind gegenwärtig höchst aktuelle Bezeichungen einer Gesellschaftsformation, die sich anscheinend derzeit ausbildet, schon ausgebildet hat oder im Entstehen begriffen ist." Der Begriff der ‚Wissensgesellschaft'[329] wird als eine Weiterentwicklung des Begriffes der ‚Informationsgesellschaft'[330] verstanden oder wird teilweise auch als dessen Synonym angesehen. Im Wesentlichen wird eine Gesellschaftsform charakterisiert, in der individuelles und kollektives Wissen und seine Organisation vermehrt zur Grundlage des sozialen und ökonomischen Zusammenlebens werden.

CASTELLS[331] postuliert dagegen auf Grundlage einer kritischen Auseinandersetzung mit der Informationsgesellschaft die ‚Netzwerkgesellschaft', dabei beschreibt der Begriff eines informationsbasierten Netzwerkes äußert anpassungsfähige Organisationen, die aus miteinander verknüpften Knoten bestehen.

Die Managementlehre schien in den Neunzigerjahren das Wissen bereits als die wichtigste Ressource von Organisationen identifiziert zu haben. Der Begriff des ‚Wissensmanagement' (engl.: Knowledge-Management) umfasst das Management von Informationen und Informationsquellen sowie deren Beziehung zueinander. Das Wissen wird dabei oft schon als vierter Produktionsfaktor neben Arbeit, Boden und Kapital angesehen.

---

327 Eigene Bearbeitung nach SCHWANINGER [1998] und eigene Ergänzungen zur Wissensentstehung.
328 Vgl. KNOBLAUCH [2005, S. 255]
329 Vgl. LANE [1966], BELL [1976], FOUCAULT [1987], RIFKIN [2000], OTT [2002], LYOTARD [2005], KÜBLER [2005, S. 16 ff.]
330 Vgl. WIENER [1948], STEINBUCH [1968], WERSIG [1985], SPINNER [1998]
331 Vgl. CASTELLS [2003]

Das von NONAKA und TAKEUCHI[332] entwickelte Modell in Abbildung 2-30 bildet einen Ansatz zur Generierung von Wissen, welches den Zusammenhang der Wissenskonversion auf der individuellen, der Gruppen- und der Organisationsebene erklärt. Der interne Lernprozess innerhalb einer Organisation wird als Spiralprozess aufgefasst, der sich einerseits auf der Achse zwischen implizitem und explizitem Wissen und andererseits über die Ebenen Individuen, Gruppe und Unternehmen vollzieht. Das Modell unterscheidet vier Varianten und eignet sich, um die Interaktionen zwischen den einzelnen Wissenstypen zu analysieren und so abzubilden bzw. zu überprüfen, unter welchen Voraussetzungen und Bedingungen die Wissensumwandlung und -generierung stattfindet. Es bildet letztlich den Rahmen für die Modellierung der dynamischen Wissenserzeugung.

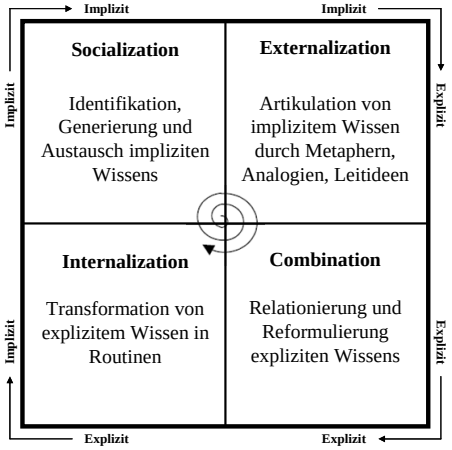


Abbildung 2-30: Varianten der Wissenskonversion - Wissenstransferprozesse (endogen / soziale Ebene)[333]

Für die vier Wissenskonversionsarten werden unterschiedliche Methoden diskutiert, die in Tabelle 2-19 zusammenfassend wiedergegeben sind.

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

Tabelle 2-19: Mögliche Methoden für die vier Arten der Wissenskonversion[334]

332 Vgl. NONAKA / TAKEUCHI [1997, S. 71 – Original 1995], Ergänzungen mit NONAKA et al. [2001]

333 Eigene Bearbeitung nach SCHWANINGER [1998] und eigene Ergänzungen zur Wissensentstehung.

334 Eigene Bearbeitung unter Verwendung von NONAKA / TAKEUCHI [1997], PREECE et al. [1994], SCHREIBER et al. [2000], HYTTINEN [2004, S.14 ff.]

Im Hinblick auf den Begriff der Regionalisierung (siehe Abschnitt 1.1.4) und dessen Bedeutung für die räumliche Modell- und Theoriebildung werden zusätzlich einige Begriffe inhaltlich ergänzt. Durch die Regionalisierung wird sowohl die Bildung von Theorien[335] als auch die Bestätigung oder Verwerfung von vorhandenen Hypothesen[336] oder Theorien unterstützt.[337]

Generell steht eine Erklärung in Zusammenhang mit den Begriffen des so genannten Explanandums (Erklärung eines Phänomens oder Ereignisses) und des Explanans (Randbedingungen, Gesetzesaussagen). Zu unterscheiden ist zwischen einer deduktiv-normologischen Erklärung und einer induktiv-statistischen Erklärung. Die deduktiv-normologische Erklärung erläutert basierend auf HEMPEL-OPPENHEIM[338] ein Ereignis dadurch, dass dieses aus einem allgemeinen Gesetz und einer Reihe spezieller Umstände (Anfangsbedingungen) erschlossen werden kann. Eine induktiv-statistische Erklärung oder auch probabilistische Erklärung unterscheidet sich dadurch, dass eine statistische Gesetzesaussage getroffen wird und kein deterministischer Zusammenhang angenommen werden kann. Abbildung 2-31 enthält Begriffe der Wissensbeschreibung und Beziehungsabläufe zur Bildung von Theorien im Allgemeinen.

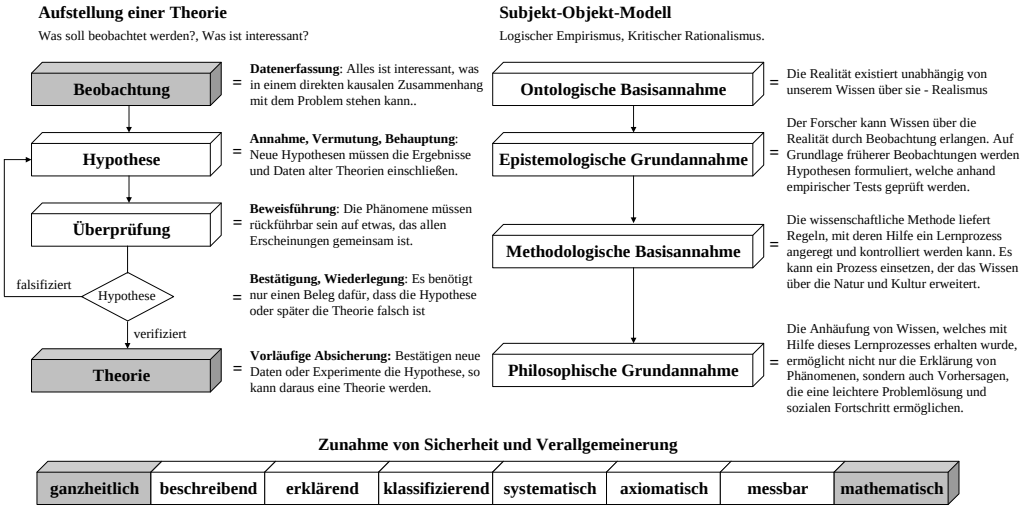


Abbildung 2-31: Wissenschaftstheorie – Begriffe und Prinzipien.[339]

---

335 Eine Theorie (gr. theoria) ist nachprüfbar, anhand einer nennenswerten Anzahl von Fällen bestätigt, nicht tautologisch (Redundanz) und nicht auf eine einzelne Instanz bezogen, sondern hat generelle Gültigkeit.

336 Eine Hypothese (gr. hypothesis) ist eine aus theoretischen Überlegungen oder aus Beobachtungen abgeleitete einzelne Schlussfolgerung, die empirisch überprüft werden soll. Zu unterscheiden ist zwischen partikulären und allgemeinen Hypothesen. Eine partikuläre Hypothese lässt sich verifizieren oder falsifizieren. Eine allgemeine Hypothese kann man jedoch nur falsifizieren und nicht verifizieren, weil die Gegenstände, die sie erfasst, nicht abzählbar sind.

337 Vgl. DROTH / FISCHER [1980] sowie FISCHER[337], der die Regionen als räumliche Geltungsbereiche empirisch gehaltvoller räumlicher Theorien mittlerer Reichweite auffasst.

338 Vgl. HEMPEL / OPPENHEIM [1953, S. 135-175], die zusätzlich Adäquatheitskriterien für eine wissenschaftliche Erklärung definieren.

339 Eigene Bearbeitung: Für das Quellenstudium diente STROEKER [1992], CHALMERS [2001].

Im Zusammenhang mit dem Data Mining stellt GAUL[340] fest, dass viele Werkzeuge davon nicht über eine Möglichkeit der Wissenskonversion verfügen. WOODS und KYRAL[341] kritisieren, dass die Begriffe des ‚Data Mining' und des ‚Knowledge Discovery in Databases' (KDD) auch inflationär für recht einfache Statistik-Werkzeuge benutzt werden, die mit einer originellen Benutzeroberfläche zur Visualisierung gekoppelt sind. ULTSCH[342] verweist darauf, dass gerade die Entdeckung von neuen bzw. verborgenen Zusammenhängen in Daten und die sich daran anschließende Extraktion von Wissen bzw. Wissensgenerierung beim Data Mining eine Berücksichtigung finden müssen und speziell darin der Unterschied zur explorativen Statistik besteht. Das Data Mining trägt unterstützend zum automatischen Erzeugen und Prüfen von Hypothesen und Modellen bei. Der Begriff der Wissensentdeckung (Knowledge Discovery) charakterisiert die Entdeckung aber gerade (natürlich-)sprachliche Darstellung von Wissen aus Datensammlungen.

Daten und Informationen sind vielfach im Überfluss vorhanden, jedoch die Einbindung in Erfahrungszusammenhänge, durch welche erst Wissen geschaffen wird, stellt sich als schwierig heraus. Eine maschinelle Verarbeitbarkeit des gewonnenen Wissens ist in der Regel zu berücksichtigen und erfolgt häufig in Form von ‚wissensbasierten Systemen'.[343] Abbildung 2-32 zeigt eine Einordnung des Data Mining in Prozessmodelle des KDD.

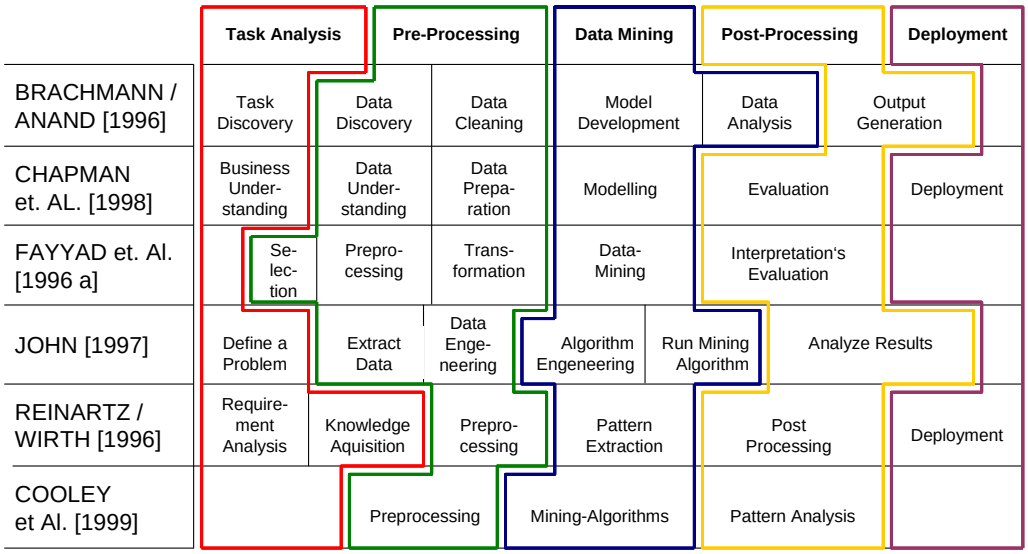


Abbildung 2-32: Vergleich der Hauptschritte im KDD-Prozess in unterschiedlichen Prozessmodellen[344]

340 GAUL, W. [1998, S.145 ff.]
341 WOODS / KYRAL [1997]
342 ULTSCH [2000]
343 Ein wissensbasiertes System ist ein intelligentes Informationssystem, in dem Wissen mit Methoden der Wissensrepräsentation und Wissensmodellierung abgebildet und nutzbar gemacht wird. Siehe ergänzend bei GRONAU / WEBER [2005], ALTENKRÜGER / BÜTTNER [1992, S. 1 ff.].
344 Eigene Bearbeitung nach GAUL [1998, S.146], zum Data Mining wird oftmals die Vorverarbeitung von Daten (Preprocessing, Feature-Extraktion) mit hinzugerechnet.

3 Datenerhebung des ‚Urban Data Mining' (Beschreibung des Status quo)

Im empirischen Teil dieser Arbeit werden anhand von ausgewählten Untersuchungsaufgaben methodische Ansätze des ‚Urban Data Mining' vorgestellt und verstärkt regionsvergleichende Untersuchungen nach raumstrukturellen Ausprägungen durchgeführt. Der Grundgedanke besteht darin, basierend auf einer möglichst hohen räumlichen Auflösung, das gesamtdeutsche Gemeindesystem in Form einer systematischen Bestandsaufnahme zu analysieren. Betrachtet werden einerseits statische Eigenschaften und andererseits jene, die Aussagen über regionale Enwicklungstendenzen ermöglichen. Folglich ergeben sich Anforderungen an die Datenerhebung und die damit in Verbindung stehenden Untersuchungsobjekte und -merkmale.

3.1 Untersuchungsobjekte

Um Aussagen über die gesamtdeutsche Raumstruktur zu treffen, ist eine räumlich-differenzierte Betrachtung anzustreben. In Anbetracht der eingeschränkten Übertragungsmöglichkeiten von einzelregionalen Fallstudien wird in dieser Arbeit mit einer sehr umfangreichen Datenbasis gearbeitet, die sich auf verschiedene räumliche und zeitliche Auflösungen bezieht, jedoch den Anspruch trägt, primär gemeindescharfe Daten einzusetzen. Die Granularität der räumlichen Auflösung wird nicht zu Beginn eindeutig festgesetzt, sondern wechselt im Hinblick auf die jeweils verfügbare Datenlage bzw. sachlich begründbaren Untersuchungskriterien.[345] Es handelt sich um einen Top-Down und Bottom-Up Ansatz, der einerseits mit Hilfe von Ähnlichkeitsuntersuchungen einen Vergleich von räumlichen Objekten vorschlägt und andererseits eine Transformation bzw. einen Austausch von Informationen zwischen räumlichen Untersuchungsebenen verfolgt (siehe Abbildung 3-1).

Da a priori kein Wissen über Ähnlichkeitsmuster auf Grundlage des hier eingesetzten Datenmaterials vorauszusetzen ist, findet eine Einschränkung der Untersuchungsmenge zu Beginn der Arbeit nicht statt. Damit wird das Ziel verfolgt, für die 12504 Gemeinden (Stand: 31.12.2004) und die 440 Kreise in Deutschland statistisches Datenmaterial zu beschaffen. Die räumliche Untersuchungsperspektive beschränkt sich infolge einer gemeindedatenbasierten Analyse nicht auf Kernstadt-Umlandbeziehungen, sondern ermöglicht die Abbildung stadtregionaler Verflechtungsbereiche. Damit wird die Abbildungsleistung der Heterogenität der Siedlungsmuster erhöht und gerade eine räumlich-differenzierte Betrachtung des suburbanen Raumes denkbar.

---

[345] Auf die Untersuchungsperspektive anderer Arbeiten sei an dieser Stelle verwiesen: Siehe Abbildung 1-3: Chronologischer Überblick zur Klassifizierung von Städten sowie BRÖCKER et al. [1998], SIEDENTOP et al. [2000], ARLT et al. [2001]. SIEDENTOP et al. [2005 c] definieren ein Schnittstellenkonzept für gemeindestatistische Daten und Daten aus Haushalts- und Personenbefragungen.

Abbildung 3-1 charakterisiert den beschriebenen Top-Down und Bottom-Up Ansatz, der bei unterschiedlich aggregiertem Datenausgangsmaterial die analytische Aussagekapazität des Untersuchungsansatzes vergrößert und Datenlücken ggf. minimiert. Eine wichtige Grundlage für den Erkenntnisgewinn über vorhandene Ähnlichkeitsmuster in den Daten wird mit Hilfe von Verfahren der Strukturerkennung und -bildung geschaffen. Die Verfahren zur Operationalisierung ermöglichen den Aufbau von Zuordnungsregeln (→ Klassifikatoren) und unterstützen eine Übertragbarkeit von Objektinformationen sowohl innerhalb der gleichen als auch zwischen verschiedenen räumlichen Ebenen (→ Datenlücken schließen).

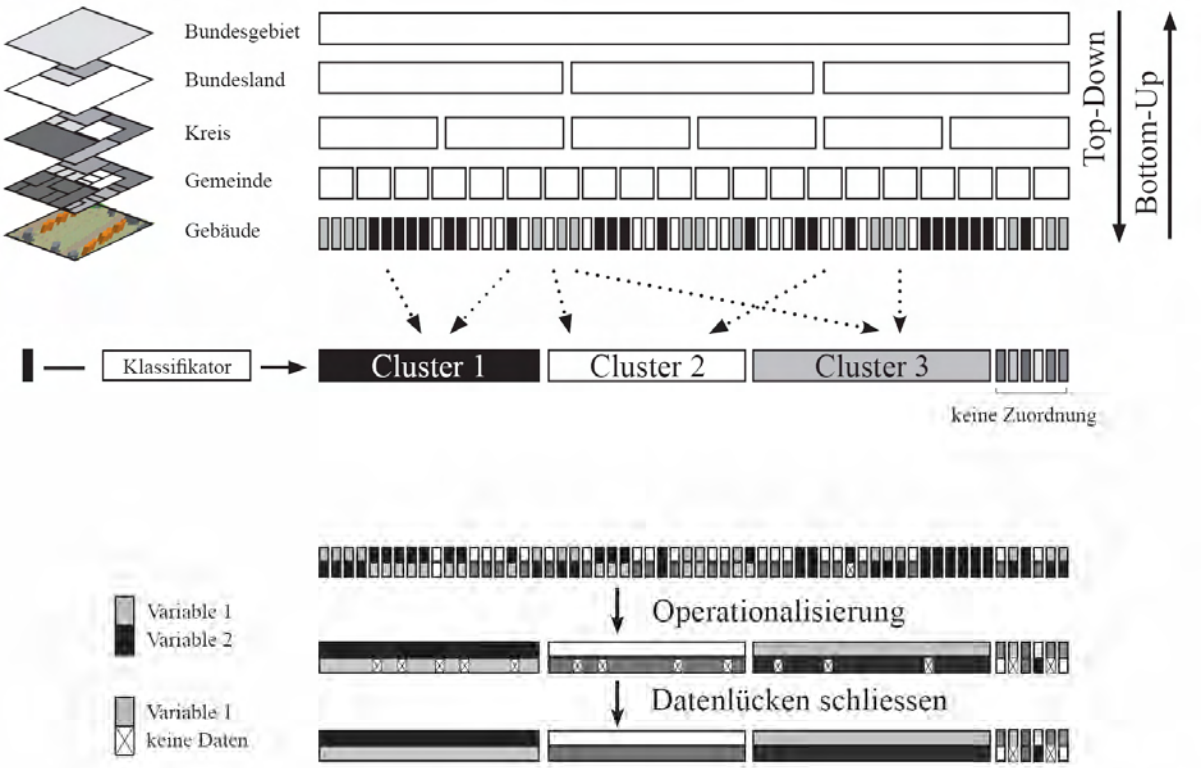


Abbildung 3-1: Top-Down-/Bottom-Up Ansatz zur Untersuchung der Siedlungs- und Gebäudestruktur[346]

Die empirisch zu ermittelnden Ähnlichkeitsmuster basieren – wie bereits erwähnt – auf administrativen Gebietseinheiten (z.B. Gemeinden), die jedoch in unterschiedlichem Umfang über regionalstatistisches Datenausgangsmaterial verfügen. Es besteht die Problematik, dass mit zunehmender räumlicher Auflösung das Datenangebot immer schwieriger in großem Umfang zu beschaffen ist. Zusätzlich ist eine möglichst konstante Gebietsstandsregelung für die zeitlich statistische Vergleichbarkeit erforderlich, die aber oftmals nicht gegeben ist. Die amtliche Statistik ist zum Beispiel aus Gründen geltender Datenschutzregelungen aber auch aufgrund eines nicht wirtschaftlich kalkulierbaren Arbeits- und Zeitaufwandes in der Lage, alle Statistiken auf Gemeindeebene aufzubereiten.

[346] Eigene Bearbeitung in Anlehnung an BEHNISCH, M. / VIEJO GARCIA, P. [2005]

Da es in manchen Fällen nicht immer möglich ist, Daten für alle räumlichen Untersuchungsobjekte zeitgleich zu ermitteln, sind Strategiekonzepte vorzuhalten, die auch eine spätere Beurteilung von Untersuchungsobjekten ermöglichen und in der Lage sind, Ähnlichkeitsmuster zusätzlich zu definieren oder wiederzuerkennen. Diesbezüglich stellt das Konzept zur Integration von Klassifikatoren eine unbegrenzte Erweiterbarkeit sicher. Abbildung 3-2 definiert einen strategischen Orientierungspfad, um mit Hilfe eines Klassifikators hochdimensionale Datenobjekte schrittweise zu beurteilen und die Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern über die gebaute Umwelt zu operationalisieren.

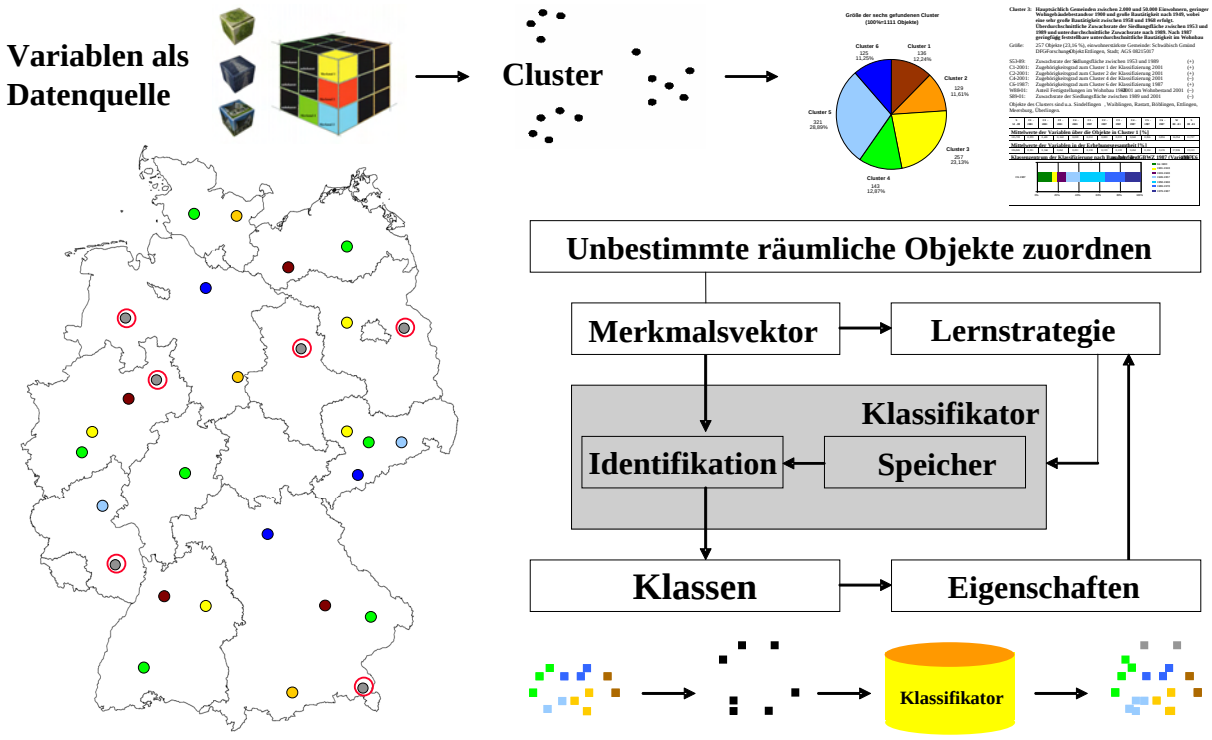


Abbildung 3-2: Strategischer Orientierungspfad für eine Ähnlichkeitsuntersuchung mit Raumbezug[347]

Um eine Untersuchung von räumlichen Objekten sicherzustellen, für die nur teilweise Daten zu einer komplexen Forschungsfrage vorliegen, ist eine Reduktion von hochdimensionalen Klassifikatoren auf niederdimensionale Einzelklassifikatoren zusätzlich ggf. denkbar. Auf diese Weise wird das Problem von fehlenden Datenausprägungen beherrschbar, und es lassen sich zumindest Teilaussagen treffen. Klassifikatoren dieser Art können sich inhaltlich auf spezielle Themengebiete (z.B. Baualtersstruktur), räumlich auf ausgewählte Untersuchungsregionen (z.B. Baden-Württemberg) und zeitlich auf relevante Zeitschnitte (z.B. das Jahr 2000) beschränken. Werden fehlende Daten später ergänzt, so lassen sich auch zunächst verwendete Einzelklassifikatoren wieder zu einem hochdimensionalen Klassifikator zusammenschalten (Tabelle 2-18: Vorteile für den Einsatz von Klassifikatornetzen).

347 Vgl. BEHNISCH, M. [2005], Erweiterte Darstellung

Das Konzept der schrittweisen Objekterweiterung bei der Untersuchung von Ähnlichkeitsmustern unterstützt die Kopplung von Resultaten aus verschiedenen räumlichen Ebenen. Mit Hilfe der Operationalisierung lassen sich bereits gewonnene Erkenntnisse über untersuchte räumliche Objekte in eine Untersuchung von tiefer oder höher aggregierten Objekten einbeziehen. Abbildung 3-3 zeigt exemplarisch für die Untersuchung von Kreis- und Gemeindedaten eine Möglichkeit zur Überlagerung von bestehenden Ergebnissen der Klassenzugehörigkeit. Das Beispiel bezieht sich auf ein Kreisobjekt, das sich aus 36 Gemeinden zusammensetzt. Dieses Kreisobjekt soll aufgrund von bereits zuvor durchgeführten Ähnlichkeitsuntersuchungen über eine definierte Zugehörigkeit zu einer Klasse verfügen und mit einer Semantik (z.B. ‚grün') belegt sein. Für einige der 36 Gemeinden wird schrittweise durch Einsatz eines Klassifikators eine Klassenzugehörigkeit auf Gemeindeebene gefunden. In diesem Beispiel wird sowohl auf der Kreis- als auch der Gemeindeebene mit den gleichen inhaltlichen Variablen eine Klassenzuordnung berechnet. Für Gemeindeobjekte, die bisher über keine Daten verfügen, besteht durch Verbindung beider Ebenen trotzdem die Möglichkeit, mit Hilfe der Kreisinformation erste Vermutungen zu äußern.

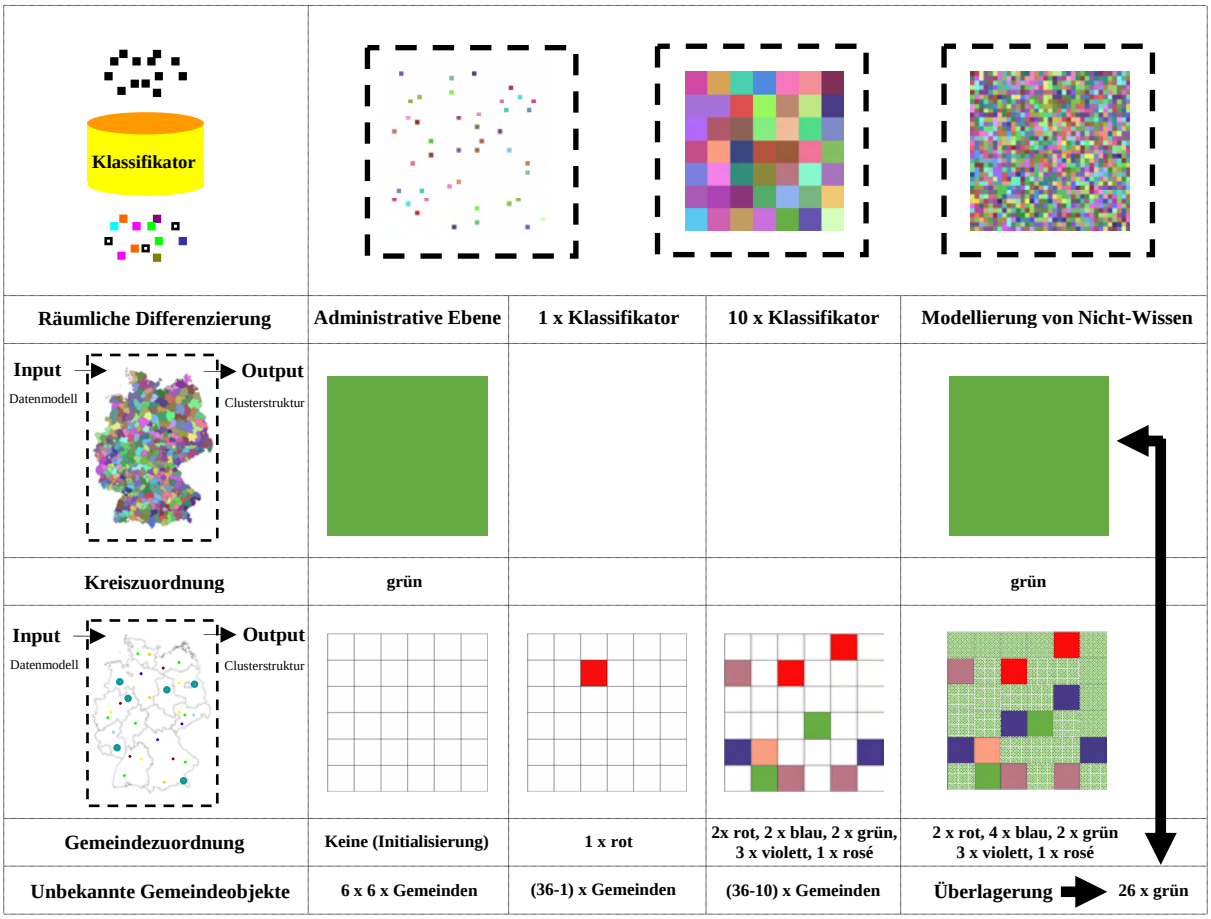


Abbildung 3-3: Erkenntnisgewinn durch Überlagerung von Klassifikationsergebnissen

Die Interpretation von Merkmalsausprägungen der Untersuchungsobjekte wird grundsätzlich erschwert aufgrund der stark voneinander abweichenden historisch bedingten länderspezifischen administrativen Gebietsgliederung.[348] Die durchschnittliche Gebietsgröße der Gemeinden variiert zwischen 5,75 km² (Rheinland-Pfalz) und 74,84 km² (Nordrhein-Westfalen). Je nach Fragestellung erhöht sich in den Bundesländern mit sehr großen Gemeinden die Gefahr, ungenaue Aussagen über die darin vorhandene Siedlungs- und Gebäudestruktur zu treffen. Oftmals bilden mehrere Ortsteile eine größere Gemeinde, und Prozesse der Dispersion oder Aussagen zur Struktur und Dynamik von Gebäudebeständen lassen sich nicht detailliert genug für die Teilgebiete einer Gemeinde abbilden. Für einzelne Untersuchungsregionen ist deshalb zukünftig die Beschaffung tiefer aggregierter Daten wünschenswert, um weitere Erkenntnisse zu gewinnen.

Für einen sinnvollen Vergleich von unterschiedlich großen und kleinen Untersuchungsobjekten, werden in dieser Arbeit Relativzahlen eingesetzt. Durch eine Relativierung werden metrische Daten erzeugt, d.h. die Messwerte der Eigenschaften von räumlichen Objekten werden für jede Gemeinde als Quotienten von Summen der empirischen Größen berechnet.[349] Es wird damit sichergestellt, dass die Ausprägungen der Variablen (siehe Abschnitt 3.2) nicht monoton mit der Objektgröße wachsen. Üblicherweise wird mit Blick auf andere Arbeiten der Stadtklassifikation (siehe Abschnitt 1.1.4) die Bevölkerung einer Gemeinde, die Gebäudeanzahl, die Gebietsfläche oder die Siedlungsfläche als räumliche Bezugsgröße[350] zur Bildung von Relativzahlen gewählt. Folglich handelt es sich um Dichtewerte, Pro-Kopf-Werte, Prozent- oder Promillezahlen.

Um zusätzlich strukturelle Verflechtungen innerhalb der administrativen Gebietsgliederung genauer erfassen zu können, werden ausgewählte Daten der so genannten ‚Geocomputation' (siehe Abschnitt 3.8) eingesetzt. Gerade unter dem Aspekt einer zunehmend feststellbaren Verfügbarkeit von raumbezogenen digitalen Daten und leistungsstarker GI-Systeme ist damit zu rechnen, dass sich weitere umfassende Lösungsmöglichkeiten für die räumliche Analyse in den nächsten Jahren ergeben werden.[351] Das Datenmaterial wird vielfach nicht auf administrative Einheiten aggregiert, so dass im Sinne der geographischen Abbildung eine genauere Wiedergabe der tatsächlich vorhandenen räumlichen Situation vorstellbar ist.

---

348 Vgl. SIEDENTOP et al. [2003, S. 27], der zusätzlich einen Bezug zu TÖNNIES [1981, S. 30] unter dem Aspekt stattgefundener Eingemeindungsprozesse herstellt.

349 Siehe zur Vertiefung Abschnitt 2.1.2 sowie VOGEL [1975, S.71]

350 Vgl. STRASSERT [1975, S. 1-11], der das Problem der Wahl einer sachlogischen Bezugsgröße diskutiert.

351 Vgl. THINH [2004 a, S.52 ff.], der zusätzlich einen Überblick von den nationalen und internationalen Programmen zur Entwicklung und Forschung im Bereich Geoinformation anführt.

	B	HB	HH	BW	BY	BB	HE	LSA	MV	NI	NRW	RP	SL	SN	SH	TH	D
I. Gebietsfläche der Bundesländer [km²], II. Bevölkerung der Bundesländer [Mio.]																	
I.	892	404	755	35752	70552	29478	21115	20446	23178	47620	34084	19847	2569	18415	15763	16172	**357042**
II.	3,39	0,66	1,73	10,72	12,44	2,57	6,10	2,49	1,72	8,00	18,08	4,06	1,06	4,30	2,83	2,36	**82,50**
III. a) Anzahl der Kreise, III. b)Anzahl der Gemeinden																	
III. a)	1	2	1	44	96	18	26	24	18	46	54	36	6	29	15	23	**439**
III. b)	**1**	**2**	**1**	**1112**	**2100**	**421**	**430**	**1118**	**873**	**1048**	**396**	**2306**	**52**	**519**	**1127**	**998**	**12504**
IV. Gebietsfläche der Landkreise [km²]																	
Min	Ausgenommen sind die Stadtstaaten			519,16	307,56	1216,62	222,40	372,46	974,66	534,74	407,09	304,87	249,21	266,49	664,21	433,38	**222,40**
Max				1860,71	1971,78	3058,04	1848,56	2422,91	2516,91	2881,40	1958,79	1626,06	555,13	1340,41	2185,90	1304,85	**3058,04**
Med.				**870,69**	**895,55**	**2107,67**	**1024,70**	**758,79**	**1989,64**	**1161,61**	**1042,52**	**748,31**	**459,08**	**769,46**	**1344,38**	**935,60**	**954,62**
Ø				985,92	964,33	2053,58	970,89	950,57	1887,52	1223,65	974,16	782,61	431,62	781,88	1391,16	902,70	**1056,57**
StA				339,99	341,97	477,48	432,04	578,62	416,88	560,14	397,98	270,02	113,40	311,92	517,02	218,36	**507,51**
V. Gebietsfläche der Gemeinden [km²]																	
Min	Ausgenommen sind die Stadtstaaten			1,80	1,40	2,80	4,05	2,07	0,31	0,42	20,50	0,40	7,61	3,58	0,45	1,34	**0,31**
Max				207,36	310,40	417,19	248,31	199,56	371,84	313,15	405,15	139,72	167,07	328,30	214,13	269,11	**417,12**
Med,				**23,2**	**26,36**	**43,78**	**41,42**	**12,63**	**20,82**	**27,76**	**74,84**	**5,75**	**41,91**	**30,03**	**10,63**	**9,53**	**16,92**
Ø				32,15	33,60	70,02	49,10	26,55	45,44	18,29	86,07	8,61	49,40	35,48	13,99	16,20	**28,4**
StA.				29,00	26,87	67,38	32,18	21,77	46,22	21,94	50,05	10,81	33,44	28,65	13,71	19,89	**33,94**
VI. a) Mittelwert zum Verstädterungsgrad [%], VI. b) Standardabweichung zum Verstädterungsgrad,																	
V. a)	69,44	54,72	58,55	14,33	11,64	10,12	16,65	9,41	7,10	12,38	21,73	12,73	23,34	11,35	11,24	7,95	**11,8**
V. b)	Ausgenommen			7,93	7,82	9,95	9,30	6,35	5,63	7,10	13,10	6,74	10,88	7,08	9,74	4,44	**8,20**
VII. a) Mittelwert zur Gebäude- und Freifläche [ha], VII. b) Standardabweichung zur Gebäude- und Freifläche																	
VI. a)	36230	6937	27504	233,76	182,10	310,44	360,15	92,35	95,61	317,57	1079,50	49,20	594,30	234,50	94,20	69,36	**191,45**
VI. b)	Ausgenommen			368,32	398,74	393,60	509,52	241,82	196,40	508,88	1309,25	145,20	511,87	573,22	235,56	180,31	**445,85**
VIII. Anteil der Gemeinden ohne zentralörtliche Funktion [%], IX. Anteil der Gemeinden mit weniger als 2000 Einw. [%], X. Anteil der Bevölkerung in Gemeinden mit weniger als 2000 Einwohnern [%], XI. Anteil der Bevölkerung in Gemeinden mit 20000 bis 50000 Einwohnern, XII. Anteil der Bevölkerung in den Gemeinden mit mehr als 50000 Einwohnern																	
VIII.																	
IX.	**0,00**	**0,00**	**0,00**	**17,00**	**36,62**	**52,26**	**3,49**	**84,79**	**84,54**	**49,24**	**0,00**	**85,13**	**0,00**	**23,31**	**81,37**	**77,96**	**57,37**
X.	0,00	0,00	0,00	1,96	7,90	8,6	0,28	24,5	28,1	6,73	0,00	29,34	0,00	4,1	21,12	20,75	**6,68**
XI.	0,00	0,00	0,00	22,20	11,34	25,82	21,10	23,53	7,02	26,60	24,27	9,68	35,73	14,93	15,12	22,10	**18,57**
XII.	100	100	100	27,15	25,35	15,27	30,24	21,75	27,68	28,50	63,35	22,36	17,06	34,00	24,10	20,15	**39,44**

Tabelle 3-1: Kennwerte zur Gebietsfläche und Bevölkerung in verschiedenen administrativen Ebenen[352]

Ergänzend zu den genannten Problemen bei der Untersuchung von räumlichen Objekten enthält Tabelle 3-1 einige Kennwerte der Bundesländer. Den Schwerpunkt bilden Daten zur Gebietsfläche, zur Bevölkerung sowie zu bestehenden Zentrale-Orte-Kategorien.[353] Unter dem Gesichtspunkt, grundlegend vergleichbare Strukturen für das Bundesgebiet zu erkennen und darzustellen, eignen sich aufgrund der insgesamt hohen räumlichen Verflechtungen eher kontinuierliche Abstufungen im regionalen Maßstab. Derartige kartografische Darstellungen, die unabhängig von den Grenzverläufen administrativer Einheiten sind, ermöglichen die räumlichen Analysewerkzeuge der GI-Systeme. Zur Erzeugung einer kontinuierlichen

352 Eigene Bearbeitung unter Einsatz von Daten der amtlichen Statistik (Gebietsstand: 31.12.2004).
Genannt seien die Abkürzungen:B=Berlin, HB=Bremen, HH=Hamburg, BW=Baden-Württemberg, BY=Bayern, BB=Brandenburg, HE=Hessen, LSA=Sachsen-Anhalt, MV=Mecklenburg-Vorpommern, NI=Niedersachsen (Region Hannover), NRW=Nordrhein-Westfalen, RP=Rheinland-Pfalz, SL=Saarland, SN=Sachsen, SH=Schleswig Holstein, TH=Thüringen, D=Deutschland. Die Spalte Deutschland bezieht sich nicht auf Werte der Bundesländer, sondern nimmt Bezug auf die verwendete Datenbasis aller Objekte.

353 Vgl. Ausweisungspraxis der Zentrale-Orte-Kategorien: Abschnitt 3.3

Oberflächendarstellung zeigt Abbildung 3-4 das Berechnungsprinzip der Kernel-Dichte.[354] Die Ansätze zur Interpolation[355] bieten zusätzliche Lösungsmöglichkeiten. Die Messung der Kernel-Dichte erfolgt für eine Vielzahl von quadratischen Rasterzellen im Raum. Für jede Rasterzelle werden in einem definierten Umgebungsumfeld die darin enthaltenen räumlichen Objekte herangezogen, um auf Grundlage ihrer Merkmalsausprägungen die Dichteniveaus zu berechnen. Die Abbildung der Dichteniveaus wird durch die gewählte Rasterzellengröße und den Suchradius beeinflusst.

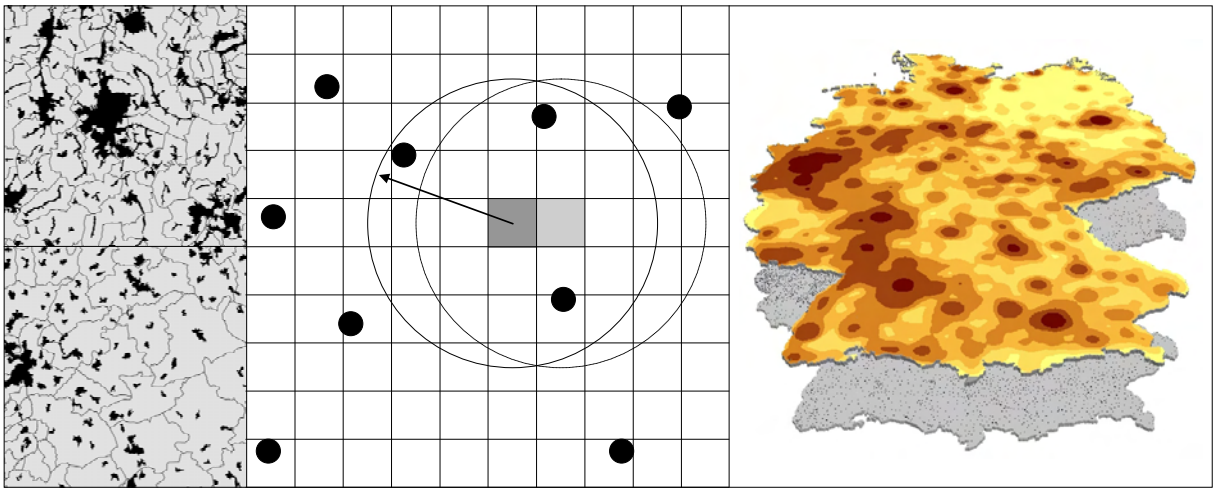


Abbildung 3-4: Abstraktion administrativer Raumbezüge bei statistischen Analysen (Kernel Dichte)[356]

Für die Darstellungen in dieser Arbeit wird bei gemeindestatistischen Daten ein 1000-Meter-Raster als Projektionsfläche gewählt. Zusätzlich zur tatsächlichen Lage von vorhandenen Siedlungsstrukturen werden abstrakte Datenstützpunkte mit Raumbezug durch Berechnung der Mittelpunkte von Gemeindegebieten vorbereitet, welche die Dichteniveaus bestimmen.[357] Um ein weich abgestuftes Bild der Oberflächendarstellung zu erzeugen, wird der Suchradius mit 20-km relativ groß gewählt. Zu begründen ist dies durch die bereits dargestellte stark variierende Gemeindegröße in Deutschland, die dazu führt, dass die abstrakt konstruierten Datenpunkte teilweise sehr weit auseinander liegen. Bei der Erstellung der hier vorgestellten gebietsstandsunabhängigen Abbildungen wäre zukünftig bei ausreichend verfügbarem Datenmaterial eine Einbeziehung von grenznahen Gemeinden im benachbarten Ausland sinnvoll. Dadurch lassen sich besondere Entwicklungen im grenznahen Bereich aufdecken bzw. genauer analysieren.

---

354 Vgl. SILVERMANN [1986], WONG / LEE [2005]

355 Folgende Interpolationsmethoden seien genannt: Kriging, vgl. OLIVER, M.A. [1990, S. 313-332], HEINE, G.W. [1986, S. 60-72] und Inverse Distance Weighted Interpolation, vgl. WATSON, D.F. / PHILIP, G.M. [1985] sowie Natural Neighbour, vgl. SIBSON [1981, S. 21-36]

356 Eigene Bearbeitung unter Kenntnis von SIEDENTOP et al. [2003, S.26]

357 Die Positionsbestimmung von Siedlungstrukturen erfolgt auf Basis der CORINE-Daten, so dass aufgrund der gegebenen Erfassungsgenauigkeit für einige Siedlungen abstrakte Datenstützpunkte gebildet werden.

3.2 Untersuchungsmerkmale und -variablen[358]

Die Auswahl der Merkmale bildet einen wichtigen Arbeitsschritt für die Entdeckung von Ähnlichkeitsmustern über die gebaute Umwelt. Nach FLOODGATE kann die Frage, ob eine bestimmte Auswahl von Merkmalen zu einem gewünschten Ähnlichkeitsmuster führt, nicht eindeutig geklärt werden: „Hence, the selection in this case has an element of trial and error, and the suitability of any selection can only be discovered empirically.“[359] VOGEL[360] bezeichnet diese Merkmale als sachlich relevant: „ [...], wenn sie die zu klassifizierenden Objekte hinsichtlich des angestrebten Ziels angemessen und ausreichend beschreiben.“ Nach VOGEL handelt es sich um „ [...] eine im wesentlichen subjektive Entscheidung“.[361] WALLACE[362] ergänzt: „ […] a highly subjective crucial element of the clustering process”.

BECHER[363] verweist auf die Definition von Merkmal und Variable. Ein Merkmal ist als eine beliebig ‚abstrakte Objekteigenschaft’ definiert, die unabhängig davon ist, ob diese durch eine empirische Größe quantifizierbar ist. Erst die Variable wird nach BECHER als ein ‚quantifizierbares’ Merkmal bezeichnet und somit basiert eine Untersuchung von räumlichen Objekten formal ausschließlich auf Variablen. Diese Unterscheidung wird in dieser Arbeit weiterverfolgt. Grundsätzlich ist die Entscheidung über die Bedeutsamkeit von Variablen aus der Verwendung der zu bildenden Ähnlichkeitsuntersuchung abzuleiten.

Seit den 90-er Jahren ist aufgrund der höheren Leistungsfähigkeit der Computerrechentechnik eine Untersuchung mit sehr vielen Variablen durchführbar. Einen genauen Überblick von möglichen Untersuchungsaspekten im ‚Urban Data Mining’ ermöglicht Abbildung 3-5. Diese strukturierte Vorgehensweise schützt vor dem Vergessen wichtiger Aspekte und drückt das theoretische Niveau aus. Der Aufbau berücksichtigt nicht die Beschaffbarkeit von einzelnen Statistiken, so dass ggf. Hierarchiezweige später aufzugeben sind und die tatsächliche Variablenauswahl einigen Einschränkungen unterliegt. Des Weiteren handelt es sich zunächst um eine generelle Darstellung prinzipieller Untersuchungsthemen. Für Folgearbeiten stehen strategische Werkzeuge zur Verfügung, die sich künftig zur Beschaffung eigener Primärstatistiken und der zukünftigen erweiterbaren Arbeitsplanung eignen.

---

358 Die hier vorliegende Arbeit wurde unter dem Gesichtspunkt angefertigt, möglichst Daten einzusetzen, die relativ leicht zu beschaffen sind und die Anwendung der erarbeiteten methodischen Vorgehensweise für die Einsatzsmöglichkeiten von Klassifikatoren demonstriert. Zusätzliche Daten lassen sich durch den Aufbau von eigenen Primärstatistiken beschaffen, und einige Daten sind aus Kostengründen nicht einzusetzen. Verwiesen sei an dieser Stelle beispielsweise auf das Digitale-Basis-Landschaftsmodell von Deutschland.

359 Vgl. FLOODGATE [1962, S. 280]

360 Vgl. VOGEL [1975, S. 51]

361 Vgl. VOGEL [1975, S. 50 ff]

362 Vgl. WALLACE, D.L. [1968]

363 Vgl. BECHER [1995, S. 55]

Abbildung 3-5 zeigt das theoretische Niveau, welches die Ausgangslage für die Ähnlichkeitsuntersuchungen bildet.

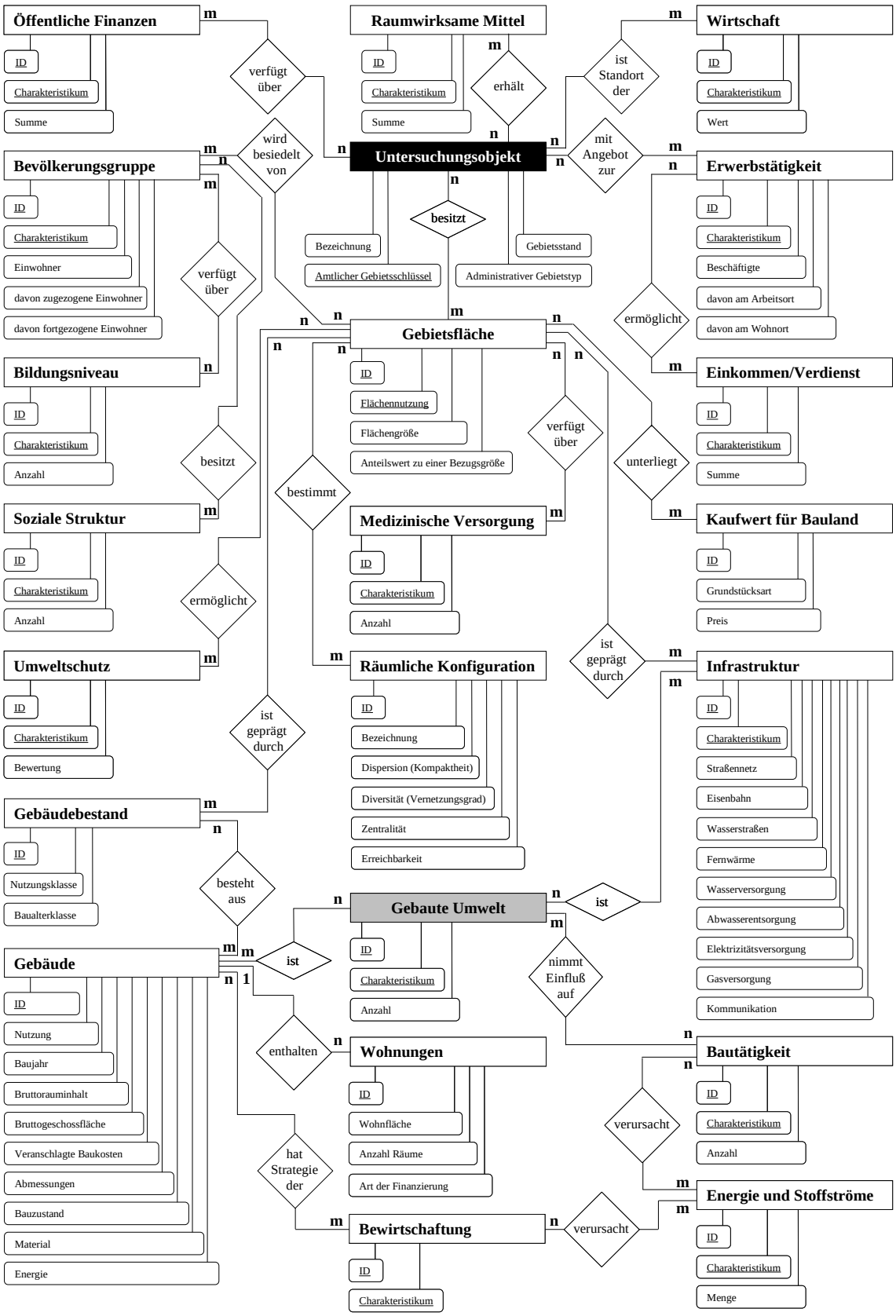


Abbildung 3-5: Darstellung von Untersuchungsmerkmalen im Kontext des ‚Urban Data Mining'

Bei der Auswahl der Variablen, die der konkreten Merkmalsbeschreibung dient, wurde im Wesentlichen darauf geachtet, inhaltliche Kriterien für die Klassifizierung zu berücksichtigen. Das Ergebnis ist jedoch elementar vom real vorhandenen Datenmaterial abhängig. Deshalb tragen die in dieser Untersuchung durchgeführten Ähnlichkeitsuntersuchungen nicht den Anspruch, die Raum- und Siedlungsstruktur in der ganzen Komplexität abbilden zu wollen. Vielmehr werden für einzelne Zeitabschnitte und gewählte räumliche Untersuchungsobjekte mit Hilfe existierender Daten verschiedene Untersuchungsperspektiven dargestellt, die zukünftig bei besserer Datenlage noch eine Erweiterung erfahren können. Die folgenden Tabellen (Tabelle 3-2 bis Tabelle 3-10) geben eine Übersicht des kreis- und gemeindestatistischen Datenbestandes. Ausgewählte Statistiken, die diesen Tabellen zu Grunde liegen, werden im Anschluss hinsichtlich ihrer Erhebungsstruktur ausführlicher beschrieben. Auf Grundlage dieses Datenbestands werden im empirischen Teil individuelle Informationen zum Aufbau der Variablenstruktur, zu Berechnungsgrundlagen und zur Vorverarbeitung gegeben.

Zeitreihen	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
X= Für 440 Kreise sind Daten in gebietsstandsbereinigter Form (Stand: 31.12.2003) vorhanden															
Bevölkerung						X	X	X	X	X	X	X	X	X	
Wanderungen						X	X	X	X	X	X	X	X	X	
Bildung						X	X	X	X	X	X	X	X	X	
Beschäftigte nach Hauptwirtschaftsabteilung								X	X	X	X	X	X	X	
Arbeitslosigkeit												X	X	X	X
Flächennutzung							X				X				
Baufertigstellungen (Wohn- und Nichtwohnbau)						X	X	X	X	X	X	X	X	X	
Wohnungsfortschreibung						X	X	X	X	X	X	X	X	X	
Baulandverkäufe						X	X	X	X	X	X	X	X	X	
Öffentl.Finanzen (Schulden/ Kassenstatistik/Einkünfte)			X			X			X			X		X	
Volkswirtschaftliche Gesamtrechnungen						X	X	X	X	X	X	X	X	X	
Umwelt (Abfall, Wasserver- und entsorgung)									X			X			

Tabelle 3-2: Kreisstatistischer Datenbestand (Statistisches Bundesamt, Statistische Landesämter)[364]

Zeitschnitt: 2000	B	HB	HH	BW	BY	BB	HE	LSA	MV	NI	NRW	RP	SL	SN	SH	TH
X= Daten sind flächendeckend vorhanden																
Messwerte zu Flächen-nutzungsmustern	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabelle 3-3: Kreisstatistischer Datenbestand (Daten der Geocomputation)[365]

[364] Datengrundlage: Querschnittsveröffentlichung ‚Statistik Regional' und ‚Genesis-Online-Regional'

[365] Datengrundlage: Ergebnisse der AML-Programme auf Basis von CORINE-Daten, vgl. THINH [2004 a], dem als Urheber dieser Methoden und der hier eingesetzten Berechnungsergebnisse besonders zu danken ist.

Zeitschnitt: 2006	B	HB	HH	BW	BY	BB	HE	LSA	MV	NI	NRW	RP	SL	SN	SH	TH
X= Daten sind flächendeckend vorhanden, X= Daten sind nur für Teilbereiche vorhanden																
Hauskoordinaten	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ALK	X			X		X	X	X	X	X	X	X	X		X	

Tabelle 3-4: Kreisstatistischer Datenbestand (Daten der Vermessungsverwaltung)[366]

Zeitreihen	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
X= Für 440 Kreise sind Daten in gebietsstandsbereinigter Form (Stand: 31.12.2003) vorhanden															
Bevölkerung						X	X	X	X	X	X	X	X	X	
Wanderungen														X	
Bildung						X	X	X	X	X	X	X	X	X	
Beschäftigte am Arbeitsort und Wohnort, inkl. Pendler														X	
Beschäftigte nach Hauptwirtschaftsabteilung						X	X	X	X	X	X	X	X	X	
Arbeitslosigkeit						X	X	X	X	X	X	X	X	X	
Flächennutzung							X				X				
Baufertigstellungen (Wohn- und Nichtwohnbau)						X	X	X	X	X	X	X	X	X	
Wohnungsfortschreibung						X	X	X	X	X	X	X	X	X	
Baulandverkäufe						X							X		
Öffentliche Finanzen (Einkünfte, Steuereinnahme)						X	X	X	X	X	X	X	X	X	
Volkswirtschaftliche Gesamtrechnungen						X	X	X	X	X	X	X	X	X	
Erreichbarkeit (Bezugspunkt Gemeindeverbände)														X	
Bevölkerungspotential														X	
Zentrale-Orte-Kategorien														X	

Tabelle 3-5: Kreisstatistischer Datenbestand (BBR-Querschnittsveröffentlichung INKAR)[367]

Zeitreihen	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
X= Für 440 Kreise sind Daten in gebietsstandsbereinigter Form (Stand: 31.12.2004) vorhanden															
Bevölkerung	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Wanderungen								X	X	X	X	X	X	X	X
Beschäftigte am Arbeitsort und Wohnort, inkl. Pendler								X	X	X	X	X	X	X	X
Flächennutzung			X				X				X				X
Raumstrukturtypen															X

Tabelle 3-6: Kreisstatistischer Datenbestand (Weitere Daten der laufenden Raumbeobachtung des BBR)

366 Datengrundlage: Amtliches Liegenschaftskataster und Hauskoordinaten (georeferenzierte Gebäudeadresse)

367 Vgl. BBR [2006], die laufende Raumbeobachtung des BBR basiert hauptsächlich auf Daten der amtlichen Statistik, selbst erarbeiteten Indikatoren und Daten des IAB (Institut für Arbeitsmarkt und Berufsforschung). INKAR ist eine Querschnittsveröffentlichung ausgewählter Daten.

Zeitschnitte	B	HB	HH	BW	BY	BB	HE	LSA	MV	NI	NRW	RP	SL	SN	SH	TH
X= Daten sind flächendeckend vorhanden																
Baualtersstruktur 1987 (Wohnbau)	X	X	X	X	X		X			X	X	X	X		X	
Baualtersstruktur 1995 (Wohnbau)						X		X	X					X		X
Baualtersstruktur 2004 (Wohnbau)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Tabelle 3-7: Kreisstatistischer Datenbestand (Daten der Gebäude- und Wohnungszählung und IBIS)[368]

Zeitreihen	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
X= Für 12504 Gemeinden sind Daten in gebietsstandsbereinigter Form (Stand: 31.12.2004) vorhanden															
Bevölkerung	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Wanderungen								X	X	X	X	X	X	X	X
Beschäftigte am Arbeitsort und Wohnort, inkl. Pendler								X	X	X	X	X	X	X	X
Beschäftigte nach Hauptwirtschaftsabteilung								X	X	X	X	X	X	X	
Flächennutzung			X								X				X
Baufertigstellungen (Wohn- und Nichtwohnbau)														X	X
Wohnbau- und Wohnungsfortschreibung					X	X	X	X	X	X	X	X	X	X	X
Erreichbarkeit (Bezugspunkt Gemeindeverbände)															X
Bevölkerungspotential															X
Zentrale-Orte-Kategorien															X
Raumstrukturtypen															X

Tabelle 3-8: Gemeindestatistischer Datenbestand (Daten aus der laufenden Raumbeobachtung des BBR)

Zeitschnitt	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
X= Daten sind flächendeckend vorhanden, X= Daten sind vom Verfasser umgeschätzt worden															
Arbeitslosigkeit														X	
Öffentliche Finanzen (Kassenstatistik, Schulden)														X	X

Tabelle 3-9: Gemeindestatistischer Datenbestand (Statistisches Bundesamt, Statistische Landesämter)[369]

Zeitschnitt: 2006	B	HB	HH	BW	BY	BB	HE	LSA	MV	NI	NRW	RP	SL	SN	SH	TH
X= Daten sind flächendeckend vorhanden, X= Daten sind nur für Teilbereiche vorhanden																
Hauskoordinaten	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ALK®	X			X		X	X	X	X	X	X	X	X	X	X	

Tabelle 3-10: Gemeindestatistischer Datenbestand (Daten der Vermessungsverwaltung)[370]

[368] Datengrundlage: Statistisches Bundesamt bzw. Landesamt und IBIS (Isoplan-Bau-Informations-System). Die Baudatenbank IBIS schreibt den Gebäude- und Wohnungsbestand jährlich auf Kreisebene fort.

[369] Datengrundlage: Querschnittsveröffentlichung ‚Statistik Lokal'

[370] Datengrundlage: Automatisierte Liegenschaftskarte und Amtliche Hauskoordinaten (georeferenzierte Gebäudeadresse)

3.3 Datenqualität und fehlende Werte

Die Qualität des empirischen Datenmaterials kann durch einige Parameter grundsätzlich beeinträchtigt sein. Zu nennen sind bewusste Verfälschungen bzw. Manipulationen (I.), systematische Fehler bei der Erfassung bzw. Berechnung (II.), mangelnde Repräsentativität (III.), ungeeignete Daten - Verletzung des Adäquationsprinzips (IV.),[371] uneinheitliche Definitionen und Erfassungsmethoden (V.), unterschiedlicher Zeitbezug (VI.) und fehlende Werte (VII.).

Das in dieser Arbeit verwendete Datenmaterial bezieht sich wie tabellarisch nachgewiesen primär auf Daten der jährlichen bzw. im Vierjahresturnus erscheinenden Statistik des Statistischen Bundesamtes bzw. der Landesämter.[372] Ergänzt wird das Datenmaterial durch die Ergebnisse von Gebäude- und Wohnungszählungen. In der Regel kann davon ausgegangen werden, dass das Datenmaterial der amtlichen Statistik[373] frei von systematischen Fehlern oder sogar beabsichtigten Manipulationen ist. Als eine der wenigen Ausnahmen ist die Fortschreibung der Gebäude- und Wohnungsbestandsstatistik zu nennen, die deshalb ausführlicher im Abschnitt 3.4.1.3 beschrieben wird, jedoch aufgrund ihrer Bedeutung nicht ausgeschlossen wird.

Die Repräsentativität der Daten ist in den meisten Fällen als gesichert zu betrachten, da es sich häufig um Ergebnisse von Totalerhebungen handelt bzw. um Daten, die auf Totalerhebungen fortgeschrieben wurden. Als Sonderfall ist die Statistik der Kaufwerte für Bauland zu nennen, die zwar auf Grundlage von Totalerhebungen der im Untersuchungszeitraum angefallenen Veräußerungsfälle erstellt wird, natürlich aber nur ein geringer Teil des Bestandes tatsächlich jährlich veräußert wird und damit von der Statistik erfasst werden kann.

Die einzelnen Fragestellungen werden mit den Objekten bearbeitet für die das statistische Datenmaterial vorhanden bzw. leicht zu beschaffen ist. Dennoch eignen sich nicht immer alle Variablen, die zuvor theoretisch denkbaren Merkmale inhaltlich treffend zu quantifizieren. Im Zuge der Dateninspektion (siehe Abschnitt 2.1) werden die gewählten Variablen auf Plausibilität und Zuverlässigkeit nochmals analysiert, um Fehlinterpretationen zu vermeiden.

Uneinheitliche Definitionen und Erfassungsmethoden können wegen der verwendeten Daten der Amtlichen Statistik genauer beurteilt und mit den Fachreferatsleitern ggf. erörtert werden.

---

371 Die Korrespondenztheorie umfasst den Begriff der Adäquation (Entsprechung, Übereinstimmung).

372 Wie bereits in Tabelle 3-8 gezeigt, bilden die Daten der laufenden Raumbeobachtung ein wichtiges Element für diese Untersuchung. Die Daten basieren im Allgemeinen auf dem Angebot des Statistischen Bundesamtes und der Statistischen Landesämter.

373 Nach LIPPE [1990, S.11] beruht die Zuverlässigkeit der amtlichen Statistik weitgehend auf der Auskunftspflicht, die entscheidend für das Prädikat ‚amtlich' ist.

Einen genaueren Eindruck von den Erhebungsanforderungen und damit verbundenen Schwierigkeiten gewinnt man zusätzlich durch die Detailbetrachtung besonders relevanter Statistiken. Da sich die Daten in der Regel auf Querschnittsveröffentlichungen bzw. Daten der laufenden Raumbeobachtung beziehen, ist bereits eine Aufbereitung bundesweiter Statistiken durch das Statistische Bundesamt erfolgt. Der Verfasser musste nicht eigenständig Daten der einzelnen Bundesländer miteinander kombinieren, so dass eine höhere Zuverlässigkeit auf jeden Fall anzunehmen ist und uneinheitliche Definitionen weitgehend auszuschließen sind.

Für den Aufbau von einzelnen Variablen werden ausnahmslos Statistiken mit einheitlichem Zeitbezug verwendet. Somit wird auf Interpolationen bzw. Extrapolationen verzichtet, die eine Vergleichbarkeit von Ausprägungen mit unterschiedlichem Zeitbezug herstellen würden. Als Voraussetzung wäre ein mittel- oder langfristiger Trend notwendig, wobei das Datenmaterial zumindest aus zwei aufeinanderfolgenden Stichtagen zur Verfügung stehen sollte.

Sowohl auf der Gemeinde- als auch der Kreisebene ist das gewählte statistische Datenmaterial der amtlichen Statistik flächendeckend vorhanden und eine uneingeschränkte Untersuchung möglich. Innerhalb der Variablenstruktur wird bei den eingesetzten Untersuchungsobjekten gesondert nach Fehlwerten gesucht. Es sind in den meisten Fällen (siehe Abschnitt 4) keine Fehlstellen vorhanden, so dass die Problematik fehlender Werte nicht unmittelbar zu lösen oder durch Schätzwerte zu ersetzen war.[374] In anderen Arbeiten der Stadtklassifikation werden fehlende Werte im einfachsten Fall durch den arithmetischen Mittelwert ersetzt.

Die vorhandenen Daten gehen nicht unmittelbar als Variable in die Ähnlichkeitsuntersuchungen ein, sondern werden zuvor in geeigneter Weise aufbereitet. In der Regel sind dazu bestimmte Rechenoperationen erforderlich, da wie bereits gezeigt, die Variablen erst durch Zusammenfassung mehrerer empirischer Größen bzw. Relativierung entstehen. Je größer die Zahl der verwendeten Variablen wird, desto wahrscheinlicher treten Gruppen von Variablen auf, die der Beschreibung eines gleichen Merkmals dienen. Im Sinne der erstmaligen Informationsgewinnung im Rahmen dieser Forschungsarbeit wird der Prozess der Gleichgewichtung der Untersuchungsvariablen für alle Untersuchungen angestrebt, um nicht von vornherein eine zusätzlich denkbare Fehlerquelle für den Klassifikationsprozess infolge verwendeter Gewichtungsfaktoren einzubeziehen. Diese Fehler könnten aufgrund mangelnder Kenntnis der Variablen entstehen, bzw. sind Vorkenntisse über die – a priori unbekannte – zu bildende Klassenstruktur nicht zu erwarten, die damit eine größere Bedeutung eines

[374] Vgl. STEINHAUSEN / LANGER [1977, S. 176 ff.], LITTLE / RUBIN [1987], BECHER [1995, S. 99 ff.], IMM [2002] und ALLISON [2002]

Merkmals in Form einer höheren Gewichtung nachvollziehbar erscheinen lassen. Fehler könnten somit jederzeit aus subjektiven Entscheidungen resultieren, die nicht auf allgemein vertretbarem objektivem Expertenwissen beruhen.

Um die Gemeinden inbesondere gebäudebestandsorientiert besser beschreiben zu können, wurde das Angebot der bereits ausführlich diskutierten amtlichen Statistik durch die Recherche nach weiteren Datenquellen und deren Nutzung verbessert. Bei der amtlichen Statistik besteht das Problem, dass der Bestand an Wohngebäuden und Wohnungen zwar erfasst wird, jedoch über den Bestand der Nichtwohngebäude nahezu keine Daten existieren.

Die Erhebung einer eigenen Primärstatistik zur Beschaffung des tatsächlich benötigten Datenmaterials war aus Gründen des damit verbundenen Aufwandes auf Gemeindeebene von vornherein auszuschließen. Theoretisch denkbar ist zwar die Schätzung des Gebäudebestandes für einzelne Gemeinden analog zu HASSLER / KOHLER.[375] Der Aufbau einer auf diese Weise ermittelten Datenbasis, die sogar eine spätere Ähnlichkeitsuntersuchung zwischen mehreren Gemeinden zulässt, ist aus Gründen des Zeitaufwandes und möglicher unkalkulierbarer Schwierigkeiten jedoch ebenfalls nicht in großem Umfang realisierbar.

In dieser Arbeit wird auf Ebene der Kreise und auch Gemeinden das statistische Datenmaterial über den tatsächlich vorhandenen Gebäudebestand und seine Nutzungsstruktur durch Abfragen auf die Automatisierte Liegenschaftskarte (ALK) und die bundesweit verfügbaren Hauskoordinaten[376] im Jahr 2006 gewonnen. Die eingesetzten Daten verstoßen nicht gegen die im Eingang dieses Abschnitts genannten Parameter zur Datenqualität (I.-VII.), so dass ebenfalls von einer hohen Datenqualität auszugehen ist. Die Art der Nutzung wird anhand von Objektschlüsselkatalogen (OSKA) durch die Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder (ADV) definiert.[377] Weiterhin ist festzustellen, dass die ALK in Deutschland gerade erst aufgebaut wird und somit nicht flächendeckend in allen Bundesländern

---

375 HASSLER, U. /KOHLER, N. [2004, S. 116 ff.]: Es werden die Arbeitsschritte erläutert, die notwendig sind, um den deutschen Gebäudebestand insbesondere die nichtlandwirtschaftlichen Betriebsgebäude zu schätzen.

376 Die Hauskoordinaten erfassen nahezu alle Gebäude in Deutschland mit einer Gebäudeadresse. Demnach existieren 19.303.874 adressierte Gebäude (Erfassungsstand 30.10.2006), wobei die Lankdkreise Parchim, Güstrow und Doberan sowie die Stadt Gotha noch nicht enthalten sind.

377 Nach Rücksprache mit den Fachdezernenten der Vermessungsverwaltung ist für das in dieser Arbeit eingesetzte Datenmaterial der ALK eine Vergleichbarkeit von Gebäudenutzungsstrukturen auf Ebene der Gemeinden bzw. Kreise in Deutschland grundsätzlich möglich. Ein flächendeckender Nachweis aller derzeit in Deutschland existierender Gebäude (100 %) lässt sich allerdings nur durch die Reproduktion von Luftbildern und Satellitenphotos garantieren. Die Objektschlüsselkataloge der Länder sind in der Regel bez. der Objektschlüssel strukturell gleich aufgebaut. Einige Unterschiede bei der Definition von Gebäudenutzungen sind aber denoch nicht auszuschließen. Diese Ungenauigkeit wird im Hinblick auf die bisher vorhandenen Kenntnisse über den Bestand im Nichtwohnbau in den Gemeinden und Kreisen nicht als Hindernis bzw. Ausschlusskriterium für die Verwendung dieser Daten betrachtet.

vorgehalten wird.[378] Als mögliche Fehlerquelle sei die manuelle Nachbearbeitung der Daten genannt.[379] Abschließend bleibt anzumerken, dass es sich zwar um Daten der amtlichen Vermessungsverwaltung handelt, jedoch auf dieser Datengrundlage bisher noch keine amtlichen Statistiken bundesweit veröffentlicht werden.[380] Generelle Informationen zur ALK und den Hauskoordinaten finden sich bei den Einzelbeschreibungen der statistischen Datenquellen.

Das Datenmaterial der amtlichen Statistik wird darüber hinaus durch Maßzahlen der Geocomputation ergänzt. Zu unterscheiden ist zwischen Daten des Leibniz-Instituts für ökologische Raumentwicklung (IÖR) und des Bundesamtes für Bauwesen und Raumordnung (BBR).

Die digitalen Bodenbedeckungsdaten aus CORINE Land Cover des Jahres 2000 bilden am IÖR die Grundlage für die Berechnungen von Maßzahlen zur Beschreibung der räumlichen Konfiguration und die Analyse von Flächennutzungsmustern. Das IÖR hat die Aufgabe, in interdisziplinärer Arbeitsweise Grundfragen einer ökologisch ausgerichteten Raumwissenschaft im nationalen, europäischen und internationalen Zusammenhang zu erforschen. Die hier eingesetzten Daten wurden am IÖR berechnet und verstoßen ebenfalls nicht gegen die im Eingang dieses Abschnittes genannten Parameter zur Datenqualität (I.-VII.). In dieser Arbeit stehen ermittelte Maßzahlen zur räumlichen Konfiguration (siehe THINH)[381] zur Verfügung und ergänzen die Datenbasis. Gerade in ländlichen Kreisen wie z.B. in Rheinland-Pfalz mit flächenmäßig sehr kleinen Siedlungsmustern wird die Problematik der Untererfassungsgrenze der baulichen Bodenbedeckung von 25 ha von THINH als Schwierigkeit für eine aussagekräftige Beschreibung der siedlungsstrukturellen Eigenschaften genannt. Nach SIEDENTOP[382] ist die Abbildung siedlungsstruktureller Merkmale auf Gemeindeebene mit Hilfe der aktuellen CORINE Land Cover Daten kaum möglich.

---

[378] Der Aufbau der ALK und die damit verbundene Erfassung von Gebäuden ist in den Bundesländern teilweise noch nicht abgeschlossen, so dass sich eine bundesweit flächendeckende Erhebung frühestens Ende 2007 realisieren lässt. In die Ähnlichkeitsuntersuchung werden die Kreise oder Gemeinden aufgenommen, für die ohne größere Schwierigkeiten die Daten beschaffbar bzw. überhaupt vorhanden sind.

[379] Die Daten von ca. 25 Mio. Gebäuden wurden nach Nutzungsart und Gebäudegrundfläche in eine eigene Datenbank manuell eingepflegt und mit Hilfe späterer Abfragen wurden einzelne Variablenwerte daraus erzeugt. Bisher handelt es sich leider nur um Teilbestände. Kontrollrechnungen und Visualisierungen haben entscheidend dazu beigetragen, eigene Fehler zu entdecken bzw. gar nicht entstehen zu lassen.

[380] Der Verfasser dieser Arbeit hat eigenständig die betreffenden amtlichen Vermessungsverwaltungen der 16 Bundesländer kontaktiert und um den Abruf von Daten bzw. die tatkräftige Unterstützung gebeten. Zukünftig wird die Beschaffung dieser Daten sicherlich wesentlich einfacher werden. Die ALK wird nach Meinung des Verfassers sicherlich eine sehr wichtige Datenquelle für das Statistische Bundesamt bzw. die Statistischen Landesämter bilden und kann die bisher durchgeführten Totalerhebungen von Gebäude- und Wohnungszählungen zukünftig ggf. sogar ersetzen.

[381] Datengrundlage: Ergebnisse der AML-Programme auf Basis von CORINE-Daten, vgl. THINH [2004 a], dem als Urheber dieser Methoden und der hier eingesetzten Berechnungsergebnisse besonders zu danken ist.

[382] SIEDENTOP et al. [2003, S. 151 und S. 163]

Die zusätzlich verwendeten Daten der Geocomputation stammen aus der laufenden Raumbeobachtung des BBR und sind deshalb als sehr vertrauenswürdig einzustufen.[383] Das Bevölkerungspotential und die durchschnittliche Fahrzeit zum nächstgelegenen Oberzentrum werden durch diese Daten ausgedrückt. Die Daten zur Fahrzeit beziehen sich auf ein Erreichbarkeitsmodell des BBR und stehen dem Verfasser nur auf Ebene der Gemeindeverbände zur Verfügung. Die Ausprägungen der einzelnen Gemeinden beziehen sich somit immer auf den jeweiligen Gemeindeverband, für den diese Werte berechnet wurden.

Als weitere Ergänzung zu den Daten der amtlichen Statistik sind die Daten aus der Baudatenbank IBIS zur Baualtersstruktur im Wohnbau in Deutschland anzusehen. Es handelt sich um eine Datenbank, die von der isoplan-Gruppe bereitgestellt wird. Die isoplan-Gruppe ist ein Marktforschungs- und Consulting-Institut und schreibt selbstständig den Gebäude- und Wohnungsbestand auf Ebene der Kreise fort. Unter Kenntnis der Probleme, die bereits mit der amtlichen Fortschreibungsstatistik des Wohngebäudebestandes verbunden sind, ist auch hier die Qualität der Daten nach fast 20 Jahren der letzen Totalerhebung gerade unter dem Parameter II (‚systematische Fehler bei der Erfassung bzw. Berechnung) zur Datenqualität als kritisch zu bewerten. Da es grundsätzlich sehr schwer ist, auf Ebene von Kreisen oder Gemeinden Aussagen über die vorhandene Baualtersstruktur zu treffen, wurden diese Daten unter Kenntnis der damit verbundenen Probleme für eine Visualisierung im Nebenteil eingesetzt, um zumindest gewisse Vermutungen zu ermöglichen, inwieweit ein Kreis über einen alten bzw. sehr neuen Wohngebäudebestand verfügt. Unterschieden werden 4 Altersklassen wie oft bei der amtlichen Statistik (bis 1948, 1949-1968, 1969-2000 und 2001-2005).

Das vom Deutschen Städtetag jährlich herausgegebene Statistische Jahrbuch Deutscher Gemeinden[384] eignet sich als Datenquelle für die hier durchgeführte flächendeckende Betrachtung nur bedingt, da es zum einen nicht alle Gemeinden und auch wenige zusätzliche Statistiken enthält, die für diese Arbeit von Interesse sind.

Eine zusätzliche merkmalsbezogene Beschreibung auf Basis der Attribute zur zentralörtlichen Funktion, die in den Landesentwicklungsplänen und Regionalplänen festgeschrieben werden, ist für diese Arbeit nicht geeignet, da erhebliche Unterschiede in der Ausweisungspraxis bestehen. SIEDENTOP[385] verweist beispielsweise auf die großen Abweichungen der Verteilungsmuster mittelzentraler Standorte zwischen den verschiedenen Bundesländern.

---

383 Vgl. BBR [2006]

384 Ab 1890 Statistisches Jahrbuch deutscher Städte – später Statistisches Jahrbuch deutscher Gemeinden. Daten auf Gemeindeebene liegen entweder erst ab 10.000 oder sogar erst ab 20.000 Einwohnern vor.

385 Vgl. SIEDENTOP et al. [2003, S. 27], Vergleich der zentralörtlichen Struktur (Ober- und Mittelzentren)

3.4 Amtliche Statistik in Deutschland[386]

3.4.1 Einzelstatistiken[387]

3.4.1.1 Bautätigkeitsstatistik

Auf Grundlage des Hochbaustatistikgesetzes (HBauStatG)[388] wird die Bautätigkeitsstatistik erhoben, um statistische Daten über die Struktur, den Umfang und die Entwicklung der Bautätigkeit im Hochbau bereitzustellen. Die Statistik ermöglicht eine Beschreibung der wirtschaftlichen Entwicklung im Bausektor und ist beispielsweise von Bedeutung für die Planung in den Gebietskörperschaften, die Wirtschaft, die Forschung und den Städtebau. Die Merkmale der Bautätigkeitsstatistik unterstützen eine Analyse von bau-, wohnungs- und energiewirtschaftlichen sowie bautechnischen Entwicklungen.

Die Datengewinnung erfolgt anhand eines vordefinierten Erhebungsbogens sowohl aus den Verwaltungsunterlagen der Bauaufsichtsbehörden nach landesrechtlicher Regelung als auch mit Hilfe der Bauherren. Erfasst werden Baumassnahmen zum Zeitpunkt der Genehmigung oder der Zustimmung bzw. dem Zeitpunkt, zu dem diese auf Grundlage landesrechtlicher Verfahrensvorschriften[389] ausgeführt werden dürfen. Weiterhin werden Baufertigstellungen und Bauzustände am Jahresende (Bauüberhänge) berücksichtigt. Als fertiggestellt gelten Wohn- bzw. Nichtwohngebäude, bei denen lediglich noch verschönernde Maßnahmen vorzunehmen sind. Bezüglich des Zeitpunktes der Fertigstellung ist die Ingebrauchnahme und nicht die baupolizeiliche Schlussabnahme entscheidend. Die Erhebungsmerkmale beziehen sich je nach Erhebungseinheit u.a. auf Gebäude, Rauminhalt, Wohnungen, Wohnräume, Wohn- bzw. Nutzfläche und veranschlagte Kosten sowie Bauherrengruppen. Nicht berücksichtigt werden Baumassnahmen für ausschließlich sonstigen Nutzraum bis zu 350 Kubikmeter Rauminhalt oder bis zu 18.000 Euro veranschlagte Kosten.[390]

Im Rahmen der Bautätigkeitsstatistik sind zusätzlich Bauabgänge entsprechend § 3 Absatz 4 des HBauStatG erfasst. Ein Abgang bezieht sich auf Gebäude oder Teile davon, die durch ordnungsbehördliche Massnahmen, Schadensfälle oder Abbruch vom Bestand abzurechnen sind. Denkbar ist auch eine Nutzungsänderung zwischen Wohn- und Nichtwohnzwecken.

---

386 Vgl. Das Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22.01.1987 (BGBl. I, S. 462, 565), zuletzt geändert durch Artikel 2 des Gesetzes vom 09.06.2005 (BGBl. I, S. 1534).

387 Der große Umfang an amtlichen Statistiken ermöglicht im Rahmen dieser Arbeit nicht eine vollständige Beschreibung. Deshalb erfolgt nur eine Auswahl von besonders relevanten Einzelstatistiken. Viele Einzelstatistiken sind Bestandteil der amtlichen Querschnittsveröffentlichungen. Zusatzinformationen finden sich unter http://www.destatis.de (10.09.2006) und http://www.bbr.bund.de/ (22.09.2006).

388 Hochbaustatistikgesetz – HBauStatG (5. Mai 1998, BGBl. I, S. 869): Gesetz über die Statistik der Bautätigkeit im Hochbau und Fortschreibung des Wohnungsbestandes.

389 Vgl. Landesbauordnungen der einzelnen Bundesländer

390 Vgl. HBauStatG §2 (05.05.1998): Erhebungseinheiten

Die jährlichen Angaben aus der Baufertigstellungsstatistik ermöglichen zusätzlich die Fortschreibung des Wohngebäude- und Wohnungsbestandes in der Zeit zwischen den Totalerhebungen (siehe Abschnitt 3.4.1.3).

Insbesondere die Baufertigstellungsstatistik erlaubt zukünftig einen Ähnlichkeitsvergleich hinsichtlich Struktur und Umfang der Bautätigkeit im Wohn-[391] und Nichtwohnbau.[392] Die erste Erhebung zur Baufertigstellung erfolgte im Jahr 1953 in den alten Bundesländern und ab 1991 begann hierzu der Aufbau von Erfassungsstrukturen in den neuen Bundesländern.[393] Die relativ lange Geltungsdauer von 20 Jahren des 2. Baustatistikgesetzes trägt dazu bei, dass die Ergebnisse der Baufertigstellungsstatistik als lange Zeitreihe geführt werden und somit eine inhaltliche Vergleichbarkeit der jährlichen Angaben gewährleistet ist. Für die räumliche Vergleichbarkeit ist aber zu beachten, dass bundesweit erst ab dem Berichtsjahr 1991 Daten vorhanden sind und gerade auf Gemeindeebene die Problematik der Gebietsstandsänderungen sehr große Schwierigkeiten bereitet.[394] Die Beschaffung von gebietsstandsbereinigten Gemeindedaten in Form längerer Zeitreihen ist für die Baufertigstellungsstatistik in großem Umfang nur schwer zu bewältigen. Die Gemeindedaten zur Baufertigstellung beziehen sich deshalb im Wohn- und Nichtwohnbau nur auf die Jahre 2003 und 2004.[395] Es wird der Anteil von fertiggestellten Gebäuden im Nichtwohnbau an der Bautätigkeit der Jahre 2003 und 2004 berechnet. Die Betrachtung von längeren Zeitreihen auf Gemeindeebene wird mit fortgeschriebenen Daten zum Wohnungs- und Wohngebäudebestand möglich (siehe Abschnitt 3.4.1.3). Weitere gebietsstandsbereinigte Daten für den Wohn- und Nichtwohnbau auf Ebene der Kreise liegen als Zeitreihe von 1993 bis 2004 vor. Für das westdeutsche Bundesgebiet werden die Daten durch die Anzahl der Baufertigstellungen ab 1983 erweitert.

---

391 Wohngebäude sind Gebäude, die mindestens zur Hälfte (gemessen an der Gesamtnutzfläche nach DIN 277) Wohnzwecken dienen. Die Wohnfläche wird nach der Wohnflächenverordnung (WoFlV) vom 25. November 2003 (BGBl. I, S. 2346) berechnet. Zu unterscheiden sind Wohngebäude mit einer Wohnung (Einfamilienhäuser), Wohngebäude mit zwei Wohnungen (Zweifamilienhäuser), Wohngebäude mit drei und mehr Wohnungen (Mehrfamilienhäuser), Wohnheime.

392 Nichtwohngebäude sind Gebäude, die überwiegend (zu mehr als der Hälfte der Nutzfläche nach DIN 277) Nichtwohnzwecken dienen. Die Nichtwohngebäude sind untergliedert in folgende Gebäudearten: Anstaltsgebäude, Büro- und Verwaltungsgebäude, Landwirtschaftliche Betriebsgebäude, Nichtlandwirtschaftliche Betriebsgebäude (u.a. Fabrik- und Werkstattgebäude, Handels- und Lagergebäude, Hotels und Gaststätten). Weitere Untergliederungen definiert die ‘Systematik der Bauwerke’ (Hrsg. Statistisches Bundesamt), die die Grundlage für die Zuordnung der Gebäude nach der Gebäudeart ist und praktisch für jedes Gebäude eine eindeutige Zuordnung regelt.

393 Es sei chronologisch auf folgende Bundesgesetzblätter verwiesen, die sich auf BauStatG (1953-1978) sowie 2. BauStatG (1979-1996) und HBauStatG (ab 1997) beziehen: 24.03.1953 (BGBl. I, S. 78), 20.08.1960 (BGBl. I, S. 704), 27.07.78 (BGBl I, S.1118), 06.06.94 (Artikel 9, BGBl I, S. 1184, 1798), 20.11.96 (Artikel 12, BGBl. I, S. 1804), 05.05.1998 (BGBl. I, S. 869), 15.12.2001 (Artikel 6, BGBl. I, S. 3762).

394 Vgl. DESTATIS [2005, S.4]: Qualitätsberichte zur Baufertigstellungsstatistik.

395 Nach Auskunft des Statistischen Bundesamtes ist eine digitale Datenspeicherung der Baufertigstellungen im Nichtwohnbau auf Ebene der Gemeinden (Stand 2004: 12504 Gemeinden) in gebietsstandsbereinigter Form für einen längeren Zeitraum nicht existent und würde deshalb eine zusätzliche Digitalisierung erfordern.

Um ergänzend zu den Untersuchungen dieser Arbeit weitere Aussagen über die Bautätigkeit zu treffen, können zukünftig Daten aus der administrativen Ebene der Bundesländer bzw. zusätzlich aus der Systematik der Gemeindegrößenklassen verwendet werden.[396]

Die Entwicklung der Baufertigstellung der Nichtwohngebäude lässt sich über einen längeren Zeitraum von 46 Jahren nachvollziehen. Die Daten existieren ab dem Jahr 1953, das als Startpunkt dieser statistischen Erhebung gilt und endet im Jahr 1999 aufgrund der Einstellung dieser Statistik durch das Statistische Bundesamt. Die Datenlage ermöglicht nur eine Untersuchung des westdeutschen Bundesgebietes und aufgrund der geschilderten Gesetzesänderungen ist der Aufbau einer durchgängig vergleichbaren und belastbaren Zeitreihe kaum möglich. Für die Abschätzung der Bautätigkeit in der Bundesrepublik Deutschland trägt diese Statistik eine große Bedeutung, weil der Nichtwohnbau durchgängig nach amtlich definierten Nutzungsklassen erfasst ist. Für das zeitliche Intervall 1953 bis 1999 existieren die folgenden Nutzungsklassen: Anstaltsgebäude, Bürogebäude, Landwirtschaftliche Betriebsgebäude, Gewerbliche Betriebsgebäude, Sonstige Nichtwohngebäude und darunter explizit aufgeführt oftmals der Schulbau.

Es wäre zusätzlich wünschenswert, die Dynamik der Gebäudeabgänge im Sinne einer gebäudebestandsorientierten Gesamtbetrachtung auf Gemeindeebene zu untersuchen. Festzustellen ist leider, dass die Abgangsstatistik einerseits erst ab 1979 bundesweit überhaupt in veröffentlichter Form einem größeren Nutzerkreis zugänglich gemacht wurde und andererseits keinesfalls Daten in einer derartig niedrigen administrativen Aggregationsstufe für eine größe Objektzahl zentral zu beschaffen sind. Die Abgangsstatistik ist auf Ebene des gesamten Bundesgebietes für Gebäude mit Wohnraum vorhanden, jedoch für primär angestrebte Ähnlichkeitsuntersuchungen von sekundärer Bedeutung. Im Anschluss an diese Forschungsarbeit sind für einzelne Gemeinden detaillierte Einzelfalluntersuchungen vorstellbar, um über die bestehende Gebäudebestandsstruktur und die Dynamik der Abgänge genaue Erkenntnise zu gewinnen.[397] SCHWAIGER,[398] aber gerade HASSLER / KOHLER[399] zeigen zur Beschaffung von historischem Datenmaterial für diese vertiefenden Untersuchungen weitere Perspektiven und Möglichkeiten auf.

---

396 Durch eigene Digitalisierung der analogen Statistikangebote wurden die Daten erzeugt. Da die Untersuchung jedoch eine eigene Forschungsarbeit repräsentiert, soll nur auf die Möglichkeit verwiesen sein.

397 Vgl. BRADLEY / FERRARA [2004]: Abschlussbericht zum DFG-Forschungsprojekt BEVAL – Neben Daten der Amtlichen Statistik sind für diese Untersuchungen Daten aus eigenen Erhebungen durch direkte Ortsinspektion oder Auswertungen der Akten zur Gebäudefeuerversicherung denkbar. Verwiesen sei auf das lange Zeit existierende Monopol der öffentlichen Sachversicherer und Gebäudefeuerversicherungen.

398 SCHWAIGER [2002, S. 119 ff.]

399 Vgl. HASSLER, U. /KOHLER, N. [2004, S. 115]

3.4.1.2 Gebäude- und Wohnungszählung

Für wohnungspolitische Überlegungen und Planungen in der Bauwirtschaft sind Angaben hinsichtlich der Qualität, Quantität und Struktur über den Bestand an Wohngebäuden und Wohnungen in möglichst tiefer regionaler Gliederung von großer Bedeutung. Umfangreiches Datenmaterial liefern die Totalzählungen bzw. so genannte Gebäude- und Wohnungszählungen, die jedoch mit hohen Kosten und großem Arbeitsaufwand verbunden sind und dadurch auch nur in sehr großen Zeitabständen durchgeführt werden.

In der Regel finden die Gebäude- und Wohnungszählungen gemeinsam mit Volkszählungen statt, so dass der hierfür vorgesehene Zählapparat, meistens ehrenamtliche Zähler, auch die Zählung der Gebäude und Wohnungen mit übernehmen kann. Aus Kostengründen[400] bleiben Nichtwohngebäude in den meisten Fällen unberücksichtigt. Zur Zeit der Erhebung bestehende aktuelle Probleme werden in die Planungsvorgaben der Zählungen aufgenommen. Vor dem zweiten Weltkrieg richtete sich die Bestimmung einer Wohnung beispielsweise nach den mietrechtlichen Verhältnissen, so dass die Summe der Hauptmietverträge der Zahl der Wohnungen entsprach. Nach dem Zweiten Weltkrieg hatte die Wohnungsnot zur Folge, dass große Wohnungen oftmals durch eine Mehrfachbelegung gekennzeichnet waren, die aber mietrechtlich gleichrangigen Wohnparteien unterlagen. Bei der Gebäude- und Wohnungszählung 1950 wurde deshalb der bautechnische Begriff ‚Wohnung' der Erhebung zu Grunde gelegt. Für so genannte Normalwohnungen galt das Kriterium des Vorhandenseins einer bauplanmäßig vorgesehenen Küche oder Kochnische. Die Gebäude- und Wohnungszählungen richten somit die damit verbundenen Erhebungspapiere auf die zu erfragenden Merkmale und Tatbestände aus und unterliegen eindeutig differenzierten Begriffsstrukturen.[401] Um das Datenmaterial im Detail zu bewerten, sind die Veröffentlichungen der jeweiligen Zählung einzusehen.

Im Folgenden wird ein Überblick zu den bundesweiten Zählungen gegeben. Uneinheitliche Definitionen und Erfassungsmethoden können bei Daten einer Totalerhebung ausgeschlossen werden. Aufgrund der definierten Zählungsstichtage können für diese Statistiken weitere Fehler durch einen unterschiedlichen Zeitbezug ausgeschlossen werden. Auf Basis der Ergebnisse der Zählung aus dem Jahr 1987 sind einige fortschreibende Rechnungen innerhalb des Statistischen Bundesamtes bzw. der Landesämter vorhanden aber auch Fremdorganisationen wie z.B. die Gruppe ‚isoplan' liefern ergänzendes Datenmaterial (vgl. Abschnitt 3.3).

---

400 Vgl. KNOP [1989, S. 483]

401 Im Nebenteil B sind Begriffsbestimmungen der jeweiligen Zählungen tabellarisch vorhanden.

Im Deutschen Reich[402] fand die erste große Wohnungszählung im letzten Jahr des ersten Weltkrieges am 28. Mai 1918 statt. Sie erstreckte sich nur auf Gemeinden mit über 5.000 Einwohnern, Hauptziel war die Feststellung der bewohnten und leerstehenden Wohnungen, der vorhandenen Küchen, der vertraglichen Jahresmieten und der Bewohner.

Am 16. Mai 1927 folgte die zweite große Reichswohnungszählung, ebenfalls nur in Gemeinden mit über 5.000 Einwohnern. Damals war es allerdings den Ländern freigestellt, auch kleinere Gemeinden mit einzubeziehen. Die Zahl der Gebäude mit Wohnungen und die Zahl und Größe der Wohnungen wurde dabei ermittelt. Das in dieser Zeit von besonderer Bedeutung existente Untermietverhältnis, das mehrfache Zusammenwohnen von Familien und die Wohndichteverhältnisse waren Ziel einer genaueren Analyse.

Zusammen mit der Volkszählung vom 17. Mai 1939 wurde die Gesamtzahl der Wohnungen ohne jede Aufgliederung und ohne eine Untersuchung der Belegungsverhältnisse ermittelt.

Anlässlich der Volkszählung am 29. Oktober 1946 wurde nach dem zweiten Weltkrieg zunächst eine Bestandsaufnahme aller Wohnungen unter dem Aspekt Mehrfachbelegung und Wohndichte durchgeführt, allerdings nur in den Ländern der amerikanisch besetzten Zone.

Mit der Volkszählung am 13.09.1950 fand die erste einheitliche und umfassende Gebäude- und Wohnungszählung im so genannten Nachkriegsdeutschland statt. Die Zählung bezieht sich auf den Gesamtbestand in West- und Ostdeutschland und enthält statisches Datenmaterial zum Nichtwohnbau. Die rechtliche Grundlage bildete das ‚Gesetz über eine Zählung der Bevölkerung, Gebäude, Wohnungen, nichtlandwirtschaftlichen Arbeitsstätten und landwirtschaftlichen Kleinbetrieben' vom 27. Juli 1950 (Volkzählungsgesetz 1950). Von 100 gezählten Gebäuden waren im Bundesdurchschnitt 61 Normalwohngebäude. Der Anteil der vor 1918 erbauten ‚Normalwohngebäude' lag im Bundesdurchschnitt bei 65 %. Die Gebäude und Wohnungen wurden nicht nur nach der Größe, sondern auch nach Eigentümer, Alter, Mietverhältnis und Ausstattung (z.B. Elektrizität, Gas und Wasser) unterschieden. Festgestellt wurden die bezahlten Wohnungsmieten und – in einer repräsentativen Nacherhebung – die gezahlten Untermieten. Die Statistik zu Kriegsschäden enthält nicht die total zerstörten Gebäude, so dass genannte Anteilswerte nicht den vollen Umfang der Zerstörung angeben.

Die am 25. September 1956 durchgeführte Erhebung ist eine reine Wohnungszählung gewesen, die aufgrund des Bundesgesetzes über Statistik der Wohn- und Mietverhältnisse (Wohnungsstatistik 1956/1957) vom 17. Mai 1956 bearbeitet wurde.

[402] Frühere Zählungen beziehen sich oft auf einzelne Bundesstaaten. Im Gebiet des preussischen Staates führte das Königlich Statistische Bureau zu Berlin in den Jahren 1861, 1878 und 1893 Zählungen durch.

Obwohl bei dieser Erhebung auch Angaben über die Gebäude erfragt wurden, galt die Gebäudeliste primär lediglich als Leitliste; die darin enthaltenen Angaben wurden nur für die Auswertung der Wohnungsergebnisse und für die Aufstellung einer Gebäudekartei als Grundlage für künftige Stichprobenerhebungen und repräsentative Zusatzerhebungen verwendet. Zahlen über die Gebäude liegen aus der Wohnungsstatistik 1956/57 in veröffentlichter Form nicht vor.

Die Zählung vom 6. Juni 1961 war primär eine Gebäudezählung. Die rechtliche Grundlage bildete das ‚Gesetz über eine Zählung der Bevölkerung und der nichtlandwirtschaftlichen Arbeitsstätten und Unternehmen im Jahr 1961 sowie einen Verkehrszensus im Jahre 1962' (Volkszählungsgesetz 1961) vom 13.April 1961 (BGBl. I, S.437). Bei der Beschreibung der Gebäude wurde das Baualter, die Bauart, die Bedachung sowie die sanitäre Ausstattung erfragt, um einen Überblick des qualitativen Zustandes der bewohnten Gebäude zu gewinnen. Die wohnungsstatistischen Ermittlungen erstreckten sich nur auf die Feststellung der Zahl der Wohnungen und „sonstigen Wohngelegenheiten" ohne Untergliederung nach der Wohnungsgröße (Raumanzahl). Zusätzlich wurde die Anzahl der Wohnparteien und Personen ermittelt.

Am 25. Oktober 1968 fand auf Grund des Gesetzes über die Gebäude- und Wohnungszählung 1968 vom 18. März 1968 (BGBl. I, S. 225) eine Gebäude- und Wohnungszählung statt. Ein Vergleich der Ergebnisse mit denen der Gebäude- und Wohnungszählung 1950, der Wohnungsstatistik 1956 und der Gebäudezählung 1961 ist nur bedingt möglich. Bei der Gebäude- und Wohnungszählung 1968 wurden alle Wohngebäude, sonstigen Gebäude, Wochenend- und Ferienhäuser sowie alle ständig bewohnten Unterkünfte erfasst, wobei in diesen Gebäuden und Unterkünften alle Wohnungen und Wohngelegenheiten gezählt wurden. Es lassen sich keine Rückschlüsse auf die Anzahl der bewohnten Nicht-Wohngebäude ziehen.

Die Volks-, Berufs-, Gebäude-, Wohnungs- und Arbeitsstättenzählung wurde nach dem Stand vom 25. Mai 1987 (Zählungsstichtag) flächendeckend durchgeführt. Das Gesetz vom 8. November 1985 (BGBl. I S.2078, Volkszählungsgesetz) bildete die Grundlage der Zählung. Erhebungsobjekt waren Gebäude mit Wohnraum und ständig bewohnten Unterkünften sowie alle dort befindlichen Wohneinheiten. Mit Blick auf die Ergebnisse der Gebäude- und Wohnungszählung von 1950 ist eine deutliche Abnahme der vor 1918 errichteten Gebäude festzustellen. Bei der Interpretation ist aber zu berücksichtigen, dass die Zählungen verschiedene Erhebungsgrundlagen besitzen.

Für Ostdeutschland wurden Zählungen über Gebäude mit Wohnraum in den Jahren 1961, 1971 und 1981 durchgeführt. Des Weiteren erfolgte eine Bestandsaufnahme im Jahr 1995.

3.4.1.3 Gebäude- und Wohnungsbestandsstatistik

Für den Zeitraum zwischen zwei Totalzählungen ermöglicht die Fortschreibung von Bestandsdaten der Wohngebäude und Wohnungen eine jährliche Bereitstellung aktueller Informationen. Diese erfolgt mit Hilfe der jeweils zuletzt durchgeführten Totalzählung und Daten aus dem jährlichen Saldo der Zu- und Abgänge an Gebäuden. Die Fortschreibung wird unter Verwendung der Ergebnisse der Bautätigkeitsstatistik realisiert, und es handelt sich somit um keine eigenständige statistische Erhebung, sondern um eine Ergebnisermittlung aus vorhandenem statistischem Datenmaterial. Die Erhebungsgrundlagen regelt das Gesetz über die Statistik der Bautätigkeit im Hochbau und die Fortschreibung des Wohnungsbestandes (Hochbaustatistikgesetz – HBauStatG).[403]

Die Hochrechnungen aus der so genannten ‚Mikrozensus-Zusatzerhebung zur Wohnsituation'[404] unterstützen zusätzlich die Fortschreibungsstatistik von Wohngebäuden und Wohnungen. Der Gesamtbestand an Nichtwohngebäuden wurde zuletzt im Rahmen der Gebäude- und Wohnungszählung aus dem Jahr 1950 für das westdeutsche Bundesgebiet ermittelt. Eine vergleichbare Fortschreibung des Bestandes von Nichtwohngebäuden ist als veröffentlichte amtliche Statistik nicht weitergeführt worden, und eine spätere Totalzählung der Nichtwohngebäude wurde aus Kostengründen nicht durchgeführt.

Aufgrund der verschiedenen Erhebungsgrundlagen von Gebäude- und Wohnungszählungen (tatsächliche Nutzung) und Bautätigkeitsstatistik (genehmigungspflichtige Erstellung) entstehen bei der Fortschreibung der Bestandszahlen oftmals höhere Werte. Beim fortgeschriebenen Wohnungsbestand entsteht zusätzlich u.a. mit wachsender zeitlicher Entfernung vom Zählungsstichtag eine Abweichung durch Wohnungszusammenlegungen. Es verringert sich zwar die Wohnungsanzahl, aber die Wohnfläche bleibt erhalten. Weiterhin bestehen unterschiedliche Ausgangsdaten für die neuen Länder und Berlin-Ost sowie das westdeutsche Bundesgebiet. Es ist davon auszugehen, dass die Aussagekraft der Bestandsdaten im Wohnungsbau für das westdeutsche Bundesgebiet mit Blick auf den langen Fort-

---

403 Hochbaustatistikgesetz – HBauStatG (5. Mai 1998, BGBl. I, S. 869), geändert durch Artikel 6 des Gesetzes vom 15. Dezember 2001 (BGBl. I, S.3762).

404 Der Mikrozensus basiert aktuell auf der europaweit größten jährlichen Haushaltsbefragung zu den Lebens- und Arbeitsbedingungen sowie der Wohnsituation in Deutschland. Es handelt sich um einen Stichprobenumfang von 820000 Personen bzw. 370000 Haushalten, wobei ca. 160000 Personen auf die neuen Bundesländer und Berlin-Ost entfallen. Das Mikrozensusgesetz 2005 vom 24.06.2004 (BGBl. I, S. 1350) bildet die Rechtsgrundlage. Vgl. DESTATIS [2006]: Qualitätsbericht zum Mikrozensus.

schreibungszeitraum seit der letzten Großzählung am 25.Mai 1987 vermindert wird. Die Daten der neuen Länder und Berlin-Ost beziehen sich für die Jahre bis 1993 auf die Zählung vom 31.12.1981 und ab 1994 auf die Gebäude- und Wohnungszählung vom 30.09.1995. Die Erstellungsmöglichkeit langer Zeitreihen ist wegen der Datenlage für Gebiete der ehemaligen DDR begrenzt und wäre überhaupt nur für den Wohnungsbau und -bestand denkbar.

Gesamtdeutsche Ergebnisse für den Hochbau werden durch die Einführung der Bautätigkeitsstatistik in den neuen Ländern ab 1991 ermöglicht. Die räumliche Vergleichbarkeit ist auf Ebene der Gemeinden durch gebietsstandsbereinigte Daten gegeben. Zum Wohnungsbestand liegen Daten für die Erhebungsmerkmale Anzahl der Wohnungen in Wohn- und Nichtwohngebäuden, deren Struktur, Anzahl der Wohnräume und die Wohnfläche vor. Ab 1994 werden diese Daten erweitert durch Ergebnisse zum Bestand an Wohngebäuden für die Erhebungsmerkmale Anzahl der Wohngebäude, Gebäudestruktur (Ein-, Zwei-, Mehrfamilienhäuser), darin befindliche Wohnungen und die Wohnfläche. Weitere theoretisch denkbare Merkmale sind aus Kostengründen nicht in die Fortschreibung einbezogen.

Eine Untersuchung des Gesamtbestandes von Wohn- und Nichtwohngebäuden ist aufgrund fehlender Fortschreibungsstatistik auf Gemeindeebene nicht möglich. Die Erzeugung von eigenen digitalisierten Primärstatistiken auf Kreis- oder Gemeindeebene ist aus Zeit- und Kostengründen nicht durchführbar. Es müssen andere Datenquellen gefunden werden, um zu dieser Thematik eine Untersuchung mit Data Mining-Verfahren durchzuführen. Die Gebäude werden nach unterschiedlichen Systematiken hinsichtlich der Nutzung strukturiert.[405]

Die in dieser Arbeit eingesetzten Klassifikationsvariablen zum Wohnungs- und Wohngebäudebestand beziehen sich auf gebietsstandsbereinigte Daten. Einbezogen werden die Veränderung der Wohnungsdichte von 1994 bis 2004 (Wohnungen je km²) und die relative Entwicklung des Wohnungsbestandes von 1994 bis 2004. Zusätzlich wird der Anteil der Einfamilienhäuser am Gebäudebestand mit gebietsstandsbereinigten Daten errechnet.

[405] Eine Systematik der Arten von Nichtwohngebäuden ist bei DUBRAL [1986, S. 524 ff.] zu finden. Zusätzlich ist auf die Systematik der Bauwerke zu verweisen, die vom Statistischen Bundesamt erarbeitet wurde. Außerdem ermöglichen der Objektschlüsselkatalog und der Objektabbildungskatalog zum Liegenschaftskataster eine Systematik für die Gebäudenutzung.

3.4.1.4 Siedlungsflächenstatistik

Das Gesetz über Agrarstatistiken[406] bildet die Grundlage für die Flächenerhebung nach Art der tatsächlichen Nutzung und die Bodennutzungshaupterhebung. Die Bodennutzungshaupterhebung erfasst als dezentrale Bundesstatistik die Siedlungsfläche. Erhebungen werden in den westdeutschen Bundesländern hierzu ab 1950 zunächst in wechselnder zeitlicher Abfolge und ab 1974 jährlich durchgeführt. Seit 1980 setzt sich die Siedlungsflächenstatistik in den westdeutschen Bundesländern aus zwei unterschiedlichen, parallel geführten Datenreihen zusammen, da zusätzlich die Flächenerhebung nach Art der tatsächlichen Nutzung alle vier Jahre veröffentlicht wird. Es handelt sich um eine Vollerhebung, die durch eine Auswertung von Daten der amtlichen Liegenschaftskataster der Länder umgesetzt wird. Ab 1992 existiert auch für das ostdeutsche Bundesgebiet diese Statistik. In der DDR basierte die statistische Erfassung der Flächennutzung auf dem Programm der Computergestützten Liegenschaftsdokumentation (COLIDO), das durch einen Ministerbeschluss vom 26.02.1981 gefordert wurde und der wirtschaftlichen Führung des Liegenschaftsbuches diente.

Ausgangspunkt der Bodennutzungshaupterhebung ist die Bodennutzungsaufnahme, die als jährliche Vorerhebung von den Landwirtschaftsämtern durchgeführt wird. Mit Blick auf das Datenmaterial der Bodennutzungshaupterhebung muss darauf verwiesen werden, dass es sich um eine Erfassung nach dem Betriebsprinzip handelt. Die Flächen werden in diesem Fall nach der Zugehörigkeit ihres Eigners (Betriebes) zu einer Gemeinde erhoben, auch wenn diese Flächen außerhalb dieses Gemeindegebietes liegen. Insbesondere bei land- und forstwirtschaftlichen Flächen können methodische Probleme bei einer Auswertung auftreten. Bei Siedlungsflächen ist nach HECKING[407] diese Problematik geringer einzuschätzen, sollte jedoch bei der Interpretation von Ergebnissen bekannt sein.

Die Flächenerhebung wird nach dem Belegenheitsprinzip (nach der tatsächlichen Lage der Flächen in einer Gemeinde) erstellt. Die geschilderten Probleme einer Bodennutzungshaupterhebung wurden damit beseitigt. Die Erhebung erlaubt eine Differenzierung im Bereich der Gebäude- und Freiflächen u.a. nach Öffentlichen Zwecken, Wohnen, Handel und Wirtschaft, Gewerbe und Industrie. Zusätzlich aufgeführt sind Betriebsflächen, Abbauland, Erholungs- und Verkehrsflächen und darunter die Straßen, Wege und Plätze. Einzelne Nutzungsarten der Flächenerhebung werden zur Siedlungs- und Verkehrsfläche zusammengefasst.[408]

---

406 Vgl. Agrarstatistikgesetz – AgrStatG in der Fassung der Bekanntmachung vom 08.August 2002 (BGBl. I, S. 3188), neugefasst durch die Bekanntmachung vom 19. Juli 2006, BGBl. I, S. 1662.

407 Vgl. HECKING et al. [1988, S. 43 ff.]: Bevölkerungsentwicklung und Siedlungsflächenexpansion.

408 Siehe Nebenteil B – Nutzungsarten und Begriffsbestimmung bei der Flächenerhebung.

Zu berücksichtigen ist, dass die Siedlungs- und Verkehrsfläche versiegelte und unversiegelte Flächen enthält. Der Anteil der tatsächlich versiegelten Flächen schwankt in Abhängigkeit der regionalen Lage. Gesicherte Zahlen hierzu liegen derzeit nicht vor. Die Bundesforschungsanstalt für Landeskunde und Raumordnung (BfLR) hat hierzu aus Auswertungen von Versiegelungskartierungen, Länderstudien und eigenen Berechnungen Schätzwerte ermittelt,[409] die der Tabelle 3-11 zu entnehmen sind. Zur Beschreibung der Versiegelungssituation reicht es streng genommen nicht aus, den Versiegelungsgrad alleine auszuweisen. Vielmehr müssten die Versiegelungsart und Eigenschaften der Baumaterialien einfließen, um die verbleibenden Bodenfunktionen einschätzen zu können.[410] Der Forschungsansatz der Städtebaulichen Strukturtypeneinteilung ermöglicht eine genauere Charakterisierung des Siedlungsraumes.[411]

Art der Fläche	**Hochverdichtet**	**Städtisch verdichtet**	**Ländlich**
Gebäude- und Freifläche	60 %	55 %	47,5 %
Verkehrsfläche	70 %	55 %	32,5 %
Erholungsfläche	35 %	30 %	22,5 %
Betriebsfläche	50 %	40 %	25 %

Tabelle 3-11: Schätzwerte der tatsächlich versiegelten Flächen (Quelle BfLR)

Die Flächenerhebung bietet wegen ihrer Erfassungsmethode aus dem Liegenschaftskataster und der damit verbundenen größeren siedlungsstatistischen Differenziertheit eine genauere Quelle für die Ähnlichkeitsuntersuchungen in dieser Arbeit. Die Einsatzmöglichkeiten der statistischen Daten zur Siedlungsfläche (siehe Abschnitt 3.2) sind jedoch je nach administrativem Gebietsstand beschränkt. Feststellbar ist, dass mit zunehmender Aggregation der Daten die Beeinträchtigung der Genauigkeit insbesondere bei räumlichen und zeitlichen Vergleichen an Bedeutung verliert.[412] In Absprache mit dem Fachreferat ‚I5 – Verkehr und Umwelt' am BBR[413] hat der Autor dieser Arbeit deshalb die Untersuchung der Gemeinden auf die Daten des Jahres 2004 beschränkt, da ein Vergleich von verschiedenen Zeitschnitten auch mit gebietsstandsbereinigten Daten als kritisch zu beurteilen wäre. Die Untersuchung von Entwicklungstendenzen kann auf der Kreisebene erfolgen. Zukünftig ist damit zu rechnen, dass durch die ALKIS-Einführung (Abschnitt 3.6.1) eine zeitliche und räumliche Vergleichbarkeit nach ersten denkbaren Umstellungsschwierigkeiten auch auf Gemeindeebene möglich wird.

409 Angaben der Bundesforschungsanstalt für Landeskunde und Raumordnung (BfLR – heute bekannt als BBR)
410 Angaben des Statistischen Landesamtes Baden-Württemberg – Referat 35: Flächenerhebung
411 Vgl. ARLT et al. [2001, S. 46], Ausführungen zur Ermittlung von Versiegelungsgraden
412 Vgl. DESTATIS [2005, S.4]: Qualitätsberichte zur Flächenerhebung.
413 Schwierigkeiten bereiten die Gemeindereformen, die Restfehler der Umstellung von COLIDO auf das in den alten Ländern geltende ALB, die Vollständigkeit der Daten sowie Aktualisierungsdefizite. Als besonders kritisch sind vergleichende Untersuchungen der Betriebs- und Erholungsflächen anzusehen.

3.4.1.5 Kaufwerte von Bauland

Rechtsgrundlage für die Statistik der Kaufwerte von Bauland ist das Gesetz über die Preisstatistik.[414] Der Kauf bzw. Verkauf von unbebauten Grundstücken mit einer Größe von 100 m² und mehr wird durch die Statistik der Kaufwerte für Bauland erfasst soweit diese in den Baugebieten der Gemeinden des Bundesgebietes liegen und Baulandeigenschaft besitzen.

Die amtliche Statistik der Kaufwerte für Bauland wird nach Auskünften der Finanzämter erstellt und unterscheidet zwischen baureifem Land, Rohbauland und sonstigem Bauland.[415] Auch wenn die Statistik als Totalerhebung konzipiert wurde, handelt es sich eher um eine ‚Grundeigentumswechselstatistik', weil nicht alle Kaufwerte für Bauland in der Statistik abgebildet werden, sondern nur die im Erhebungsjahr tatsächlich erfolgten Veräusserungsfälle.[416] Deshalb sind insbesondere kurz- bzw. mittelfristige Datenvergleiche als erschwert zu bewerten, auch wenn eine durchgängige Struktur der Erhebungsmerkmale vorliegt. Aktuell besteht eine weitere Einschränkung, da zur Verbesserung der Statistik die Gutachterausschüsse als zusätzliche Berichtsquelle herangezogen werden und Erfassungsunterschiede denkbar sind. Weiterhin ist keine Differenzierung nach den von den Gemeinden festgesetzten Nutzungsarten bzw. Baugebieten möglich, so dass der räumliche Vergleich ebenfalls nur bedingt möglich ist.[417] Ab 1991 liegen Statistiken für die Kaufwerte für Bauland in ganz Deutschland vor, davor nur Ergebnisse für das frühere Bundesgebiet.

Da der überwiegende Teil der Baulandumsätze mit baureifem Land getätigt wird,[418] sind im Anschluss an die Berechnungen der Ähnlichkeitsuntersuchung bei Bedarf weitere Untersuchungen denkbar. In der Regel beziehen sich die Werte der veröffentlichten Statistik auf Gemeinde- und Kreisebene auf einen Dreijahresdurchschnitt. Die Entwicklung der Kaufwerte je m² Bauland lässt sich mit höher aggregierten Daten durchaus nachverfolgen, und es ergeben sich Hinweise über die langfristige Preisentwicklung für Baulandgrundstücke.

---

414 Vgl. Gesetz über die Preisstatistik vom 29. Mai 1959 (BAnz. Nr. 104 S.1), zuletzt geändert durch Art. 16 des Gesetzes vom 19. Dezember 1997 (BGBl. III 720-9). Die Statistik steht in Verbindung mit dem Gesetz über die Statistik für Bundeszwecke (BstatG) vom 22. Januar 1987 (BGBl. I S.462, 565), Grunderwerbssteuergesetz (GrEStG), Baugesetzbuch (BauGB) und der Wertermittlungsverordnung. Vgl. DESTATIS [2005, S.4]: Qualitätsberichte zur Statistik für Kaufwerte von Bauland und Gesetzestext.

415 „Zum sonstigen Bauland gehören Industrieland (unbebaute Grundstücke, die Gewerbezwecken dienen oder dienen sollen), Land für Verkehrszwecke (als Straßen, Parkplätze, Flugplätze, für Eisenbahnen u. ä. genutzte oder vorgesehene Grundstücke) und Freiflächen (als Gartenanlagen, Spielplätze, Erholungsplätze dem öffentlichen Gebrauch dienende oder als solche ausgewiesene Flächen)" EPPING [1977, S. 79].

416 Vgl. DESTATIS [2005]: Qualitätsbericht zur Statistik für Kaufwerte von Bauland.

417 Nach der Baunutzungsverordnung (BauNVO) werden Wohnbauflächen, gemischte Bauflächen, gewerbliche Bauflächen und Sonderbauflächen unterschieden, die jeweils weiter untergliedert sind. Bei HEUER [1979, S. 377] ist eine vollständige Übersicht zu finden. Siehe zusätzlich Nebenteil B.

418 Die Umsätze mit Rohbauland und sonstigem Bauland sind in vielen Gemeinden so gering, dass die daraus berechneten durchschnittlichen Quadratmeterpreise nicht repräsentativ sind.

3.4.1.6 Bevölkerungsstatistik

Die Ergebnisse der jeweils letzten Volkszählung[419] bilden die Basis für die Fortschreibung der Bevölkerungszahlen, die Ergebnisse der Statistik der natürlichen Bevölkerungsbewegung (Geburten- und Sterberate) und der Wanderungsstatistik (Zu- und Fortzüge, Wechsel der Hauptwohnung sowie Änderung der Staatsangehörigkeit). Als Rechtsgrundlage ist das Gesetz über die Statistik der Bevölkerungsbewegung und die Fortschreibung des Bevölkerungsstandes[420] zu nennen. Der Mikrozensus[421] ergänzt die Fortschreibung der Bevölkerungsentwicklung und der räumlichen Bevölkerungsbewegung. Durch den relativ langen Fortschreibungszeitraum seit 1987 ist die Zuverlässigkeit der statistischen Daten mit gewissen Genauigkeitsverlusten behaftet, auf die im Rahmen dieser Arbeit nur verwiesen sein soll.[422]

Die Erhebungsunterlagen für die Wanderungsstatistik sind die An- und Abmeldungsscheine, die gemäß den Landesvorschriften bei einem Wohnungswechsel in den Einwohnermeldeämtern geführt werden. Die Bevölkerungszahl trägt sowohl direkt bei der Variablenbildung im Rahmen der Ähnlichkeitsuntersuchung eine Bedeutung als auch indirekt als Datenquelle für die Berechnung der Bevölkerungspotentiale des BBR. Zusätzlich werden die Zu- und Fortzüge der Bevölkerung als Variable erfasst.

3.4.1.7 Beschäftigtenstatistik

In der Beschäftigtenstatistik sind die sozialversicherungspflichtig Beschäftigten, also in der Regel alle Arbeiter/innen und Angestellten (einschl. Auszubildenden) mit Ausnahme der meisten Selbstständigen und Beamten erfasst. Es handelt sich um über 75 % von allen Erwerbstätigen.[423] Ausgangspunkt dieser Statistik ist das am 01. Januar 1973 eingeführte ‚integrierte Meldeverfahren' zur Sozialversicherung (gesetzliche Kranken- und Rentenversicherung), wobei die aktuelle gesetzliche Grundlage für die Aufbereitung der Statistik die Meldungen nach §28a des Vierten Buches Sozialgesetzbuch – Sozialversicherung – SGB IV vom 23. Dezember 1976 bilden.[424]

---

419 Die Volkszählung in der Bundesrepublik Deutschland wurde am 25. Mai 1987 zuletzt durchgeführt und hat umgerechnet rund 500 Millionen Euro gekostet. Obwohl 1990 ca. 16 Millionen weitere Bürger mit anderen infrastrukturellen Voraussetzungen zum Staatsgebiet hinzukamen, wurden keine weiteren Daten ermittelt. In Ostdeutschland beziehen sich die Daten deshalb auf das Jahr 1981.

420 Vgl. Gesetz über die Statistik der Bevölkerungsbewegung und die Fortschreibung des Bevölkerungsstandes vom 04.07.1957 (BGBl. I, S.694) in der Fassung der Bekanntmachung vom 14.03.1980 (BGBl. I, S.308), zuletzt geändert durch Artikel 2 des Gesetzes zur Änderung des Melderechtsrahmengesetzes (MRRG) vom 25.03.2002 (BGBl. I, S.1191).

421 Vgl. DESTATIS [2006]: Qualitätsbericht zum Mikrozensus. Das Mikrozensusgesetz 2005 vom 24.06.2004 (BGBl. I, S. 1350) bildet die Rechtsgrundlage.

422 Statistisches Bundesamt: Statistisches Jahrbuch 2001 der Bundesrepublik Deutschland

423 Vgl. DESTATIS [2005, S. 5]: Qualitätsbericht zur Beschäftigtenstatistik.

424 Vgl. Viertes Buch Sozialgesetzbuch – Sozialversicherung – SGB IV,23. Dezember 1976 (BGBL. I, S. 3845)

Die Bundesagentur für Arbeit (BA) stellt nach § 282 a Abs. 1 Sozialgesetzbuch III (24.03.1997: BGBl. I, S. 594) dem Statistischen Bundesamt anonymisiertes Datenmaterial zu sozialversicherungspflichtig Beschäftigten für Auswertungen und zur Weitergabe an die Statistischen Ämter der Länder zur Verfügung. Seit dem 30.06.1974 liegen die Daten für das frühere Bundesgebiet und ab dem 31.03.1992 für die neuen Bundesländer und Berlin Ost vor, wobei in den Anfangsjahren einige Einschränkungen bestehen.

Die Statistik ermöglicht heute Angaben zu folgenden Merkmalen: Geschlecht, Staatsangehörigkeit, allgemeiner und beruflicher Ausbildungsabschluss, ausgeübte Tätigkeit, Auszubildende, Stellung im Betrieb als Facharbeiter, Meister oder Polier und andere Vollzeitbeschäftigte, Voll-/Teilzeitbeschäftigung, Wirtschaftszweig des Betriebes, Rentenversicherungsträger als Arbeiter, Angestellte bis 2004 sowie Arbeits- und Wohnort. Aus den Angaben zum Arbeits- und Wohnort werden die Ein- und Auspendlerstatistiken erstellt.[425]

Durch die Statistik entsteht ein repräsentatives Bild der räumlichen und strukturellen Verteilung von Arbeitsplätzen und deren Entwicklung. Es ist jedoch auf einige Zeitreihenbrüche hinzuweisen, da in den letzten Jahren einige Aktualisierungen bei der Klassifikation der Wirtschaftszweige stattgefunden haben. Der Nachweis der sozialversicherungspflichtig Beschäftigten unterliegt seit dem 31.03.1998 der europaweit eingeführten ‚Klassifikation der Wirtschaftszweige' (WZ 93), so dass ein Vergleich nach wirtschaftlichen Zweigen mit früheren Zeitreihen keinesfalls erfolgen sollte. Ab dem Stichtag 30.06.2003 gilt die neue WZ 2003. Des Weiteren wird der zeitliche Vergleich durch die eingeführte DEÜV (Datenerfassungs- und -übermittlungsverordnung)[426] erschwert, da es sich dabei um ein neues Meldeverfahren handelt und eine derartige Umstellung sich auf die direkten Bezüge zu früheren Stichtagen stark auswirkt. Die räumliche Vergleichbarkeit ist wie bei den meisten Statistiken durch die Gebietsstandsänderungen nur teilweise gegeben, da der inländische Arbeits- und Wohnort bei der Beschäftigtenstatistik nach dem gültigen amtlichen Gemeindeschlüssel erfasst wird. Für diese Arbeit existieren auf Ebene der Gemeinden gebietsstandsbereinigte Daten für die Jahre 1997 bis 2003, wobei unter Beachtung der oben genannten Probleme der Aufbau diesbezüglicher dynamischer Variablen nur für den Zeitraum 1999 bis 2003 sinnvoll ist. Grundsätzlich bilden die Zahlen der sozialversicherungspflichtig Beschäftigten an Stelle nicht verfügbarer aktueller Daten der Arbeitsstättenzählung eine wichtige Grundlage zur Bewertung des Arbeitsmarktes und der Bedeutung einzelner Gemeinden als Arbeitsstandort.

---

425 Siehe Nebenteil B – Informationen zur Begriffsbestimmung.

426 Vgl. Gesetzestext zur DEÜV-Verordnung, gültig seit 01.01.1999 und neugefasst durch die Bekanntmachung vom 23.01.2006, (BGBl I, S. 152).

3.4.2 Querschnittsveröffentlichungen[427]

3.4.2.1 Statistik regional[428]

Die so genannte ‚Statistik regional' ist eine von den Statistischen Ämtern des Bundes und der Länder gemeinsam veröffentlichte Datenbank. Es werden Basisdaten aus der amtlichen Statistik[429] zu hauptsächlich wirtschaftlichen und soziodemografischen Fragen jährlich auf Ebene der kreisfreien Städte und Kreise in digitaler Form veröffentlicht. Es werden ca. 1100 Variablenausprägungen bereitgestellt. Zu einigen ausgewählten Statistiken (z.B. Bevölkerung, Bautätigkeit, Öffentliche Finanzen) werden zusätzlich längere Zeitreihen aufbereitet, die auf Grund der Gebietsstandsbereinigung einen multitemporalen Datenvergleich ermöglichen.

3.4.2.2 Statistik lokal[430]

Von den Statistischen Ämtern des Bundes und der Länder wird gemeinsam eine Datenbank unter dem Namen ‚Statistik lokal' herausgegeben. Diese Datenquelle bezieht sich auf alle Gemeinden in ganz Deutschland und enthält ebenfalls Daten der veröffentlichten Basisdaten der amtlichen Statistik. Der Umfang des Statistikangebotes ist aber etwas reduzierter als bei ‚Statistik Regional'. Es handelt sich um rund 300 Variablenausprägungen, die analysiert und verglichen werden können. Es existieren Daten für einen festen Berichtszeitpunkt.

3.4.2.3 GENESIS-Online[431]

Als Auskunftsdatenbank bietet ‚GENESIS-Online' die Möglichkeit, das Datenangebot der üblichen amtlichen Statistik über das Internet zu nutzen. Zu unterscheiden ist zwischen ‚GENESIS-Online' und ‚GENESIS-Online Regional', wobei einerseits Daten auf Bundesebene und andererseits Daten auf Ebene der kreisfreien Städte und Kreise abrufbar sind. Es ist ein kostenfreies und ein kostenpflichtiges Datenangebot vorhanden. Insgesamt werden mehr als 100 Variablenausprägungen der amtlichen Statistik bereitgestellt.

---

427 Die Querschnittsveröffentlichungen enthalten Daten aus wichtigen Bereichen der amtlichen Statistik und ermöglichen einen schnellen selbstständig durchgeführten Datenabruf infolge vorbereiteter Datenbankstrukturen.

428 Datenquelle dieser Arbeit: Ausgabe 2005, Berichtszeitpunkt: 31.12.2003, Zeitreihen ab 31.12.1995

429 Wenn in dieser Arbeit von Basisdaten der amtlichen Statistiken gesprochen wird, so sind damit die seit geraumer Zeit veröffentlichten Statistikangebote des Bundes und der Länder gemeint. Zusammenfassend sind hier die Hauptkategorien kurz aufgeführt, die sich auf die regional tiefer gegliederten Ebenen beziehen: Gebiet, Bevölkerung, Gesundheit, Bildung, Beschäftigung, Gewerbemeldungen, Landwirtschaft, Produzierendes Gewerbe, Gebäude und Wohnungen, Bautätigkeit, Baulandverkäufe, Tourismus, Verkehr, Sozialwesen, Wohngeld, Öffentliche Finanzen, Volkswirtschaftliche Gesamtrechnungen, Umwelt, Wahlen.

430 Datenquelle dieser Arbeit: Ausgabe 2005, Berichtszeitpunkt: 31.12.2003, keine Zeitreihen

431 Es wurden Daten mit Hilfe von ‚GENESIS-Online Regional' vereinzelt beschafft. Da für diese Arbeit im Wesentlichen Daten aus dem veröffentlichten Datenangebot der ‚Statistik Regional' verwendet wurden, bilden diese Daten nur eine weitere Ergänzung. (Gemeinsames neues statistisches Informationssystem):
http://www-genesis.destatis.de/genesis/online/logon (10.09.2006),
http://www.regionalstatistik.de/genesis/online/logon (10.09.2006).

3.4.2.4 INKAR Raumbeobachtung[432]

Das Bundesamt für Bauwesen und Raumordnung bietet Indikatoren und Karten zur Raumentwicklung an, die im Wesentlichen auf Daten der amtlichen Statistik beruhen. Es handelt sich um ein Datenangebot, welches Informationen zu den Standort- und Lebensbedingungen in Deutschland für unterschiedliche Raumbezugsebenen (Länder, Kreise und Gemeinden bzw. Gemeindeverbände) enthält. Es werden zusätzlich für Deutschland auch Daten für den Europäischen Vergleich auf NUTS-Ebene[433] 0, 1 und 2 zur Verfügung gestellt. Insgesamt handelt es sich um ca. 800 Variablenausprägungen zu 23 zentralen Themenbereichen.[434] Viele der Variablen sind als Zeitreihe mit Werten für Einzeljahre abrufbar. Es handelt sich um gebietsstandsbereinigte Daten, so dass Vergleiche über einen längeren Zeitraum möglich sind.

3.5 Landesentwicklungs- und Regionalpläne in Deutschland[435]

Generell dient die Regionalplanung unterhalb der staatlichen Raumordnung einer Konkretisierung der fachlichen Integration und Umsetzung landesplanerischer Ziele. Mit Bezug zum Raumordnungsgesetz § 1 (BGBl. 1997 I S. 2102) gilt das Zentrale-Orte-Konzept als eines der zentralen Instrumente der Raumordnung zur Erreichung einer nachhaltigen Entwicklung. BLOTEVOGEL[436] verweist darauf, dass die räumliche Ausrichtung der Siedlungsentwicklung auf zentrale Orte als Voraussetzung für den sozial gerechten und ökonomischen Einsatz von Ressourcen sowie einer Begrenzung des Verbrauches von Naturgütern gilt.

Die Kenntnis über die Attributierung der Gemeinden nach den in den Landesentwicklungs- und Regionalplänen festgelegten Zentrale-Orte-Kategorien ermöglicht in dieser Arbeit eine Bewertung der damit verbundenen Ausweisungspraxis und eine vertiefende Betrachtung von gefundenen Ähnlichkeitsmustern im empirischen Teil dieser Arbeit.

---

[432] Für diese Forschungsarbeit wurden Daten der aktuellen Datenbank ‚INKAR' eingesetzt (Ausgabe 2005, Berichtszeitpunkt: 31.12.2003, Zeitreihen jährlich, beginnend mit 1995).

[433] Die Nomenklatur der Territorialen Einheiten für Statistik ist ein Klassifikationsgefüge, welches ein einheitliches territoriales System für die Statistikbereitstellung ermöglicht.

[434] Die zentralen Themenbereiche sind: Arbeitslosigkeit, Bauen und Wohnen (Baulandmarkt und Bautätigkeit, Gebäude- und Wohnbestand), Beschäftigung und Erwerbstätigkeit (Struktur, Sektoren u. Berufsbereiche), Bevölkerung (Altersstruktur, Bevölkerungsprognose, Bevölkerungsstruktur, Mobilität, Natürliche Bevölkerungsbewegungen), Bildung (Ausbildungsangebot, Schulische Bildung), Einkommen und Verdienst, Flächennutzung, Medizinische Versorgung und Infrastruktur, Öffentliche Haushalte, Raumwirksame Mittel, Siedlungsstruktur, Sozialstruktur, Verkehr und Erreichbarkeit, Wirtschaft (Landwirtschaft, Wirtschaftliche Leistung, Fremdenverkehr).

[435] Die Landesentwicklungs- und Regionalpläne enthalten Informationen zur Attributierung der Gemeinden nach Zentrale-Orte-Kategorien. Die Daten dieser Arbeit wurden einerseits dem Autor als Datei vom BBR bis zum Zentralitätsgrad 40 zur Verfügung gestellt und andererseits eigenständig für die Unter- und Kleinzentren in mühsamer Einzelarbeit nacherfasst. Als Quelle diente dazu das Planmaterial der dafür zuständigen regionalen Planungsstellen.

[436] BLOTEVOGEL, H. [2002, S. 19-40]

Tabelle 3-12 enthält die gültige Klassifikation von zentralen Orten in Deutschland und ermöglicht ein Verständnis der einzelnen Klassen.

Zentralitätsgrad (Typ)		Begriffsdefinitionen der Länder nach Programmen, Plänen usw.
10 (11)	◉	Oberzentrum (Teil eines Oberzentrums), Oberzentraler Städteverbund, mit Teilfunktionen eines Oberzentrums
20 (21)		Mögliches Oberzentrum (Teil eines möglichen Oberzentrums), zum Oberzentrum auszubauen, Mittelzentrum mit Teilfunktion[en] eines Oberzentrums bzw. mit oberzentralen Teilfunktionen
30 (31)	◉	Mittelzentrum (Teil eines Mittelzentrums), Mittelzentraler Städteverbund, Mittelzentrum im Grundnetz bzw. im Ergänzungsnetz oder im Verdichtungsraum, Mittelzentrum in Funktionsergänzung bzw. in gegenseitiger Funktionsergänzung
40 (41)		Mögliches Mittelzentrum (Teil eines möglichen Mittelzentrums), Unter-/Grundzentrum mit Teilfunktion[en] eines Mittelzentrums, Stadtrandkern I. Ordnung mit Teilfunktionen eines Mittelzentrums, Mittelzentrum mit Teilfunktion, teilfunktionales Mittelzentrum mit gegenseitiger Funktionsergänzung
50 (51)	⊙	Unterzentrum (Teil eines Unterzentrums), Grundzentrum, Stadtrandkern 1. Ordnung, Siedlungsschwerpunkt in großen Verdichtungsräumen
60 (61)		Mögliches Unter-/Grundzentrum (Teil eines möglichen Unter-/Grundzentrums)
70 (71)	○	Kleinzentrum (Teil eines Kleinzentrums), Ländlicher Zentralort, Stadtrandkern II. Ordnung
90	•	Gemeinden [bisher] ohne zentralörtliche Funktion

Tabelle 3-12: Attributierung der Zentrale-Orte-Kategorien[437]

Die Aufstellung und Umsetzung der Regionalpläne obliegt den einzelnen Bundesländern, so dass sich dadurch ggf. eine Unterschiedlichkeit in der Ausweisungspraxis der Zentralen-Orte-Kategorien ergibt. SIEDENTOP[438] stellt fest, dass mit Einschränkungen für vergleichende Aussagen unter dem Aspekt der Zentrale-Orte-Kategorien zu rechnen ist. Der statistische Vergleich von Untersuchungsobjekten und speziell eine sich daran anschließende Interpretation bezüglich der Zentrale-Orte-Kategorien hat die räumlichen Verteilungsmuster zu prüfen und einzubeziehen. Abbildung 3-6 zeigt ein Beispiel für unterschiedliche räumliche Verteilungsmuster der Zentralen-Orte-Kategorien in Deutschland.

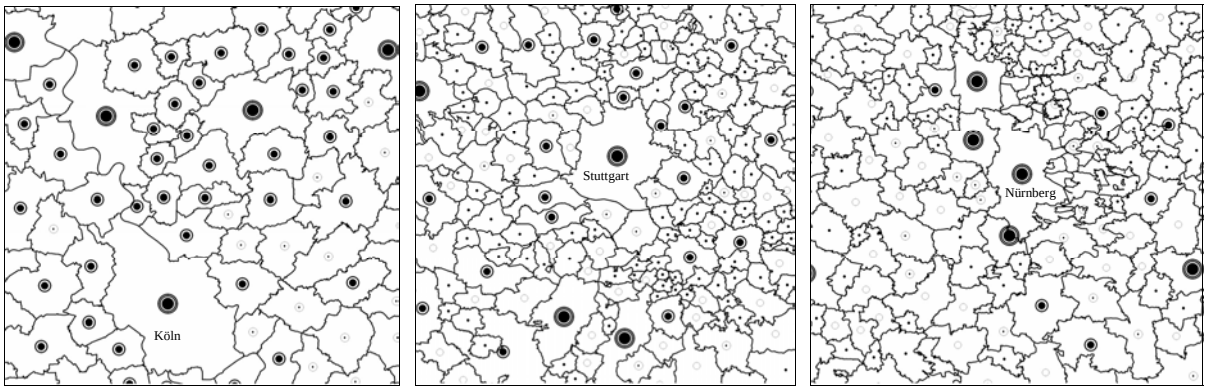


Abbildung 3-6: Ausgewählte Beispiele für Verteilungsmuster von Zentrale-Orte-Kategorien[439]

437 Eigene Bearbeitung nach SIEDENTOP et al. [2003, S.15]
438 SIEDENTOP et al. [2003, S.16]
439 Eigene Bearbeitung unter Verwendung der verfügbaren Daten des BBR und eigener Nacherhebung.

3.6 Vermessungsverwaltung in Deutschland

In Deutschland sind die Vermessungs- und Katasterangelegenheiten Ländersache, so dass es darüber hinaus keine einheitliche gesetzliche Regelung auf Bundesebene gibt. Die Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV) wurde jedoch dazu gegründet, um eine Einheitlichkeit im Vermessungswesen herzustellen und das amtliche deutsche Vermessungswesen im Ausland zu vertreten. Die Aufgaben des amtlichen Vermessungswesens werden von den Katasterbehörden innerhalb der Kreise und kreisfreien Städte sowie von den Landesvermessungsämtern und der Bezirksregierung wahrgenommen.

Die Daten des amtlichen Vermessungswesens werden als Geobasisdaten bezeichnet, die in einem Geobasisinformationssystem geführt werden und auch historisch gewordene Daten aufrechterhalten. Der einheitliche geodätische Raumbezug und die Geobasisdaten bilden die Grundlage für alle raum- und bodenbezogenen Informationssysteme, Planungen und Maßnahmen der Landesverwaltung und der Kommunen. Die Daten können auch von öffentlichen und privaten Stellen genutzt werden. Das amtliche Vermessungswesen umfasst als öffentliche Aufgabe die Erhebung, Führung und Bereitstellung der Daten der Landesvermessung und des Liegenschaftskatasters. Es stellt den einheitlichen geodätischen Raumbezug her und erhebt dazu Festpunktdaten, unterhält Positionierungsdienste sowie Daten aller Liegenschaften.

Um eine einheitliche Führung des Liegenschaftskatasters zu gewährleisten, unterstützen die Landesvermessungsämter und die Bezirksregierung die Erstellung, Pflege und Weiterentwicklung von Programmsystemen für automatisierte Verfahren und Erneuerungsarbeiten einer Katasterbehörde, welche eine überörtliche Bedeutung haben bzw. deren Leistungskraft übersteigen. Die Fortführung des Liegenschaftskatasters erfolgt durch Liegenschaftsvermessungen, die von öffentlich bestellten Vermessungsingenieuren und den Kataster- und Vermessungsverwaltungen selbst durchgeführt werden.

Als Bestandteile des Liegenschaftskatasters gelten die Liegenschaftskarte (Katasterkarte, Flurkarte) und das Liegenschaftsbuch. Die Funktion des Liegenschaftskatasters hat sich im Laufe der Jahre stark gewandelt. Anfang des 19. Jahrhunderts diente es als reines Steuerkataster, während es heute als Mehrzweckkataster verschiedene Anforderungen und Funktionen erfüllt. Es beinhaltet nach wie vor die Ergebnisse der amtlichen Bodenschätzung (Steuerkataster), dient als amtliches Verzeichnis der Grundstücke im Sinne des §2 Abs. 2 der Grundbuchordnung (Eigentumskataster) und ist Basisfunktion für die Bereiche Wirtschaft, Verwaltung, Naturschutz, Umweltschutz, Landesplanung, Bauleitplanung und Bodenordnung.

Im Liegenschaftskataster der jeweiligen Länder sind für die dazugehörigen Landesgebiete alle Liegenschaften (Flurstücke,[440] Gebäude[441]) enthalten und gemäß Tabelle 3-13 beschrieben.

Hauptelement	Zusätzliche Beschreibung
Liegenschaftsangaben	Geometrische Form, Lage und Größe der Liegenschaften einschließlich der bestimmenden Koordinaten, den Angaben zu Flurstücksnummern, Straßennamen, Hausnummern und Lagebezeichnungen
Eigentümerangaben	Name und Geburtsdaten der Eigentümer und Erbbauberechtigten in Übereinstimmung mit dem Grundbuch einschließlich bekannter Anschriften, Anteilsverhältnisse, Verwalterangaben und Grundbuchbezeichnungen
Nutzungsangaben	Definition der Nutzung nach amtlichen Objektschlüsselkatalogen
Charakteristische Topographie	-
Öffentlich-Rechtliche Festlegungen	Auf den Grund und Boden bezogene Bewertungen, Rechte, Beschränkungen, Belastungen oder andere Festlegungen
Bodenschätzungsergebnisse	-

Tabelle 3-13: Aspekte zur Beschreibung und Darstellung der Liegenschaften

Die digitale Führung der Liegenschaftsinformationen in einem automatisierten Liegenschaftsbuch (ALB) und einer automatisierten Liegenschaftskarte (ALK) bietet gegenüber der klassischen analogen Führung von Liegenschaftsinhalten zusätzliche Möglichkeiten, insbesondere die Auswertung und Verknüpfung mit Daten anderer Themen und Fachbereiche (z.B. Hauskoordinaten und deutsche Grundkarte M 1:5000). Die Verfahrenslösungen ALK und ALB zur Führung der Liegenschaftsinformationen basieren jedoch noch auf Standards der elektronischen Datenverarbeitung aus den 70er- und 80er-Jahren, so dass eine äußerst begrenzte Leistungsfähigkeit anzumerken ist. Aufgrund der getrennt entwickelten und betriebenen Systeme ALK und ALB und der daraus resultierenden notwendigen doppelten Erfassung und Fortführung von Daten besteht gerade die Gefahr von Dateninkonsistenz. Zukünftig werden nach Beschluss der AdV die bisherigen Komponenten ALK und ALB in einem System, dem amtlichen Liegenschaftskataster-Informationssystem (ALKIS®) integriert und der Aufbau eines vollwertigen Geoinformationssystems verfolgt. ALKIS® ist in ein so genanntes ‚AAA-Referenzmodell'[442] eingebettet, dass internationale Normen (ISO/TC 211, OGC) und eine einheitliche Modellierung berücksichtigt.

---

440 Ein Flurstück (Katastergrundstück, Parzelle) ist ein begrenzter Teil der Erdoberfläche, der im Liegenschaftskataster unter einer Bezeichnung geführt und auf Antrag bzw. direkt durch Amtsverwaltungen gebildet wird.

441 Als Gebäude werden dauerhafte, selbstständig benutzbare, überdeckte bauliche Anlagen defniert, die wegen ihrer Bedeutung im Liegenschaftskataster nachzuweisen sind. Die Gebäude sind geeignet, dem Schutz von Menschen, Tieren, Sachen oder der Produktion von Wirtschaftsgütern zu dienen. Wann ein Gebäude im Liegenschaftskataster nachzuweisen ist, regeln die Vermessungs- und Katastergesetze der Länder.

442 AAA-Konzept (AFIS®-ALKIS®-ATKIS®): ATKIS® = Amtliches-Topographisches-Karteninformationssystem, AFIS® = Amtliches Festpunkt-Informationssystem. Die Länder BW, HH, NI, RP und SH sind maßgeblich für den Pilotbetrieb verantwortlich. Vgl. http://www.alkis.info/ (10.10.2006)

3.6.1 Automatisierte Liegenschaftskarte[443]

Das Vorhaben ‚Automatisierung der Liegenschaftskarte' wurde zur automatisierten Führung des Zahlen- und Kartennachweises des Liegenschaftskatasters entwickelt. Das Gesamtsystem setzt sich aus einem ALK-Verarbeitungsteil und einem ALK-Datenbankteil zusammen.

Der ALK-Datenbankteil erfüllt im Wesentlichen die Aufgabe, so genannte Primärdateien (Grundrissdatei, Punktdatei, Datei der Messungselemente) fortzuführen als auch zu benutzen und ermöglicht eine ALK-Auftragsverwaltung (Überwachung und Verwaltung der Aufträge des Auftragsbuches). Es werden die Daten des Verarbeitungsteils entgegengenommen, entsprechende Dateien eingerichtet, fortgeführt und benutzt sowie Ergebnisse dem Verarbeitungsteil wieder zur Verfügung gestellt (Ergebnisdaten).

Im ALK-Verarbeitungsteil erfolgt die Antragsbearbeitung und graphische Verarbeitung. Auf den Datenbankteil kann dabei zentral oder auch dezentral durch einen oder mehrere Verarbeitungsteile zugegriffen werden. Bei der Nutzung der ALK als Basis für grundstücksbezogene Informationssysteme wird vorausgesetzt, dass alle Anwender und Nutzer den digitalen Kartennachweis lesen und weiterverarbeiten können. Als Verarbeitungsschnittstelle wurde zwischen dem Datenbankteil und dem Verarbeitungsteil die ‚Einheitliche Datenbankschnittstelle' (EDBS)[444] festgelegt. Die Anwender sind in der Lage, die Geometrie von Flurstücken, Straßenverläufen und Bebauung durch weitere Attribute oder Geometrieinformationen zu ergänzen. Als Bezugssystem wird das World Geodetic System von 1984 (WGS'84) festgelegt und die Koordinaten sind im Gauß-Krüger-Meridianstreifen definiert.

Die Automatisierung der Liegenschaftskarte berücksichtigt eine elementargeometrische Sicht (punktförmige, linienförmige, flächenförmige, textförmige Einheiten) und eine Objektsicht (fachspezifische Menge von elementargeometrischen Einheiten). Als Objekte geführt werden beispielsweise ein Flurstück, ein Gebäude oder eine Nutzungsfläche. Um die Objekte zu bilden, werden spezielle Objektverschlüsselungen der Einzelinformationen für jeden Anwender- bzw. Funktionsbereich definiert. Die Verschlüsselung berücksichtigt Informationen aus der Liegenschaftskarte und dem Liegenschaftsbuch (ALB). Die Informationen aus dem Liegenschaftskataster und der topographischen Kartographie werden verknüpft. Der Grundrissnachweis des Liegenschaftskatasters ist so aufgebaut, dass eine integrationsfähige Datenbasis von Informationssystemen geschaffen wird. Schließlich wird der Datenaustausch mit Informationssystemen sonstiger Nutzer des Liegenschaftskatasters vorgehalten.

---

[443] Die Automatisierte Liegenschaftskarte wird durch ‚Automatic Real Estate Map' ins Englische übersetzt.
[444] Die Übergabe der Daten wird durch einen 7-Bit-Code nach DIN 66003 umgesetzt.

Grundsätzlich ist bei der Verschlüsselung zwischen dreistelligen Folienschlüsseln und vierstelligen Objektschlüsseln zu unterscheiden. Die von der AdV als verbindlich geltenden Objektschlüsselkataloge (OSKA) beschreiben in einheitlicher Form die Grundrissinformationen des Liegenschaftskatasters. Die Anwender sind zusätzlich in der Lage, Sonderschlüsselkataloge (Sonder-OSKA) zu definieren und diese bestimmten Folien zuzuordnen. Die Folien sind als fachliche Gruppen zu verstehen, denen die Objektschlüssel zugeordnet werden. Gleiche Objektschlüssel können deshalb in mehreren Folien existieren. Die Objektabbildungskataloge (OBAK) beschreiben jedes Objekt mit seiner Geometrie und seinen fachlichen Funktionen (Attributen). Tabelle 3-14 gibt einen Überblick der Schlüsselbereiche.

Schlüssel	Folienanzahl	Funktionsbereich
000-099	40	Liegenschaftskataster
001	1	Flurstücke (OS z.B. 0233=Flurstück / -sgrenze; 0239=Flurstück / -sgrenze im Verkehrsweg)
002	1	Flur- und Gemarkungsgrenzen (OS z.B. 0231=Gemarkungsgrenze, 0232=Flur / -grenze)
003	1	Verwaltungsgrenzen (OS z.B. 0214=Kreisgrenze, 0215=Gemeindegrenze)
011	1	Gebäude (OS z.B. 1113=Postamt, 1011=Hochhausbegrenzungslinie)
021	1	Nutzungsarten (OS z.B. 0241=Nutzungsartengrenze, 4000=Erholungsfläche)
081	1	Verkehrsflächen (OS z.B. 5111=Autobahn, 5241=Radweg)
100-199	22	Landesvermessung
200-699	100	Kommunaler Funktionsbereich
700-998	30	Funktionsbereich anderer Stellen
999		Folienübergreifende Objektteile bzw. Kartenzeichen

Tabelle 3-14: Verschlüsselung der Grundrissinformationen der ALK[445]

Mittlerweile ist die ALK in den meisten Bundesländern eingeführt. Einen Sonderfall stellt das Bundesland Bayern dar, in dem die Aufgaben der ALK durch die digitale Flurkarte (DGF) übernommen werden. Der Bearbeitungsstand der ALK variiert sehr stark zwischen den einzelnen Bundesländern und das beeinflusst erheblich die Datenbeschaffung und erschwert aktuell noch die Ähnlichkeitsuntersuchungen. In Bezug auf die Gesamtfläche der BRD lag im Jahr 2005 eine digitale Liegenschaftskarte zu 86 % vor.[446] Jedoch existiert nicht in allen Bundesländern ein objektstrukturierter Nachweis auf Basis des AdV-Standards.

Die Daten dieser Arbeit beziehen sich auf die Folie 11 (Gebäude). Da die Verschlüsselungssystematik einige Freiräume für länderspezifische Anwendungen bietet, sind die Angaben über den Gesamtbestand und die Nutzungsarten (Nebenteil B) genau zu prüfen.[447]

---

445 Eigene Bearbeitung auf Basis von Informationen der AdV

446 Vgl. STÖPPLER (LVERMA-NRW), unveröffentlichte Präsentation zum Stand der ALK in Deutschland.

447 Will man eine Erfassungsgenauigkeit von 100 % voraussetzen, so kann dies nach Auskunft der Vermessungsverwaltung nur über eine Auswertung von Satellitenaufnahmen garantiert werden. Das Liegenschaftskataster bildet aber eine sehr gute Grundlage, um den deutschen Gebäudebestand genauer zu analysieren. Die Gebäudenutzungsarten können unter Kenntnis der länderspezifischen Objektschlüsselkataloge (AdV-Standard) und den damit verbundenen Eigenarten nahezu miteinander verglichen werden.

Abbildung 3-7 definiert die Bundesländer, welche in dieser Arbeit eine Datenbereitstellung ermöglichten. Gezeigt werden Teilbestände des deutschen Gebäudebestandes und die Nutzungsarten in aggregierten Darstellungen auf der Kreis- und Gemeindeebene.

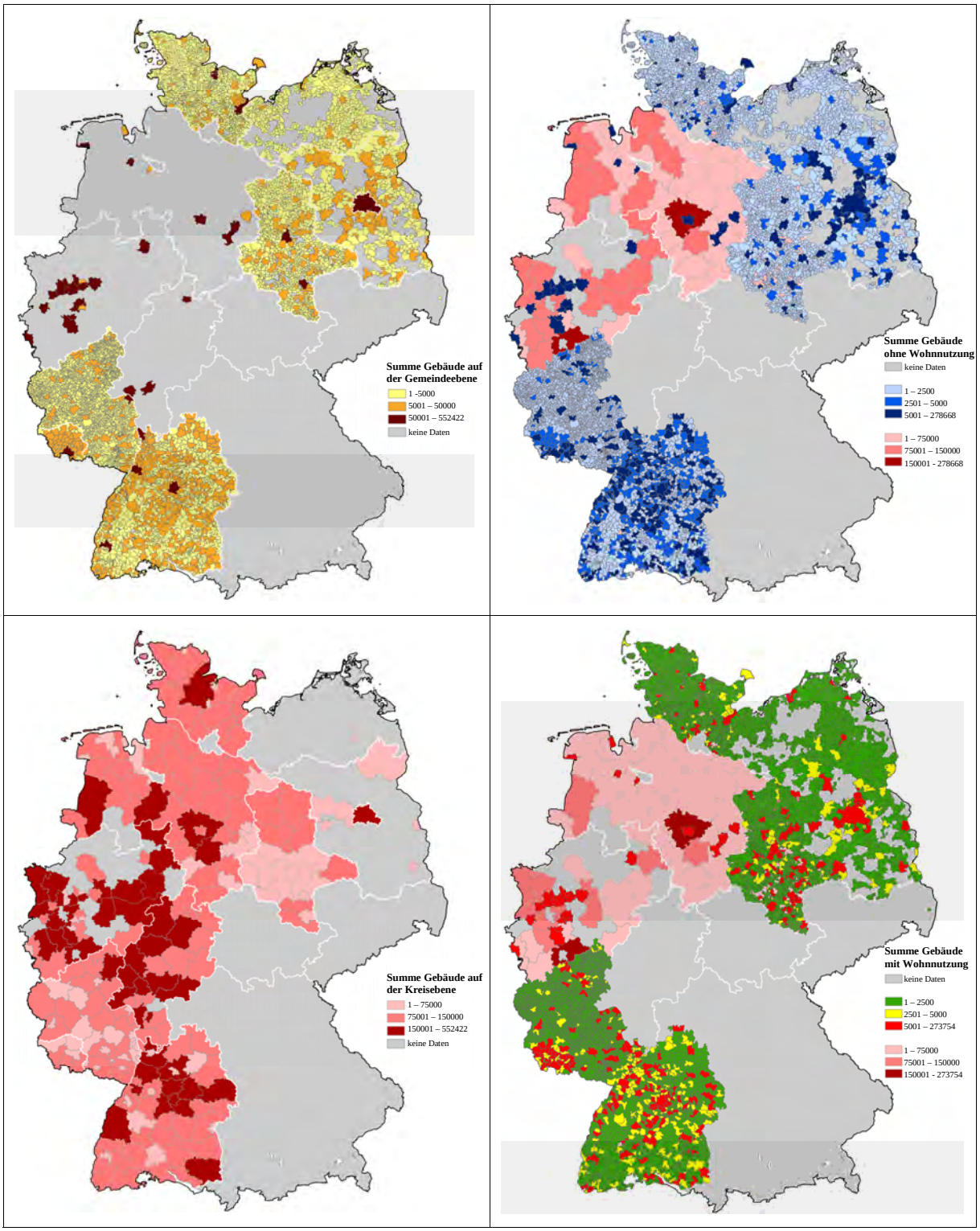


Abbildung 3-7: Teilbestände des Gebäudebestandes nach Nutzungsstruktur und Hierarchieebene[448]

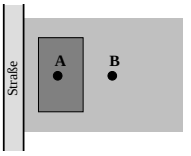
[448] Eigene Bearbeitung unter Verwendung der verfügbaren Daten aus der ALK

3.6.2 Hauskoordinaten[449]

Die Hauskoordinaten (georeferenzierte Gebäudeadressen) werden aus einer Verknüpfung der Adresse (Lagebezeichnung) eines Gebäudes und seiner exakten Lage (Verortung einer Lagebezeichnung) charakterisiert. Grundelemente bilden das Gebäudekennzeichen (Schlüssel, Verwaltungseinheit, Lagebezeichnung) und die zugehörige Gebäudekoordinate. Das von den Vermessungsverwaltungen der Länder geführte Liegenschaftskataster dient als Datenquelle für die Hauskoordinaten (z.B. Gauß-Krüger, UTM oder geografische Koordinaten). Da das Liegenschaftskataster kontinuierlich gepflegt und durch die zuständigen Katasterämter aktualisiert wird, handelt es sich um eine zuverlässige und sehr genaue Datenquelle, um den Gebäudebestand flächendeckend genauer zu beschreiben.

Es ist darauf hinzuweisen, dass die Hauskoordinaten lediglich die Gebäude erfassen, die über eine postalische Anschrift verfügen. Die Gebäudekoordinate befindet sich dabei innerhalb des Hausumrings und die Flurstückskoordinate innerhalb des Flurstücks. Nicht mit erfasst durch eine Abfrage auf die verfügbaren Hauskoordinaten und eine daraus resultierende statistische Auswertung wird der Bestand an Nebengebäuden. Gerade bei größeren Industriebaubeständen wirkt sich dieses sicherlich negativ aus. Im Anschluss an diese Arbeit können aber zukünftige Einzelfalluntersuchungen noch genauere Korrelationen zwischen den Hauskoordinaten und dem tatsächlichen Gebäudebestand aufdecken. Es sind darüber hinaus keine zusätzlichen Unterscheidungen nach der Nutzungsart möglich. Der Straßennamen und die Hausnummer werden über einen Schlüssel mit der jeweiligen Hauskoordinate verbunden. Tabelle 3-15 zeigt das Format der Hauskoordinaten für einige ausgewählte Beispiele.

Schemadarstellung



A= Koordinate innerhalb des Hausumrings

B= Koordinate innerhalb des Flurstücks

Gebäudeanzahl* in Deutschland: 19.303.874, erfasst am 31.10.2006

Berlin (332177), Bremen (166404), Hamburg (264548), Baden-Württemberg (2676786), Bayern (3102050), Brandenburg (593937), Hessen (1445097), Sachsen-Anhalt(596051), Mecklenburg-Vorpommern (244208), Niedersachsen (2278673), NRW (3961595), Rheinland-Pfalz (1160299), Saarland (313236), Sachsen (799640), Schleswig-Holstein (808837), Thüringen (560336)

*) Daten aus den Landkreisen Parchim, Güstrow und Bad Doberan sowie der Stadt Gotha sind noch als fehlend zu beklagen.

Gebäudekennzeichen																
Daten(bank)spez.		Verwaltungseinheit						Adresse			Gebäudekoordinate		Postalische Anschrift			
M	[B]BNNNNNNNN	Q	LL	R	KK	GGGG	OOOO	SSSS	Hnr.	ZzHnr.	yyyyyyy,yyy	xxxxxxx,xxx	SN	PPPPP	PON	ZzPON
N;	501909171	;A	;05	;3	;15	;000	;0000	;01608	;43	;	;2570033,600	;5641995,700	;In der Gracht	;51105	;Köln	;
N;	501975454	;A	;05	;3	;15	;000	;0000	;04338	;14	;	;2558220,000	;5645747,800	;Braugasse	;50859	;Köln	;
N;	501975455	;A	;05	;3	;15	;000	;0000	;04338	;14	;a	;2558233,300	;5645772,900	;Braugasse	;50859	;Köln	;
N;	503248064	;A	;05	;3	;15	;000	;0000	;15260	;9	;	;2572011,900	;5620434,200	;Moselweg	;53347	;Alfter	;(Rheinld.)
Datensatzkennung Datensatznummer		Qualitätsmerkmal Land		RegBez	Kreis Stadt	Gemeinde	Ortsteil	Straße	HNr.	Zusatz	Rechtswert Gauß-Krüger-Koordinaten	Hochwert	Straßenname	PLZ	Postalischer Ortsname	Zusatz

Tabelle 3-15: Hauskoordinaten-Format[450]

449 Das LVA-NRW übernimmt die Vermarktung der Hauskoordinaten für die Vermessungsverwaltungen der 16 Bundesländer. Im Jahre 2003 wurde für den länderübergreifenden Vertrieb von georeferenzierten Gebäudeadressen die GVHK (Gemeinschaft zur Verbreitung der Hauskoordinaten) gegründet.

450 Eigene Bearbeitung unter Verwendung der Daten der Hauskoordinaten im AdV-Format, Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV)

Abbildung 3-8 zeigt die Dichte der durch Hauskoordinaten erfassten Gebäude in einer von den Gemeindegrenzen losgelösten Darstellung.

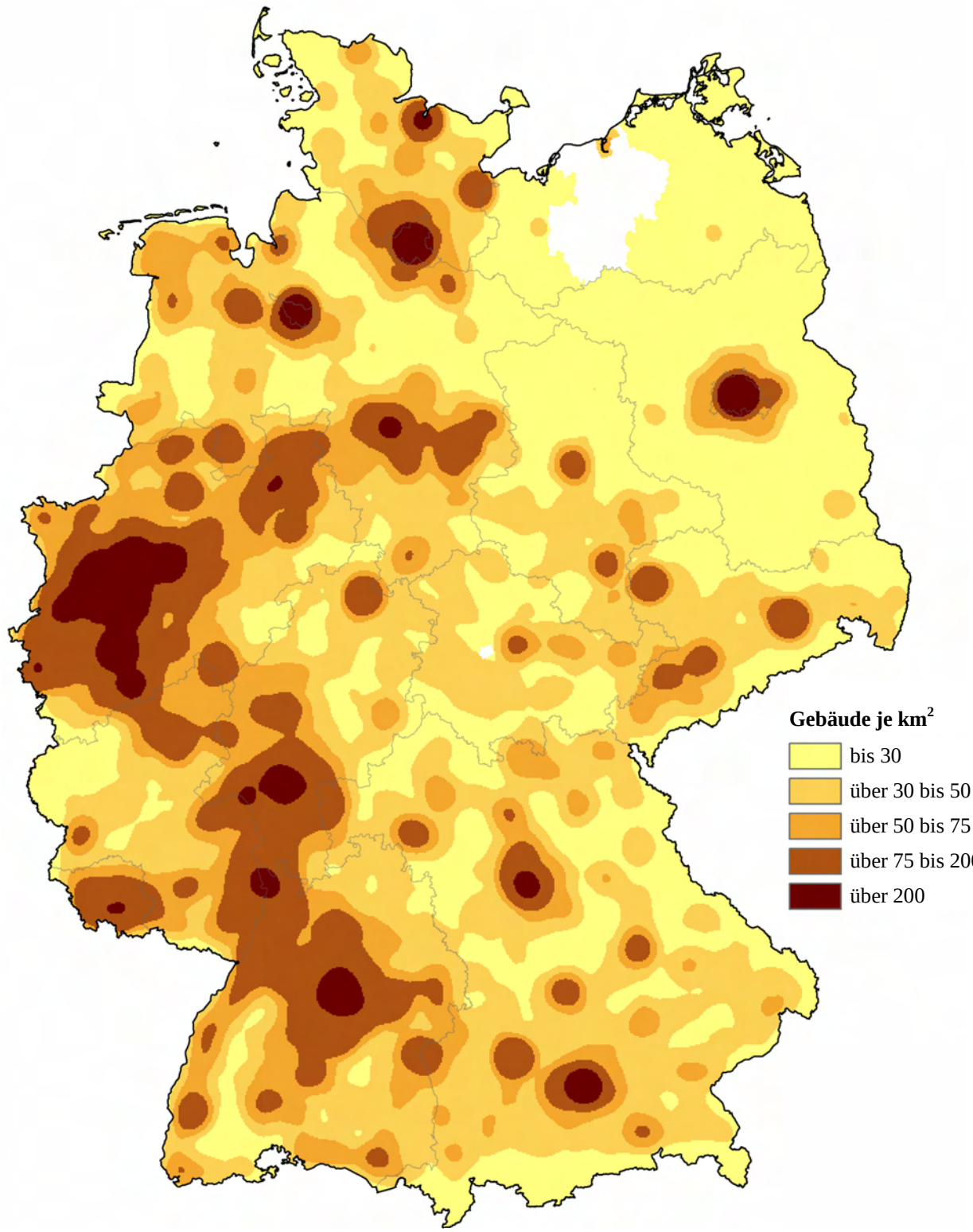


Abbildung 3-8: Gebäudedichte in Deutschland[451]

[451] Eigene Bearbeitung unter Verwendung von Daten der Hauskoordinaten. In Deutschland existieren an die 19.303.874 adressierten Gebäude (Stand 30.10.2006). Es ist zu berücksichtigen, dass die Daten aus den Landkreisen Parchim, Güstrow und Bad Doberan sowie der Stadt Gotha noch als fehlend zu beklagen sind.

3.7 Deutsches Fernerkundungsdatenzentrum (DFD)

Das europäische Projekt CORINE LAND COVER 2000 wurde in Deutschland im Auftrag des Umweltbundesamtes vom DFD bearbeitet. Die europäische Federführung hat die EEA (European Environment Agency)[452] übernommen. Die Satellitenbildgrundlage entstand aus dem Projekt ‚Image2000' unter Verantwortung des Joint Research Centre (JRC).[453]

3.7.1 Projektbeschreibung – CORINE Land Cover

Das Programm CORINE Land Cover 2000 (CLC-2000)[454] beinhaltet die europaweite Aktualisierung der Landnutzung und Bodenbedeckung. Das Ziel bestand darin, eine gemeinsame Datenbasis aufzubauen, welche Aussagen zur Bodenbedeckung, Landnutzung und deren Veränderung in den letzten 10 Jahren ermöglicht, und darüber hinaus im nationalen als auch europäischen Kontext Vergleichsmöglichkeiten anzubieten.

Die Datengrundlage für die Erfassung bilden europaweit so genannte einheitlich orthorektifizierte[455] Landsat-7-Daten[456] der Jahre 1999 bis 2001. Annähernd 330 Szenen für eine vollständige Abdeckung der damals zugehörigen EU-Staaten waren notwendig, wobei für Deutschland ca. 32 Landsat-Szenen erforderlich waren. Zur Erfassung wurde eine einheitliche Methode eingesetzt, die mit Hilfe eines durchgängigen Klassifizierungsschlüssels erfolgt, der auf der Systematik der Ersterfassung im Jahr 1990 beruht.[457] Die Erfassung der Landnutzung und den damit verbundenen Änderungen seit 1990 erfolgte in einem GIS-gestützten System mit Hilfe visueller Interpretationen, die durch automatische Verfahren unterstützt wurden. Zunächst wurden am DFD die Vektordatenbasis von 1990 sowie die Satellitendaten der Erfassung auf die orthorektifizierten Daten angepasst. Anschließend wurden die aktuelle Kartierung der Bodenbedeckung 2000 und eine Kartierung der Veränderungen gegenüber 1990 erzeugt. Zusätzlich konnte der Datensatz von 1990 infolge eines Interpretationsprozesses eine Korrektur und damit Überarbeitung erfahren. Es ist festzustellen, dass das Potential der Landsat-7-Daten für die Landnutzungskartierung mit dem Ansatz von CLC-2000 nicht ausgeschöpft und zukünftig mit weiteren Verbesserungen der Auflösung zu rechnen ist. Komplexere Forschungsfragen lassen sich zukünftig bearbeiten.[458]

---

452 Vgl. http://dataservice.eea.europa.eu/dataservice/ (Stand: 31.10.2006),

453 Vgl. http://jrc.cec.eu.int (31.10.2006)

454 Vgl. KEIL et al. [2002, S. 95-104] und UBA [2004]

455 Bei der Orthorektifizierung handelt es sich um ein Verfahren der geometrischen Entzerrung.

456 Seit dem Start von Landsat 1 (ERST-1) liefert das Landsat-Programm kontinuierlich Daten der Landesoberfläche und der Küstenregionen der Erde. Seit Juli 1999 ist Landsat 7 ETM+ in Betrieb. Die Initiatoren des Programms sind NASA und U.S. Geological Survey (USGS). Das DFD betreibt eine Empfangsstation.

457 Vgl. DEGGAU et al. [1998]

458 Vgl. RIDD / LIU [1998, S. 95 ff.] ; NETZBAND [1998], KRESSLER / STEINHOCKER [2001, S. 140 ff.], EU Project MURBANDY (Monitoring Urban Dynamics): http://murbandy.sai.jrc.it/ (Stand: 31.10.2006)

3.7.2 Datenbeschreibung – CORINE Land Cover

Die Erstauslieferung von Daten der CORINE Land Cover Erfassung erfolgte im Januar 2005. Es stehen für das Jahr 2000 und korrigiert für das Jahr 1990 sowie als Darstellung der Veränderungen zwischen 1990 und 2000 Vektor- und Rasterdatensätze zur Verfügung. Die Vektordaten werden für ganz Deutschland und auch für die einzelnen Bundesländer in Einheiten der topographischen Karte TK100 bereitgestellt. Die Rasterdatensätze stehen in den Auflösungen 100m x 100m, 250m x 250m und 1km x 1km als Datensätze zur Verfügung. Der Erfassungsmaßstab beträgt 1:100.000, wobei unter Berücksichtigung der Erfassung von 1990 die Neuflächen erst ab einer Größe von 25 ha und Veränderungen von Landnutzungsgrenzen erst ab 5 ha aufgenommen wurden. Die Flächen, welche eine linienförmige Ausprägung aufweisen (z.B. Gewässerläufe), wurden ab einer Breite von 100 m erfasst. Mögliche Projektionen sind Gauß-Krüger Zone 3, Gauß-Krüger Zone 4 oder UTM Zone 32.

Die Daten von CLC-2000 basieren unverändert auf dem Klassifikationssystem der Bodenbedeckungsarten der Kartierung CLC-1990. Eine zusätzliche Ergänzung bilden Erläuterungen zu einzelnen Klassendefinitionen.[459] Generell sind 44 Bodenbedeckungs- und Nutzungskategorien in drei Hierarchieebenen zu unterscheiden. In Deutschland sind insgesamt 37 von 44 Kategorien des europäischen Umfeldes anzutreffen. Die Hauptkategorien sind wie folgt zu unterscheiden: ‚bebaute Fläche', ‚landwirtschaftliche Flächen', ‚Wälder und naturnahe Flächen' sowie ‚Feucht- und Wasserflächen'. Auf Grundlage der CLC-2000 Daten werden die Maßzahlen zur räumlichen Konfiguration von Siedlungsstrukturen (siehe Abschnitt 3.9.2) berechnet. Zusätzlich denkbar sind im Anschluss an diese Arbeit Datenvergleiche zwischen verschiedenen Zeitschnitten der CLC-Erhebung (z.B. 1990 und 2000) und umfangreiche Einzelfallstudien von Repräsentanten der Ähnlichkeitsuntersuchung. Abbildung 3-9 zeigt einen CLC-Datenvergleich der Siedlungsstruktur von Berlin, die auf der Klassifikationssystematik (Bodenbedeckungs- und Nutzungskategorien) der Abbildung 3-10 basiert.

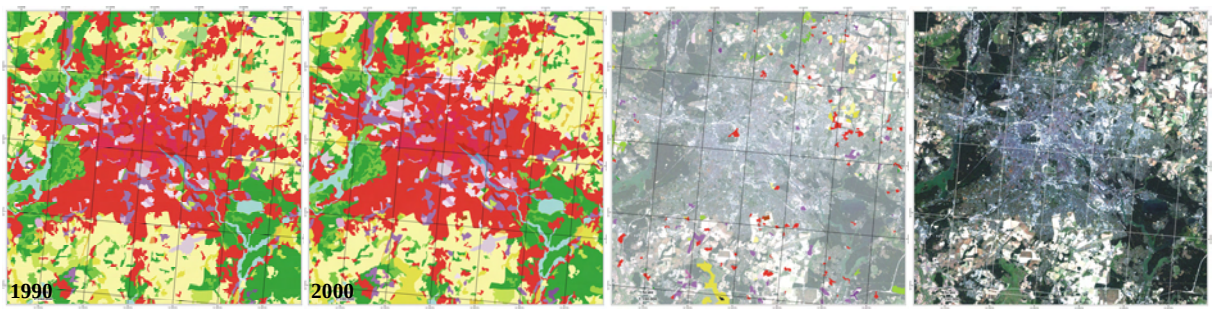


Abbildung 3-9: CLC-Datenvergleich für Siedlungsstrukturen am Beispiel Berlin (1990 und 2000)[460]

[459] Vgl. EEA [1997], BOSSARD et al. [2000]
[460] Eigene Bearbeitung, Abbildungen der EEA: http://dataservice.eea.europa.eu (Stand 15.10.2006).

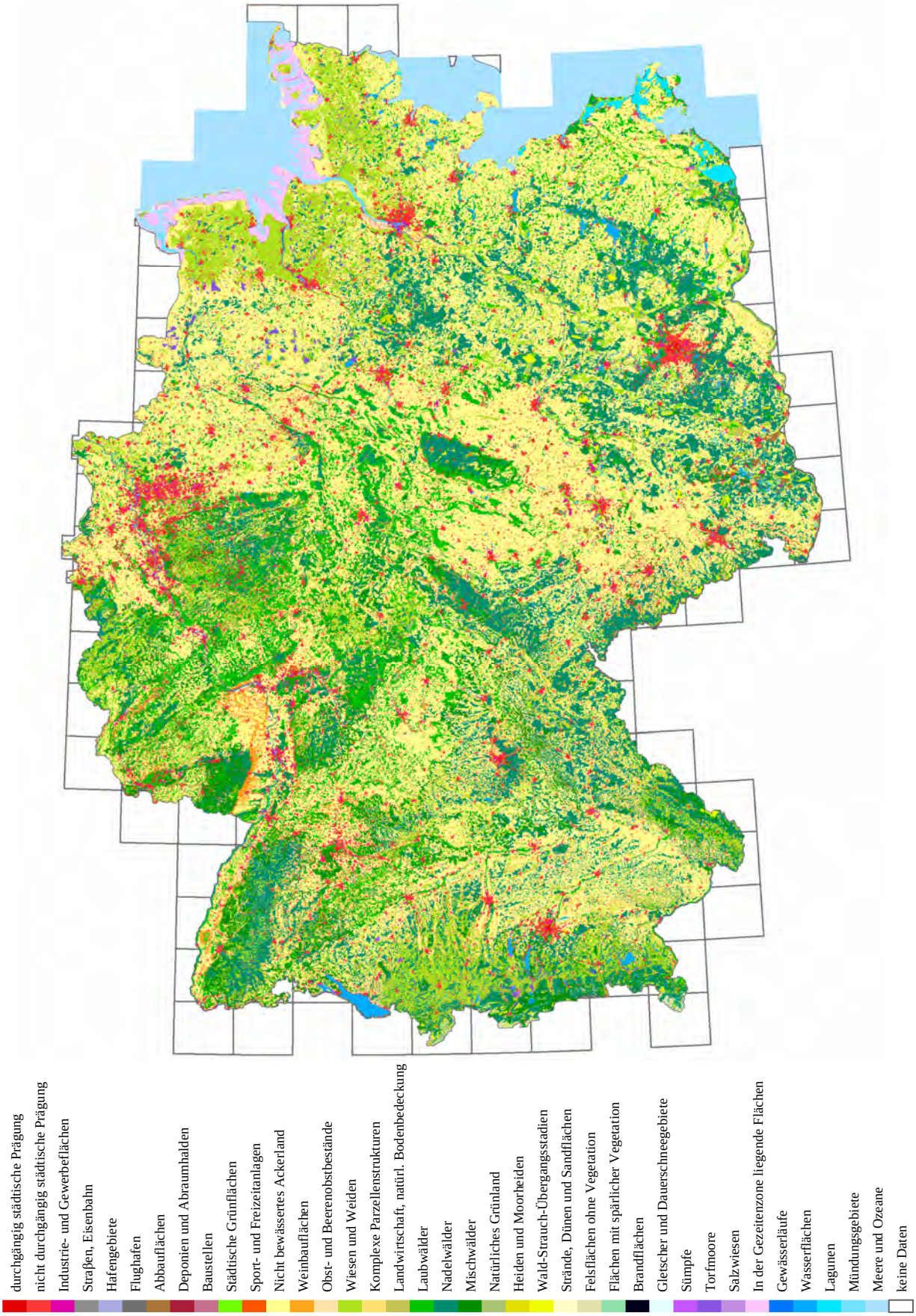


Abbildung 3-10: Klassifikationssystem der CLC-Bodenbedeckungsarten in Deutschland[461]

461 Eigene Bearbeitung unter Verwendung der CLC-2000 Daten.

3.8 Bundesamt für Kartographie und Geodäsie (BKG)[462]

Das BKG übernimmt in Zusammenarbeit mit der AdV die Aufgabe, topographische Geobasisdaten in Form digitaler Vektor- (Digitale Landschafts- und Geländemodelle, Verwaltungsgrenzen und Geographische Namen) und Rasterdaten (Digitale Topographische Karte) anzubieten. Zusätzlich werden analoge Kartenwerke in verschiedenen Maßstäben und zeitlichen Relationen veröffentlicht. Darüber hinaus ist das BKG Mitglied von ‚EuroGraphics',[463] so dass auch Europäische Datensätze bereitgestellt werden. Das Datenmaterial dient zu Visualisierungszwecken und bildet eine Datenquelle für die Geocomputation (Abschnitt 3.9).

3.8.1 Verwaltungsgrenzen

In dieser Arbeit erfolgt die Visualisierung von statistischem Datenmaterial und Ergebnissen der Ähnlichkeitsuntersuchung auf Grundlage der so genannten VG250 (Verwaltungsgrenzen, M 1:200.000), welche die Verwaltungseinheiten von allen hierarchischen Verwaltungsebenen vom Staat bis zu der untersten Verwaltungsebene der Gemeinden (NUTS 5 Level) enthält.

3.8.2 Digitales Geländemodell

Die Höhendaten, welche als statistische Zahlenwerte in dieser Arbeit verwendet werden, sind aus Kostengründen nur mit Hilfe des DGM 1000 (Gitterweite ca. 1 km x 1 km) bestimmt worden. Weitere 3D-Visualisierungen von Repräsentanten der Ähnlichkeitsuntersuchung erfolgen auf Grundlage des DGM 1000. Das Digitale Geländemodell Deutschland verfügt über eine sehr feinteilige Erfassung (Gitterweite 25 m bzw. 50 m in Gauß-Krüger-Abbildung). Während beim DGM 1000 nur die Topographische Karte (TK50) als Datenquelle dient, basiert das DGM Deutschland direkt auf Höhendaten der Landesvermessungsämter. Als Erfassungsmethoden sind hier das Laserscanning, Photogrammetrie und Digitalisierung von Höhenlinien zu nennen. Die Genauigkeit beim DGM 1000 liegt bei einer mittleren Höhe, die aus dem 1 km x1 km Gitter bestimmt wird. Das DGM Deutschland weist dagegen eine Lagegenauigkeit von ±1 bis 5 m und eine Höhengenauigkeit von ± 1 bis 8 m auf.

3.8.3 Digitale Topographische Karte

In dieser Arbeit wird die Digitale Topographische Karte als zusätzliche räumliche Bezugsgrundlage eingesetzt, um im Rahmen der Ergebnisinterpretation ggf. Hintergrundinformationen (z.B. Verkehrsnetz) über einzelne Gemeinden abzufragen. Die Rasterdaten werden in unterschiedlichen Maßstabsbereichen angeboten (1 : 25.000 bis 1 : 1.000.000).

---

[462] Eine Abfrage von Statistischen Daten aus dem Digitalen Basis-Landschaftsmodell wurde leider nicht vom BKG kostenfrei unterstützt. Das Basis-DLM kann als Informationsquelle über attributierte und georeferenzierte topographische Objekte dienen. Vgl. http:///www.adv-online.de (Stand 25.10.2006)

[463] Vgl. http://www.eurographics.org/eng/00_home.asp (Stand 28.10.2006)

3.9 Datengenerierung durch Geocomputation[464]

Die Forschungsintensität der GIS-basierten Modellierung, Analyse und Simulation wurde zuletzt infolge der zunehmenden Verfügbarkeit von Geodaten stark beschleunigt und förderte die Entwicklung von untersuchungsmethodischen Instrumenten, insbesondere die verbesserten Berechnungsansätze zur Neugenerierung von Daten aus bereits existenten raumbezogenen Datenbeständen. In den letzten Jahren hat sich der GIS-Einsatz in vielen Fachdisziplinen durchgesetzt, jedoch ist das Potential der GI-Systeme bei weitem noch nicht ausgeschöpft.[465]

Der vor einigen Jahen neu begründete Wissenschaftszweig des Geowissenschaftlichen Berechnens[466] verfolgt die Anwendung von innovativen Methoden und Werkzeugen aus berechnungsintensiven Fachdisziplinen wie der Mathematik und Informatik auf explizit räumliche Fragestellungen und setzt sich mit deren Integration in GI-Systeme auseinander. Dazu gehören Multiagentensysteme, Such- und Optimierungsverfahren, Genetische Verfahren und Neurocomputing sowie weitere Bereiche des Datenaufbaus im räumlichen Kontext.

Obwohl eine rasante Entwicklung der GIS-Technologie bereits stattgefunden hat, zeigen die heute kommerziell verfügbaren GIS noch eine Reihe von Defiziten und Schwächen bei der Modellierung und Simulation in den Umwelt- und Geowissenschaften. Als größtes konzeptionelles Hindernis der kommerziellen GIS ist das Fehlen der zeitlichen Dimension der geometrischen GIS-Daten zu nennen, welches aber schon lange einem breiten Forschungsinteresse unterliegt.[467] Darüber hinaus sind die heute existierenden GI-Systeme aufgrund verschiedener Entwicklungskonzepte, Datenmodelle und Marktstrategien oft abgeschlossene Systeme.[468] Dadurch wird der Geodatenaustausch, die Implementierung neuer Analysemethoden und die Kopplung von Modellen erschwert. Gerade die Kopplung erweitert das Anwendungsspektrum der GI-Systeme, weil sich weitere Berechnungs- und Modellierungsfunktionen hinzufügen lassen, die ein kommerzielles GIS nur schwer oder gar nicht ausführen kann. Es existieren verschiedene Lösungsansätze und eine Gewinnung weiterer Daten ist möglich.[469] Im Folgenden werden die für diese Forschungsarbeit zusätzlich relevanten Ansätze der Neugenerierung von Daten aus bereits bestehenden Datenbeständen kurz erläutert.

---

464 Vgl. LONGLEY et al. [1998], ATKINSON / MARTIN [2000], FISCHER / LEUNG [2001], FISCHER [2006], OPENSHAW et al. [2000]

465 BILL [1999, S.209]

466 Seit 1996 wird die International Conference on Geocomputation jährlich veranstaltet. Genannt sei die Ankündigung 2005: Geocomputation: The Art and Science of Solving Complex Spatial Problems with Computers, http://igre.emich.edu/geocomputation2005/ (Stand: 18.10.2006).

467 Vgl. STEYART / GOODCHILD [1994], STREIT / WIESMANN [1996, S. 161-173], LANGRAN [1992], PEUQUET / QUIAN [1995], YUAN [1996], WORBOYS [1998], EGENHOFER / GOLLEDGE [1998]

468 Vgl. STREIT [2000, S. 2-3]

469 Vgl. THINH [2004 a, S. 52 ff.]

3.9.1 Berechnung des Bevölkerungspotentials[470]

Grundsätzlich erlaubt das Potentialkonzept[471] eine Regionalisierung von regionalstatistischen Daten und führt durch die Bildung von räumlichen Aggregationen zu einer von den Verwaltungsgrenzen losgelösten Darstellung. Der Potentialansatz ist für die Regionalwissenschaft ein Hilfsmittel zur Analyse räumlicher Interaktionen. Es handelt sich um eine Verallgemeinerung des Gravitationsansatzes, der auf dem Ansatz von NEWTON basiert.[472]

In der Bevölkerungsgeographie wird die Möglichkeit räumlicher Interaktionen mit Hilfe des Bevölkerungspotentials gemessen. Je mehr Bevölkerung in der Umgebung eines Ortes erreichbar und je geringer der dabei zur Raumüberwindung benötigte Aufwand ist, desto höher ist dessen Kontaktpotential. Das physikalische Gesetz wird dabei auf die Regionalwissenschaft übertragen, wobei in diesem speziellen Fall die Massen als Bevölkerung an verschiedenen Orten aufgefasst werden. Jede dieser Massen übt eine bestimmte Anziehungskraft auf andere Orte aus, die auch als Gravitationspotential eines Ortes bezeichnet werden kann. Aus der Anziehungskraft mit der ein bestimmter Ort auf die Bewohner eines anderen Ortes wirkt, ergibt sich ein partielles Gravitationspotential, welches im übertragenen Sinne an diesem Ort die Bevölkerung entstehen lässt. Dieser Ansatz lässt eine Übertragung auf eine ganze Region zu, die aus n Orten besteht, so dass sich das gesamte Potential eines Ortes i aus der Summe über alle n partiellen Potentiale berechnet.[473]

Das Bevölkerungspotential (P_i) berechnet sich für einen Ort wie folgt:

$$P_i = \gamma \cdot \sum_{j=1}^{n} \frac{M_j}{r_{ij}^a} \; ; \qquad r_{ij} \neq 0, i \neq j \tag{78}$$

Dabei ist das Potential P an einem bestimmten Ort i direkt von den in den anderen Orten j vorhandenen Massen (M_j) abhängig, d.h. der dort ansässigen Bevölkerung. Es ist umgekehrt proportional zu den Entfernungen vom Ort i zu den Orten j (r_{ij}). Der Entfernungsexponent a ist als Maßzahl zu verstehen, um den Umfang der durch geographische Hindernisse vorhandenen räumlichen Widerstände abzubilden. Als feste Größe dient der Faktor γ.[474]

---

470 Vgl. SPANGENBERG [2003], Erklärung des BBR-Indikators „Regionales Bevölkerungspotential“

471 Vgl. STEWART, J.Q. [1947, S. 461-548]: Das Konzept des Potenzials wurde bereits im Jahr 1947 von der Naturwissenschaft auf die Sozialwissenschaft übertragen.

472 Vgl. MISNER [2000]: Das Gravitationsgesetz wurde von Newton entdeckt und besagt, dass die Kraft (F), mit der sich zwei Massen, m_1 und m_2 anziehen, dem Produkt dieser Massen direkt und dem Quadrat des Abstandes r der Massenschwerpunkte umgekehrt proportional ist: $F = \frac{\gamma \cdot m_1 \cdot m_2}{r^2}$, γ ist eine Naturkonstante

473 Vgl. ISARD, W. [1960, S.493 ff.] und KAU [1970, S. 11 ff.], STAACK [1995, S. 104], BRESSLER [2001]

474 Vgl. LAUSCHMANN [1973, S. 56 ff.]: Diese Größe ist zu schätzen. Erfasst werden die Einwirkungen der Wirtschafts- und Sozialstruktur, des wirtschaftlichen Entwicklungsstandes, der Konjunkturlage u.ä.

Als Bezugspunkte zur Berechnung der Bevölkerungspotentiale werden häufig die geometrischen Mittelpunkte der betrachteten Gebietseinheiten verwendet. Unter Umständen kann jedoch die Potentialbestimmung sehr willkürlich werden, falls es sich um sehr große bzw. heterogene räumliche Bezugseinheiten handelt. Als zusätzliches methodisches Problem ist das zu berücksichtigende Eigenpotential des Ortes zu nennen.

In neueren Studien und dem hier beschriebenen Ansatz des BBR wird das regionale Bevölkerungspotential im Gegensatz zu der Newtonschen Funktion mit Hilfe einer Wachstumsfunktion in Form einer negativen Exponentialfunktion berechnet, die der Gewichtung von Wegelängen oder Reisezeiten dient.[475] In der Physik dienen diese Funktionen der Modellierung des radioaktiven Zerfalls oder der Absorption von Licht. Mit Hilfe der gewählten Distanzgewichtung werden die Aktionsräume der interagierenden Individuen ausgedrückt. Anhand der Pendlerstatistiken wurden beim BBR Rückschlüsse auf diese Aktionsräume gezogen, da Bewegungen zwischen dem Wohnort und der Arbeitsstätte grundsätzlich eine große Bedeutung beizumessen ist und auf diese Weise annähernd eine Bestimmung der Obergrenze für die Raumüberwindung möglich wurde.[476] Für den regionalen Bezug wurde die Halbwertdistanz β mit 10 km verwendet. Die Umkreisgemeinden wurden auf eine maximale Entfernung von 50 km beschränkt. Der Datenumfang konnte dadurch reduziert werden und der BBR beziffert den Gewichtungswert mit unter 3,6 %. Für die Gemeinden in Deutschland ist unter diesen Annahmen das regionale Bevölkerungspotential für eine Gemeinde i im Umkreis von 50 km (Ø 350 Gemeinden) wie folgt berechenbar:

$$P_i = \sum_{j=1}^{n} M_j \cdot e^{\beta \cdot r_{ij}} \tag{79}$$

mit

$\beta = -0{,}0693\,(\approx \text{Halbwertdistanz} = 10\,\text{km}),\ e = 2{,}7183,\ r_{ij} = \text{Luftliniendistanz (Gemeindemittelpunkte)}$

Für die Eigenbevölkerung der Gemeinden wurde als Eigendistanz der halbe Radius der kreisförmig angenommenen Siedlungs- und Verkehrsfläche verwendet. Das BBR geht davon aus, dass in Gemeindemittelpunktnähe überwiegend höhere Siedlungsdichten dominieren und die besiedelte Fläche real nicht konzentrisch angeordnet ist.

[475] Vgl. BRÖCKER [1984, S. 55-99] und [1989] sowie NIEBUHR [2000, S.33]

[476] Der BBR beruft sich auf die laufende BBR-Umfrage zur Entfernung zwischen Wohn- und Arbeitsort

Die Potentialwerte ermöglichen eine Darstellung in einer Potentialkarte. Abbildung 3-11 stellt das regionale Bevölkerungspotential für 2004 auf Gemeindebasis gemessen in Einwohnern als interpolierte statistische Oberfläche dar. Es handelt sich um ein aus der räumlichen Verteilung der Bevölkerung generiertes eindimensionales Zentralitätsmaß.

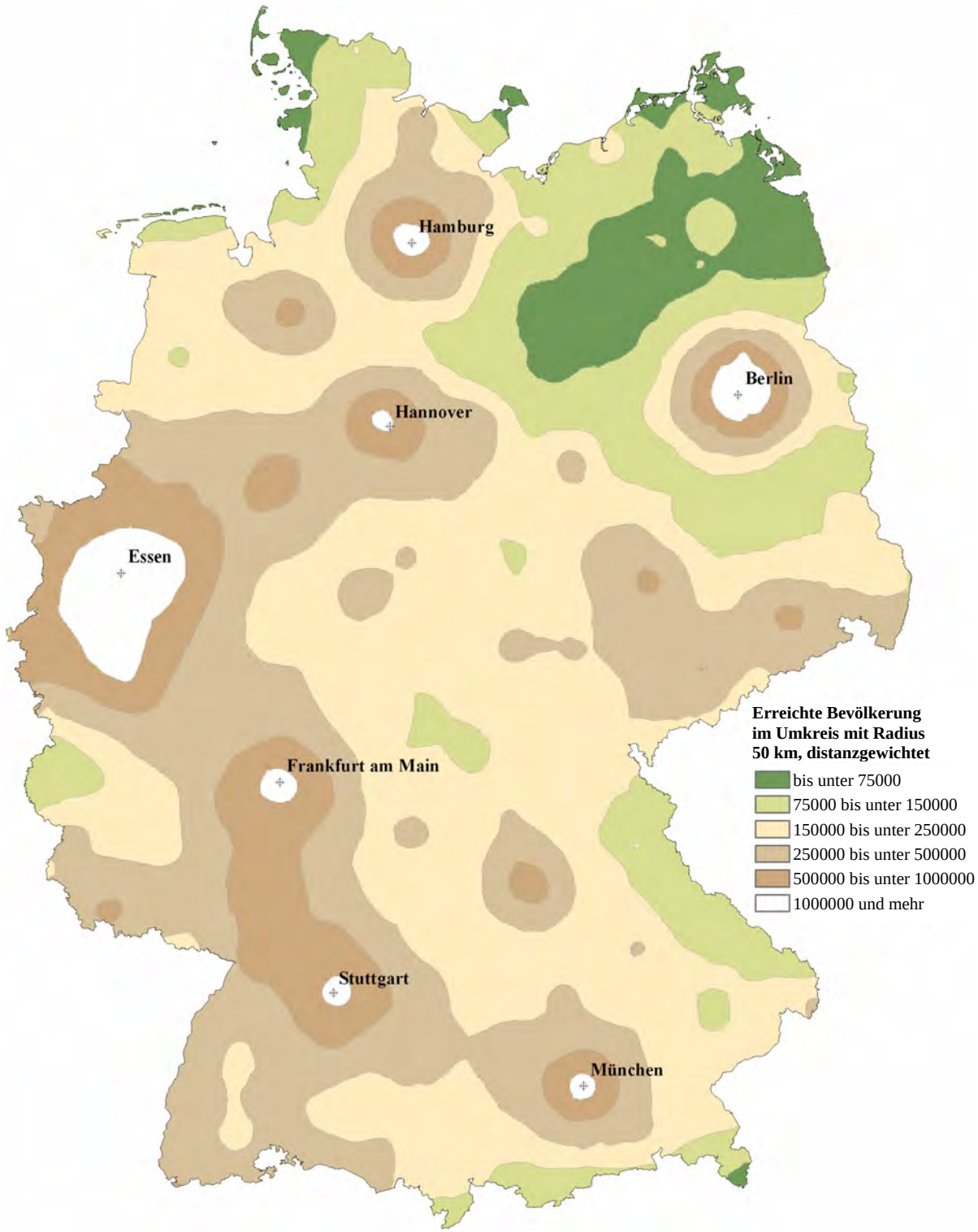


Abbildung 3-11: Potentialkarte auf Grundlage der Einwohner in den Gemeinden in Deutschland[477]

[477] Eigene Bearbeitung unter Verwendung der Daten zum Bevölkerungspotential des BBR.

3.9.2 Messung der räumlichen Konfiguration[478]

Eine Beschreibung von raumbezogenen Eigenschaften der urbanen Muster ist in der Regel weitgehend unscharf oder qualitativ definiert. Vielfach werden Abstraktionen vorgenommen und die Verteilung der Siedlungsfläche in einer Region als so genannter Schwarzplan abgebildet. Aus jüngster Zeit ist als Beispiel für eine derartige Vorgehensweise die Darstellung der Stadt-Region Stuttgart hervorzuheben, die bei SIEVERTS[479] für verschiedene Zeitschnitte (1850, 1950, 1995) genutzt wird, um das Phänomen der ‚Zwischenstadt' als eigene Entwicklungsform anhand dieser Schwarzpläne aufzuzeigen.

Unter dem Gesichtspunkt der Bedeutung der Begriffe ‚kompakte Stadt' und ‚durchmischte Stadt' sowie der eingeführten Metapher ‚Zwischenstadt'[480] wird es zunehmend wichtiger, neben einer auf den üblichen statistischen Kenngrößen basierenden Charakterisierung der Stadtregion Untersuchungskriterien zu verwenden, die die räumliche Anordnung von Flächennutzungsmustern genauer zu beschreiben helfen. Forschungsarbeiten, die eine Messung der raumbezogenen Eigenschaften von Stadtregionen zum Ziel haben oder die Simulation der Siedlungsmuster ermöglichen, sind noch sehr selten. THINH[481] diskutiert die Vor- und Nachteile der existierenden mathematischen Verfahren zur Messung der Kompaktheit von urbanen Mustern und stellt neue Methoden und Untersuchungsansätze vor.

Mit Hilfe der CORINE Land Cover Daten und Daten des Digitalen Landschaftsmodells sowie topografischer Karten ist am IÖR, Dresden, auf Basis des dort entwickelten Strukturtypenansatzes[482] eine umfangreiche Geo-Datenbasis der Flächennutzungsstruktur in Deutschland aufgebaut worden. Diese bezieht sich im Wesentlichen auf die kreisfreien Städte bzw. Stadtkreise und den Stadtverband Saarbrücken. Im Jahr 2006 wurde durch THINH auf Basis der CORINE Land Cover Daten am IÖR, Dresden, ein erweiterter Datensatz für alle 440 Landkreise und kreisfreien Städte entwickelt, der dann die Grundlage für die vom IÖR, Dresden, bereitgestellten Maßzahlen ‚Zerklüftungsgrad' und ‚Vernetzungsgrad' bildete.

---

478 Durch enge Zusammenarbeit mit dem IÖR, Dresden und insbesondere durch die ungewöhnlich große Hilfsbereitschaft von Herrn NGUYEN XUAN THINH, PD Dr.-rer.nat. habil., Dipl.-Mathematiker, existieren durch seine umfangreichen Berechnungen zusätzliche Maßzahlen, um die räumliche Konfiguration der 440 Kreise in dieser Arbeit zu charakterisieren. Vgl. THINH [2004 a, S. 10 ff. und 68 ff].

479 Vgl. STREIT [2005, S. 513]: „Trotz solcher Allgemeinaussagen können wir aber feststellen, dass das Beziehungsgeflecht zwischen ‚Kernstadt' und ‚Stadtregion' nicht geringer geworden ist, sondern vielmehr deutlich zugenommen hat […]. Auch die von Thomas SIEVERTS angestoßene Diskussion, die sich um die Metapher ‚Zwischenstadt' rankt, steht in Bezug zu diesem regionalen Kontext."

480 Vgl. Abschnitt 1.1.2

481 Vgl. THINH [2002, S. 409-422], THINH et al. [2002, S. 475-492] und THINH [2004 a] sowie THINH [2004 b, S. 221-232]

482 Die räumliche Stadtgliederung in Strukturtypenflächen wurde von HEBER / LEHMANN [1996] entwickelt und durch ARLT et al. [2001] qualifiziert.

3.9.2.1 Zerklüftungsgrad zur Messung der Kompaktheit

Die hier vorgestellten Maßzahlen zur Messung der Kompaktheit unterliegen der Annahme, dass das Muster der Flächennutzungsstruktur einer regionalen Gebietseinheit aus einer Menge von Polygonen unterschiedlicher Form besteht. Die Messung der Kompaktheit bezieht sich auf das räumliche Gefüge dieser Polygone und nicht auf die Form oder Kompaktheit der einzelnen Polygone.

Zu Beginn ist eine Generalisierung der Flächennutzung für alle Polygone vorzunehmen, d.h. die verschiedenen Polygone sind entweder Bestandteil des Siedlungsraumes oder des Freiraumes. THINH[483] verwendet zusätzlich den Begriff Freiflächen und spricht dabei von Flächen im Siedlungsraum, die nicht bebaut sind.

Die Berechnung des Zerklüftungsgrades erfordert die Definition von p_i als Umfang und a_i als Flächeninhalt der Polygone. Wenn man die gesamte Siedlungsfläche eines Untersuchungsgebietes in einem Kreis vereinigt, so entspricht dieser äquivalente Kreis dem Flächeninhalt der aufsummierten Flächeninhalte der Einzelpolygone der Siedlungsstruktur. Für die Berechnung des Umfanges $P_{\min}$ ergibt sich daraus die folgende Gleichung:

$$P_{\min} = 2\sqrt{\pi \sum_{i=1}^{n} a_i} \tag{80}$$

Weiterhin ist die Gesamtrandlänge eines Siedlungsgebietes wie folgt definiert:

$$P = \sum_{i=1}^{n} p_i \tag{81}$$

Das Verhältnis $\frac{P}{P_{\min}} = 1$, wenn die Stadt aus einer einzigen Siedlungsfläche bestehen würde, d.h. am kompaktesten wäre. Ansonsten gilt stets $\frac{P}{P_{\min}} > 1$. Je größer die Zerklüftung in einem Untersuchungsgebiet bezogen auf die Anordnung der Einzelpolygone ist, desto größer ist somit das Verhältnis $\frac{P}{P_{\min}}$. Der Grad der Zerklüftung einer Siedlungsstruktur wird als Verhältniswert folgendermaßen berechnet:

$$\text{Zerklüftungsgrad} = \frac{P}{P_{\min}} = \frac{\sum_{i=1}^{n} p_i}{2\sqrt{\pi \sum_{i=1}^{n} a_i}}; \quad \text{[dimensionslos]} \tag{82}$$

[483] THINH [2004 a, S. 16] und THINH et al. [2007]

Abbildung 3-12 gibt ein Beispiel zu den Ausführungen über den Zerklüftungsgrad. Dargestellt ist die Siedlungsfläche des Stadtkreises Karlsruhe (ohne die Grünflächen im Siedlungsraum) als abstrakter Schwarzplan sowie der dazugehörige äquivalente Kreis. Da für den Stadtkreis Karlsruhe und den Stadtkreis Remscheid jeweils ein Zerklüftungsgrad von 5,96 berechnet wurde, sind beide urbanen Flächenmuster abgebildet. Betrachtet man die berechneten Werte aller Landkreise und kreisfreien Städte, so existieren Werte im Intervall von 2,98 und 23,35. Je höher der Wert liegt, desto zerklüfteter ist das vorhandene Siedlungsmuster. Ein Wert von 1 entspricht wie gezeigt der ‚kompaktesten' Siedlungsfläche.

Abbildung 3-12: Die Siedlungsfläche des Stadtkreises Karlsruhe und der dazugehörige äquivalente Kreis sowie die Siedlungsfläche des Stadtkreises Remscheid als Objekt mit gleich großem Zerklüftungsgrad

Als Einschränkung bei der Verwendung derartig berechneter Zerklüftungsgrade ist die geringe Sensitivität gegenüber dem Auseinanderdriften von Siedlungsmustern zu nennen. Das Berechnungsverfahren ist nicht in der Lage, räumliche Distanzen zwischen den verschiedenen Siedlungsmustern zu erfassen. Hierzu ist ein von THINH[484] entwickeltes Kompaktheitsmaß auf der Grundlage der Rasteranalyse und des Gravitationsansatzes einsetzbar. Bei diesem Ansatz ist es möglich, die Stadtausdehnung (Dispersion) und die lakunäre[485] Beschaffenheit der städtischen Struktur zu messen. Als Nachteil dieses Verfahrens ist aber ein sehr hoher Rechenaufwand zu nennen. Zukünftig können mit diesen berechneten Maßzahlen optimierende bzw. vertiefende Betrachtungen stattfinden.

In der folgenden Abbildung 3-13 werden die Zerklüftungsgrade der Siedlungsfläche für die Kreisdaten dargestellt. Da sich die Verteilungsdichte eines Merkmalsvektors x in der Regel nur ungenau mit Hilfe einer einzelnen Normal- bzw. Gaußverteilung beschreiben lässt, wurde eine genauere Approximation mit Gaußschen Mischverteilungen vorgenommen und zwei Entscheidungsgrenzen gemäß Abschnitt 2.3.5 bestimmt. Lokalisiert werden auf diese Weise

484 Vgl. THINH [2002 a, S.414-415]

485 Vgl. Begriff aus der Medizin: lakunär = schwammig, buchtig, höhlenartig. Im urbanen Kontext ist unter diesem Begriff die Löchrigkeit einer städtischen Struktur zu verstehen (siehe THINH [2004, S.24]).

drei Klassen: ‚Kompaktheit', ‚Kompromiss zwischen kompakter und disperser Entwicklung' und ‚Zersiedelung bzw. Dispersion'. Die Region Hannover ist in den Grenzen des ehemaligen Land- und Stadtkreises dargestellt.

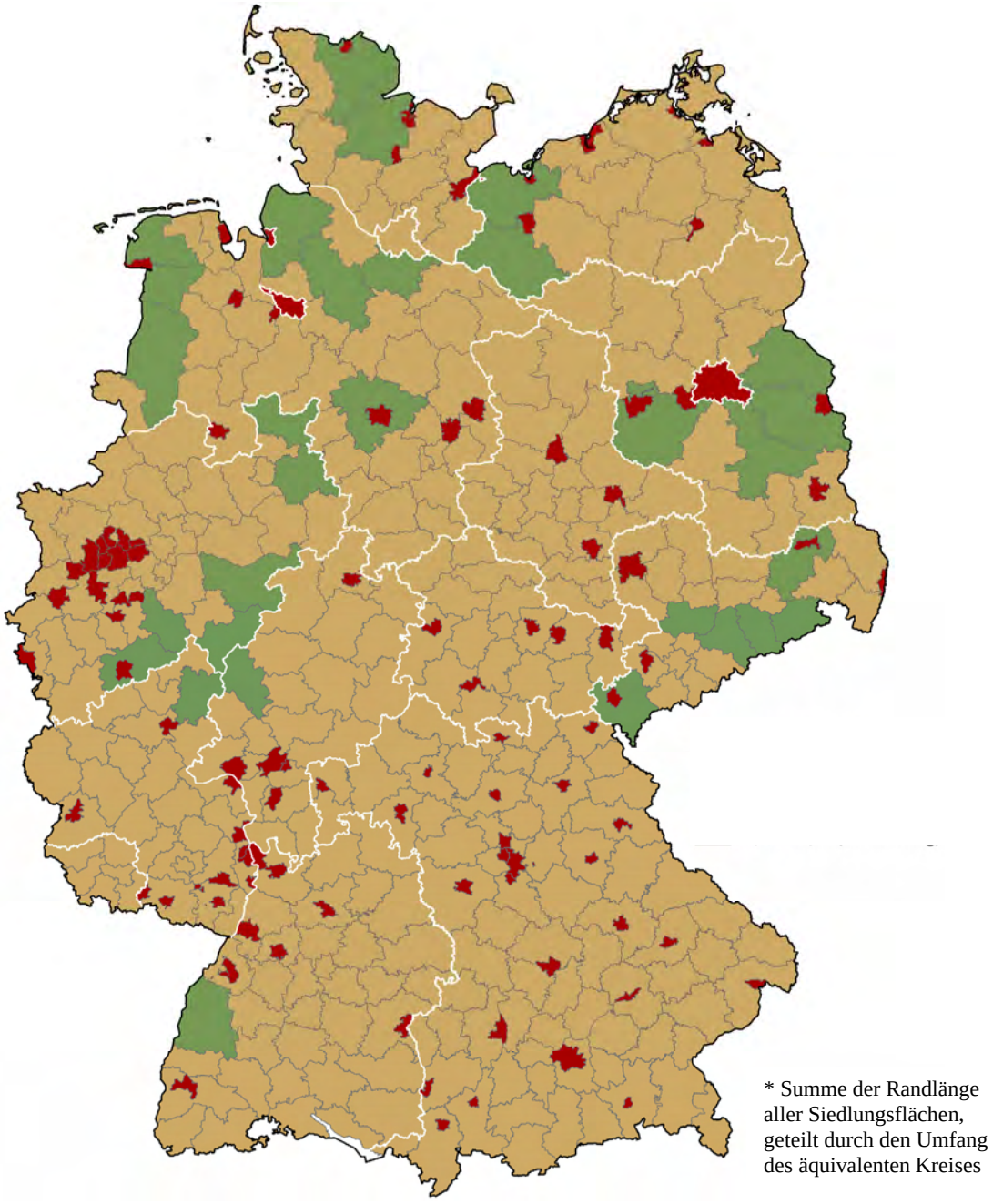


Abbildung 3-13: Klassifikation der Zerklüftungsgrade der Landkreise und kreisfreien Städte[486]

[486] Eigene Bearbeitung unter Einsatz bereitgestellter Werte des Leibniz-Instituts für ökologische Raumentwicklung e.V. (IÖR). Die Klasseneinteilung wurde mit Hilfe des Gauss-Ansatzes vorgenommen. Das hierzu aufgestellte Gauss-Mixtur-Modell ist im Nebenteil B hinterlegt.

3.9.2.2 Vernetzungsgrad zur Messung der Nutzungsmischung

Um die Nutzungsmischung innerhalb einer Region messen zu können, lassen sich sogenannte Vernetzungsgrade einsetzen. Diese werden berechenbar, indem das räumliche Nachbarschaftsgefüge der Flächen unterschiedlicher Nutzungen in einer Region mit Hilfe des GIS abgebildet wird. Dabei werden die Flächennutzungen zunächst vordefinierten Strukturtypen zugewiesen. Die räumliche Nachbarschaftsbeziehung zwischen zwei gewählten Strukturtypen wird durch die Gesamtheit aller gemeinsamen Grenzlinien zwischen den damit verbundenen Flächen charakterisiert. Die Vernetzung ist als aufsummierte Länge der Grenzlinien zu verstehen, die anhand der Flächen eines gewählten Strukturtyps zu Flächen von benachbarten Strukturtypen gemessen wird. Mit Hilfe einer GIS-Analyse lässt sich eine Grenzlinienmatrix aufbauen, die sich dazu eignet, unterschiedliche Vernetzungsmuster zu untersuchen.

Als Beispiel seien die Strukturtypen ‚Wohnbau' und ‚Industrie- und Gewerbe-Verkehrsfläche' gewählt, die zusammen eine gemeinsame Grenzlinie in einer Untersuchungsregion (SK Karlsruhe) aufweisen. Tabelle 3-16 zeigt Nachbarschaftsbeziehungen der Flächen.

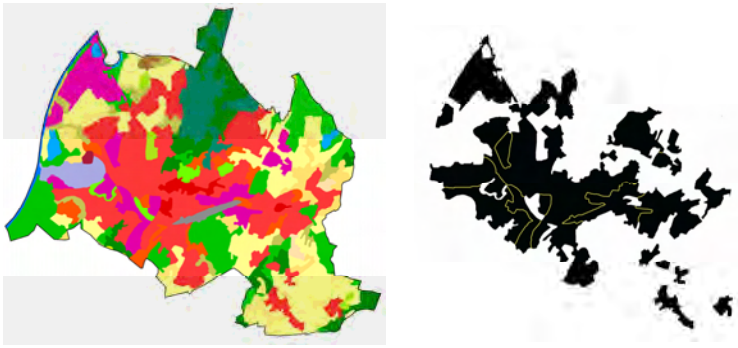
Stadtkreis Karlsruhe: Nachbarschaftsgefüge Wohnbau und Industrie-Gewerbe-Verkehrsfläche, Wert = **5,91 %**

Bodenbedeckungsart (CLC-2000)

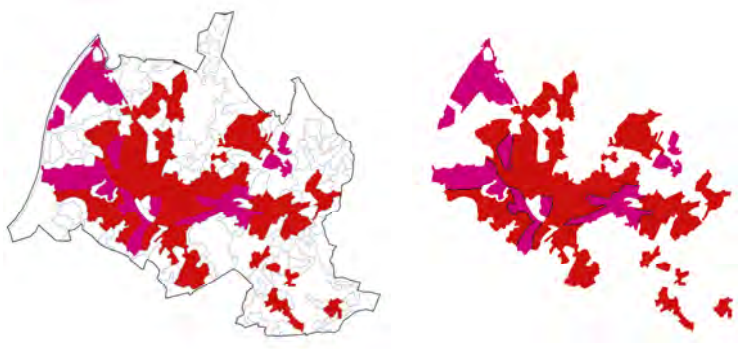
- 1) Durchgängig städtische Prägung [111]
- 2) Nicht durchgängig städtische Prägung [112]
- 3) Industrie- und Gewerbeflächen [121]
- 4) Straßen, Eisenbahn [122]
- 5) Hafengebiete [123]
- 6) Abbauflächen [131]
- 7) Deponien und Abraumhalden [132]
- 8) Baustellen [133]
- 9) Städtische Grünflächen [141]
- 10) Sport- und Freizeitanlagen [142]
- 11) Nicht bewässertes Ackerland [211]
- 12) Weinbauflächen [221]
- 13) Obst- und Beerenobstbestände [222]
- 14) Wiesen und Weiden [231]
- 15) Komplexe Parzellenstrukturen [242]
- 16) Landwirtschaft und natürliche Bodenbedeckung [243]
- 17) Laubwälder [311]
- 18) Nadelwälder [312]
- 19) Mischwälder [313]
- 20) Wald- Strauch- Übergangsstadien [324]
- 21) Gewässerläufe [511]
- 22) Wasserflächen [512]

Selektion (CLC-2000 bei 22 Klassen)

- CLC-Code: 111 bis 112
- CLC-Code: 121 bis 123



Konfiguration der Bodenbedeckungsarten und Grenzlinienschwarzplan



Grenzlänge zwischen Wohnbau und Industrie-Gewerbe-Verkehrsfläche

Tabelle 3-16: Prinzipdarstellung zur Berechnung des Vernetzungsgrades nach THINH[487]

[487] Eigene Bearbeitung, siehe auch THINH [2004, S.30 ff.]

Abbildung 3-14 zeigt eine Klassenaufteilung der Vernetzungsgrade für die 440 Landkreise und kreisfreien Städte. Die Region Hannover wurde in den Grenzen des ehemaligen Land- und Stadtkreises untersucht.

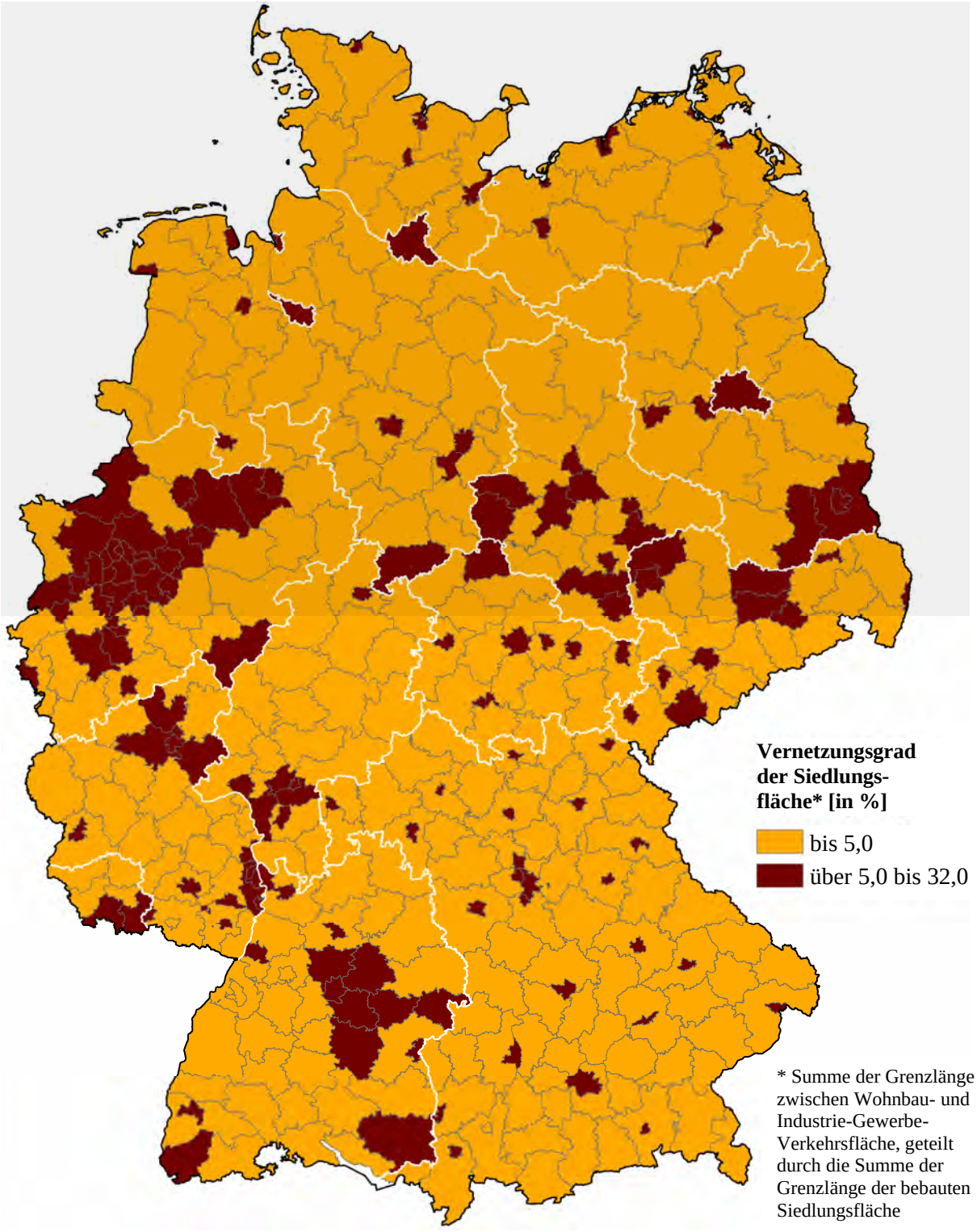


Abbildung 3-14: Klassifikation der Vernetzungsgrade (Wohnbau-, Industrie-Gewerbe-Verkehrsfläche)[488]

[488] Eigene Bearbeitung unter Einsatz bereitgestellter Werte des Leibniz-Instituts für ökologische Raumentwicklung e.V. (IÖR). Die Klasseneinteilung wurde mit Hilfe des Gauss-Ansatzes vorgenommen. Das hierzu aufgestellte Gauss-Mixtur-Modell ist im Nebenteil B hinterlegt.

Im Allgemeinen ist der Vernetzungsgrad wie gezeigt eine Messgröße zur Beurteilung von Nachbarschaften unterschiedlich genutzter Flächen auf mittelmaßstäblichem gesamtstädtischem Niveau. Anhand von Gaußschen Mischverteilungen wurde eine Entscheidungsgrenze bei 5,0 durch eine genauere Approximation des vorliegenden Verteilungsverlaufes ermittelt. Die resultierende Klassenbildung lokalisiert auf diese Weise Regionen der ‚Nutzungsmischung' bzw. ‚Nutzungstrennung'. Generell deuten niedrige Vernetzungsgrade auf eine hohe Verzahnung, während ein hoher Wert auf die räumliche Funktionstrennung verweist.

Das Aufstellen einer Grenzlinienmatrix wird an einem weiteren Vernetzungsmuster gezeigt, welches sich eignet die ökologische Qualität einer Stadt zu beurteilen. Dabei wird die Vernetzung von Freiflächen und Freiräumen und bebauter Fläche berücksichtigt und eine Grenzlinienmatrix (Tabelle 3-17) beispielhaft auf der Grundlage von 9 definierten Strukturtypen gemäß THINH[489] aufgebaut. Der Strukturtyp 1, 2, 3, 4 charakterisiere bebaute Flächen und unbebaute Flächen seien durch den Strukturtyp 5, 6, 7, 8 und 9 definiert.

[2,1]							
[3,1]	[3,2]						
[4,1]	[3,2]	[4,3]					
[5,1]	[3,2]	[5,3]	[5,4]				
[6,1]	[3,2]	[6,3]	[5,5]	[6,5]			
[7,1]	[3,2]	[7,3]	[5,6]	[6,6]	[7,6]		
[8,1]	[3,2]	[8,3]	[5,7]	[6,7]	[7,7]	[8,7]	
[9,1]	[3,2]	[9,3]	[5,8]	[6,8]	[7,8]	[8,8]	[9,8]

Tabelle 3-17: Beispiel einer Grenzlinienmatrix für bebaute Flächen und die Freiflächen und Freiräume [490]

Der Vernetzungsgrad wird als Prozentwert der Längensumme aller Grenzlinien der bebauten Flächen mit den unbebauten Flächen zur Gesamtlängensumme aller Grenzlinien zwischen den bebauten Flächen und Flächen aller anderen 9 Strukturtypen definiert und wie folgt berechnet.

$$\text{Vernetzungsgrad} = \frac{\sum_{x=5}^{9}\sum_{y=1}^{4}[x,y]}{\sum_{y=1}^{4}\sum_{x=y+1}^{9}[x,y]} \cdot 100 \qquad [\%] \qquad (83)$$

489 Vgl. THINH [2004 b, S. 31]

490 Hierbei bedeutet [x, y] die Längensumme aller gemeinsamen Grenzlinien zwischen Flächen von zwei beliebigen unterschiedlichen Strukturtypen x und y (x=2(1)9, y=1(1)x-1)), siehe THINH et al. [2000].

3.9.3 Berechnung der PKW-Erreichbarkeit

Beim Erreichbarkeitsmodell des BBR handelt es sich um eine modellhafte Abbildung des gesamten europaweiten Straßennetzes zum Stand 2003. Über ca. 20.000 Messpunkte wird anhand der PKW-Fahrzeiten zu allen Zielorten mit Hilfe einer Netzanalyse im Erreichbarkeitsmodell die PKW-Erreichbarkeit berechnet. Es ist möglich, eine von administrativen Einheiten losgelöste Auswertung durchzuführen. Die Daten sind geeignet, die Raumstruktur zu beschreiben und hinsichtlich der Lagegunst zu zentralen Orten als Träger wichtiger Raumfunktionen zu bewerten. Einen wichtigen Einflussfaktor für die Zugänglichkeit von Zentren und damit für die Vermittlung räumlicher Standortattraktivität stellt die Qualität der Verkehrsinfrastruktur dar. Bei Abbildung 3-15 handelt es sich um eine Darstellung der Reisezeitisochronen im motorisierten Individualverkehr (MIV) zum nächsten Oberzentrum.

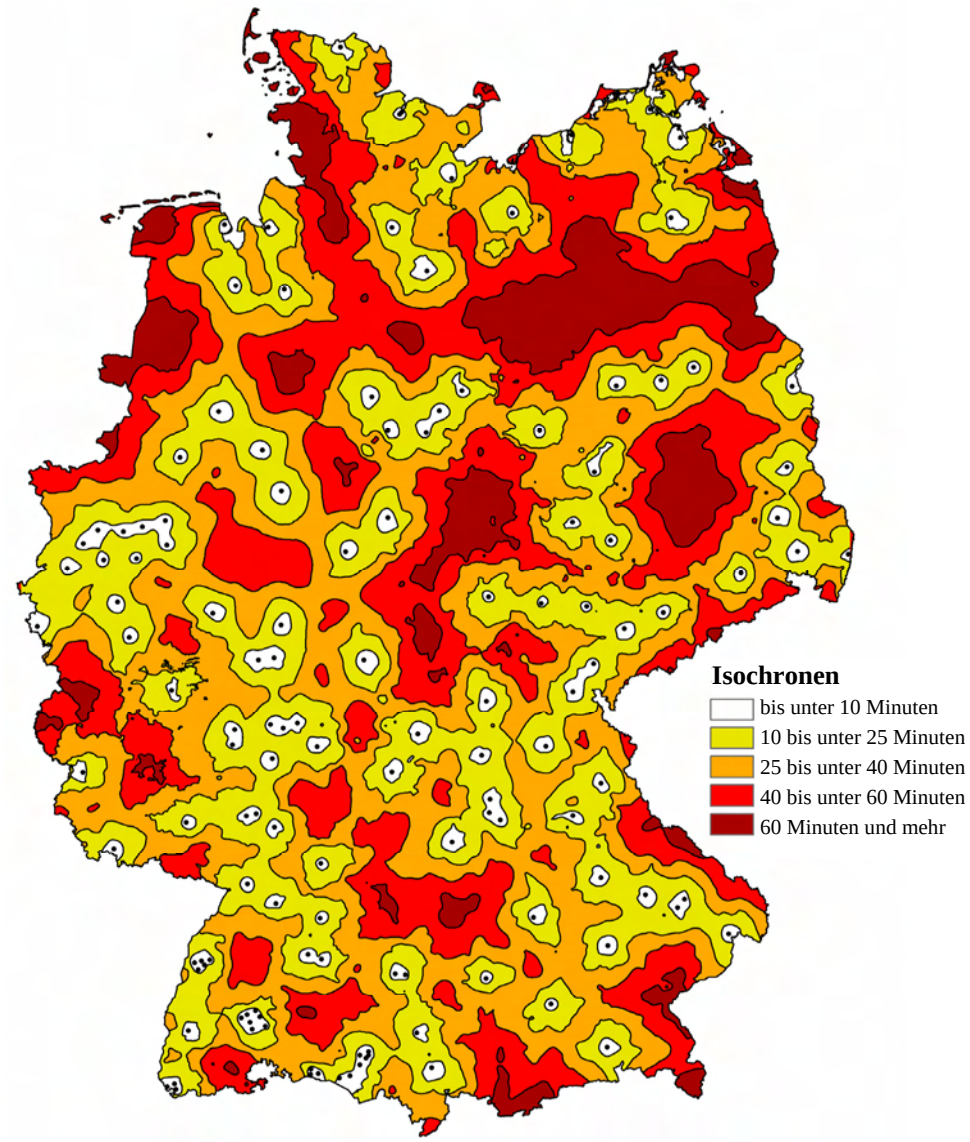


Abbildung 3-15: Reisezeitisochronen im motorisierten Individualverkehr (MIV)[491]

[491] Eigene Bearbeitung auf Grundlage von Daten des BBR – Referat I 1 ‚Raumentwicklung'.

4 Datenaufbereitung des ‚Urban Data Mining' (Referenzbeispiel zur Regularisierung)

Für die eingesetzten Variablen erfolgt die Verteilungsuntersuchung in dieser Arbeit nach einem definierten Arbeitsprinzip, um den Grundvoraussetzungen üblicher Methoden des Data Mining gerecht zu werden,[492] d.h. die verwendeten Variablen werden vorverarbeitet, indem darauf geachtet wird, dass die Variablen einem gleichen Verteilungsverlauf (z.B. Normalverteilung) folgen. Das an dieser Stelle dargestellte Referenzbeispiel der Datenregularisierung ermöglicht die Rekapitulation der relevanten Teilschritte.

4.1 Aufstellung einer Verteilungshypothese

Zunächst ist eine Hypothese über die zu erwartende Verteilung[493] einer Variablen aufzustellen, die dann entweder verifiziert oder falsifiziert wird. Wenn es sich um eine zusammengesetzte Messgröße handelt, sind sowohl die Messgröße selbst als auch die damit verbundenen Ausgangsgrößen zu untersuchen, um Verständnis für die Einzelverteilungen zu erhalten. Es ist dabei zu überlegen in welchem Wertebereich die Daten liegen könnten, um ggf. Fehler in den Rohdaten oder Ausreißer frühzeitig zu identifizieren.

Gegeben: Eine Berechnungsgröße (‚**AuslastungGebFreiflaeche**') und damit verbundene Ausgangsgrößen (‚**Bevoelkerung**', ‚**SVBeschaeftigte**' und ‚**FlaecheGebFrei**').

Es sei als zusammengesetzte Berechnungsgröße die Variable Auslastung der Gebäude und Freifläche (‚AuslastungGebFreiflaeche') gemäß der folgenden Berechnungsformel definiert:

$$AuslastungGebFreiflaeche = \frac{(Bevoelkerung + SVBeschaeftigte)}{FlaecheGebFrei} \qquad [km^2] \qquad (84)$$

Es handelt sich um einen siedlungsräumlich baulichen Dichtewert, wobei der Zähler aus der Summe der Einwohner in einer Gemeinde und der dort vorhandenen sozialversicherungspflichtig Beschäftigten am Arbeitsort gebildet wird. Der Nenner wird durch eine Flächengröße in Form der Gebäude- und Freifläche beschrieben, die im Wesentlichen Wohn-, Misch- und gewerbliche Nutzflächen umfasst. Die Variable ermöglicht Messungen von Konzentrationsphänomenen, wobei der Wertebereich durch positive Werte (>0) und einem vorab nicht bekannten Maximalwert charakterisiert wird. Bei den Ausgangsgrößen dürfen ebenfalls nur positive Werteausprägungen (>0) vorliegen.

---

[492] Vgl. ULTSCH [2006 c], siehe zusätzlich auch bei HARTUNG [2005, S. 832 ff.] oder ERB [1990, S. 57 ff.]: Verfahren des Data Mining (z.B. Clusteralgorithmen) setzen voraus, dass die verwendeten Variablen vorverarbeitet sind und dem gleichen Verteilungsverlauf folgen, d.h. beispielsweise normalverteilte Variablen.

[493] Es sei darauf verwiesen, dass Verteilungen Modelle für die Wahrscheinlichkeitsdichte der Daten sind. Das bedeutet, dass zu jeder Ausprägung ein Wert benannt werden kann, der die Wahrscheinlichkeit angibt, an dieser Stelle einen Datenpunkt zu finden (siehe Abschnitt 2.1.1.1).

Für die Verteilung der Berechnungsgröße ‚AuslastungGebFreiflaeche' wird eine linkssteile bzw. rechtssteile Verteilung vermutet, da wahrscheinlich nur einige wenige Gemeinden über eine sehr große Auslastung der Gebäude- und Freifläche verfügen und eine sehr große Anzahl der Gemeinden eine im Vergleich dazu geringere Dichte aufweisen.

Für die Ausgangsgrößen ‚Bevoelkerung', ‚SVBeschaeftigte' und ‚FlaecheGebFrei' werden ebenfalls linkssteile bzw. rechtsschiefe Verteilungen erwartet, d.h eine große Anzahl von Gemeinden mit niedrigen und einige wenige Gemeinden mit hoher Werteausprägung. Bei den Gemeinden mit einer großen Anzahl Beschäftigten als auch einer hohen Einwohnerzahl wird es sich im Wesentlichen um die Mittel- und Oberzentren handeln, die als Orte der Beschäftigung und Infrastruktur die wesentlichen Funktionen gemäß der Raumordnung beinhalten und nicht den Hauptanteil der Gemeinden in Deutschland bilden.

Um eine schiefe Verteilung charakterisieren zu können, kann nach einer Möglichkeit zur Modellierung des Verteilungsverlaufs gesucht werden, bzw eine nichtlineare Transformation auf die Daten angewendet werden. Der Zweck nichtlinearer Transformationen besteht darin, eine Verteilung so zu transformieren, dass die transformierte Verteilung einer theoretischen (einfachen) entspricht. Zunächst ist deshalb eine Vermutung über die Art der mit der Messgröße verbundenen Phänomene (z.B. Wachstums- oder Konzentrationsprozesse) zu formulieren, um ggf. eine Transformation aus der ‚Ladder of Power' zielorientiert auszuwählen (siehe zur Vertiefung Tabelle 2-2 im methodischen Teil dieser Arbeit).

Es wird vermutet, dass es sich bei der Berechnungsgröße ‚AuslastungGebFreiflaeche' um eine Wachstumsgröße handelt, da bei steigender Einwohner- und Beschäftigungszahl die Auslastung der Gebäude- und Freifläche zunimmt. Demzufolge könnten die Daten einer Lognormalverteilung folgen.

Bei den Teilgrößen ‚Bevoelkerung', und ‚SVBeschaeftigte' ist in erster Annäherung aufgrund von möglichen regionalen Konzentrationseffekten eine Lognormalverteilung zur Modellierung vorstellbar. Es wird davon ausgegangen, dass in den Agglomerationsräumen weitere Konzentrationsprozesse stattfinden.

Die Teilgröße ‚FlaecheGebFrei' bezieht sich auf die Fläche (km²), so dass eine Wurzeltransformation zur Modellierung der Daten sich ggf. eignen könnte. Darüber hinaus könnte die Hypothese einer Lognormalverteilung der Daten aufgrund möglicher Wachstumseffekte zutreffen, d. h. man geht davon aus, dass in den Regionen mit großen Bevölkerungs- oder Beschäftigtenzahlen auch mit größeren Gebäude- und Freiflächen zu rechnen ist.

4.2 Prüfung der Verteilungshypothese

Zur Prüfung einer zunächst aus inhaltlicher Überlegung gewonnenen Vorstellung über die Verteilung einer Variablen eignen sich Maßzahlen sowie vor allem visuelle Werkzeuge.

Die Ausprägungen der zuvor beispielhaft gewählten Messgröße werden in statistischer Form anhand von Lage- und Streuungsmaßen untersucht. Diese erlauben es, einen Eindruck von den Wertebereichen zu gewinnen, wobei zur korrekten Beurteilung die Art der Verteilung immer eine wichtige Rolle einnimmt. Beispielsweise erweisen sich für mehrgipflige Verteilungen der Mittelwert und Median zur Charakterisierung der Verteilung in der Regel nicht als sinnvolle Größen. Zusätzlich werden deshalb grafische Hilfsmittel eingesetzt, wie z.B. Histogramme, Box-Plots, QQ-Plots oder die Darstellungen der Pareto Density Estimation (Wahrscheinlichkeitsdichte). Mit den Quantil/Quantil-Plots werden zwei Verteilungen bildhaft miteinander verglichen. Die Pareto Density Estimation eignet sich dazu, Verteilungen zu prüfen, indem gemessene Werte gegen eine bekannte Verteilung aufgetragen werden.

Tabelle 4-1 enthält Lage- und Streuungsmaße der Variable (‚AuslastungGebFreiflaeche') und ihrer Ausgangsgrößen (‚Bevoelkerung', ‚SVBeschaeftigte' und ‚FlaecheGebFrei').

Untersuchungsgröße	Min	Mean	Median	Max	Anteil Zeros	Anteil NaN	NonZero Mean	NonZero Median	StdAbw
‚AuslastungGebFreiflaeche'	54,13	2964	2734	15332	0	0			1470,6
‚Bevoelkerung'	5	6637	1483	3387806	0	0			43505
‚SVBeschArbeitsort'	0	2134	195	1042300	2,3	0	2184	210	17111
‚SummeBevSVBesch'	5	8771	1701	4430100	0	0			60193
‚FlaecheGebFrei' [km²]	0,89	192,6	71,54	36230	0	0			612,3

Tabelle 4-1: Lage- und Streuungsmaße der Variable , AuslastungGebFreiflaeche'

Unter der Voraussetzung, dass es sich bei der zu beschreibenden Verteilung um eine eingipflige Verteilung handelt, lassen sich mit Hilfe der Kombination von Mittelwert und Median bereits weitere Schlüsse auf die Form der Verteilung ziehen: Man spricht von einer rechtsschiefen bzw. linkssteilen Verteilung, wenn der Median kleiner als der Mittelwert ist. Mit den Anteilswerten der fehlenden Daten (‚AnteilNaN') und der Nullwerte (‚AnteilZeros') besteht die Möglichkeit, die Rohdatenqualität zu beurteilen. Aus den Minimal und Maximalwerten lassen sich ggf. Unplausibilitäten bzw. Datenfehler identifizieren und beseitigen.

Die Verteilungen und Eigenschaften der Untersuchungsgrößen sind in Tabelle 4-2 zusammengefasst. Es wird die Vermutung bestätigt, dass es sich um linkssteile bzw. rechtsschiefe Verteilungen handelt und zeigt die Modellierung unter der Annahme einer Log-Normalverteilung. Im Folgenden werden einige Darstellungen genauer diskutiert.

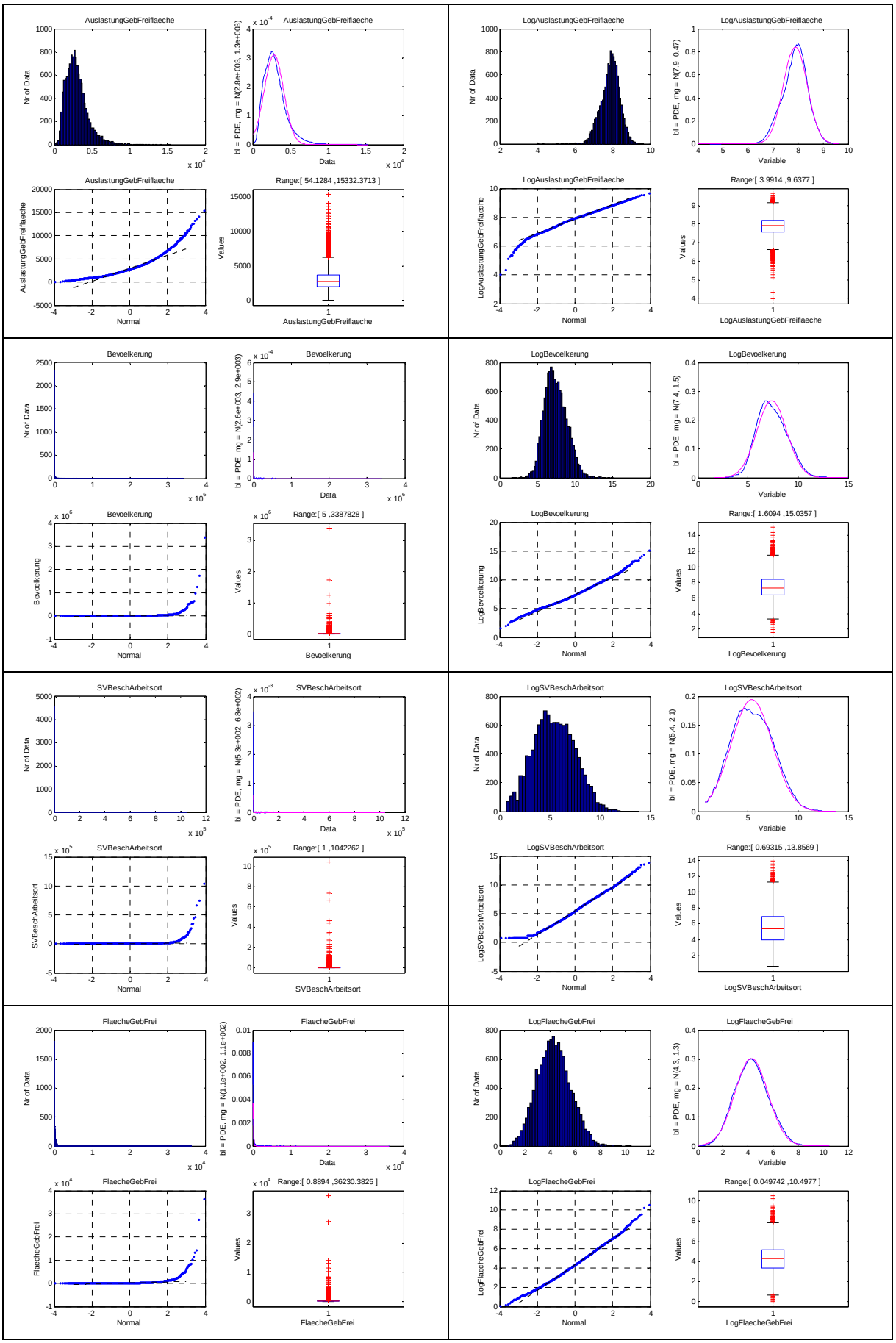


Tabelle 4-2: Einzeluntersuchung der Variable ‚AuslastungGebFreiflaeche' und Berechnungsgrößen

Die QQ-Plots und Histogramme der Rohdaten bestätigen die erste Vermutung, dass es sich um schiefe (linkssteile) Verteilungen handelt. Bei den QQ-Plots ist die Quantile der jeweils beobachteten Verteilung gegen die Quantile der Standardnormalverteilung aufgetragen. Die Quantile der zu untersuchenden Merkmale ist auf der Y-Achse aufgetragen, so dass an dieser Achse auch das Ablesen des Wertebereichs des Merkmals möglich ist. Die QQ-Plots belegen, dass durch Auftragung der genannten Verteilungen in einem Koordinatensystem die entstandenen Punkte keine deutliche Gerade bilden und keine Normalverteilung vorliegt.

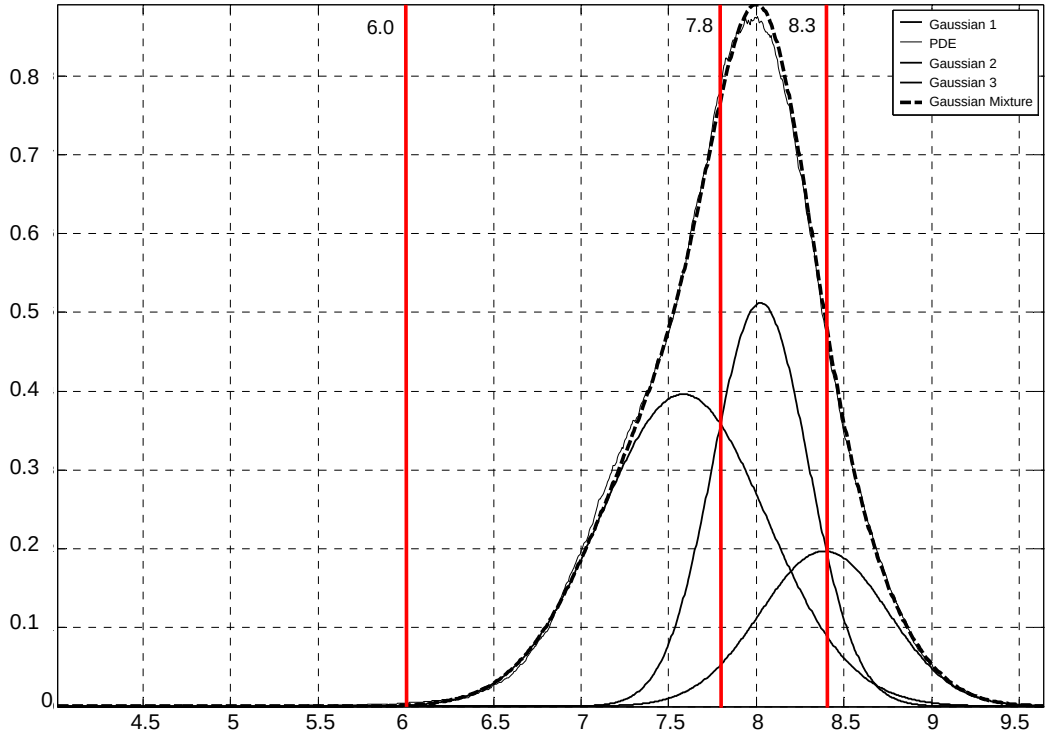
Weiterhin wurde geprüft, ob es sich um Daten handelt, deren Logarithmus normal verteilt ist. In erster Annäherung ermöglicht die Logarithmierung der vorhandenen Daten eine Modellierung des Verteilungsverlaufes.

Der QQ-Plot der Berechnungsgröße ‚AuslastungGebFreiflaeche' zeigt hierzu, dass die logarithmierten Daten zwar einer Geraden folgen, doch gewisse Unebenheiten bestehen. Mit Blick auf die dazugehörige PDE der logarithmierten Daten (blaue Kurve) und der Normalverteilung (rosa Kurve) wird deutlich, dass gerade im Flankenbereich des Kurvenverlaufes keine deutliche Überdeckung vorliegt. Um die besondere Komplexität des Verteilungsverlaufes der Variable ‚AuslastungGebFreiflaeche' zu berücksichtigen, bietet sich ein Modellierungsansatz auf Grundlage von Gauss-Mixturen zusätzlich an. Die Untersuchung der drei Ausgangsgrößen deutet ebenfalls auf die Anwendungsmöglichkeit eines komplexeren Modellierungsansatzes (siehe PDE und QQ-Plot). Die Gebäude- und Freifläche (‚FlaecheGebFrei') wurde darüber hinaus mit wurzeltransformierten Daten (ohne Abbildung) untersucht, jedoch führte dies zu keiner sinnvollen Modellierung der Verteilung.

Das Regularisierungsprinzip dieser Arbeit sei wie folgt zusammengefasst: Wenn die Verteilung einer Variablen nicht a priori bekannt ist, muss die Verteilung der Daten empirisch festgestellt werden. Die Sichtung der Daten erfolgt im Wesentlichen anhand von statistischen Verfahren, die eine vorgelegte Verteilung mit einer standardisierten Verteilung, z.B. der Normalverteilung vergleichen. Das Ziel besteht darin, dass die Untersuchungsvariablen einem gleichen Verteilungsverlauf folgen. Werden beispielsweise Abweichungen von Standardnormalverteilungen festgestellt, so gilt es entsprechende Umformungen (Transformationen) festzulegen, mit denen die Daten in eine bekannte Verteilung transformiert werden können. Durch Transformation kann auf die empirische Verteilung geschlossen werden. Es empfiehlt sich, die Plausibilität der Verteilung auch aus der Datenquelle heraus zu verifizieren. In dieser Arbeit wird die Verteilungsuntersuchung durch den Gauß-Ansatz ergänzt, der dazu dient, einen als komplex erkannten Verteilungsverlauf genauer zu beschreiben.

Gaußs-Mixtur-Modelle (siehe zusätzlich Abschnitt 2.3.5) verwenden Kombinationen von Gaußverteilungen bzw. es werden einzelne Gaußsche Normalverteilungen additiv zusammengesetzt, um einen Verteilungsverlauf abzubilden. Mit Hilfe einer derartigen Modellierung ist es möglich, Entscheidungsgrenzen für Klassifizierungsvorgänge aufzubauen. Mit Hilfe des EM-Algorithmus lassen sich Verteilungen automatisiert durch Gaußverteilungen modellieren. Da diese Lösung besonders von den Initialisierungsparametern abhängt, handelt es sich immer nur um ein lokales Optimum. Es werden deshalb stets mehrere Lösungen (2 bis 5 Gaußverteilungen) automatisiert berechnet. Darüber hinaus besteht die Möglichkeit einer manuellen Zusammensetzung von Gaußnormalverteilungen. Zur Qualitätskontrolle dient die Pareto-Dichte-Schätzung, die einen geeigneten Schätzer für Wahrscheinlichkeitsdichten von Mischungen von gaußverteilten Daten darstellt.

Die Verteilung der Berechnungsgröße (‚AuslastungGebFreiflaeche') wurde beispielhaft mit einem Gauss-Mixtur-Ansatz (gestrichelte Linie) modelliert. Die Entscheidungsgrenzen werden durch den Schnittpunkt der Kurvenverläufe der einzelnen Gaußverteilungen (Gaussian 1 bis 3) gebildet. Es werden mit drei Entscheidungsgrenzen (vertikale Markierung) 4 unterschiedliche Konzentrationsniveaus von Personen je km² Gebäude- und Freifläche in einer Gemeinde unterschieden (siehe Abbildung 4-1).



Klasse 1 (sehr geringe Konzentration): ‚AuslastungGebFreiflaeche' ≤500 Personen/km²
Klasse 2 (geringe Konzentration): 500 < ‚AuslastungGebFreiflaeche' ≤ 2500 Personen/km²
Klasse 3 (mittlere Konzentration): 2500 < ‚AuslastungGebFreiflaeche' ≤ 4000 Personen/km²
Klasse 4 (hohe Konzentration): ‚AuslastungGebFreiflaeche' > 4000 Personen/km²

Abbildung 4-1: GMM der Berechnungsgröße ‚LogAuslastungGebFreiflaeche' und Klasseneigenschaften

5 Deskriptive Erfassung räumlicher Struktureigenschaften mit Gauß-Mixturmodellen

5.1 Untersuchungsaufgabe: Polyzentralität und räumliche Vielfalt

Im Vergleich zu seinen europäischen Nachbarn hat Deutschland ein historisch gewachsenes polyzentrisches Städtesystem. Zwar konnte sich zu Beginn des 19. Jahrhunderts Berlin als Hauptstadt gegenüber den anderen Städten deutlich absetzen und in der Folgezeit zu den damaligen Weltstädten London, Paris und Wien aufschließen, jedoch brach diese Entwicklung mit dem Zweiten Weltkrieg und der nachfolgenden Teilung Deutschlands ab.[494]

Im föderalen System der Bundesrepublik Deutschland bildete die Polyzentralität in den nachfolgenden Jahrzehnten eine wichtige Voraussetzung, um auf Basis der analytischen Zentrale-Orte-Theorie von WALTER CHRISTALLER[495] ein flächendeckendes normatives Konzept zur Raumordnung dauerhaft umzusetzen. Grundsätzlich dient das Zentrale-Orte-Konzept als Lenkungsansatz, um die Voraussetzungen für einen sozial gerechten und ökonomisch effizienten Einsatz von Ressourcen sowie eine Begrenzung des Verbrauchs von Naturgütern zu schaffen.[496] Im zentralistischen Osten ist zunächst der Polyzentralität eine geringere Bedeutung im Rahmen der Raumordnungspolitik beigemessen worden. Nach 1990 ist auch für die neuen Länder das Zentrale-Orte-Konzept als zentrales Element der Raumordnung in Verwendung. Im Jahr 2003 existieren ca. 160 zentrale Orte in Deutschland, die einen unterschiedlichen Einfluss auf den umgebenden Bezugsraum ausüben.

Neben dem Ziel der Konvergenz der Lebensverhältnisse in den Teilräumen stand als weiteres Ziel der Raumordnungspolitik in Deutschland der Erhalt der räumlichen Vielfalt und Landschaften im Blickpunkt, um besondere Qualitäten langfristig zu stärken und hervorzuheben. An dieser Stelle sei auf das Grundgesetz verwiesen, welches neben dem Prinzip des regionalen Ausgleichs auch den regionalen Wettbewerb berücksichtigt.[497]

Die Aspekte Konvergenz und regionale Vielfalt ermöglichten es in der Vergangenheit, unterschiedliche raumordnungspolitische Schwerpunkte zu setzen, wobei oftmals gerade die Angleichung von Lebensverhältnissen als besonders relevant erkannt wurde. Historisch sei in Deutschland beispielsweise auf die Debatte um die Überwindung des Stadt-Land-Gefälles (60er- und 70er-Jahre), die Problematisierung des Nord-Süd-Gefälles (80er-Jahre) oder die Angleichung der Lebensverhältnisse in Ost- und West (90er-Jahre) verwiesen.

---

494 Vgl. ARING [2005, S. 7]
495 Vgl. CHRISTALLER [1933]
496 Vgl. BLOTEVOGEL [2002, S. 21], ARL [2002], BLOTEVOGEL [2005]
497 Vgl Grundgesetz für die Bundesrepublik Deutschland: Artikel 107 (Länderfinanzausgleich), Artikel 91a (Gemeinschaftsaufgaben) und Artikel 104a (Finanzhilfen des Bundes).

Gerade durch die Wiedervereinigung ist aus Sicht der Raumplanung einerseits ein größeres Betrachtungsgebiet entstanden und andererseits wurde auch aufgrund der in der Vergangenheit unterschiedlichen politischen Entwicklung ein Gebiet mit größerer Vielfalt geschaffen. ARING[498] verweist darauf, dass sich die Vorstellungsbilder von Zentrum und Peripherie sowie die Vorstellung von räumlicher Ausgewogenheit ändern und regionale ‚Spreizungen' zwischen Zentrum und Peripherie, wie sie in Frankreich oder Südschweden immer selbstverständlich waren, auch in Deutschland zukünftig genauer einzubeziehen sind.

Veränderte Rahmenbedingungen, wie der demographische Wandel oder die Globalisierung, nehmen zusätzlich Einfluss auf die räumlichen und funktionalen Erscheinungsbilder. Eine kritische Auseinandersetzung mit dem Zentrale-Orte-Konzept findet zunehmend Einzug in die politische Diskussion und Vorschläge für eine behutsame Weiterentwicklung entstehen.[499]

Oft sind Veränderungen erst über längere Zeit genauer einzuordnen und zu beurteilen. Teilweise sind hierzu gewohnte Sichtweisen zu verlassen, um Erkenntnisse aus bestehenden Raumstrukturen zu gewinnen. Die großen Veränderungen im Umland der großen Städte in den 90er-Jahren stellen dafür ein Beispiel dar. Diese können einerseits als fortschreitende Entwicklung eines stattfindenden Suburbanisierungsprozesses betrachtet werden und andererseits aber auch als eine neue Qualität der stadtregionalen Entwicklung mit den Begriffen ‚Zwischenstadt' und ‚Postsuburbania' belegt werden.[500] SIEDENTOP[501] verweist in 2003 darauf, dass die post-suburbanen Räume sich noch nicht von den Kernstädten umfassend entkoppelt haben und nur wenige Gemeinden von einer Dezentralisierung der Beschäftigung partizipieren. MOTZKUS und ARING[502] heben ein sich sowohl in quantitativer als auch qualitativer Hinsicht verbessertes Beschäftigungsangebot im suburbanen Raum hervor. Neben der bereits vertrauten Suburbanisierung des Wohnens, der Arbeitsstätten und der Einzelhandelsstandorte wurden in den 90er-Jahren Bildungs-, Kultur- und Freizeiteinrichtungen von Dezentralisierungstendenzen erfasst.[503]

In diesem Kapitel ist zu klären, inwieweit eine Ausdifferenzierung des Gemeindesystems nach Beschäftigungsschwerpunkten möglich ist und eine Analyse der Gemeinden nach ausgewählten raumstrukturellen Eigenschaften die Betrachtungsschärfe vergrößert.

---

498 Vgl. ARING [2005, S. 7]

499 Vgl. BMVBW [2006]: Eine Leitbilddiskussion einer Expertenkommission ist im Hinblick auf Handlungsstrategien für die Anpassung des Zentrale-Orte-Konzeptes von der 32. Ministerkonferenz für Raumordnung aufgegriffen und in einem raumordnungspolitischen Orientierungsrahmen benannt worden.

500 Vgl. SIEVERTS [1997] und ARING [1999]

501 Vgl. SIEDENTOP et al. [2003, S.96]

502 Vgl. MOTZKUS [2000, S.269] und ARING [2001, S. 117]

503 Vgl. BURDACK/HERFERT [1998, S. 30] und HESSE/BRUNS [2000, S.4]

5.2 Herleitung und Darstellung der Ergebnisse

Mit Blick auf die zuvor geschilderte Untersuchungsaufgabe wird ein Ansatz verfolgt, der einerseits eine Einteilung von Gemeinden nach Beschäftigungsschwerpunkten und dem Grad der Diversität ermöglicht, andererseits werden siedlungsraumbezogene Kenngrößen auf ihre Eignung zur Klassenbildung geprüft. Die Objekte innerhalb einer Klasse sollen sich dabei möglichst ähnlich sein und zu Objekten anderer Klassen deutliche Unterschiede aufweisen.

5.2.1 Beschäftigungsschwerpunkte und Diversität nach Wirtschaftszweigen

Es ist die Frage zu klären, wie viele Beschäftigungsschwerpunkte (Diversität) in einer Gemeinde existieren und von welcher Art diese sind. Ausgangspunkt zum Aufbau der Variablenstruktur bildet die Statistik der sozialversicherungspflichtig Beschäftigten. Es ist darauf hinzuweisen, dass die Statistik die Beamten und Selbstständigen nicht erfasst und dadurch eine gewisse Untererfassung der Erwerbstätigkeit bei dieser Untersuchung vorliegt.[504] Die Statistik unterscheidet 17 Wirtschaftszweige[505] gemäß Tabelle 5-1, die folgende Anteilswerte an den sozialversicherungspflichtig Beschäftigten insgesamt besitzen.

WZ	Bezeichnung	Aggregation	Anteil
A	Land- und Forstwirtschaft	(A,B)	1,19 %
B	Fischerei und Fischzucht	(A,B)	
C	Bergbau und Gewinnung von Steinen und Erden		0,45 %
D	Verarbeitendes Gewerbe		25,74 %
E	Energie- und Wasserversorgung		0,93 %
F	Baugewerbe		6,48 %
G	Handel, Instandhaltung, Reparatur von Kraftfahrzeugen, Gebrauchsgütern		14,95 %
H	Gastgewerbe		2,84 %
I	Verkehr und Nachrichtenübermittlung		5,53 %
J	Kredit- und Versicherungsgewerbe		3,90 %
K	Grundstücks- / Wohnungswesen, Vermietung beweglicher Sachen		11,56 %
L	Öffentliche Verwaltung, Verteidigung, Sozialversicherung	(L,Q)	6,46 %
M	Erziehung und Unterricht	(M,N,O,P)	19,96 %
N	Gesundheits-, Veterinär- und Sozialwesen	(M,N,O,P)	
O	Erbringung von sonstigen öffentlichen und persönlichen Dienstleistungen	(M,N,O,P)	
P	Private Haushalte mit Hauspersonal	(M,N,O,P)	
Q	Exterritoriale Organisationen und Körperschaften	(L,Q)	s.o.
Summe der Anteilswerte der Beschäftigungsgruppen an den Beschäftigten insgesamt (A-Q)			100,00 %

Tabelle 5-1: Klassifikation der Wirtschaftszweige des Statistischen Bundesamtes (WZ2003)[506]

[504] Vgl. Statistische Berichte (Bevölkerung u. Erwerbstätigkeit): Erfasst sind ca. 80 % der Erwerbstätigen. Ein vollständiger Datensatz zur Erwerbstätigkeit ist derzeit auf Gemeindeebene nicht erhältlich.

[505] Aus Datenschutzgründen wurden die gemeindebasierten Daten um Zahlenwerte kleiner 3 je Gemeinde von der Bundesanstalt für Arbeit bereinigt. In kleineren Gemeinden mit wenigen Beschäftigten wirkt sich dies auf eine Erfassung der Beschäftigungsstruktur besonders deutlich aus. Die Fehlstellen wurden im Ausgangsdatenmaterial jedoch nicht markiert, so dass der Versuch der Rückrechnung z.B. anhand der vollständigen Ländersummen oder Techniken der Fehlstellenbehandlung nicht zielgerichtet durchführbar ist. Zu vermuten ist aus einer Sicht auf die Rohdaten, dass die Wirtschaftszweige Fischerei und Fischzucht, Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung, Kredit-/Versicherungsgewerbe sowie Private Haushalte mit Hauspersonal besonders davon betroffen sind.

[506] Eigene Bearbeitung auf Basis von Informationen der Klassifikation der Wirtschaftszweige (WZ2003) des Statistischen Bundesamtes. Die Aggregation ist beim Statistischen Bundesamt enthalten.

Das Statistische Bundesamt unterscheidet in einer höher aggregierten Form die Gruppen Land- und Forstwirtschaft, Fischerei (Gruppe 1), Produzierendes Gewerbe (Gruppe 2), Handel, Gastgewerbe und Verkehr (Gruppe 3) sowie sonstige Dienstleistungen (Gruppe 4). In dieser Arbeit wird die Beschäftigungsstruktur mit acht Variablen abgebildet. Die Variablen berechnen sich für definierte Beschäftigungsgruppen als Anteilswerte an der Gesamtbeschäftigung und addieren sich zu 100 %. Die Definition der Beschäftigungsgruppen orientiert sich an der dargestellten Systematik des Statistischen Bundesamtes. Da in Deutschland ca. 65 % der sozialversicherungspflichtig Beschäftigten im Tertiären Sektor arbeiten, wird für diesen Bereich im Hinblick auf eine differenzierte Betrachtung nach funktionalen Eigenschaften der Gemeinden eine größere Variablenanzahl gebildet, während der Primäre Sektor (Wirtschaftszweige A-B) nur mit einer Variablen abgebildet wird.

Das Untersuchungsmodell wird mit Hilfe eines Gauß-Ansatzes (siehe Abschnitt 2.3.4) entwickelt, der es ermöglicht, Entscheidungsgrenzen für jede einzelne Variable aufzustellen. Der EM-Algorithmus wird eingesetzt, um das Mixturenmodell an die empirischen Verteilungen anzupassen. Modelliert werden die Variablenwerte mit einer Mischung von jeweils zwei Gauss-Verteilungen. Da der EM-Algorithmus nur ein lokales Optimum entdeckt, werden die Lösungen mehrfach berechnet, um gefundene Entscheidungsgrenzen zu bestätigen. Die Plausibilitätsprüfung erfolgt zusätzlich mit der Pareto Density Estimation (Wahrscheinlichkeitsdichte). Abbildung 5-1 zeigt an der Variable ‚Sonstige Beschäftigte im Produzierenden Gewerbe' die Festlegung der Entscheidungsgrenze von 22,8.

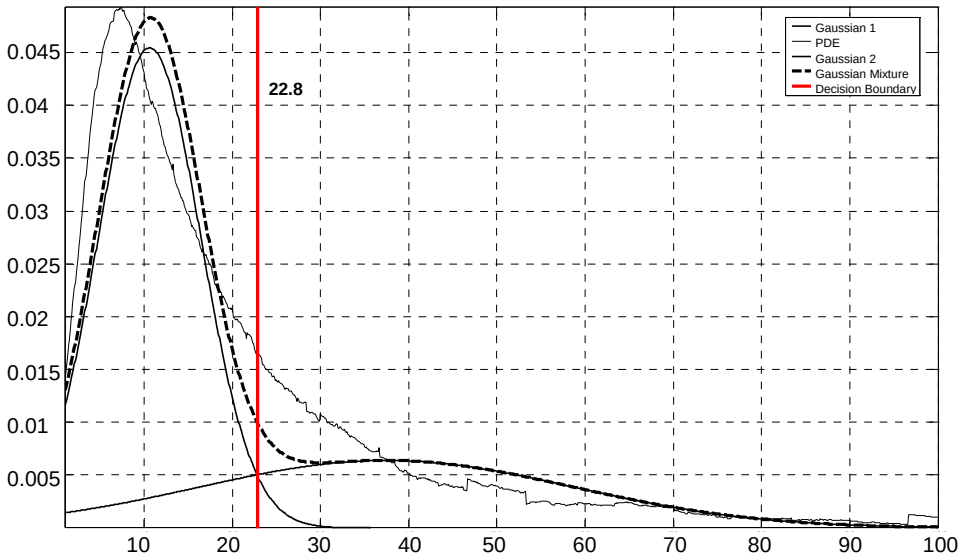


Abbildung 5-1: GMM der Variable ‚Sonstige Beschäftigte im Produzierenden Gewerbe'[507]

[507] Quelle: Berechnungsergebnis (MATLAB® 7.0.1). Im Nebenteil B sind die für diese Untersuchung erzeugten Gauß-Mixtur-Modelle für alle Variablen hinterlegt.

Die Entscheidungsgrenze bildet das Kriterium, um besonders stark ausgeprägte Beschäftigungsverhältnisse in einer Gemeinde zu identifizieren und damit zwischen ‚großen' und ‚weniger großen' Beschäftigungsanteilen innerhalb eines Wirtschaftszweiges zu unterscheiden. Tabelle 5-2 enthält ermittelte Entscheidungsgrenzen für sieben Untersuchungsvariablen. ‚Beschäftigte ohne Angaben' (Null in Daten = 99,3 %) sind nicht berücksichtigt.

Variable	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**
Semantik	‚Land- / Forstwirtschaft und Fischerei'	‚Verarbeitendes Gewerbe'	‚Sonstiges Produzierendes Gewerbe'	‚Handel, Gastgewerbe, Verkehr'	‚Kredit-/Versicherungs-, Grundstückswesen'	‚Öffentliche Verwaltung'	‚Öffentliche, private Dienstleistung'
Aggregation (WZ)	A-B	D	C, E, F	G, H, I	J, K	L, Q	M, N, O, P
Entscheidungsgrenze	**9,4**	**24**	**22,8**	**38,2**	**16,6**	**13,4**	**27,5**
Klassifizierungsschlüssel	5	5	1	1	1	1	1

Tabelle 5-2: Übersicht zu den Untersuchungsvariablen der Beschäftigungsstruktur (WZ-2003)

In der Tabelle ist zusätzlich ein Beispiel zu einem Klassifizierungsschlüssel aufgeführt, der auf den ermittelten Entscheidungsgrenzen basiert. Der Klassifizierungsschlüssel besteht aus einer siebenstelligen Ziffer, die sich von links nach rechts auf die jeweilige Beschäftigungsgruppe bzw. Variable bezieht. In dem siebenstelligen Ziffernblock sind die Stellen für eine Gemeinde mit einer besonders ausgeprägten Beschäftigungsgruppe durch ein Symbol markiert. Es werden zwei Symbole verwendet, wobei die ‚1' einen Wert unterhalb und die ‚5' einen Wert oberhalb der Entscheidungsgrenze ausdrückt. Das Symbol ermöglicht somit eine Aussage, ob der Anteilswert einer Beschäftigungsgruppe in einer Gemeinde oberhalb oder unterhalb der durch den Gauß-Ansatz bestimmten Entscheidungsgrenze liegt.

Insgesamt wurden anhand des Klassifizierungsschlüssels 72 verschiedene Kombinationen entdeckt. Wie sich anhand der berechneten Ergebnisse gezeigt hat, ist es möglich, Gemeinden mit einem eindeutigen Beschäftigungsschwerpunkt zu ermitteln und nach der Beschäftigungsart zu strukturieren. Darüber hinaus sind Gemeinden auf Grundlage dieses Klassifizierungsansatzes ermittelbar, die keinen eindeutigen Beschäftigungsschwerpunkt aufweisen bzw. solche, die durch eine besondere Diversität charakterisierbar sind. In Folge eines Aggregationsprozesses wurden aus inhaltlichen Überlegungen sieben Klassen mit eindeutigem Beschäftigungsschwerpunkt gebildet. Bei zusätzlichem Erkenntnisbedarf ist mit Blick auf weitere Klassen mit zwei oder drei Beschäftigungsschwerpunkten eine feinteiligere Betrachtung möglich. Zu beachten ist allerdings, dass es sich bei der gewählten Aggregation bereits um die objektstärksten Klassen handelt.

Abbildung 5-2 zeigt das verortete Ergebnis von sieben Klassen mit eindeutigem Schwerpunkt. Auf regionaler Ebene sind deutliche Unterschiede erkennbar. Die Mehrzahl der Gemeinden besitzt eindeutige Beschäftigungsschwerpunkte. Des Weiteren existieren sowohl in Ost- als auch in Westdeutschland Gemeinden, die Diversität oder keinen Schwerpunkt zeigen.

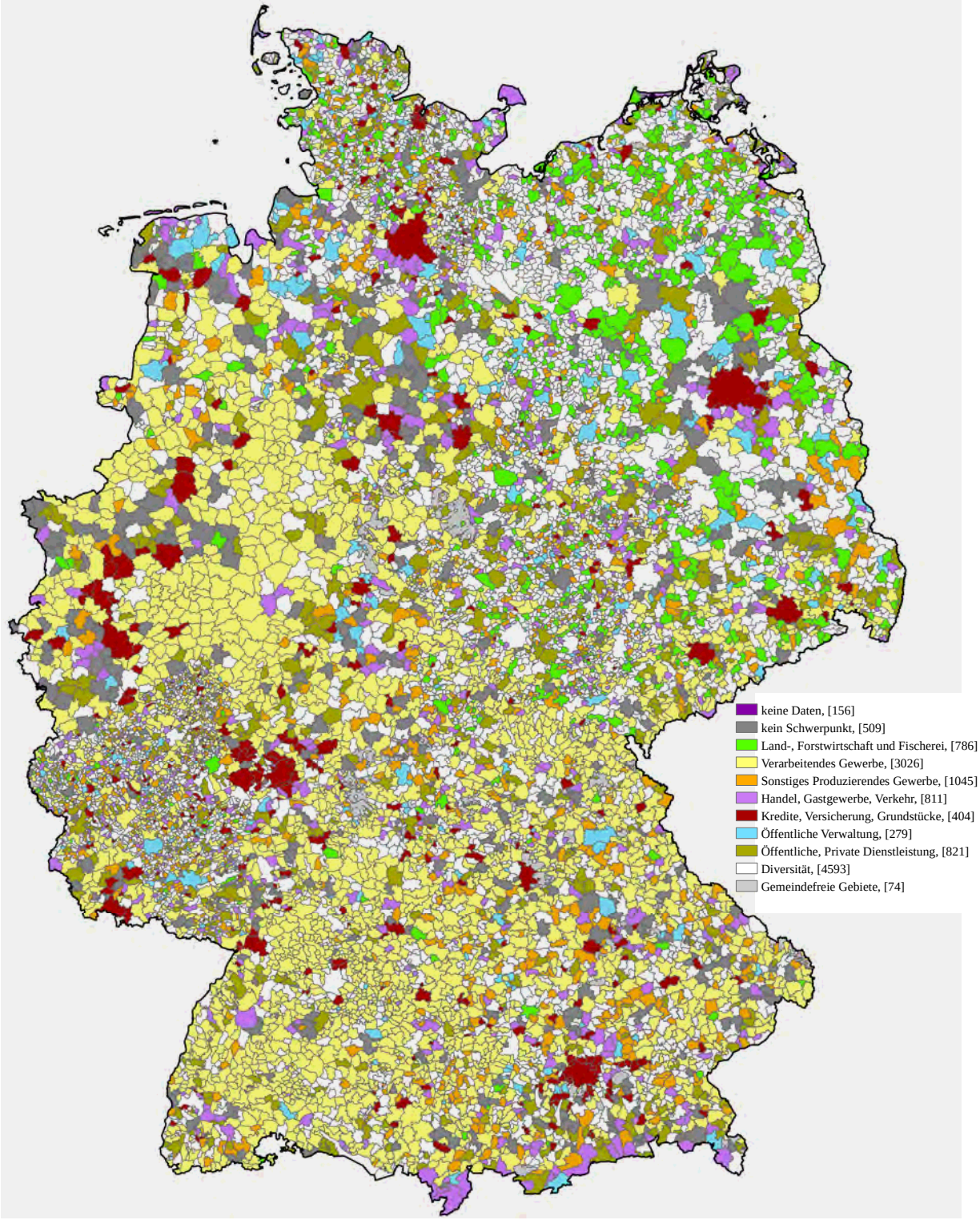


Abbildung 5-2: Klassifizierung nach Wirtschaftszweigen (Beschäftigungsschwerpunkte) [508]

[508] Die Legende zeigt zusätzlich die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Gesondert unterschieden nach der Anzahl der Beschäftigungsschwerpunkte (z.B. 1, 2, oder 3) werden die Gemeinden in Abbildung 5-3, um auf diese Weise gezielt Gemeinden nach dem Grad der Diversität bzw. Gemeinden mit keinem eindeutigen Schwerpunkt abzubilden.

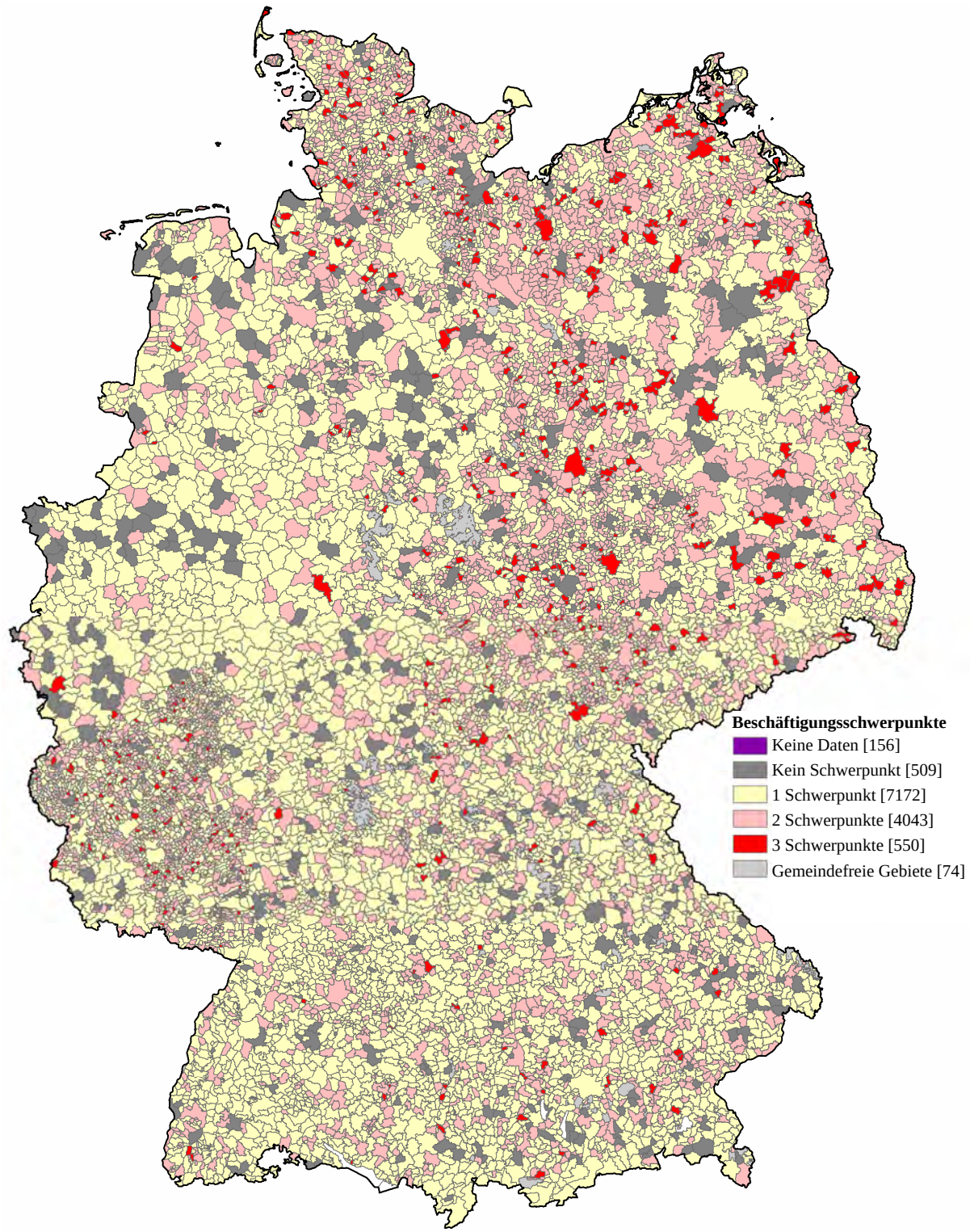


Abbildung 5-3: Klassifizierung nach Wirtschaftszweigen (Beschäftigungsdiversität)[509]

[509] Die Legende zeigt zusätzlich die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Das verortete Ergebnis der sieben Klassen mit eindeutigem Beschäftigungsschwerpunkt lässt erkennen, dass Schwerpunkte im Primären Sektor bei 786 Gemeinden existieren und ungefähr zwei Drittel dieser Gemeinden auf das ostdeutsche Untersuchungsgebiet entfallen. Ca. 22 % der 786 Gemeinden befinden sich in Mecklenburg-Vorpommern, weitere 17 % liegen in Sachsen-Anhalt und 15 % der Gemeinden sind Schleswig-Holstein zuzuordnen. In Mecklenburg-Vorpommern hat ca. ein Fünftel der Gemeinden einen Beschäftigungsschwerpunkt im Primären Sektor.

Einen Schwerpunkt im Verarbeitenden Gewerbe besitzen ca. 25 % (3026 Gemeinden) der 12430 Gemeinden in Deutschland. In diesem Wirtschaftszweig sind darüber hinaus auch 25 % der sozialversicherungspflichtig Beschäftigten tätig. Es fällt auf, dass Beschäftigungsschwerpunkte in diesem Bereich verstärkt im westdeutschen Bundesgebiet existieren (2529 Gemeinden), wobei in den Bundesländern NRW und Baden-Württemberg sogar fast 60 % der dortigen Gemeinden diesen Schwerpunkt aufweisen. In Hessen und Bayern sowie dem Saarland sind mehr als ein Drittel der Gemeinden durch diesen Schwerpunkt zu charakterisieren. Im ostdeutschen Bundesgebiet zeigen im Bundesland Sachsen ca. 30 % der 519 Gemeinden einen Schwerpunkt im Verarbeitenden Gewerbe.

Mit einem Beschäftigungsschwerpunkt im sonstigen produzierenden Gewerbe wurden 1045 Gemeinden ermittelt. Die Bundesländer Bayern, Rheinland-Pfalz und Hessen zeigen eine besondere Ausprägung. Exemplarisch hervorzuheben ist der Aspekt, dass die Gemeinden mit einer traditionell bekannten Prägung durch den Wirtschaftszweig ‚Bergbau, Gewinnung von Steinen und Erden' auch dem Beschäftigungsschwerpunkt im sonstigen produzierenden Gewerbe zugeordnet werden, falls nach wie vor eine hohe Beschäftigung in diesem Bereich besteht.

Betrachtet man die Gemeinden mit einem eindeutigen Beschäftigungsschwerpunkt im Dienstleistungsbereich, so zeigen insgesamt 20 % der 12430 Gemeinden in Deutschland einen eindeutigen Schwerpunkt, wobei bei diesem Untersuchungsansatz vier Arten unterschieden werden. Der erste Beschäftigungsschwerpunkt bezieht sich auf ‚Handel, Gastgewerbe, Verkehr' und wurde bei 811 Gemeinden erkannt. Die Umlandgemeinden der Kernstädte besitzen innerhalb der Regionen Rhein-Main, Hannover, Berlin, München und Hamburg vermehrt einen derartigen Beschäftigungsschwerpunkt. Darüber hinaus sind in süddeutschen Grenzregionen vermutlich aufgrund verstärkter Handelsbeziehungen und einem besonders starken Gastgewerbe mehrere Gemeinden durch diesen Schwerpunkt zu beschreiben.

Der Beschäftigungsschwerpunkt im Bereich ‚Kredite, Versicherung, Grundstückswesen' ist insbesondere in den Kernstädten (z.B. Frankfurt, Köln, München oder Berlin) anzutreffen. Fast 35 % der 404 ermittelten Beschäftigungsschwerpunkte beziehen sich auf das Kernstadtgebiet bzw. kernstadtnahe Umland. In Hessen zeigt im Vergleich zu den anderen Bundesländern ein relativ großer Anteil der Gemeinden diesen Beschäftigungsschwerpunkt. Im Jahr 2000 verweist MOTZKUS[510] darauf, dass die Kernstädte eine große Bedeutung in Bezug auf das Kredit- und Versicherungsgewerbe in Metropolregionen Westdeutschlands haben.

Der Beschäftigungsschwerpunkt ‚Öffentliche Verwaltung' konnte bei 279 Gemeinden erkannt werden. Es ist bei diesem Schwerpunkt kein spezifisches Verteilungsmuster der Gemeinden z.B. auf die Kernstädte oder das kernstadtnahe Umland festzustellen. Bezogen auf die Gemeindeanzahl der jeweiligen Bundesländer insgesamt existieren mit dem Schwerpunkt ‚Öffentliche Verwaltung' anteilsmäßig im bundesweiten Vergleich relativ viele Gemeinden in Brandenburg, Niedersachsen und Sachsen-Anhalt.

Die Beschäftigungsschwerpunkte zu sonstigen öffentlichen und privaten Dienstleistungen wurden in 821 Fällen ermittelt. In den dazugehörigen Wirtschaftszweigen arbeiten ca. 20 % der sozialversicherungspflichtig Beschäftigten. Es handelt sich damit um den zweitstärksten Beschäftigungszweig in Deutschland. Die Gemeinden mit diesem Schwerpunkt verteilen sich bundesweit in einem unregelmäßigen Muster. Bezogen auf die Gesamtzahl der Gemeinden in den jeweiligen Bundesländern zeigen das Saarland, Hessen, Niedersachsen und Sachsen am deutlichsten diesen Beschäftigungsschwerpunkt.

Zwei Klassen mit besonderen Eigenschaften sind im Folgenden genauer zu charakterisieren. Eine Klasse dient der Ermittlung von Gemeinden, die nicht durch einen Beschäftigungsschwerpunkt geprägt sind und dadurch eher Eigenschaften der Diversität aufweisen. Hierbei handelt es sich um 4593 Gemeinden, die entweder zwei oder drei verschiedene Beschäftigungsschwerpunkte aufweisen. Eine weitere Klasse erfasst darüber hinaus 509 Gemeinden, die entweder eine sehr große Diversität besitzen (>3 Beschäftigungsschwerpunkte) oder gar keinen Beschäftigungsschwerpunkt zeigen. In Bezug auf die Diversität sind im Verhältnis auf die Gesamtzahl der Gemeinden in den einzelnen Bundesländern in Baden-Württemberg, Bayern und Hessen weniger Gemeinden mit dieser Eigenschaft aufzufinden. Dagegen besitzt ungefähr die Hälfte der Gemeinden in Schleswig-Holstein, Rheinland-Pfalz, Brandenburg, Mecklenburg-Vorpommern und Sachsen-Anhalt sowie Thüringen mindestens zwei Beschäftigungsschwerpunkte.

[510] Vgl. MOTZKUS [2000, S. 270 ff.]

Betrachtet man das bevölkerungsreichste Bundesland Nordrhein-Westfalen gesondert, so sind hier im bundesweiten Vergleich sehr wenige Gemeinden zu identifizieren, die über eine Diversität mit zwei oder drei Beschäftigungsschwerpunkten verfügen. Im Verhältnis zu den anderen Bundesländern sind hier aber sehr viele Gemeinden dadurch beschreibbar, dass sie gar keinen eindeutigen Beschäftigungsschwerpunkt aufweisen (53 von 396 Gemeinden insgesamt). Bei diesen Gemeinden handelt es sich oftmals um Gemeinden mit ehemals Bergbau oder Schwerindustrie, die heutzutage bezogen auf Eigenschaften in den Wirtschaftssektoren (siehe Abbildung 5-7) zunehmend eine tertiäre Prägung erfahren. Ähnliches kann für das Saarland festgestellt werden. Weitere Gemeinden ohne eindeutigen Beschäftigungsschwerpunkt, sind im Umland von Hannover und Berlin in größerer Anzahl zu finden.

Abbildung 5-4 zeigt die Gemeinden mit eindeutigem Beschäftigungsschwerpunkt, wobei diese nach ihrer Häufigkeit innerhalb der gültigen Zentrale-Orte-Kategorie aufgetragen sind.

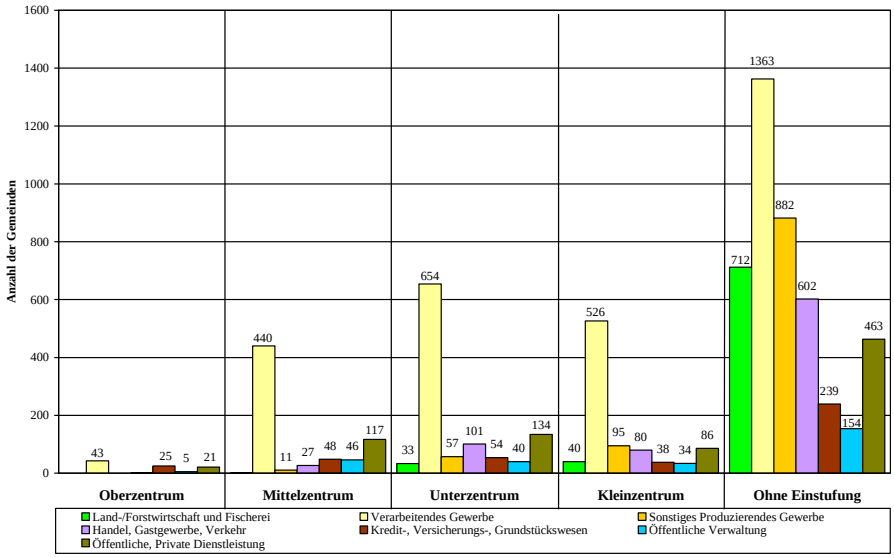


Abbildung 5-4: Gemeinden in Deutschland mit Beschäftigungsschwerpunkten je Zentrale-Orte-Kategorie

Gemeinden mit Beschäftigungsschwerpunkt im Verarbeitenden Gewerbe sind in allen fünf Raumordnungskategorien in großer Häufigkeit vorhanden. Die Beschäftigungsschwerpunkte des Primären Sektors und des sonstigen Produzierenden Gewerbes wurden in größerer Zahl bei Gemeinden ohne zentralörtliche Einstufung ermittelt. Die bereits in der Verortung erkannte Bedeutung des Kredit-, Versicherungs- und Grundstückswesens für die Kernstädte ist mit Blick auf die Kategorie Oberzentrum wiedererkennbar. Ein großer Anteil der Oberzentren verfügt über diesen eindeutigen Beschäftigungsschwerpunkt. Bezogen auf die Beschäftigungsschwerpunkte ist keine Unterscheidung zwischen Kernstadt und Umland in einem überproportionalen Maße möglich.

Abbildung 5-5 zeigt die Beschäftigungsschwerpunkte der Gemeinden nach Raumstrukturtypen (siehe Abschnitt 10.2.2). Im Hinblick auf das Verarbeitende Gewerbe existiert bei sehr vielen Gemeinden außerhalb der Agglomerationsräume und ihrer Kerne ein Beschäftigungsschwerpunkt in diesem Wirtschaftszweig. Bei den Gemeinden des Äußeren Zentralraums lassen sich nahezu 60 % der Gemeinden durch einen Beschäftigungsschwerpunkt im Verarbeitenden Gewerbe kennzeichnen. Im Zwischenraum mit Verdichtungsansätzen und auch im Peripherraum mit Verdichtungsansätzen liegt der Anteil der Gemeinden dieses Beschäftigungsschwerpunktes in ähnlicher Größenordnung. Im Inneren Zentralraum verfügen ca. 55 Prozent der Gemeinden über dienstleistungsorientierte Beschäftigungsschwerpunkte.

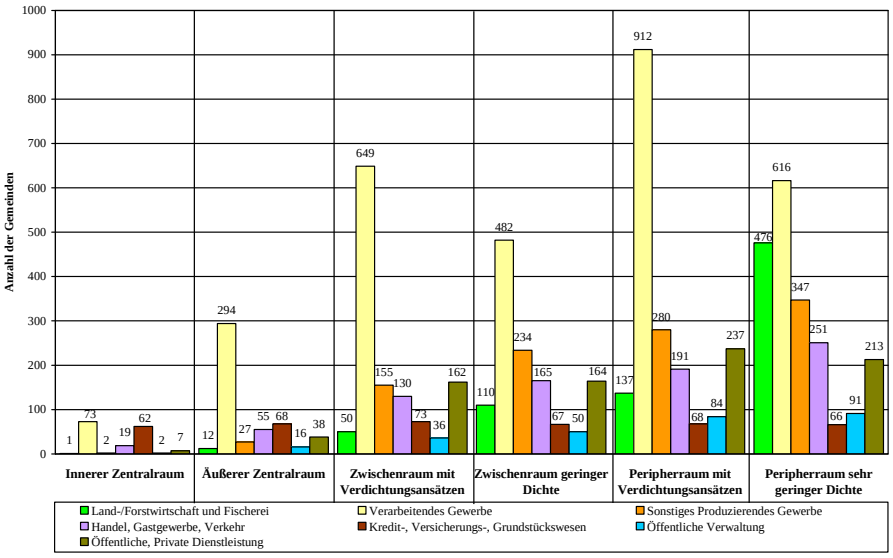


Abbildung 5-5: Beschäftigungsschwerpunkte je Raumstrukturtyp und Wirtschaftszweig

Zusammenfassend zeigt Abbildung 5-6 die Gemeinden im Hinblick auf ihre Diversität, jedoch nun differenziert nach Raumstrukturtypen. Die meisten Gemeinden mit mehr als einem Beschäftigungsschwerpunkt liegen im Zwischenraum oder Peripherraum.

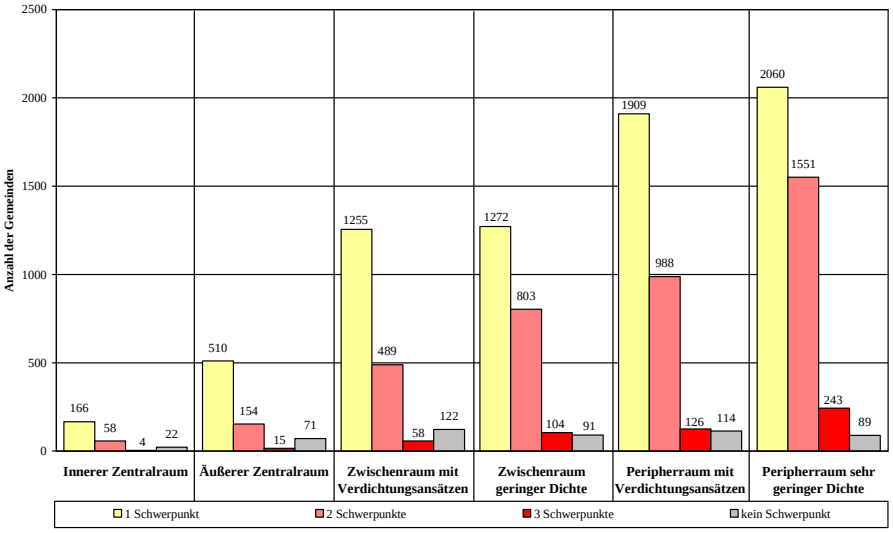


Abbildung 5-6: Beschäftigungsdiversität in den Gemeinden je Raumstrukturtyp

5.2.2 Beschäftigungsschwerpunkte nach Wirtschaftssektoren

Eine ergänzende Betrachtung zu den bereits berechneten Beschäftigungsschwerpunkten nach Wirtschaftszweigen bildet die hier dargestellte Untersuchung. Gefragt wird ebenfalls nach möglichen Beschäftigungsschwerpunkten von Gemeinden, jedoch basiert die Variablenstruktur auf der Gliederung nach den drei Wirtschaftssektoren.[511] Das Ziel besteht somit darin, in einer abstrakteren Weise Gruppen zu bilden, deren Objekte einen gleichen Beschäftigungsschwerpunkt besitzen und deutliche Unterschiede zu anderen Gruppen zeigen.

Die statistischen Daten der sozialversicherungspflichtig Beschäftigten werden nach den folgenden drei Wirtschaftssektoren gemäß Tabelle 5-3 gegliedert. Für das Jahr 2003 gelten in den Sektoren die dargestellten Anteilswerte der sozialversicherungspflichtig Beschäftigten.

Sektor	Bezeichnung	Anteilswert	Aggregation (WZ2003)	Entscheidungsgrenze
I.	Primärer Sektor	1,19 %	**(A-B)**	9,4 %
II.	Sekundärer Sektor	33,60 %	**(C-F)**	35,4 %
III.	Tertiärer Sektor	65,21 %	**(G-Q)**	29,9 %

Tabelle 5-3: Klassifikation der Wirtschaftszweige nach Wirtschaftssektoren (3-Sektorenmodell)[512]

Die Beschäftigungsstruktur wird mit vier Variablen abgebildet, wobei in gleicher Weise zu der Untersuchung der Wirtschaftszweige die vierte Variable eingesetzt wird, um die Beschäftigungsverhältnisse ohne Angabe zu selektieren. Die Variablen addieren sich in Bezug auf die Gesamtbeschäftigung zu 100 %. Mit Blick auf die Variablenstruktur und infolge der Dateninspektion, wird mit Hilfe des Gauß-Ansatzes (Abschnitt 2.3.4 und Berechnungsergebnisse im Nebenteil B) erneut eine Unterscheidung der Gemeinden nach ‚viel' bzw. ‚wenig' Beschäftigten in einem Wirtschaftssektor vorgenommen.

Der Beschäftigungsschwerpunkt im Primären Sektor deutet auf besonders ausgeprägte landwirtschaftliche Beschäftigungsverhältnisse hin. Anhand eines Beschäftigungsschwerpunktes im Sekundären Sektor, lässt sich ein hoher Besatz mit Industrie und Gewerbe in einer Gemeinde vermuten. Die im sekundären Sektor eingesetzten Arbeitskräfte verzeichnen im Durchschnitt eine höhere Produktivität als die Beschäftigten anderer Sektoren.

Ein Beschäftigungsschwerpunkt im Tertiären Sektor ist als Indiz für den Umfang der Versorgungsleistungen einer Gemeinde für ihre eigene Bevölkerung und für die Bevölkerung ihres Umlandes anzusehen.

---

511 JEAN FOURASTIÉ hat 1949 als erster ein Klassifizierungsschema der wirtschaftlichen Aktivitäten vorgeschlagen. Vgl. FOURASTIÉ [1952, S. 70 ff.]

512 Eigene Bearbeitung auf Basis von Informationen des Statistischen Bundesamtes

Abbildung 5-7 zeigt das Klassifizierungsergebnis nach Beschäftigungsschwerpunkten in den Wirtschaftssektoren. Zwischen Ost- und Westdeutschland existieren deutliche Unterschiede für Beschäftigungsschwerpunkte im Primären Sektor sowie bei der räumlichen Verteilung der Gemeinden, die keinen eindeutigen Schwerpunkt aufweisen.

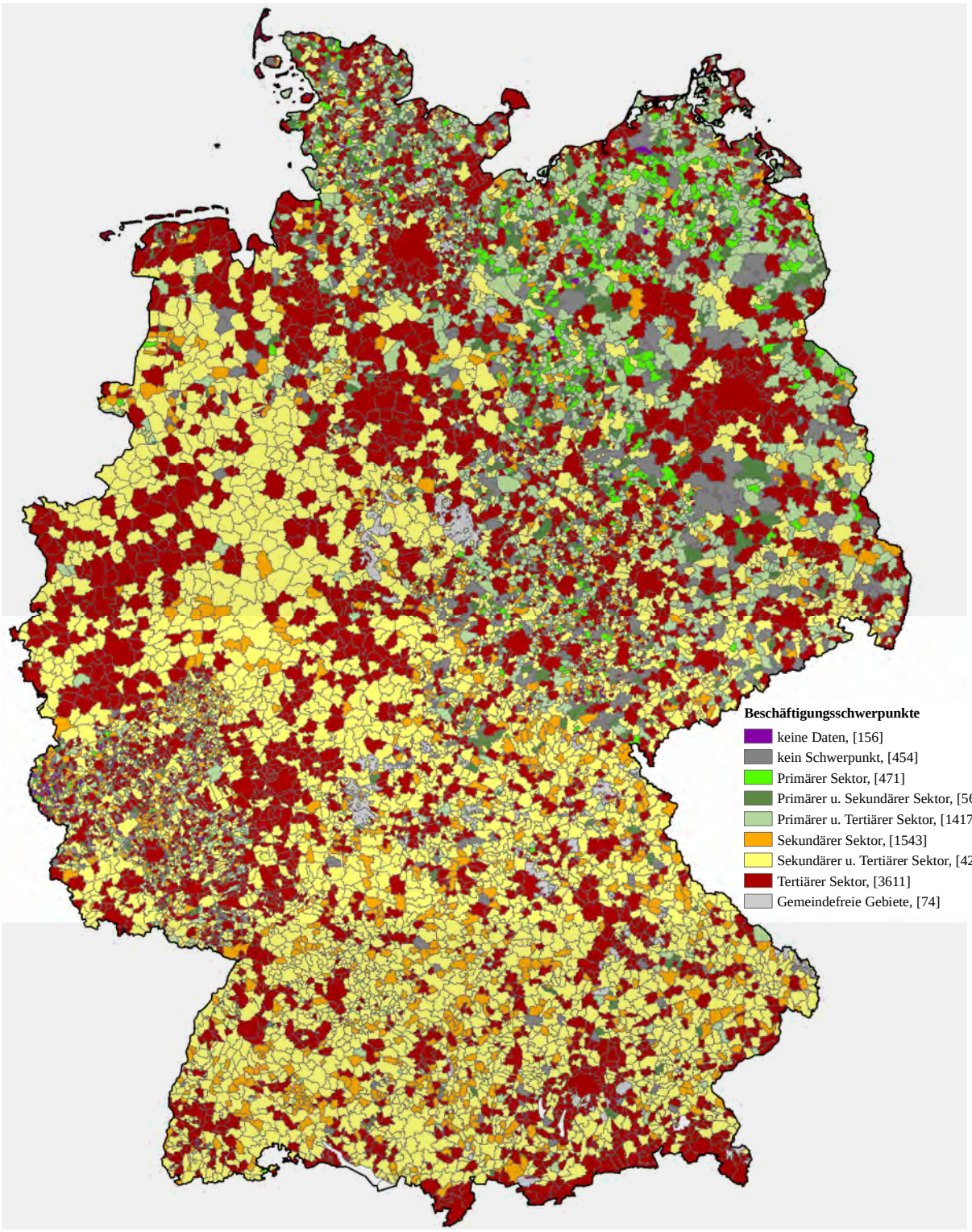


Abbildung 5-7: Klassifizierung nach Wirtschaftssektoren (Beschäftigungsschwerpunkte)[513]

513 Die Legende zeigt zusätzlich die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Ca. 30 % der 12430 Gemeinden weisen einen eindeutigen Beschäftigungsschwerpunkt im Tertiären Sektor auf. Bei ungefähr 12,5 % der Gemeinden ist ein eindeutiger Beschäftigungsschwerpunkt im Sekundären Sektor festzustellen und weniger als 5 % der Gemeinden lassen sich mit einem eindeutigen Beschäftigungsschwerpunkt im Primären Sektor identifizieren.

In Hinblick auf eindeutige Beschäftigungsschwerpunkte entfallen im Primären Sektor 323 der 471 Gemeinden auf den ostdeutschen Raum, wobei ca. 75 % dieser Gemeinden in Mecklenburg-Vorpommern und Sachsen-Anhalt liegen. Von den 148 Gemeinden in Westdeutschland entfallen 81 Gemeinden auf Schleswig-Holstein und 59 Gemeinden auf Rheinland-Pfalz. Die 1543 Fälle mit eindeutigen Beschäftigungsschwerpunkten im sekundären Sektor lassen sich nicht einem spezifischen Verteilungsmuster auf Gemeindeebene zuweisen. Auf Ebene der Bundesländer entfallen ca. 950 der insgesamt 1166 Gemeinden auf Baden-Württemberg, Bayern und Rheinland-Pfalz. Aus den zuvor verortet dargestellten Beschäftigungsschwerpunkten nach Wirtschaftssektoren ist ersichtlich, dass sich vermehrt die Beschäftigungsschwerpunkte im Tertiären Sektor auf die Agglomerationsräume und ihre Kerne konzentrieren. Die Räume Rhein-Main, Rhein-Ruhr sowie Berlin, Hamburg, Hannover und München sind durch eine größere Anzahl von Gemeinden mit Beschäftigungsschwerpunkten im Tertiären Sektor geprägt. Darüber hinaus enthalten aber auch agglomerationsferne bzw. ländliche Gebiete einige Gemeinden mit einem deutlich tertiärisierten Beschäftigungsprofil.

Betrachtet man die Gemeinden ohne Beschäftigungsschwerpunkt, so entfallen 207 Gemeinden auf den westdeutschen und 247 Gemeinden auf den ostdeutschen Untersuchungsraum. Diese Gemeinden zeigen in ihrem Beschäftigungsprofil für alle drei Wirtschaftsektoren einen Beschäftigungsanteil, der oberhalb der ermittelten Entscheidungsgrenzen liegt (Sektor I > 9,4 %, Sektor II > 35,4 % und Sektor III>29,9 %). Gerade im Großraum Berlin sind in den umgebenden Bundesländern vermehrt Gemeinden mit diesen Eigenschaften zu erkennen. Inwieweit konjunkturelle Veränderungen sich stärker oder schwächer auf Gemeinden ohne Beschäftigungsschwerpunkt auswirken, kann an dieser Stelle nicht behandelt werden.

Einen Beschäftigungsschwerpunkt in jeweils zwei der drei Wirtschaftssektoren zeigen 6195 Gemeinden. Davon besitzen 4218 Gemeinden Beschäftigungsschwerpunkte im sekundären und im tertiären Sektor. Ca. 80 % dieser Gemeinden liegen in Westdeutschland.

Zusätzlich zu der kartografischen Abbildung wird das Ergebnis der Beschäftigungsschwerpunkte im Folgenden in Relation zu den bestehenden Zentrale-Orte-Kategorien und den erst in jüngster Zeit eingeführten Raumstrukturtypen des BBR (siehe Abschnitt 10.2.2) gesetzt.

Abbildung 5-8 und Abbildung 5-9 zeigen die Verteilung der Gemeindeergebnisse auf die jeweiligen Raumordnungskategorien. Festzustellen ist, dass der Tertiäre Sektor als eindeutiger Beschäftigungsschwerpunkt nicht nur in den Ober- und Mittelzentren identifiziert wurde. Die Gliederung nach Innerem und Äußerem Zentralraum zeigt, dass die in diesen Raumordnungstypen enthaltenen Gemeinden mit großer Häufigkeit einen Beschäftigungsschwerpunkt im Tertiären Sektor besitzen. Auch der Zwischen- und der Peripherraum zeigen Beschäftigungsschwerpunkte im Tertiären Sektor und damit eine Prägung einer dienstleistungsorientierten Wirtschaft. Bei den 6195 Gemeinden, die mehr als einen eindeutigen Beschäftigungsschwerpunkt in den drei Sektoren aufweisen, ist festzustellen, dass es sich bei diesen Fällen mit großer Häufigkeit um Gemeinden ohne eine zentralörtliche Einstufung handelt. Viele der Beschäftigungsschwerpunkte im Primären Sektor entfallen auf diese Kategorie.

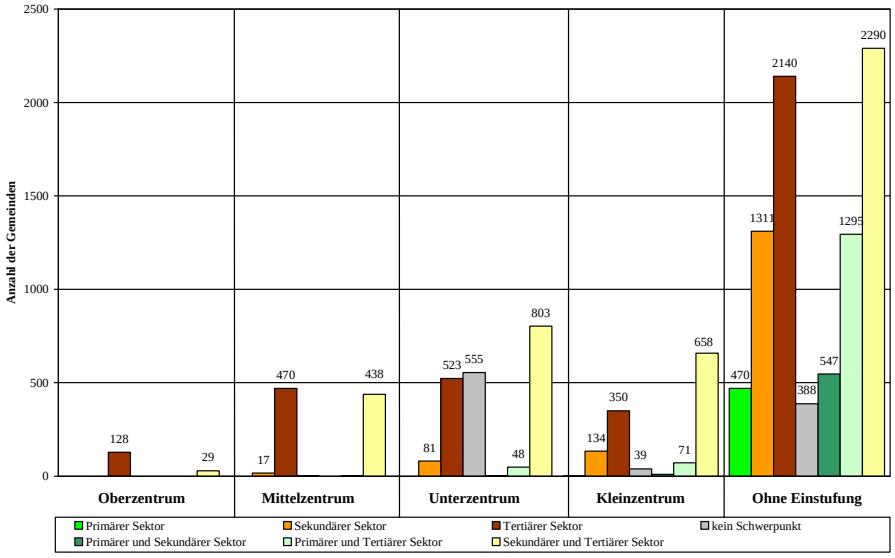


Abbildung 5-8: Beschäftigungsschwerpunkt je Zentrale-Orte-Kategorie und Wirtschaftssektor

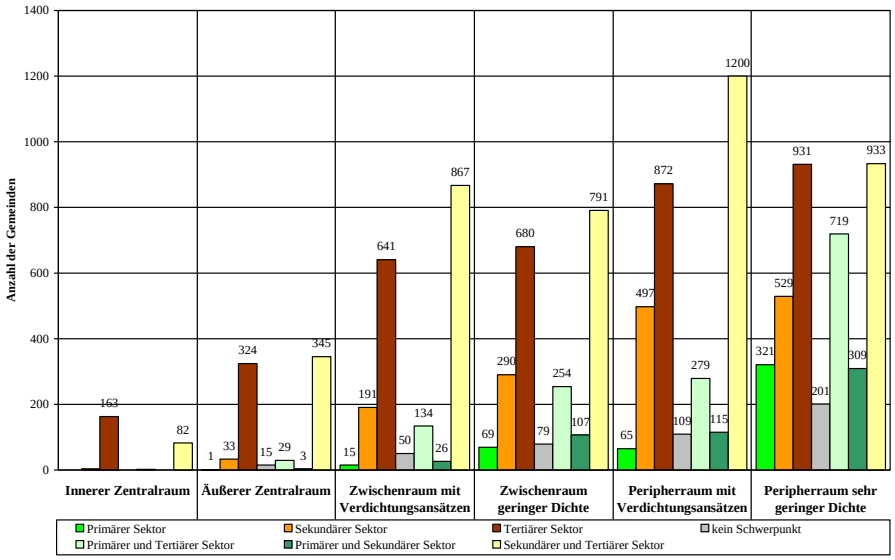


Abbildung 5-9: Beschäftigungsschwerpunkt je Raumstrukturtyp und Wirtschaftssektor

5.2.3 Einzeluntersuchung raumstruktureller Kenngrößen

Im Spannungsfeld von Polyzentralität und räumlicher Vielfalt und vor dem Hintergrund einer sachlich-räumlichen Differenzierung des deutschen Gemeindesystems werden im folgenden zusätzlich zu den bereits ermittelten Beschäftigungsschwerpunkten raumstrukturelle Kenngrößen einer Einzeluntersuchung unterzogen. Erörtert werden die Eigenschaften der Gemeinden in Bezug auf gewählte gemeindestatistische Messgrößen. Geprüft wird, ob und in welcher Weise eine Klassenbildung auf Grundlage dieser Daten überhaupt plausibel und sinnvoll ist. Angestrebt wird die Durchdringung struktureller Beziehungen von aggregierten Messgrößen und dazugehörigen Ausgangsgrößen. Erklären lassen sich etwaige Klassen zukünftig ggf. durch weitere Mess- und Untersuchungsergebnisse.

Tabelle 5-4 zeigt die sechs Variablen, die einer genaueren Einzeluntersuchung unterzogen werden. Neben der Messgröße sind die dazugehörige Messvorschrift, die Einheit sowie die Variablennotationen aufgeführt, die bei den Verteilungsuntersuchungen relevant sind.

Messgröße	Messvorschrift	Einheit	Variablenname
(1) Verstädterung[514]	Anteil der Siedlungs- und Verkehrsfläche an der Katasterfläche	[%]	‚GradVerstaedterung'
(2) Nutzungsproportion[515]	Anteil der Gebäude- und Freifläche an der Siedlungs- und Verkehrsfläche	[%]	‚FlaecheAntGebFreiSiedVerk'
(3) Konzentration[516]	Einwohner und Arbeitsplätze je km² Gebäude- und Freifläche	[Personen je km²]	‚AuslastungGebFreiflaeche'
(4) Entdichtung[517]	Anteil der Ein- und Zweifamilienhäuser am Wohnbaubestand	[%]	‚WBauAnt12FamHaus'
(5) Beschäftigungsdisparität[518]	Quotient von sozialversicherungspflichtig Beschäftigten am Arbeitsort und sozialversicherungspflichtig Beschäftigten am Wohnort*100	[ ]	‚Arbeitsplatzausstattung'
(6) Erreichbarkeit[519]	Fahrzeit zum nächsten Oberzentrum (PKW)	[Minuten]	‚Fahrzeit'

Tabelle 5-4: Übersicht zu ausgewählten raumstrukturellen Kenngrößen

514 Vgl. ARLT et al. [2001, S.5]
515 Vgl. Laufende Raumbeobachtungen, BBR [2006]: Flächenerhebung nach Art der tatsächlichen Nutzung
516 Vgl. STAACK [1995, S. 128], SIEDENTOP et al. [2005, S. 77], BEHRENS / MARHENKE [1997]
517 Vgl. SIEDENTOP et al. [2005, S. 76]
518 Vgl. SIEDENTOP et al. [2003, S. 102] und [2005, S. 79]
519 Es ist darauf hinzuweisen, dass es sich um verbandsgemeindebezogene Daten handelt, so dass eine gewisse Unschärfe bei diesen Daten vorliegt. (siehe BECHER [1995, S.117], SIEDENTOP et al. [2005, S. 81])

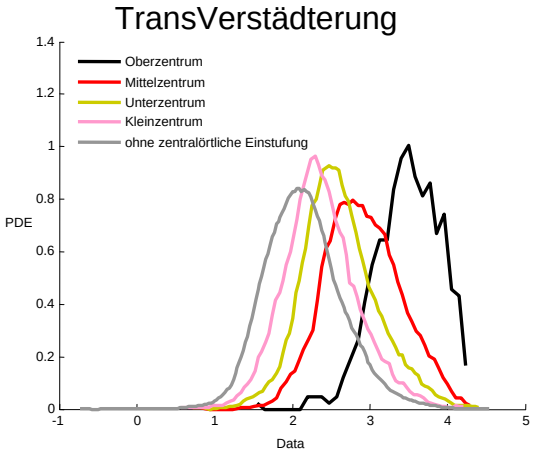
Tabelle 5-5 enthält die Ergebnisse der Verteilungsuntersuchungen und die durch den Gauss-Ansatz ermittelten Entscheidungsgrenzen.[520] Diese erlauben eine Klassifizierung jeder einzelnen Messgröße und bilden die Grundlage für einen möglichen mehrdimensionalen Untersuchungsansatz.

Messgröße	**Verteilungshypothese**	**Verteilung (grob inspiziert)**	**GMM/Grenzen (Modellierung)**	**Entscheidungsregel**	**Klassenstärke**
(1) Verstädterung	Linkssteile Verteilung Wenige Gemeinden hoch verstädtert, viele geringer verstädtert	**Log(Data)** folgt Normalverteilung	Bimodal, 2 Gauß-Verteilungen **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) Nutzungsproportion	Normalverteilung	**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	Linkssteile Verteilung Viele Gemeinden mit geringer Konzentration Wenige Gemeinden mit hoher Konzentration	**Log(Data)** folgt Normalverteilung	Multimodal, 3 Gauß-Verteilungen **Grenzen: 2500, 4000** **Log(Data): 7.9, 8.3**	Klasse 1: **Data ≤2500** Klasse 2: **Data >2500 ≤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	Linkssteile Verteilung wenige Gemeinden mit Wohnungsbauschwerpunkt viele Gemeinden mit hohem Eigenheimanteil	Umkehransatz der Variablendaten: y=log(100-Data) **Log((100-Data)+1)** folgt Normalverteilung	Multimodal, 3 Gauß-Verteilungen **Grenze: 2 %, 15 %** **LOG(Data): 1,1 und 2,8** **invertierte Grenze: 98 %, 80 %**	Klasse 1: **Data <85** Klasse 2: **Data ≥85 und <98** Klasse 3 **Data ≥98**	Klasse 1, [1420] Anteil: 11,42 % Klasse 2, [8869] Anteil: 71,35 % Klasse 3, [2141] Anteil 17,22 % Davon 100 % Einfamilienhausgemeinden [518]
(5) Beschäftigungsdisparität	Linkssteile Verteilung wenige Beschäftigungszentren Viele Gemeinden mit hauptsächlich Wohnstandsortfunktion	**Log(Data+1)** folgt Normalverteilung	Multimodal, 3 Gauss-Verteilungen **sachlogisch erzwungen: 100**	Klasse 1: **Data <100** Klasse 2: **Data ≥100**	Klasse 1, [10808] Anteil: 86,95 % Klasse 2, [1622] 13,05 %
(6) Erreichbarkeit	Linkssteile Verteilung Wenige Gemeinden mit kurzer Fahrzeit Viele Gemeinden mit großer Fahrzeit	**Log(Data+1)** folgt Normalverteilung	Bimodal, 2 Gauß-Verteilungen Grenze: **30 Minuten Log (Data): 3.3**	Klasse 1, **Data=0** Klasse 2, **Data >0 und ≤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]

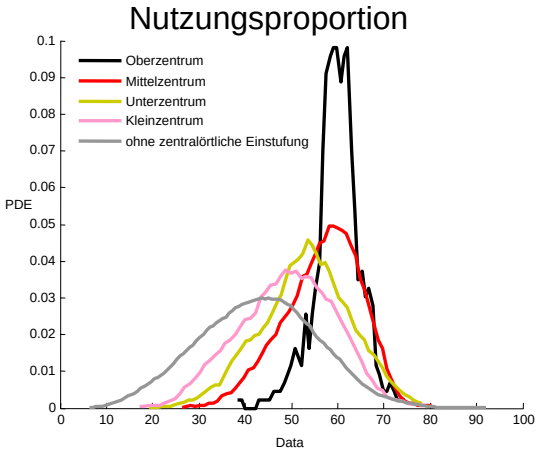
Tabelle 5-5: Zusammenfassung der Datenaufbereitung von raumstrukturellen Variablen

[520] Im Abschnitt 4 ist exemplarisch an einer Variablen die gewählte Vorgehensweise ausführlich beschrieben. Im Nebenteil B sind die Ergebnisse der durchgeführten Verteilungsuntersuchungen vollständig hinterlegt.

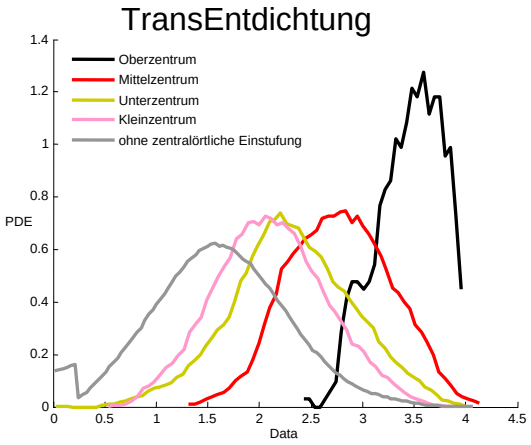
Die vorverarbeiteten Daten der Messgrößen wurden vor dem Hintergrund der Polyzentralität und räumlichen Vielfalt in ihrer strukturellen Zusammensetzung mit Hilfe von PDE-Mischungen analysiert. Aus Abbildung 5-10 ist zu erkennen, dass die Messgrößen in ihren Ausprägungen den Zentrale-Orte-Kategorien folgen und sich dadurch eine eigene Zusammensetzung der Verteilung ergibt. Die Messgröße ’Konzentration’ wurde mit den Beschäftigungsschwerpunkten nach Wirtschaftssektoren in Beziehung gesetzt. Je geringer die Konzentration ist, desto deutlicher prägt hier der Primäre Sektor die Beschäftigungsstruktur.



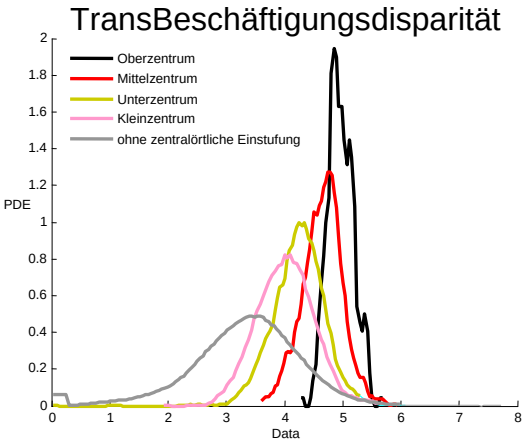
Basis: Anteil der Siedlungs- und Verkehrsfläche an der Katasterfläche



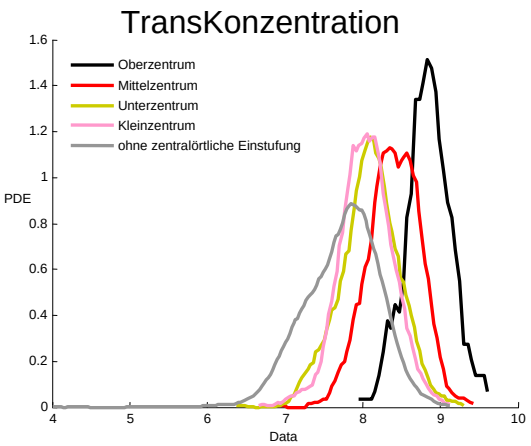
Basis: Anteil der Gebäude- u. Freifläche an der Siedlungs- u. Verkehrsfläche



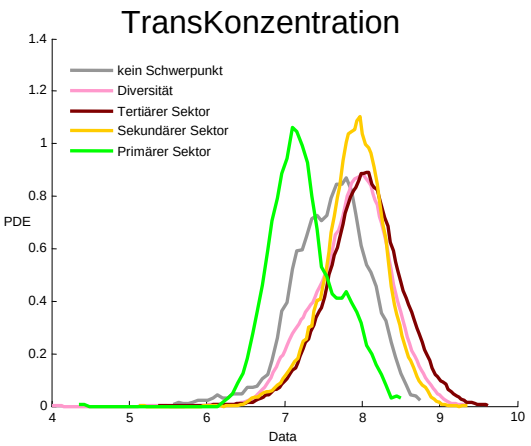
Basis: Anteil der Mehrfamilienhäuser am Wohnbaubestand



Basis: Quotient SV-Beschäftigte am Arbeitsort u. SV-Beschäftigte am Wohnort*100



Basis: Einwohner und Arbeitsplätze je km² Gebäude- und Freifläche



PDE-Mischung nach Beschäftigungsschwerpunkten in Wirtschaftsektoren

Abbildung 5-10: PDE-Untersuchung nach Zentrale-Orte-Kategorien mit ausgewählten Messgrößen

Abbildung 5-11 zeigt die Lokalisierung der Messgröße ‚Verstädterung'. Klasse 1 repräsentiert geringer verstädterte Gemeinden, dabei beträgt der Median 9,0 %. Klasse 2 enthält hoch verstädterte Gemeinden (Median: 26 %). Messbar werden in der Regel konzentrische Verstädterungsmuster bzw. eine flächenhafte Verstädterung im Agglomerationsraum.

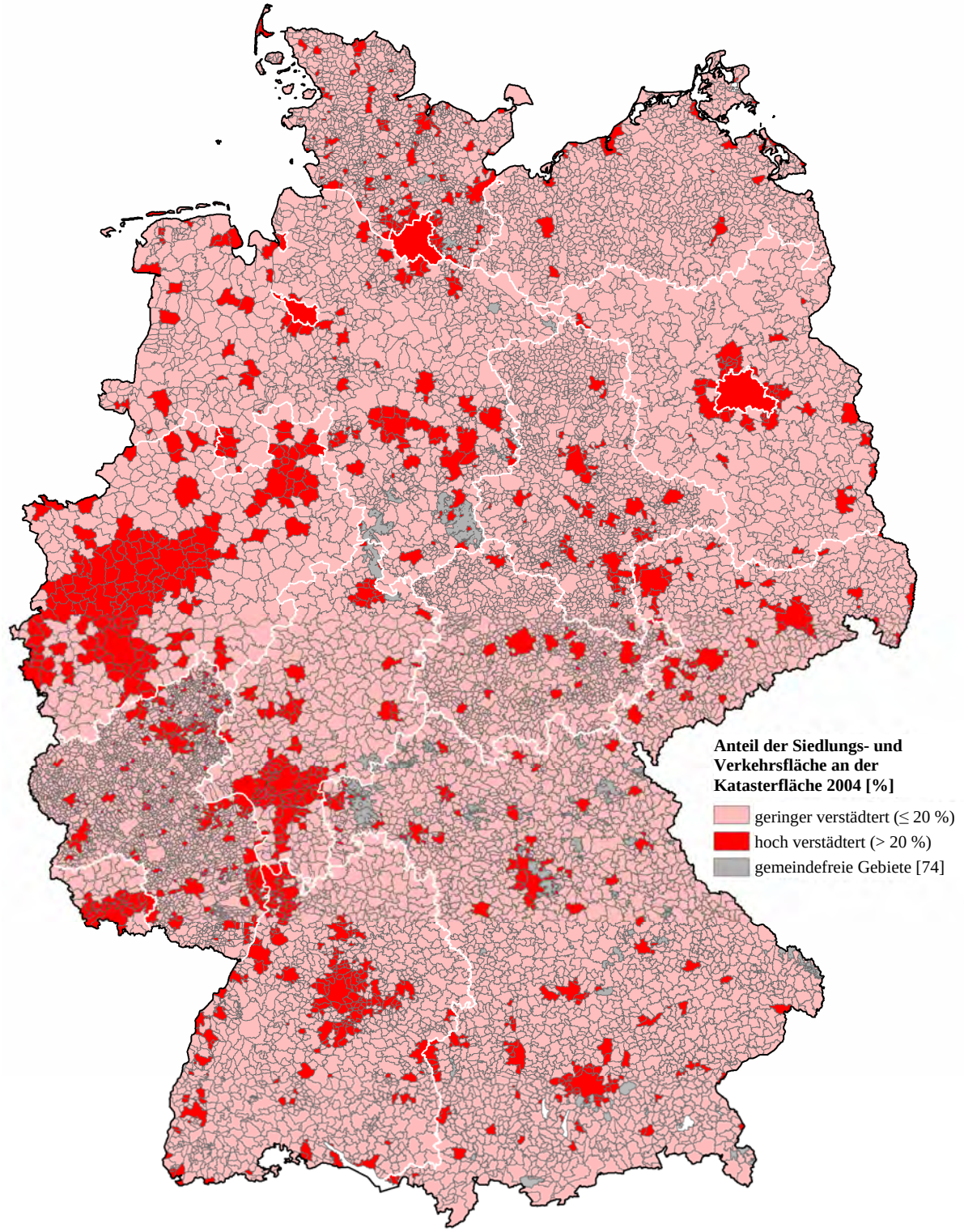


Abbildung 5-11: ‚Verstädterung' (geringer verstädtert [11029 Objekte], hoch verstädtert [1401 Objekte])[521]

[521] Die Bildunterschrift zeigt die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Abbildung 5-12 zeigt die verortete Klassenstruktur der Messgröße ‚Nutzungsproportion'. Klasse 1 repräsentiert Gemeinden, die durch einen kleinen Verhältniswert von Gebäude- und Freifläche und der Siedlungs- und Verkehrsfläche (kleine Nutzungsproportion) beschreibbar sind. Es handelt sich hauptsächlich um ländliche Gemeinden. Klasse 2 weist dagegen große Nutzungsproportionen auf und umfasst u.a. Kernstadtregionen und verstädterte Gebiete.

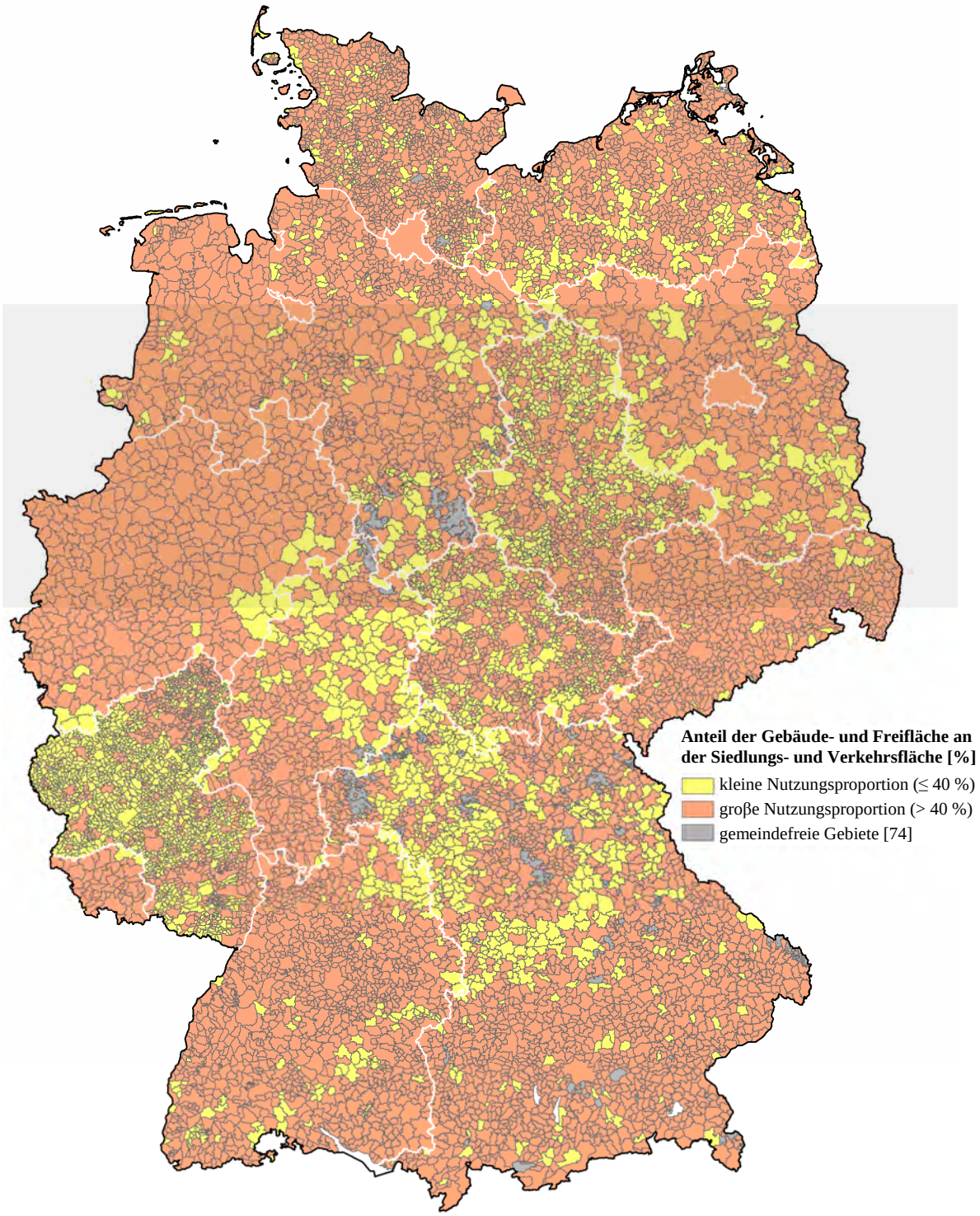


Abbildung 5-12: ‚Nutzungsproportion' (klein [3938 Objekte], groß [8492 Objekte])[522]

[522] Die Bildunterschrift zeigt die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Abbildung 5-13 zeigt das Ergebnis der Klasseneinteilung zur Messgröße ‚Konzentration', wobei drei Konzentrationsausprägungen ermittelt wurden. Es zeigen sich deutliche flächenhafte Konzentrationsprozesse innerhalb der (Groß-)Stadtregionen. Die räumliche Ausdehnung der Kernstädte ist ein die Stadtentwicklung in Deutschland seit Jahrzehnten dominierender Trend.

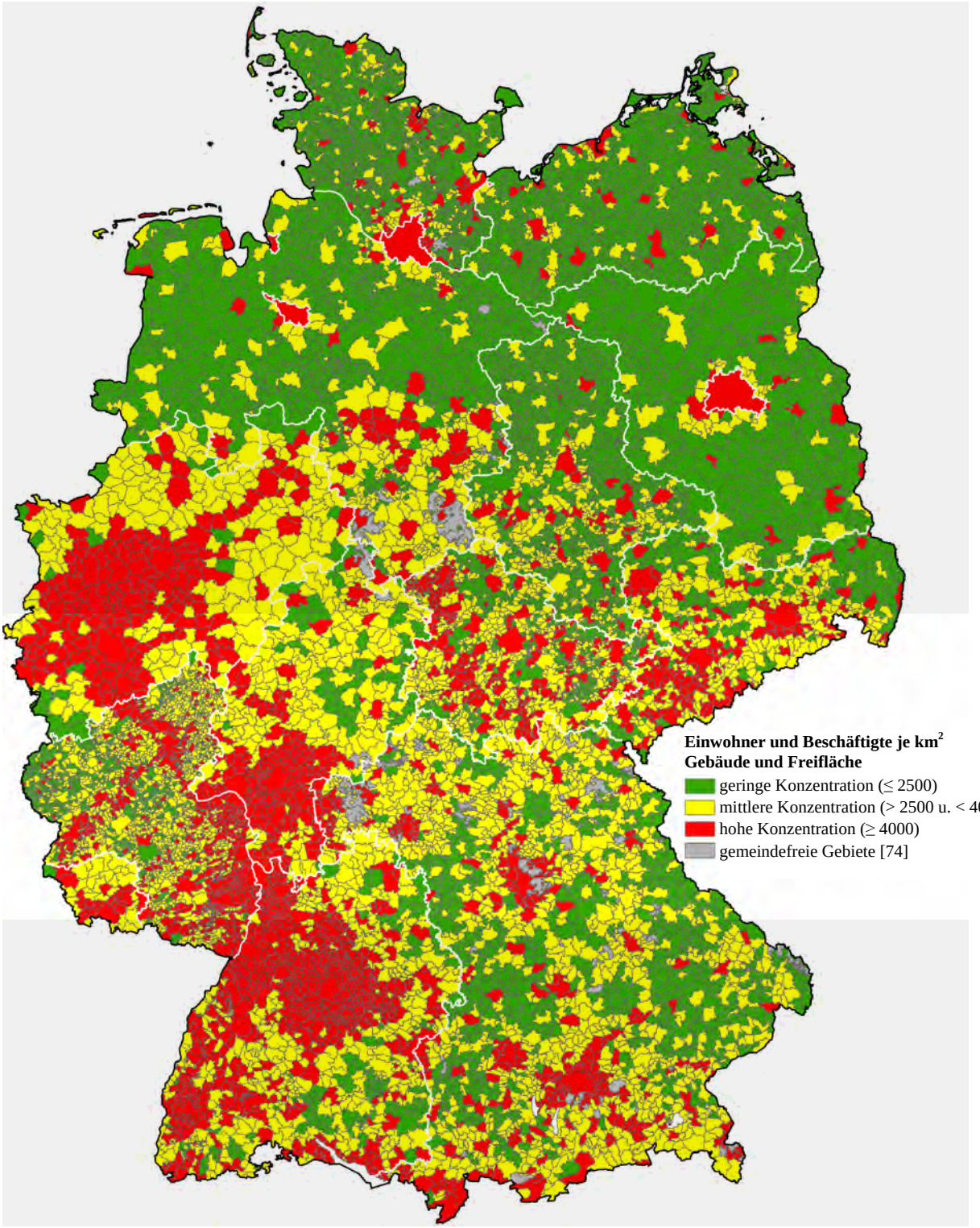


Abbildung 5-13: ‚Konzentration' (gering [5263], mittel [4802], hoch [2365])[523]

[523] Die Bildunterschrift zeigt die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Gemäß Abbildung 5-14 handelt es sich um die Ausdifferenzierung der Gemeinden nach dem Anteilswert der Ein- und Zweifamilienhäuser (‚Entdichtung'). Unterschiede bestehen zwischen kernstadtnahen bzw. Kernstadtgemeinden und dem weiteren Umland. Gesondert erfasst werden Gemeinden, die über sehr große Anteilswerte verfügen (>98 %).

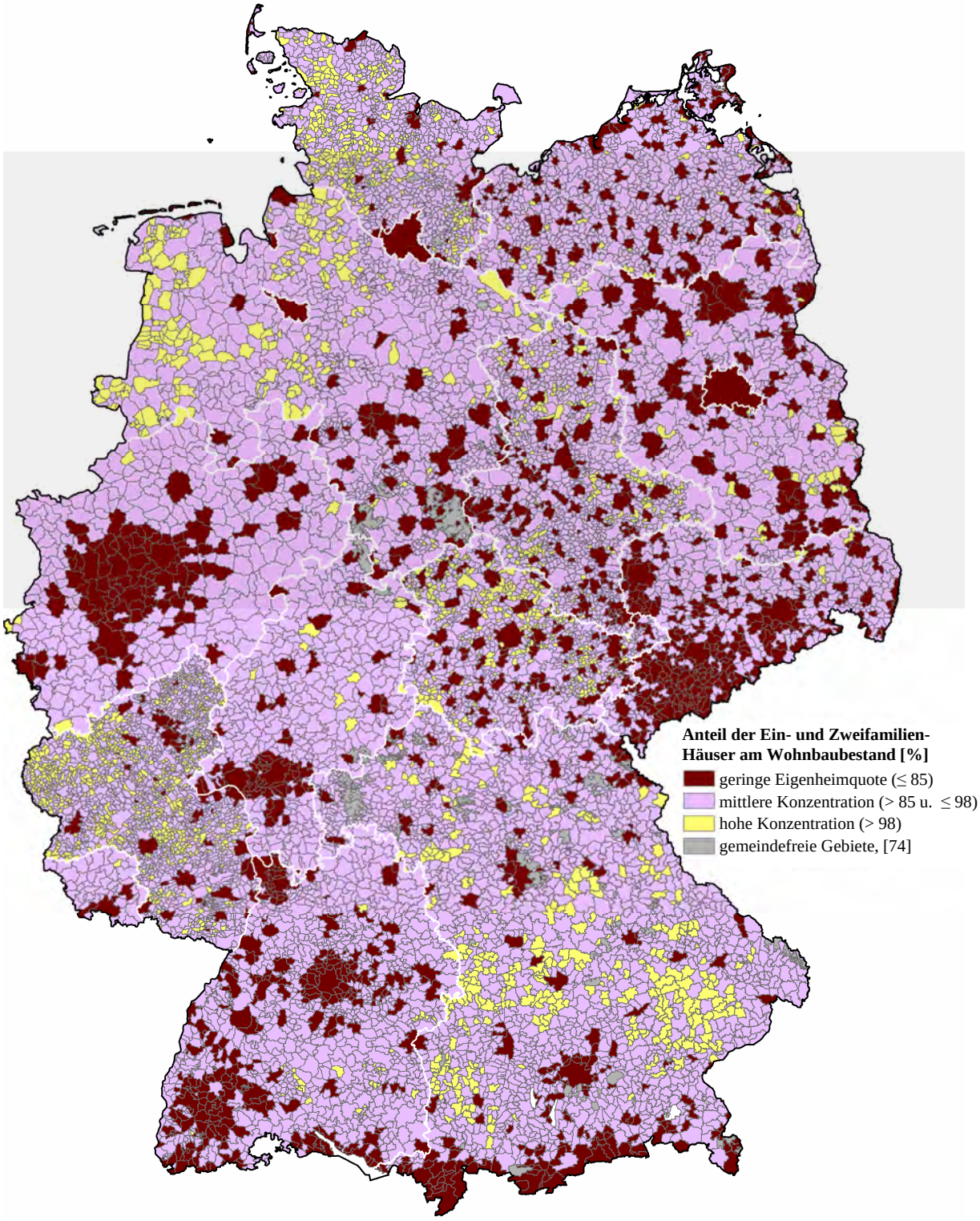


Abbildung 5-14: ‚Entdichtung' (geringe [1420], mittlere [8869], hohe [2141] Eigenheimquote)[524]

[524] Die Bildunterschrift zeigt die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Abbildung 5-15 zeigt die Verortung der Klassen für die Messgröße ‚Beschäftigungs-disparität'. Erkannt werden Gemeinden, die als Arbeitsmarktzentren bzw. Einpendlermagnet wirken. Es handelt sich primär um Ober- und Mittelzentren. Gemeinden des Umlandes sind in hohem Maße auf den kernstädtischen und kernstadtnahen Arbeitsmarkt ausgerichtet.

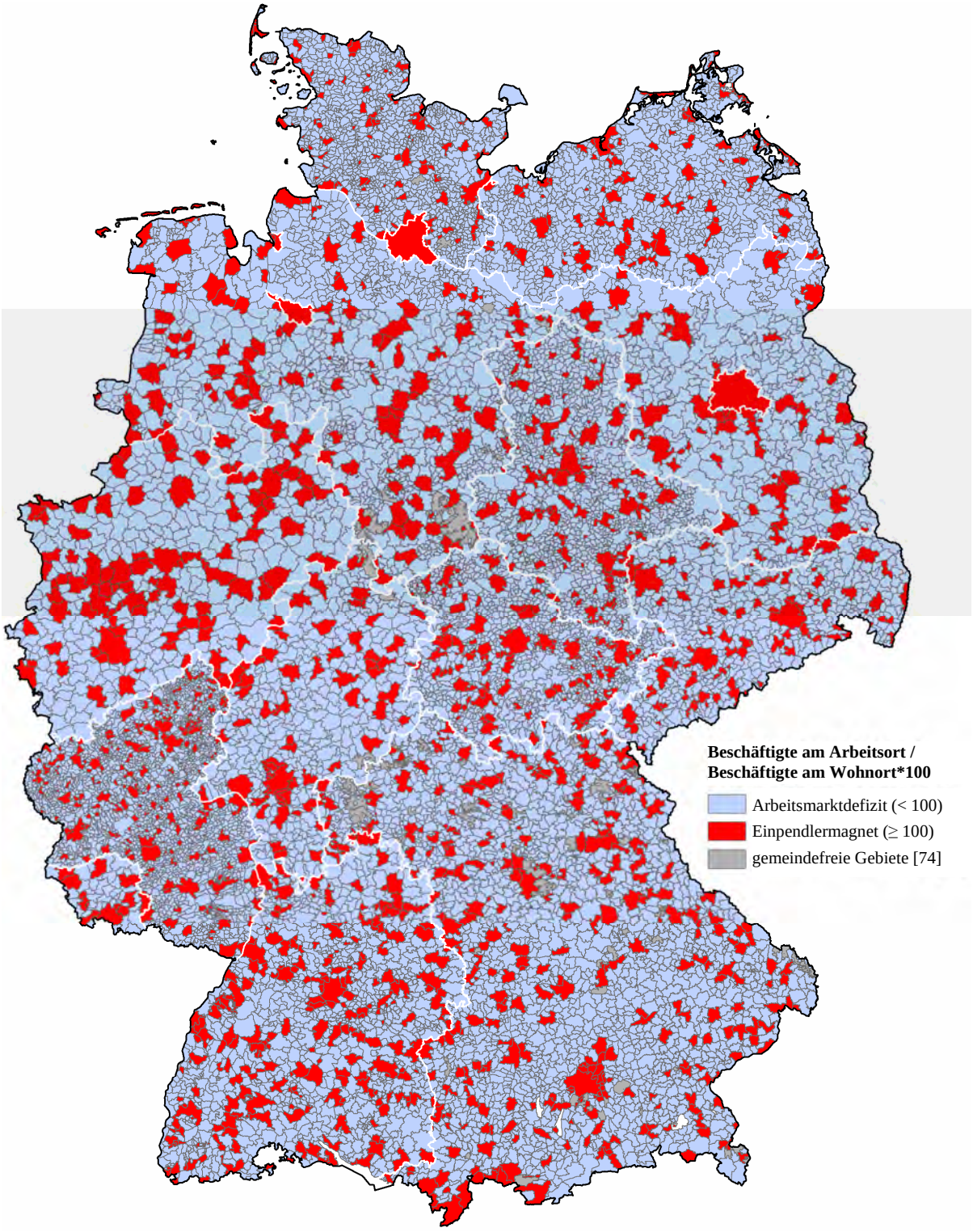


Abbildung 5-15: ‚Beschäftigungsdisparität' (Arbeitsmarktzentren [1622], Arbeitsmarktdefizit [10808])[525]

[525] Die Bildunterschrift zeigt die Klassenstärke bzw. Gemeindeanzahl je Klasse.

Abbildung 5-16 zeigt die Klassenstruktur der Messgröße ‚Erreichbarkeit'. Unterschieden sind oberzentrennahe (Median: 21 Minuten) und oberzentrenferne Gemeinden (Median: 43 Minuten) sowie die Oberzentren. Ablesbar ist der Erschließungsgrad des Umlandes zu den nächstgelegenen Oberzentren im motorisierten Individualverkehr.

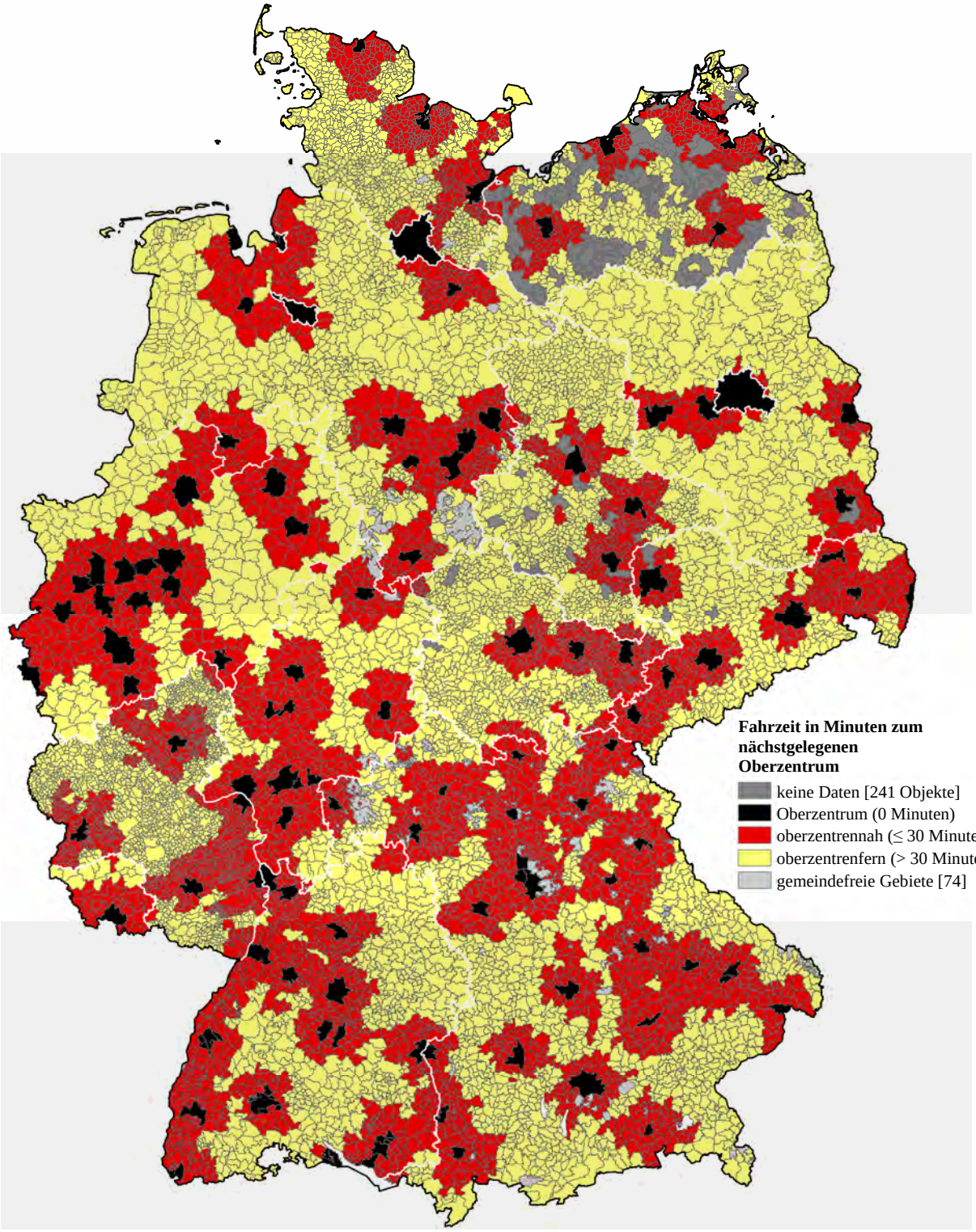


Abbildung 5-16: ‚Fahrzeit' (oberzentrennah [5092], oberzentrenfern [6964], Oberzentrum [133])[526]

[526] Gezeigt ist die Klassenstärke bzw. Gemeindeanzahl je Klasse. Es sind Datenlücken zu beklagen.

Die Untersuchung der Messgröße Verstädterung führte dazu, dass die deutschen Gemeinden nach dem Grad ihrer Verstädterung (Entscheidungsgrenze: 20%) in zwei Klassen aufteilbar sind. Es sind Bereiche mit einem höheren Siedlungs- und Verkehrsflächenanteil erkennbar, die sich weit über die Kernstädte und das engere Umland hinaus erstrecken (Klasse 2). Vielfach sind die Agglomerationsräume kaum noch von geringer verstädterten Gebieten getrennt. Konzentrisch flächenhafte Verstädterungsmuster zeichnen sich beispielsweise in der Region Stuttgart ab. Bei 60 % der Gemeinden mit einer geringeren Verstädterung liegt der Verstädterungsgrad sogar unterhalb 10 %. Diese Gemeinden besitzen mit Blick auf die Raumstrukturtypen vielfach auch eine geringere Bevölkerungsdichte. In Bezug auf die PDE-Untersuchungen nach Zentrale-Orte-Kategorien und Raumstrukturtypen (siehe Nebenteil B) sind bei den Oberzentren bzw. im Inneren Zentralraum in der Regel die größten Verstädterungsgrade festzustellen.

Betrachtet man die zwei Klassen der Messgröße Nutzungsproportion (Anteil der Gebäude- und Freifläche an der Siedlungs- und Verkehrsfläche), so ergibt sich zunächst ein etwas ungewohntes Bild, da sich die Gemeinden im kernstadtnahen Umland nicht deutlich von den Gemeinden des weiteren Umlandes abgrenzen. Identifiziert wird eine Gruppe von Gemeinden überwiegend im ländlichen Raum mit einem geringeren Anteilswert der Gebäude- und Freifläche. Es befinden sich ungefähr 70 % der Gemeinden im Peripherraum.

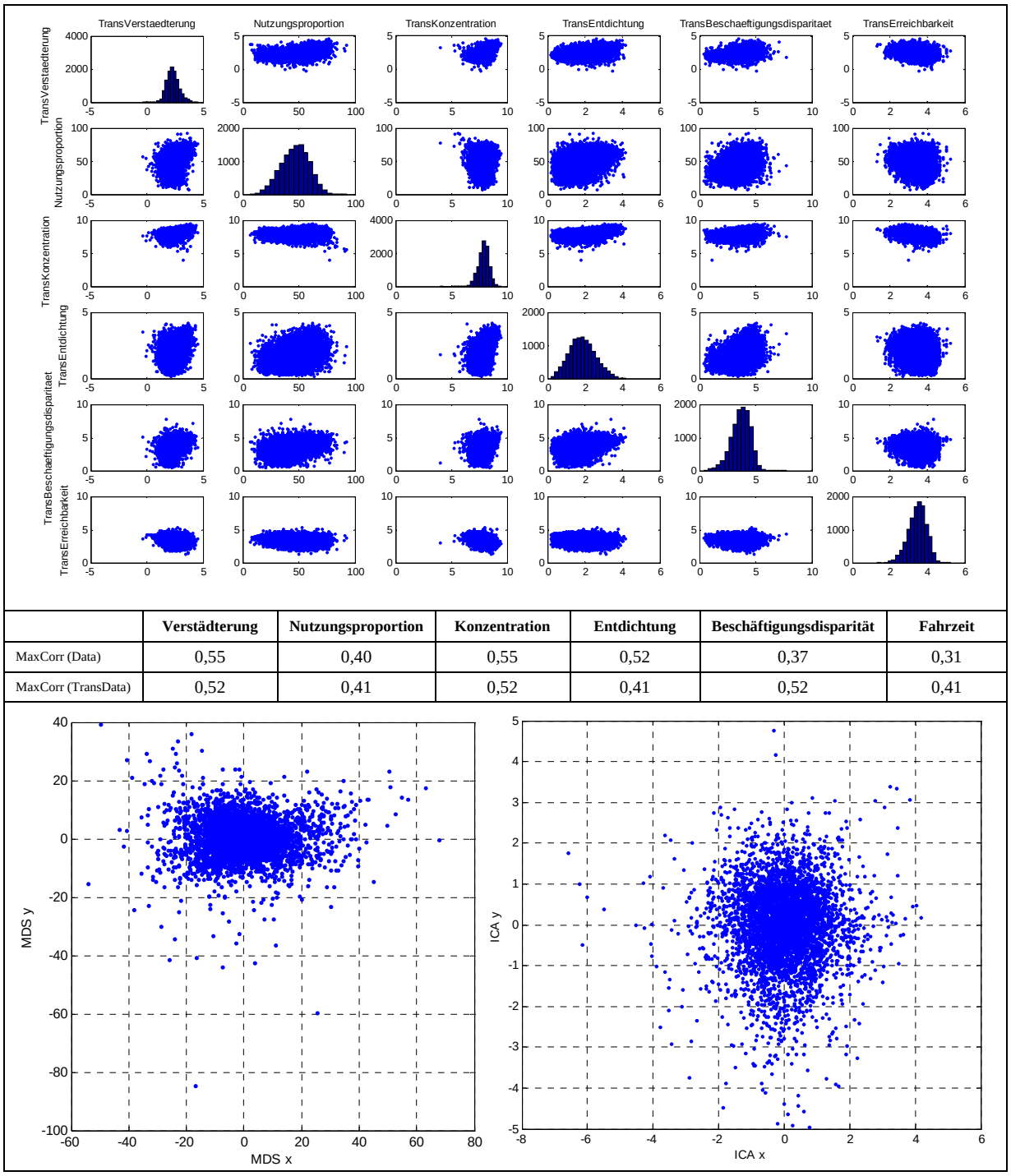
Die Messgröße Konzentration charakterisiert Gemeinden nach der Auslastung der Gebäude- und Freifläche durch Einwohner und Arbeitsplätze. Es handelt sich damit um eine Messung der Siedlungsdichte in einer modifizierten Form. Unterschieden werden Gemeinden infolge der Verteilungsuntersuchung nach geringer, mittlerer und hoher Konzentration. Besonders hervorzuheben sind die hohen Dichtewerte auch im Umland der Agglomerationskerne, die sich flächenhaft abzeichnen. Mit Blick auf die Ergebnisse der PDE-Untersuchung nach Zentrale-Orte-Kategorien bestehen größere Unterschiede zwischen den Oberzentren und den sonstigen Kategorien. In Bezug auf die Wahrscheinlichkeitsdichte der Ausprägungen von Unter- und Kleinzentren sowie Gemeinden ohne zentralörtliche Einstufung bestehen jedoch kaum Unterschiede. Die PDE nach Raumstrukturtypen (Nebenteil B) zeigen charakteristische Unterschiede zwischen den Kategorien des Zentralraums und den anderen vier Raumstrukturtypen. Anhand der Klasseneinteilung dieser Messgröße lassen sich im westdeutschen Bundesgebiet die Gemeinden mit einer hohen Auslastung der Gebäude- und Freifläche zu größeren Zentralräumen zusammenfassen. In dieser Hinsicht haben die Regionen Stuttgart, Rhein-Main und Rhein-Ruhr besonders charakteristische Eigenschaften.

Im Rahmen der Verteilungsuntersuchung wurde erkannt, dass der invertierte Anteilswert der Mehrfamilienhäuser am Wohnbaubestand einer Lognormalverteilung folgt. Mit Hilfe von Gauss-Mixtur-Modellen wurden drei Klassen anhand der ermittelten Entscheidungsgrenzen bei 85 % und 98 % (Anteilswerte der Ein- und Zweifamilienhäuser im Wohnbau) aufgebaut. Der räumliche Schwerpunkt des Mehrfamilienhausbaus wird dabei in den Agglomerationsräumen erkannt (siehe Region Stuttgart und Rhein-Ruhr). Gerade die Kernstädte, aber auch der engere suburbane Raum verzeichnen die größten Anteilswerte. In Ostdeutschland zeichnet sich ein etwas anderes Bild aufgrund einer anderen historischen Entwicklung und früherer Wohnungsbaupolitik ab. Es sind größere Anteilswerte der Mehrfamilienhäuser am Wohnbaubestand auch in Gemeinden des weiteren Umlandes, entfernt von den Kernstädten, festzustellen. Mit Blick auf die Zentrale-Orte-Kategorien sind weitere Unterschiede erkennbar. Die Gemeinden ohne zentralörtliche Einstufung, die Kleinzentren und die Unterzentren sind verstärkt durch den Eigenheimbau geprägt (siehe auch PDE-Darstellung). Von den 2141 Gemeinden mit einer hohen Eigenheimquote sind 2054 Gemeinden (96 %) ohne zentralörtliche Einstufung.

Aufgrund der Eigenschaften der Messgröße ‚Beschäftigungsdisparität' wurde die Entscheidungsgrenze von 100 nicht durch Modellierung des Verteilungsverlaufes ermittelt, sondern nimmt Bezug auf den Aspekt, dass dieser Wert einen theoretischen optimalen Grad der Nutzungsmischung ausdrückt. Ein ausgeglichenes Verhältnis der Beschäftigten am Arbeitsort und der Beschäftigten am Wohnort ist nicht mit einer Abdeckung der Nachfrage nach Beschäftigungsmöglichkeiten gleichzusetzen. Die Gemeinden werden dahingehend unterschieden, ob ein Teil der vorhandenen Arbeitsplätze von Einpendlern eingenommen wird (Wert > 100) oder ob ein Teil der Bevölkerung das Gemeindegebiet zum Aufsuchen einer Beschäftigung verlassen muss (Wert < 100). Feststellbar ist eine deutliche Orientierung auf den kernstädtischen Arbeitsmarkt. Die Oberzentren und Mittelzentren sind wichtige Standorte des Arbeitsmarktes. Darüber hinaus wurden Gemeinden erkannt, die zwar keine zentralörtliche Einstufung besitzen aber dennoch durch Einpendler aufgesucht werden. Insgesamt ist eine deutliche Beschäftigungsdisparität zwischen den Kernstädten und dem Umland festzustellen.

Die Messgröße ‚Erreichbarkeit' berücksichtigt die Fahrzeit zum nächstgelegenen Oberzentrum im motorisierten Individualverkehr und trennt oberzentrennahe und oberzentrenferne Gemeinden. Das bestehende Autobahnnetz ist in vielen Fällen indirekt erkennbar und führt erheblich zu einer Reduktion der Fahrzeit zu den Oberzentren.

Tabelle 5-6 ermöglicht eine Beurteilung von Variablenzusammenhängen in grafischer und rechnerischer Form (Scatter-Plot / Korrelationswerte). Die Histogramme verdeutlichen den einheitlichen Verteilungsverlauf infolge gewählter Transformationen. Die Darstellung der mehrdimensionalen Skalierung verweist auf die Schwierigkeit, klar unterscheidbare Gruppen mit Hilfe von Cluster-Algorithmen mit diesen Variablen bilden zu können (Punktwolke).



	Verstädterung	Nutzungsproportion	Konzentration	Entdichtung	Beschäftigungsdisparität	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

Tabelle 5-6: Scatter-Plot und MDS-Plot sowie ICA-Plot der 6 raumstrukturellen Variablen

Die Vermutung, dass es nur schwer möglich ist, deutlich unterscheidbare Klassen mit Hilfe eines Clusteralgorithmus bilden zu können, wird durch eine ESOM-Karte in Abbildung 5-17 bestätigt (siehe Theorie in Abschnitt 2.2.2). Klare Grenzen bzw. eindeutige Gruppierungen wie sich diese bei anderen Datensätzen mit ESOM-Karten ausbilden, zeichnen sich bei den hier vorliegenden Daten der 12430 Gemeinden in keiner Weise ab.

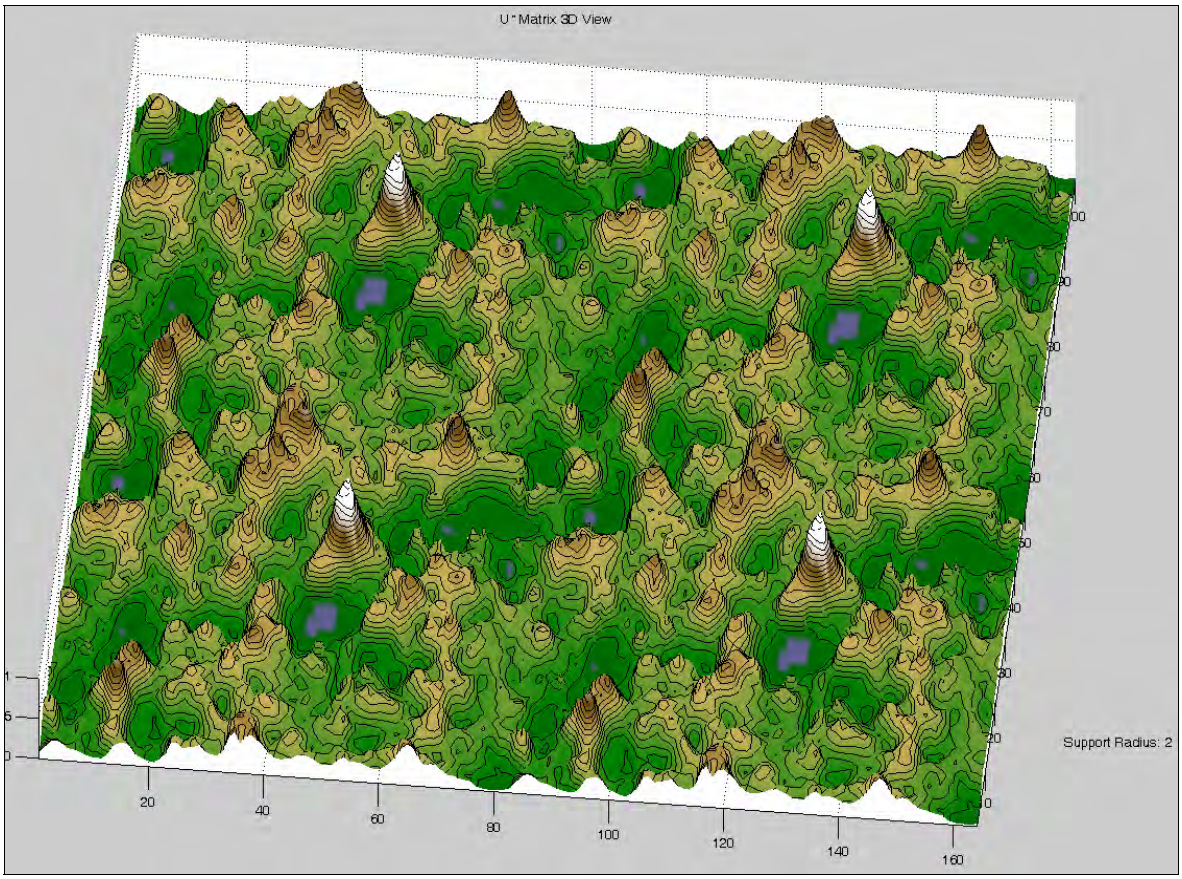


Abbildung 5-17: 3-D-Ansicht der U*-Matrix (N=12430, D=6, 50x82 Neurons)[527]

Aufgrund erkannter Gruppierungsschwierigkeiten gemäß MDS und U*-Matrix wird entschieden, dass für eine mehrdimensionale Betrachtung der Gemeinden die bereits gewonnne Einzelklassifizierung im Rahmen der Verteilungsuntersuchung besser geeignet ist, um zumindest auf Basis dieser ermittelten Klassen eine Gesamtbetrachtung durchzuführen. Es lassen sich beispielsweise mit den sechs Variablen und dazugehörigen Einzelklassifizierungen etwa 200 Klassen ermitteln, wobei diese vielfach zu schwach besetzt sind. Um mit Blick auf die große Anzahl der Kombinationsmöglichkeiten eine schwer zu interpretierende Klassenanzahl zu vermeiden, findet eine zusätzliche Variablenselektion statt. Gewählt werden drei der sechs Variablen, die gute Eigenschaften zur Objekttrennung aufweisen (Erreichbarkeit, Verstädterung sowie Einwohner- und Beschäftigungskonzentration).

527 Siehe zusätzliche Grafiken im Nebenteil B.

Die erste Messgröße ‚Erreichbarkeit' unterstützt die Klassifikation in dem Sinne, dass eine Aussage über die Fahrzeit im motorisierten Individualverkehr zum nächstgelegenen Oberzentrum in Deutschland getroffen werden kann. Charakterisiert wird dadurch die Verflechtung zwischen Umland und Zentrum. Infolge der Verteilungsuntersuchung wurde eine Entscheidungsgrenze von 30 Minuten gefunden, die im Sinne eines abstrahierten gravitationstheoretischen Verständnisses dazu eingesetzt wird, um die Peripherie und das direkt durch das Oberzentrum beeinflusste Umland zu unterscheiden.

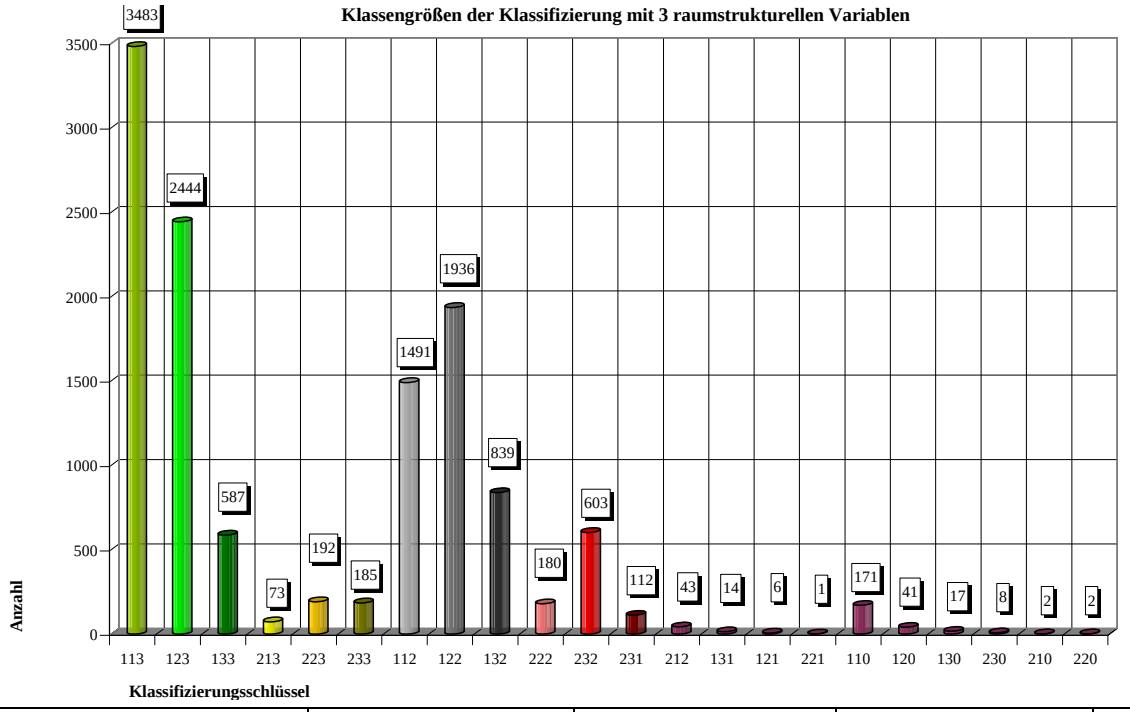
Die zweite Messgröße ‚Konzentration' wird für die Klassifikation verwendet, um eine Differenzierung nach der vorhandenen Auslastung der Gebäude- und Freifläche durch Bewohner und Beschäftigte vornehmen zu können. Das deutsche Gemeindesystem wird dadurch in Anlehnung an die von BOUSTEDT[528] bereits formulierte Einwohner- und Arbeitsplatzdichte regional ausdifferenziert. Diese in der Vergangenheit häufig eingesetzte Dichtegröße bezieht sich noch auf die Gemeindefläche insgesamt, so dass eine gewisse Unschärfe entsteht, da nicht die tatsächlich bebaute Fläche berücksichtigt wird und damit das Messergebnis erheblich von der Gebietsgröße einer Gemeinde beeinflusst wird. Aus diesem Grunde wird insbesondere mit Bezug auf SIEDENTOP[529] die Gebäude- und Freifläche als genauere räumliche Bezugsgröße herangezogen. Infolge der Verteilungsuntersuchung und der Modellierung der Verteilung mit Gauss-Mixtur-Modellen wurden zwei Entscheidungsgrenzen erarbeitet (2500 und 4000 Einwohner und Arbeitsplätze je km² Gebäude- und Freifläche), die die Gemeinden nach dem Grad der Konzentration identifizieren.

Die dritte Messgröße ‚Verstädterung' wird deshalb in die geplante Klassifikation integriert, da vor dem Hintergrund des technologischen Wandels und der sich ändernden wirtschaftlichen Bedingungen nicht nur Umstrukturierungsprozesse innerhalb eines Oberzentrums selbst stattfinden, sondern auch gerade im Umland eine Ausdehnung der Verstädterung erfolgt bzw. Auswirkungen der Schrumpfung sichtbar werden. Monozentrische Regionen lösen sich teilweise auf, und es entstehen polyzentrale Regionen (agglomerieren). Neue Zentren des demographischen und wirtschaftlichen Wachstums entstehen sowohl durch Prozesse der Suburbanisierung als auch durch standörtlich differenzierte Bedingungen. Zukünftig werden angesichts vermehrter Schrumpfungsprozesse auch Phasen einer Disurbanisierung auftreten. Die Verteilungsuntersuchung ermöglichte die Unterscheidung der Gemeinden nach dem Grad der Verstädterung in geringer und hoch verstädterte Gemeinden (Entscheidungsgrenze: 20%).

---

528 Vgl. BOUSTEDT [1975 b]
529 Vgl. SIEDENTOP et al. [2005, S. 77]

Mit Blick auf das Ziel einer raumstrukturellen Klassifizierung wurde anhand der Entscheidungsregeln der Messgrößen ‚Erreichbarkeit', ‚Konzentration' und ‚Verstädterung' die Klassenstruktur aufgebaut. Tabelle 5-7 zeigt die Klassengrößen mit dazugehöriger Definition der Klassifizierungsschlüssel. Die Datenlücken der Messgröße ‚Fahrzeit' erklären einige Sonderfälle der Klassifizierung.



Semantik	Verstädterungsgrad			Konzentration			Erreichbarkeit			Anteil
	Median	∅	StAbw	Median	∅	StAbw	Median	∅	StAbw	[in %]
113 = Peripherie, geringe Dichte	7,30	7,82	3,1	1736,2	1709	495,8	46,20	49,3	15,4	28,02
123 = Peripherie, mittlere Dichte	9,70	10,1	3,8	3064,1	3120	411,3	41,17	44,1	11,8	19,66
133 = Peripherie, hohe Dichte	10,97	11,2	4,2	4565,3	4781	730,5	39,44	43,5	11,4	4,72
213 = Regionalzentrum, geringe Dichte	25,17	27,1	7,4	2033,8	1843	618,4	41,50	47,3	16,1	0,59
223 = Regionalzentrum, mittlere Dichte	24,19	27,6	9,6	3198,5	3244	430,1	40,13	43,8	13,6	1,54
233 = Regionalzentrum, hohe Dichte	25,71	28,2	8,5	4942,0	5229	997,8	36,73	41,7	12,9	1,49
112 = Inter-Agglomerativ, geringe Dichte	7,60	8,2	3,1	1948,0	1863	457,5	22,46	21,6	5,8	12,00
122 = Inter-Agglomerativ, mittlere Dichte	10,59	10,9	3,7	3147,4	3183	416,6	21,95	21,3	5,9	15,58
132 = Inter-Agglomerativ, hohe Dichte	14,11	13,6	4,1	4664,9	4891	797,3	20,72	20,4	6,0	6,75
222 = Agglomerationseinheit, mittlere Dichte	23,47	26,3	7,9	3431,9	3366	403,0	20,49	19,5	6,1	1,45
232 = Agglomerationseinheit, hohe Dichte	27,68	30,9	10,2	5622,9	5880	1444,8	18,54	18,3	6,0	4,85
231 = Agglomerationszentrum	37,57	40,0	12,6	7557,6	7874	2073,8	0	0	0	0,90
212 = Sonderfälle: Agglomerationseinheit	27,65	30,0	8,8	1710,6	1624	660,1	19,24	19,8	5,97	0,35
131 = Sonderfälle: Agglomerationszentrum	17,48	15,9	3,6	4859,4	5057	846,1	0	0	0	0,11
121 = Sonderfälle: Agglomerationszentrum	9,64	12,2	4,7	3439,0	3504	375,4	0	0	0	0,05
221 = Sonderfälle: Agglomerationszentrum	22,07	22,0		3899,1	3899		0	0		0,01

Ziffer 1	Ziffer 2	Ziffer 3
1 = hoch verstädtert (>20), [%] 2 = geringer verstädtert (≤20), [%]	1 = Niedrige Konzentration (≤ 2500), [P./km²] 2 = Mittlere Konzentration(>2500 ≤ 4000), [P./km²] 3 = Hohe Konzentration > 4000, [P./km²]	0 = keine Daten, 'Nan' 1 = Oberzentrum, Fahrzeit = „0" 2 = Fahrzeit ≤ 30 Minuten 3 = Fahrzeit > 30 Minuten

Tabelle 5-7: Klassenbedeutung und -größen der mehrdimensionalen Variablenbetrachtung

Im Spannungsfeld von Polyzentralität und räumlicher Vielfalt lässt sich gemäß Abbildung 5-18 mit den Variablen ‚Verstädterungsgrad', ‚Konzentration' und ‚Fahrzeit' eine statische Erfassung und Strukturierung von Agglomerations- und Verdichtungsprozessen vornehmen.

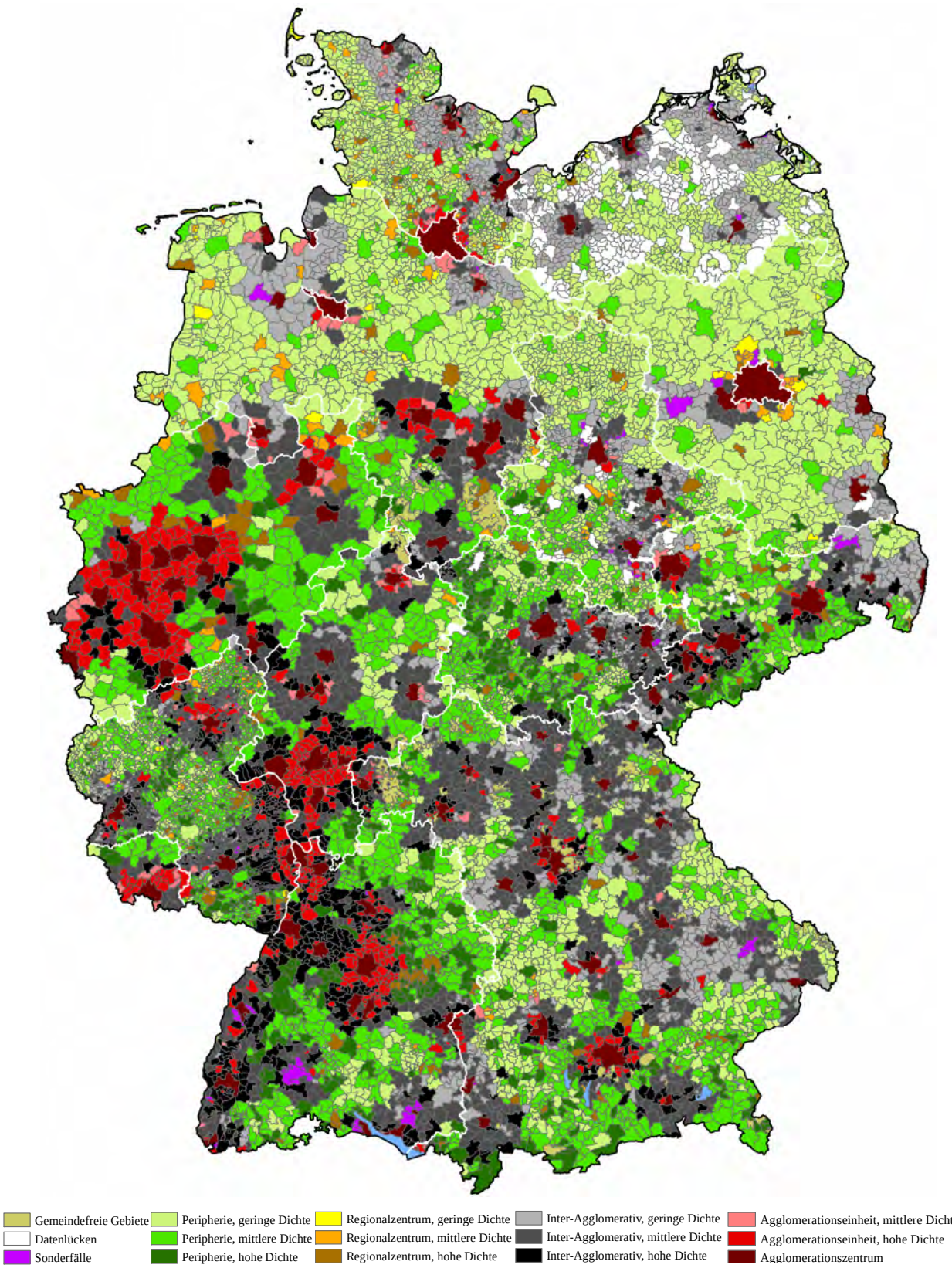


Abbildung 5-18: Verortung des Klassifizierungsergebnisses mit drei raumstrukturellen Variablen

Das Ergebnis der Klassenbildung ermöglicht eine Raumbeobachtung, die der Grundidee folgt, die heutige Situation der Verstädterung bzw. die entstandenen Agglomerationen abzubilden.

In Bezug auf ein gewähltes Agglomerationszentrum[530] 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 zur Beschreibung der Gemeinden sowohl innerhalb als auch außerhalb des Interaktionsraums mit Hilfe von drei Dichteklassen.

Der Grad der Verstädterung als Anteil der Siedlungs- und Verkehrsfläche an der Gemeindefläche insgesamt ermöglicht die Identifizierung von besonders verstädterten Gemeinden. In diesem Ansatz werden die Gemeinden, welche innerhalb des Interaktionsraums liegen, als Agglomerationseinheiten bezeichnet. Diese tragen bei ähnlichen Verstädterungseigenschaften der Nachbargemeinden zum Wachstum der Agglomeration insgesamt bei. Gemeinden, die in der Peripherie liegen und einen besonders großen Verstädterungsgrad aufweisen, werden als Regionalzentrum bezeichnet. Eine Gemeinde gilt als Inter-Agglomerativ, wenn diese sich im Interaktionsraum befindet, jedoch nicht über einen Verstädterungsgrad oberhalb von 20 % verfügt. Abbildung 5-19 zeigt zusammenfassend die Grundstruktur der Klassen.

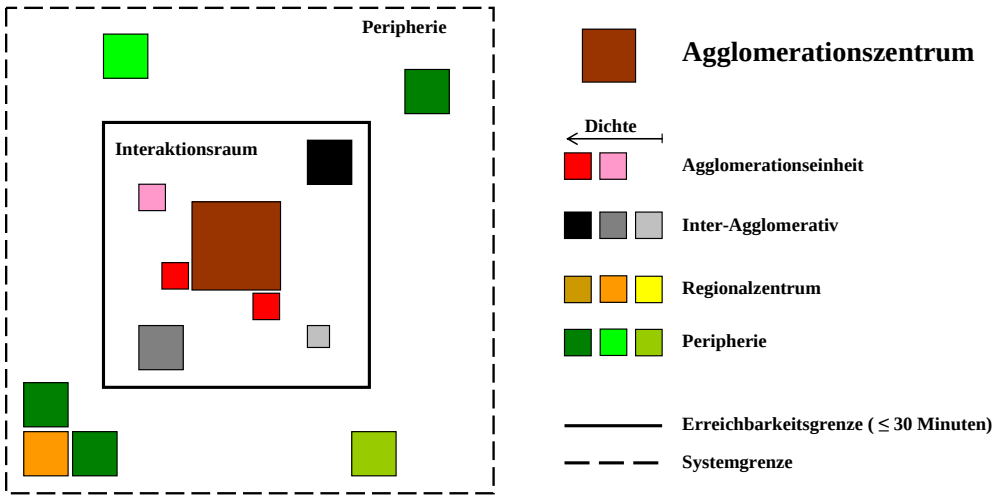


Abbildung 5-19: Strukturierung des Agglomerations- und Verdichtungsprozesses

[530] Der Begriff ‚Agglomerationszentrum' ist ein abstrakt gewählter Begriff, der den Ausgangspunkt für eine größere Agglomeration in der Zukunft bilden kann bzw. bereits Bezugspunkt einer Agglomeration ist.

Mit Blick auf die Verortung der Klassen (siehe Abbildung 5-18) werden in Deutschland spezifische Regionalstrukturen erkennbar.

Die einzelnen Agglomerationszentren werden untereinander mit ihrem dazugehörigen Interaktionsraum vergleichbar. Hinsichtlich der Einwohner und Arbeitsplätze je km² Gebäude- und Freifläche und dem Verstädterungsgrad wurden verschiedene Verflechtungsmuster erfasst. Die Regionen Rhein-Main, Rhein-Ruhr sowie die Region Stuttgart repräsentieren hochverdichtete und deutlich verstädterte Agglomerationen. Es ist zu vermuten, dass in Zukunft bei weiter fortschreitendem Wachstum der Bevölkerung und der Beschäftigten eine noch stärkere Agglomeration entstehen wird und damit die Übergänge zwischen einzelnen Verflechtungsgebieten noch schwieriger abgrenzbar sein werden.

In Bezug auf einige ostdeutsche Agglomerationszentren (z.B. Chemnitz, Neubrandenburg oder Schwerin) lassen sich deutlich geringere Dichte- und Verstädterungseigenschaften bei den Gemeinden im Verflechtungsgebiet erkennen. Hier kann nicht von einer homogen dichten städtischen Agglomeration gesprochen werden, sondern eher von einem Kerngebiet, welches als ausgewiesenes 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 Dichte und Verstädterungsgrad messbar. Die als Regionalzentren definierten Gemeinden sind durch einen höheren Verstädterungsgrad gekennzeichnet. Es ist ggf. möglich, dass diese Gemeinden weitere besondere Eigenschaften haben und auch einen stärkeren Einfluss auf die Nachbargemeinden ausüben könnten. Es ist aufgefallen, 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. Als Gemeinden mit hoher Dichte in Baden-Württemberg seien genannt: Schwäbisch-Hall, Bad Mergentheim oder Crailsheim sowie in Niedersachsen Papenburg, Meppen oder Aurich. Die unabhängig von Dichte und Verstädterung zusätzlich vorhandenen spezifischen Eigenschaften von Gemeinden der gleichen Klasse könnten zukünftig mit methodischen Ansätzen der Klassenerklärung genauer analysiert werden.

In Abbildung 5-20 sind Regionalstrukturen in einem größeren Detailausschnitt ersichtlich. Es handelt sich um die Region Stuttgart und den Großraum Berlin. In der Region Stuttgart kann bereits 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 (Oberzentren) Pforzheim, Karlsruhe und Heilbronn sowie Tübingen grenzen diese aneinander und überlagern sich in Teilbereichen. Es entsteht eine spezifische Regionalstruktur, die insbesondere durch eine sehr große Anzahl von sogenannten inter-agglomerativen Gemeinden mit hoher Einwohner- und Beschäftigtendichte entlang der Verkehrsachsen gekennzeichnet ist. In dieser Region liegt das durch SIEDENTOP[531] definierte Basisniveau der Verstädterung als Maß für einen großflächig nicht unterschrittenen Siedlungs- und Verkehrsflächenanteil besonders hoch im Vergleich zu anderen Regionen (siehe Halle-Leipzig). Außerhalb des Interaktionsraumes Stuttgart liegen weitere sogenannte Regionalzentren (z.B. Göppingen), die sich entlang der Bundesstraße B10 nach Osten ausdehnen. Diese Gemeinden zeigen ebenfalls eine hohe Dichte und sind hoch verstädtert, jedoch liegt die Fahrzeit zum nächstgelegenen Oberzentrum oberhalb 30 Minuten. Auf eine Besonderheit ist im Großraum Berlin einzugehen, da dort in direkter Nachbarschaft zum Agglomerationszentrum eine größere Anzahl Gemeinden als Regionalzentrum erkannt wurde. Diese zeigen oftmals eine geringe bis mittlere Dichte. Es handelt sich um ein frühes Suburbanisierungsstadium.[531] Das sich in der Industrialisierung heraus gebildete Siedlungssystem mit einem Ring größerer Mittelstädte um das Zentrum ist weitgehend noch erhalten.

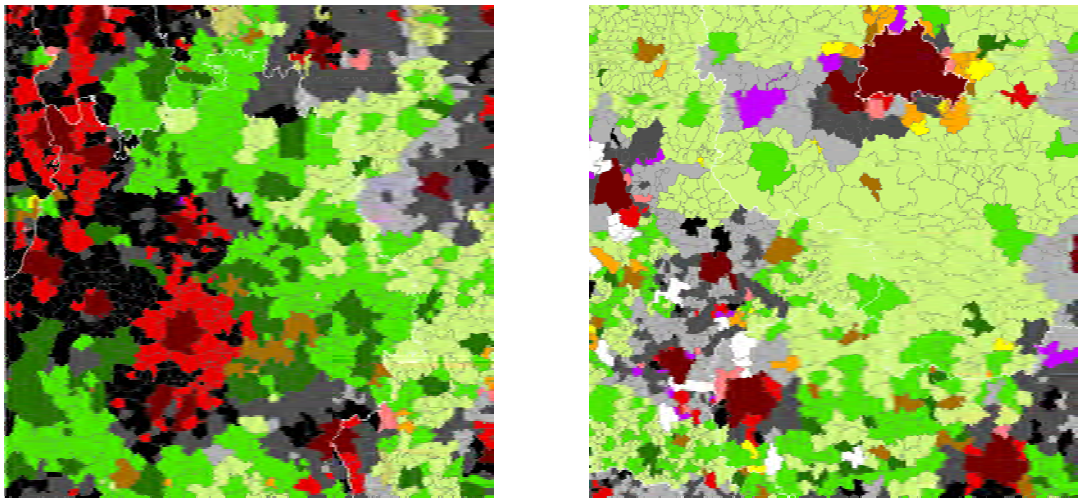


Abbildung 5-20: Detailansicht zu Agglomerations- und Verdichtungseigenschaften (Stuttgart und Berlin)

[531] Vgl. SIEDENTOP [2003, S. 186], siehe Ausführungen zur Region Berlin bei SIEDENTOP [2003, S. 46]

5.3 Diskussion der Ergebnisse

Es wurden zwei Untersuchungsansätze verfolgt, um das Gemeindesystem im Kontext von räumlicher Vielfalt und Polyzentralität zu beschreiben. Im Folgenden sollen beide methodischen Ansätze und das Datenmaterial diskutiert werden. Anschließend wird auf bestehende Ansätze der Raumbeobachtung eingegangen, um das im Wesentlichen empirisch gefundene Ergebnis in bestehende Ansätze einzuordnen. Vorab sei darauf hingewiesen, dass analytische Ansätze der Regionalisierung im Gegensatz zu synthetischen zu einer Abgrenzung von Verflechtungsgebieten geführt haben. Verflechtungsgebiete bestehen häufig aus einem Zentrum oder Kerngebiet und einem mehr oder weniger abgestuften Umland. Ideale Beispiele von Ansätzen zu Verflechtungsgebieten zeigen CHRISTALLER (1933), BURGESS (1925) und BOUSTEDT (1953).[532]

Der erste gewählte methodische Ansatz dieser Arbeit basiert auf der Verwendung von Gauß-Mixtur-Modellen zur Bestimmung von Beschäftigungsschwerpunkten. Es handelt sich um einen Untersuchungsvorschlag, um Gemeinden nach Beschäftigungsschwerpunkten genauer zu charakterisieren und Aussagen über die räumliche Verteilung zu ermöglichen.

Dieser Ansatz stellt eine innovative Methode dar, mit deren Hilfe man sich in kürzester Zeit einen Gesamtüberblick von Gemeinden verschaffen kann, ohne für jeden Wirtschaftszweig oder Wirtschaftssektor getrennt eine in diesem Fall stark subjektiv bedingte Interpretation vornehmen zu müssen. Der Untersuchungsansatz hängt in hohem Maß von der Festlegung der Entscheidungsgrenze ab, die bestimmt, bei welcher Ausprägung ein Beschäftigungs-schwerpunkt definiert wird. Zu beachten ist, dass zu diesem Zweck die Gauß-Mixtur-Modelle mit Hilfe des sogenannten EM-Algorithmus (Abschnitt 2.3.5.2) aufgebaut wurden, die festlegen, ob Gemeinden große oder weniger große Beschäftigungsanteile in einem Wirtschaftszweig oder Wirtschaftssektor aufweisen. Da der Algorithmus selbst immer nur ein lokales Optimum bestimmt, wurden die Ergebnisse mehrfach berechnet, und das hier dargestellte Resultat konnte für diese Fälle bestätigt werden.

Die Gauß-Mixtur-Modelle wurden nicht dazu eingesetzt, um den Verteilungsverlauf genau zu modellieren. Dies wurde im zweiten Untersuchungsansatz durchgeführt und dient dort als Grundlage für eine anschließende Klassenbildung. Zukünftig ist es denkbar, nach weiteren methodischen Untersuchungsansätzen zu suchen, um die gefundenen Ergebnisse und Entscheidungsgrenzen zu validieren.

---

[532] Vgl. BURGESS [1925], CHRISTALLER [1933] und BOUSTEDT [1953, S. 47-62]

Mit Blick auf die Qualität der gemeindescharfen Daten ist darauf hinzuweisen, dass es sich nur um Anteilswerte der sozialversicherungspflichtig Beschäftigten in einem Wirtschaftszweig oder Wirtschaftssektor handelt. Dadurch existiert eine gewisse Untererfassung der Gesamtbeschäftigung, was aber aus Gründen der Datenverfügbarkeit aktuell nicht zu umgehen ist. Dies wurde bereits im Zusammenhang mit anderen Forschungsarbeiten diskutiert (Abschnitt 3.4.1.7). Weiterhin besteht bei besserer Datenlage die Möglichkeit, die Gesamtstruktur der Erwerbstätigen zu untersuchen. Es sei darauf hingewiesen, dass der Ansatz nach Beschäftigungsschwerpunkten nicht die Anzahl der Beschäftigten in einer Gemeinde berücksichtigt, so dass keine Aussagen zu Konzentrationen von Beschäftigten oder einer Beschäftigungssuburbanisierung getroffen werden können. Besondere Bedeutung erhält der gewählte Ansatz im Kontext der Diversität bzw. Diversifizierung.[533] Messbar werden Eigenschaften einerseits in statischer Hinsicht und andererseits zukünftig in dynamischer Hinsicht in Bezug auf erfolgte Veränderungen von Beschäftigungsschwerpunkten. In Zukunft sind ähnliche Untersuchungen mit Nutzungsarten der Gebäude denkbar.

Der zweite zuvor gezeigte Untersuchungsansatz vergleicht die Gemeinden in Deutschland anhand von sechs ausgewählten raumstrukturellen Messgrößen. Es wurde geprüft, inwieweit mit den gewählten Variablen überhaupt eine Klassenbildung möglich ist. Infolge von Verteilungsuntersuchungen und durch eine Modellierung von Verteilungsverläufen mit Gauß-Mixtur-Modellen (siehe Abschnitt 4) konnten Entscheidungsgrenzen für jede Variable gefunden und zu einer Klassenbildung eingesetzt werden. Auf die Wichtigkeit der Datenaufbereitung wurde bereits in Abschnitt 2.1 hingewiesen, doch sei an dieser Stelle nochmals auf HARTUNG[534] verwiesen, der wichtige Grundvoraussetzungen für den Einsatz der multivariaten Verfahren nennt und die in dieser Arbeit berücksichtigt werden. Die Pareto Density Estimation bildet in Anlehnung an die Arbeitsmethodik von ULTSCH[535] ein fundamentales Element bei der Verteilungsuntersuchung und dient der Untersuchung der Klassifizierungsvariablen. Weiterhin werden Untersuchungsmöglichkeiten von Verteilungen mit PDE-Mischungen vorgestellt. Es ist möglich, die Ausprägungen von Variablen in Verbindung mit bereits bestehenden Klassenstrukturen genauer zu prüfen und weitere Informationen über die Struktur von Variablen zu erhalten. Die Einzeluntersuchung der Messgrößen umfasst eine Überprüfung der Daten mit Zentrale-Orte-Kategorien und Raumstrukturtypen des BBR. In den Daten konnten spezifische Strukturen erkannt werden.

---

533 Vgl. ULLMANN / DACEY [1962], Diversifikationsindex
534 Vgl. HARTUNG [2005]
535 Vgl. ULTSCH [2006 c]

Die Verortung von einzelnen Messgrößen erfolgt nicht auf subjektiv festgelegten Klassengrenzen oder aus Expertenwissen. Vielmehr bilden die Gauß-Mixtur-Modelle eine Charakteristik in den Daten ab und liefern eine zusätzliche Entscheidungsgrundlage.

In diesem Kapitel wurde auf die Bedeutung der Integration von Methoden der Strukturerkennung hingewiesen, um anhand dieser überhaupt erst über methodische Möglichkeiten für eine mehrdimensionale Klassifizierung im weiteren Verlauf entscheiden zu können. Gerade mit Hilfe einer Emergenten SOM (siehe Abschnitt 6.2) wurde erkannt, dass keine deutliche Struktur in den Daten der sechs Messgrößen vorhanden ist. Die Anwendung eines Clusteralgorithmus hätte daher nicht zu einer klar unterscheidbaren Gruppenstruktur geführt. Im Hinblick auf eine Eindeutigkeit der Gruppenstruktur erfolgte in dem gezeigten Untersuchungsfall mit sechs raumstrukturellen Messgrößen eine mehrdimensionale Gemeindebetrachtung auf Basis von Entscheidungsgrenzen.

Es ist anzumerken, dass bei einer Klassifizierung anhand von Entscheidungsgrenzen sehr große Klassenzahlen mit steigender Anzahl der Untersuchungsvariablen entstehen können. Im Sinne der Interpretierbarkeit erfolgte eine weitere Variablenreduktion durch visuelle Überprüfung. Insgesamt wurden bei diesem Untersuchungsansatz mit drei Variablen aus den Gemeindedaten 22 Klassen ermittelt, was der theoretisch zu erwartenden Klassenzahl (24) sehr nahe kommt.

Das Klassifizierungsergebnis vermittelt einen Eindruck von der Verstädterung, der Erreichbarkeit der nächstgelegenen Oberzentren und der Konzentration von Einwohnern und Beschäftigten innerhalb eines Untersuchungsraumes. Aus inhaltlicher Sicht besteht die Herausforderung gerade darin, nur mit einer begrenzten Variablenanzahl eine große Aussagetiefe zu erzielen. Zusätzlich sei auf Techniken der Klassenerklärung (Operationalisierung) verwiesen, die zukünftig eine Klassifizierung mit weiteren Eigenschaften belegen und so das Verständnis für eine Klasse vergrößern können. Generell ermöglicht die raumstrukturelle Abgrenzung eine Systematik mit vergleichbaren Raumeinheiten, so dass weitere Analysen von Strukturen und Entwicklungsprozessen im Anschluss möglich sind.

Im methodischen Sinne hat die dargelegte Einzeluntersuchung der sechs raumstrukturellen Variablen eine problemorientierte Sicht auf die Notwendigkeit einer Verteilungsuntersuchung geliefert. Zusätzlich wurde auf den Einsatzvorteil von Verfahren der Strukturerkennung hingewiesen. Die Modellierung der Verteilung unterstützt den Aufbau von Entscheidungsgrenzen sowie den Aufbau von üblichen Schwellwerten im urbanen Forschungsfeld.

Im Kontext von raumstrukturellen Abgrenzungen sei auf die sogenannte BOUSTEDT-Systematik verwiesen, die von OLAF BOUSTEDT[536] ausgehend von den in den USA definierten Standard Metropolitan Areas (SMA) Anfang der fünfziger Jahre erarbeitet und in der Folgezeit vor allem in der Landesplanung und Raumordnung angewandt wurde. In diesem Ansatz wurde die Berufspendlerquote dazu verwendet, die Einzugsbereiche der Kernstädte zu bestimmen. Im Vergleich zu dem in den USA entwickelten Modell mit einem Kernstadtkreis und einem Umlandkreis, wird das Umland der Kernstädte weiter differenziert, um auf diese Weise ländlich geprägte Bereiche von Vorstädten und Subzentren trennen zu können. Als Messgrößen wurden damals die Bevölkerungsdichte und die Agrarerwerbsquote eingesetzt, um eine Gliederung einer sogenannten Stadtregion in drei Zonen vorzunehmen: Ergänzungsgebiet, verstädterte Zone und Randzone. Das Modell der Stadtregion nach BOUSTEDT ist in Abbildung 5-21 aufgeführt. BOUSTEDT[537] definiert als Stadtregion: „denjenigen Umlandbereich im Agglomerationsraum einer (großen) Stadt, dessen Einwohner überwiegend nichtlandwirtschaftliche Berufe ausüben und von denen der überwiegende oder zumindest ein erheblicher Teil seine Existenzgrundlage in den Arbeitsstätten der Kernstadt hat.“

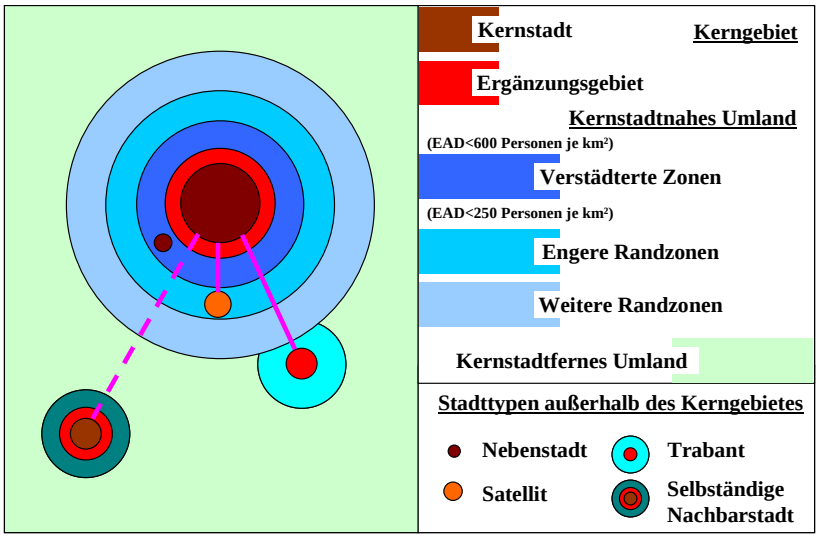


Abbildung 5-21: Modell der Stadtregion nach BOUSTEDT (EAD=Einwohner- und Arbeitsplatzdichte)[538]

Das Modell wurde für die Planungszwecke und als Instrument zur Beobachtung des Agglomerationsprozesses im Jahr 1953 entwickelt und 1971 nochmals durch BOUSTEDT an die geänderten Lebens- und Arbeitsverhältnisse angepasst. Die zonale Differenzierung basierte ursprünglich auf der Bevölkerungsdichte, jedoch wurde diese im weiteren Verlauf durch die Einwohner- und Arbeitsplatzdichte (EAD) als exaktere Messgröße ersetzt. Die

536 Vgl. BOUSTEDT [1953], BOUSTEDT [1970], BOUSTEDT [1975 b]
537 Vgl. BOUSTEDT [1975 a, S.432]
538 Eigene Bearbeitung (Entwurf: LICHTENBERGER [1998], angelehnt an BOUSTEDT [1975 a, S. 343])

engere Randzone wurde durch die Pendlerquote in das Kernstadtgebiet definiert und liegt 1971 bei mindestens 25 %. Die weitere Randzone wurde anhand des Anteilswertes der landwirtschaftlichen Berufstätigkeit vom Umland abgegrenzt, wobei der Schwellenwert von 50 % nicht überschritten werden durfte.

Als Mindestgröße zur Definition einer Agglomeration wurde eine Einwohnerzahl von 80.000 festgelegt, wobei die Kernstadt bzw. der ‚Agglomerationskern' nur in Ausnahmefällen eine Bevölkerungszahl von 40.000 Einwohnern unterschreiten durfte. Die Auswahl an Indikatoren hat sich im Verlauf der nachfolgenden Arbeiten kaum geändert. Nach BOUSTEDT wurden zusätzlich vier Stadttypen außerhalb des Kerngebietes definiert. Eine Nebenstadt übernimmt die zentralen Funktionen geringer Reichweite. Innerhalb der Randzone wird ein Satellitenort durch eine Auspendlerquote in das Kernstadtgebiet von über 90 % charakterisiert. Bei einer Trabantenstadt, die außerhalb der Randzone liegt, pendeln mehr als 50 % in das Kernstadtgebiet. Insgesamt wird in einer Trabantenstadt ein Einpendlerüberschuss erzielt, und es existiert ein eigenständiger Wirtschaftsbereich. Eine Nachbarstadt zeigt nur noch eine lockere Bindung an die Kernstadt und ist weiter entfernt als die Trabantenstadt.

ISENBERG[539] führte zusätzlich im Jahr 1957 den Begriff der Ballungsräume ein, um die polyzentrischen Verstädterungsformen zu kennzeichnen (damals mindestens 500.000 Menschen auf einer Fläche von 500 km²).

Der Begriff Verdichtungsraum wird in der Folgezeit aus einer Kombination von Ballungsräumen und Stadtregionen aufgebaut, wobei als zusätzliche zonale Differenzierungskriterien die EAD und die relative Bevölkerungszunahme nach BORCHERDT[540] verwendet werden.

Im Jahr 2000 wurden sogenannte BIK-Regionen[541] für das wiedervereinigte Deutschland im Anschluss an eine bereits erneut erfolgte BOUSTEDT-Revision aus dem Jahr 1987 aufgebaut. Die Gemeinden werden hier über ein zielgerichtetes Pendlerverhalten analytisch an die Zentren angebunden, so dass eine fast flächendeckende Struktur von Verflechtungsgebieten unterschiedlicher Größe definiert werden kann. Berechnet wird die Pendlerquote, indem die Zielpendlerquote auf eine gemeinsame Kernstadt gemessen wird (Entscheidungskriterium: Mindestens 7 % der Wohnbevölkerung pendeln als sozialversicherungspflichtig Beschäftigte

539 Vgl. ISENBERG [1957]

540 Vgl. BORCHERDT [1983]

541 Der Begriff BIK-Regionen (BIK Aschpurwis+Behrens GmbH) ist aus den Stadtregionen nach BOUSTEDT entstanden. Auf der Basis der Volkszählung von 1987 wurde für die alten Bundesländer eine Aktualisierung des Ansatzes aus dem Jahr 1970 erarbeitet. Durch die Wiedervereinigung wurde eine erste methodische Anpassung erforderlich, und im Jahr 2000 sind die BIK-Stadtregionen für ganz Deutschland gemeindescharf nochmals überarbeitet worden. Vgl. BEHRENS / MARHENKE [1997, S. 165-186]

in diese Kernstadt). Die Pendlerquote ermöglicht die Unterscheidung von Gemeinden, die entweder einer ‚BIK-Region' angehören oder außerhalb dieser liegen. Die Gemeinden in einer BIK-Region werden mit einer Vier-Klassen-Systematik anhand der Einwohner- und Arbeitsplatzdichte zusätzlich gegliedert: Kernbereich, Verdichtungsbereich, Übergangsbereich, Peripherer Bereich.

Im Gegensatz zu der in dieser Arbeit gewählten Gebäude- und Freifläche als Bezugsgröße wird bei den BIK-Regionen nach wie vor die Gemeindefläche insgesamt verwendet. Folgende Bereiche sind als Strukturtyp in einer BIK-Region definiert: Kernbereich (EAD $\geq$ 1000), Verdichtungsbereich (500 $\geq$ EAD < 1000), Übergangsbereich (150 $\geq$ EAD < 500) und peripherer Bereich (EAD $\leq$ 150) sowie Gemeinden ohne Zuordnung. Das Konzept der BIK-Regionen verwendet keine Messgröße zur Art der Beschäftigung wie dies noch im ursprünglichen BOUSTEDT-Ansatz vorgesehen war. Die BIK-Regionen sind heutzutage beispielsweise in der Umfragenforschung ein gebräuchliches Instrument, um auf dieser Grundlage weitere vertiefende Auswertungen durchzuführen.

Der anhand von Entscheidungsgrenzen aufgebaute Klassifizierungsansatz dieser Arbeit führt im Ergebnis auf die zu Beginn verwiesenen Verflechtungsgebiete (Kern-Umland-Beziehung).

CHRISTALLER definiert 1933 zentrale Orte und charakterisiert deren Einzugsbereiche. Als Verflechtungsmaß wird die ‚Reichweite' definiert, über die ein ‚Zentrum' von den Nutzern aus dem Umland in Anspruch genommen wird. Bei BURGESS wird eine Stadtregion dadurch beschrieben, dass sich ein konzentrisches Modell auf ein Oberzentrum ausgerichtet hat (Central Business District). Die Eigenschaften der konzentrischen Ringe werden über verschiedene Kriterien der Dichte und der Nutzung formuliert. Als Verflechtungsmaß wird die Ausrichtung auf das Zentrum verwendet. Wie bereits gezeigt, dient die Pendlerbeziehung bei Ansätzen nach BOUSTEDT als Verflechtungsmaß zwischen Kern und Umland.

Den Ausgangspunkt für den gewählten Klassifizierungsansatz dieser Arbeit bilden die im Jahr 2003 ausgewiesenen Oberzentren (Zentrale-Orte-Konzept). Die Raumordnung und Landesplanung verwendet den Begriff des Oberzentrums für die höchste Zentralitätsstufe. Dieses deckt den höher und spezialisierten Bedarf, übernimmt überregionale Verwaltungs- und Wirtschaftsfunktionen und besitzt vielfach Universitäten und Fachhochschulen, Spezialkrankenhäuser und Großkliniken, Theater, sowie Museen und Banken. Viele der Oberzentren besitzen eine Einwohnerzahl oberhalb von 100.000 Einwohnern. Die Oberzentren werden dazu verwendet, um einen ersten Bezugspunkt in einem Verflechtungsgebiet zu definieren.

Vor dem Hintergrund, dass die Bewohner und Beschäftigten einer Gemeinde häufig zu einem Zentrum in Interaktion treten, wurde die Fahrzeit zum nächstgelegenen Oberzentrum als Verflechtungsmaß gewählt. Der Interaktionsraum, der von den Gemeinden und dem dazugehörigen Oberzentrum aufgespannt wird, entsteht durch Untersuchung der Eigenschaften der Messgröße ‚Erreichbarkeit', so dass als maximale Verflechtung eine Fahrzeit von 30 Minuten identifiziert wurde. Gemeinden außerhalb dieses Interaktionsraumes werden durch den Begriff der Peripherie charakterisiert.

Im Vergleich zu dem Gliederungsansatz der BIK-Regionen, der aus dem BOUSTEDT-Ansatz hervorgegangen ist und sich an der Zielpendlerquote zur Kernstadt orientiert (Schwellenwert mindestens 7 % der Wohnbevölkerung), wurde hier der Interaktionsraum ermittelt, da in den vergangenen Jahren zunehmend interkommunale Austauschsprozesse neben den üblichen Kernstadt-Umland-Beziehungen große Relevanz erlangt haben und fortschreitende Dekonzentrationsprozesse zu erwarten sind.[542]

Abbildung 5-22 zeigt drei strukturräumliche Typen von Pendlerbewegungen, wobei zwischen Gemeinden im axialen (Anschluss an den Schienenverkehr vorhanden) und Gemeinden im interaxialen Raum (kein Anschluss an den Schienenverkehr) unterschieden wird. Die axial-radialen Pendlerbewegungen erfolgen zwischen Kernstadt und den Achsengemeinden oder zwischen den Achsengemeinden selbst. Die interaxial-radiale Pendlerbewegung ist dadurch gekennzeichnet, dass der Quell- bzw. Zielort eine Kernstadt bzw. eine interaxiale Gemeinde ist. Als tangentiale Pendlerbewegungen werden Verflechtungen bezeichnet, deren Quell- und Zielorte im interaxialen oder axialen Raum liegen.

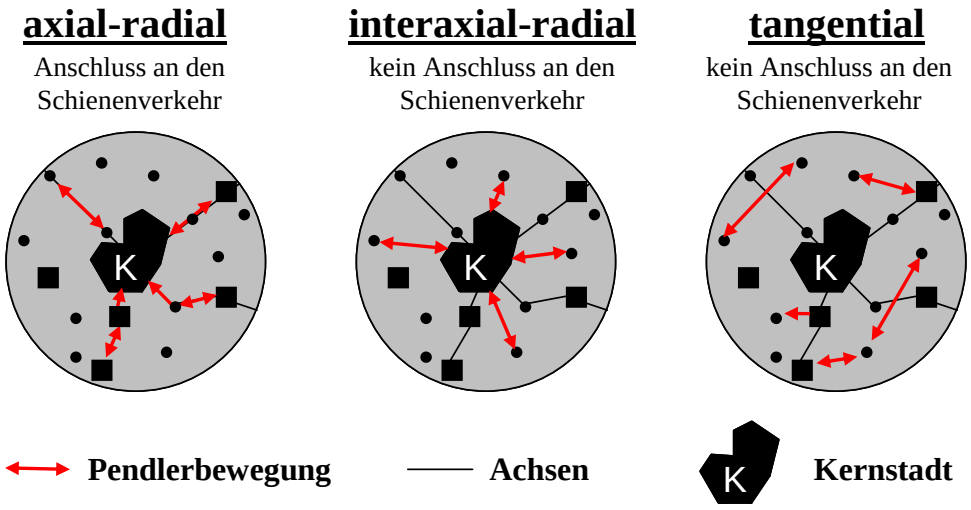


Abbildung 5-22: Idealtypische Verflechtungsmuster (Pendlerbewegung) [543]

[542] Vgl. SIEVERTS [1997] und BRAKE et al. [2005]
[543] Eigene Bearbeitung (Entwurf von SIEDENTOP [2003])

Nach SIEDENTOP[543] lebt und arbeitet mehr als die Hälfte der Bevölkerung des suburbanen Raumes in Gemeinden mit unmittelbarem Zugang zu einer Autobahn, wobei dabei deutliche regionale Unterschiede bestehen. Des Weiteren beziehen sich tangentiale Verflechtungsmuster in einigen Regionen Deutschlands sogar auf mehr als die Hälfte aller Pendlerbewegungen.[544] Da für diesen Klassifizierungsansatz lediglich Daten zur Erreichbarkeit im motorisierten Individualverkehr der Verbandsgemeinden vorliegen, sollte der Ansatz zukünftig durch Daten zum Schienenverkehr oder dem öffentlichen Nahverkehr verfeinert werden.

Der Klassifizierungsansatz verfolgt wie die bestehenden Ansätze den Gedanken, dass Verdichtungs- und Agglomerationsprozesse aus der Siedlungstätigkeit von privaten Haushalten und Personen sowie auch von Betrieben und Beschäftigten resultieren, die dabei konsumtive und produktive Tätigkeiten ausüben. BOUSTEDT[545] hat in der ersten Revision seines Ansatzes dazu die Bevölkerungsdichte durch die Einwohner- und Arbeitsplatzdichte ersetzt. Der Untersuchungsansatz der vorliegenden Arbeit verwendet ebenfalls die Summe der Einwohner und Arbeitsplätze, jedoch dient als Bezugsgröße die Gebäude- und Freifläche im Sinne der tatsächlich ausgelasteten Flächen einer Gemeinde und nicht die Katasterfläche.

Während im ersten Ansatz von BOUSTEDT noch der Anteil der Beschäftigten in der Landwirtschaft Berücksichtigung fand, um das Umland genauer abzugrenzen, ist in den Folgearbeiten aus Gründen von wirtschaftlichen Umstrukturierungsprozessen darauf verzichtet worden. Die Untersuchung der Beschäftigungsschwerpunkte hat in meiner Arbeit die Betrachtungsschärfe auf das deutsche Gemeindesystem zusätzlich erhöht und gezeigt, dass der Primäre Sektor bzw. der Beschäftigungsschwerpunkt ‚Landwirtschaft, Fischerei und Fischzucht' nur vereinzelt im westdeutschen Bundesgebiet identifiziert vorliegt. In der Regel sind auch im Umland der Oberzentren Beschäftigungsschwerpunkte im sekundären und tertiären Sektor zu finden.

Der Verstädterungsgrad wurde in den bisherigen vergleichbaren Ansätzen nicht direkt als Abgrenzungskriterium einbezogen. Der Verstädterungsgrad wird durch die Bewertung des Anteils der Siedlungs- und Verkehrsfläche an der Gemeindefläche insgesamt gemessen und bezieht sich auf die statistischen Daten der Flächenerhebung nach Art der tatsächlichen Nutzung. SIEDENTOP[544] stellt fest: „Insgesamt können Siedlungsstrukturkonzepte mit Bezug zu Dichte und Zentralität weiterhin empirische Legitimität beanspruchen".

---

544 SIEDENTOP [2003, S. 126 ff.], siehe zur Vertiefung: SIEDENTOP et al. [2005 c]
545 Vgl. BOUSTEDT [1975 a, S. 344]

Im Gegensatz zu den aktuellen Vorgehensweisen zum Aufbau der BIK-Regionen werden in dieser Arbeit für den Aufbau der Verflechtungsgebiete lediglich die Oberzentren als Bezugspunkt verwendet. Der BIK-Ansatz geht dagegen von einer wesentlich differenzierteren Mindestgröße der Kernstadt aus, die bei 5000 Einwohnern angesetzt ist und im weiteren Verlauf in Verbindung mit den angebundenen Gemeinden zur Regionsabgrenzung dient. Je nach Größe des Regionstyps werden Ballungsräume, Stadtregionen, Mittelzentrengebiete sowie Unterzentrengebiete unterschieden.

Den Ausgangspunkt zum Aufbau der Verflechtungsgebiete dieser Arbeit bildete der Aspekt, dass das überproportionale Wachstum der nichtzentralen Orte in Westdeutschland und die überproportionalen Bevölkerungsverluste zentraler Orte in Ostdeutschland in den meisten Agglomerationen zu einer ansatzweise erkennbaren Nivellierung der Zentrale-Orte-Hierarchie geführt haben.[546] Es entstand deshalb der Gedanke, Oberzentren zunächst als Agglomerationszentrum zu begreifen und den mit Hilfe der ‚Fahrzeit' ermittelten Interaktionsraum und darüber hinaus den Peripherraum hinsichtlich der Dichte und Verstädterungscharakteristika zu beobachten. Das Ergebnis ermöglicht einerseits die Identifizierung von polyzentrischen Regionalstrukturen im Verflechtungsbereich von Oberzentren und andererseits die Aufdeckung von monozentrisch geprägten geringer besiedelten Regionalstrukturen. Die Gemeinden in der sogenannten Peripherie wurden nicht wie bei den bereits dargestellten BIK-Regionen vertiefend auf ihre Anbindung an Mittelzentren oder Unterzentren untersucht. Jedoch wurde eine Gemeinde in der sogenannten Peripherie als Regionalzentrum bezeichnet, falls ein größerer Verstädterungsgrad gemessen werden konnte. Mit Blick auf die fortschreitenden komplexen Einsatzmöglichkeiten der geografischen Informationssysteme wird zukünftig auch eine Anbindung der statistischen Daten an die tatsächliche Siedlungsstruktur in Form eines rasterzellenbasierten Ansatzes denkbar.[547] Die zonale Abgrenzung nach der Auslastung von Gebäude- und Freiflächen kann dann mit kleinräumigerem Vorgehen weiter verfeinert werden. Es wird damit außerdem möglich, bisherige Ansätze zu überprüfen.

Als weitere Untersuchungskriterien können darüberhinaus die Daten der Geocomputation in den Raumstrukturansatz integriert werden. Verwiesen sei hier beispielsweise auf die im Abschnitt 3.9 diskutierten Daten zur räumlichen Konfiguration nach THINH[548] (z.B. Zerklüftungsgrad, Vernetzungsgrad), um Verdichtungs- und Agglomerationsprozesse weiter zu präzisieren.

---

546 Vgl. SIEDENTOP [2003, S. 192] und ZECK [2003, S. 725 ff.]
547 Vgl. GIFFINGER et al. [2006]
548 Vgl. THINH [2004 a]

5.4 Fazit

Durch die Ermittlung von Beschäftigungsschwerpunkten auf der Grundlage von Entscheidungsgrenzen erfährt die deskriptive Raumbeobachtung eine zusätzliche Betrachtungsperspektive. Es wurde ein Ansatz gezeigt, um Gemeinden zu ermitteln, die entweder über einen eindeutigen Beschäftigungsschwerpunkt verfügen oder durch Diversität bzw. keinen Beschäftigungsschwerpunkt gekennzeichnet sind. Ermittelt wurden die Beschäftigungsschwerpunkte der Gemeinden in den Wirtschaftszweigen (z.B. Kredit- und Versicherungsgewerbe) und in den Wirtschaftssektoren (z.B. Tertiärer Sektor).

Des Weiteren wurde gezeigt, dass die Klassenbildung auf Basis von drei ausgewählten Messgrößen in Anlehnung an BOUSTEDT[549] zur Erfassung und Strukturierung des Agglomerations- und Verdichtungsprozesses besonders geeignet ist. Generell ist eine regional sehr unterschiedliche Charakteristik von Gemeindestrukturen festzustellen, insbesondere im Vergleich der Agglomerationsgebiete und der Peripherie. Mit Blick auf die Verstädterung ist der Rückgang des Freiraums am Rande der Kernstädte im engeren suburbanen Raum deutlich zu bemerken. Die Suburbanisierung hat dazu geführt, dass sich in Teilbereichen flächenhaft bebaute Stadtregionen entwickelt haben, die sich weit in das Umland ausdehnen und im Kontext der Zwischenstadtdiskussion in der jüngsten Vergangenheit eine große Relevanz erlangt haben. Eine vertiefende Klassenerklärung ist zukünftig im Sinne der Operationalisierung zusätzlich vorstellbar, um z.B. weitere Erkenntnisse über die von Agglomerations- und Verdichtungsprozessen betroffenen Gemeinden zu erhalten. Im Hinblick auf das in der Raumordnungspolitik verwendete normative Zentrale-Orte-Konzept ist zusätzlich zu den hier dargestellten Ergebnissen ergänzend anzumerken, dass von den bestehenden zentralen Orten in der Vergangenheit kaum lenkender Einfluss auf die Problematik der Suburbanisierung ausgegangen ist. Es wurde nicht der Dispersion von Einzelhandelsgroßbetrieben und Beschäftigtenzentren im produzierenden Gewerbe und zunehmend auch im Tertiären Sektor entgegengewirkt, so dass vielfach statt eines prägenden Oberzentrums polyzentrische Regionalstrukturen entstanden sind. Diese sogenannten ‚zentralörtlichen Kooperationsräume'[550] werden nicht von den klassischen Zentrale-Orte-Kategorien repräsentiert, die innerhalb des Zentrale-Orte-Konzeptes der Raumordnungspolitik entwickelt wurden. Im Verflechtungsbereich von Oberzentren wird die Ausweisungspraxis zentraler Orten diskutiert, als reformbedürftig betrachtet und die zusätzliche Zentralitätsstufe ‚Metropolregion' debattiert.[551]

---

549 Vgl. BOUSTEDT [1953], BOUSTEDT [1975 b]
550 Vgl. ARL [2002]
551 Vgl. ARL [2002], SIEDENTOP [2003, S. 192 ff.], siehe ‚Metropolregion' bei BLOTEVOGEL [2005]

6 Charakterisierung räumlicher Entwicklungstendenzen durch Visual Mining

6.1 Untersuchungsaufgabe: Schrumpfung, Wachstum oder Stagnation

Als Herausforderung an eine zukunftsbeständige Stadtentwicklung hat sich ein Paradigmenwechsel seit geraumer Zeit vollzogen, da nicht mehr das Wachstum allein die Stadtentwicklung in Deutschland bestimmt, sondern eine Vielzahl der Städte und Gemeinden zunehmend mit Problemen der Stagnation, des Nachfragerückgangs und der Schrumpfung als dauerhafte Entwicklung konfrontiert ist.[552] Generell sind Schrumpfung bzw. Wachstum als systemische Prozesse zu verstehen, die dadurch beschreibbar sind, dass rückläufige negative bzw. wachsende positive Entwicklungen dominant werden.

Wachstum und Schrumpfung in der Raumentwicklung sind in unterschiedlichen räumlichen Zusammenhängen zu finden.[553] Innerhalb von Stadtregionen ist eher die regional bedeutsame Suburbanisierung die Triebfeder für die kleinräumige Verteilung von Wachstum und Schrumpfung. Ganze Regionen werden durch die eher überregional bedeutsame Schrumpfung mit allen Folgewirkungen vor das Problem der langfristigen Sicherung der öffentlichen Daseinsvorsorge gestellt. Des Weiteren lässt sich Wachstum auch regional weiträumig feststellen. Dieses erfordert Strategien zur Sicherung und nachhaltigen Unterstützung, um weiteren Nutzen für die Wachstumsregionen und die gesamte Volkswirtschaft zu sichern.

Im historischen Rückblick ist nachzuverfolgen, dass die Entwicklung des Stadtwesens in Deutschland keineswegs linear verlief. Mit der einsetzenden Industrialisierung zu Beginn des 19. Jahrhunderts enstand zwar die normative Gleichsetzung von Stadtentwicklung und Wachstum, die jedoch keineswegs begründbar ist, da bei weitem nicht alle Städte vom demographischen und ökonomischen Aufschwung profitierten.[554] Darüber hinaus sind im Mittelalter z.B. durchaus sogenannte Wüstungen oder städtischer Verfall im Zuge von Kriegen oder Epidemien bekannt.[555] Im Gegensatz zu stadtplanerischen Utopien eines urbanen Wachstums zeigte sich in den 50er- und spätestens zu Beginn der 70er-Jahre des vorigen Jahrhunderts in allen Industrienationen, dass Schrumpfung und Wachstum von Städten gleichzeitig auftretende Phänomene paralleler Normalität einer urbanen Entwicklung darstellen.[556] Der Beirat für Raumordnung mahnte 1972 zu einer differenzierten Betrachtung der Verdichtungsräume und trennte expandierende, stagnierende und schrumpfende Räume.[557]

---

552 Vgl. Programme: ‚Stadtumbau Ost' [2002] / ‚Stadtumbau West' [2004], MÜLLER / SIEDENTOP [2004]
553 LIEBMANN / ROBISCHON [2003], MÜLLER / SIEDENTOP [2003], NAGLER et al. [2004].
554 Vgl. REINBORN [1996], REULECKE [1985, S.49]
555 Vgl. BENKE [2004, S. 7-14], KÜNTZEL [2006]
556 Vgl. OSWALT [2004], [2005], [2006]
557 Vgl. ARL [1995, S. 1010]

In den 80er-Jahren widmete sich eine Debatte dem Wachstum, der Stagnation und der Schrumpfung, welche im Kontext der vor allem in Industrieregionen einsetzenden Deindustrialisierung (z.B. Abbau der Schwerindustrie) in der früheren Bundesrepublik geführt wurde.[558] In der Diskussion stand zusätzlich eine Ausdifferenzierung der Großstädte nach Wachstums- oder Schrumpfungseigenschaften.[559]

In der ehemaligen DDR lässt sich die Gleichzeitigkeit von Schrumpfung und Wachstum ebenso beobachten, da beispielsweise die Hauptstadt Berlin-Ost als kontinuierlicher Anziehungspunkt der Binnenwanderung wirkte, während der ländliche Raum bis zum Ende der DDR durch Schrumpfungsprozesse gekennzeichnet war.[560]

Nach der Wiedervereinigung wurde der Osten Deutschlands stärker von Schrumpfungsproblemen erfasst als Westdeutschland. Ursachen liegen in den nicht bewältigten Folgen der gesellschaftlichen Transformation und den fortschreitenden negativen Auswirkungen eines komprimiert auftretenden strukturellen Wandels.[561] HANNEMANN[562] verweist auf einen Entwicklungsprozess in einigen Regionen Ostdeutschlands, insbesondere im ländlichen Raum, der zu einer Herausbildung von sogenannten ‚Deinvestitionsgebieten' geführt hat. Die sogenannte ‚Deökonomisierung' wird begrifflich zur Beschreibung der geringen Effizienz und Konkurrenzfähigkeit der lokalen ostdeutschen Wirtschaftskreisläufe eingeführt.

LANG et al.[563] setzen sich mit der Komplexität einer Stadtentwicklung unter Schrumpfungsbedingungen im Hinblick auf ostdeutsche Entwicklungstrends auseinander und verwenden das Bild der ‚Lean City' als Entwurf der Stadtentwicklung unter Schrumpfungsbedingungen.

Bis zum Jahr 2050 wird in der Bundesrepublik Deutschland ein Rückgang der Bevölkerung von 82,5 Mio. auf 78,9 bis 67,1 Mio. Einwohner erwartet.[564] Aufgrund von konzentrierten Suburbanisierungsprozessen wird bis 2025 ein Bevölkerungsrückgang von bis zu 25 Prozent in ostdeutschen Städten erwartet sowie die Entstehung von Stadtregionen mit mehr oder weniger zusammenhanglosen Teilgebieten. In Zukunft ist aufgrund dieser Prognosen über die mittel- und langfristige demografische Entwicklung und die weiterhin andauernden wirtschaftlichen Veränderungsprozesse damit zu rechnen, dass neben wachsenden und weiter prosperierenden Regionen weite Teile Deutschlands von Schrumpfung betroffen sein werden.

---

558 Vgl. GÖB [1977, S. 149], HÄUßERMANN / SIEBEL [1983], [1987, S. 111], STREICH [1987, S. 128]
559 Vgl. HÄUSSERMANN / SIEBEL [1987, S. 44.] und [1988, S. 84 ff.]
560 Vgl. KRESS [2006]
561 Vgl. WUTTKE [2003, S. 5 ff.]
562 Vgl. HANNEMANN [2002], HANNEMANN et al. [2002], vgl. ländlicher Raum: HANNEMANN [2004]
563 Vgl. LANG / TENZ [2003]
564 Vgl. Vorausschätzung der Bevölkerung: EUROSTAT [2000], DIW [2004], DESTATIS [2003]

Die zentrale Frage dieses Kapitels bezieht sich auf die Möglichkeit zur Feststellung von regionalen Wachstums- oder Schrumpfungstendenzen in Deutschland. Gesucht werden für die Städte und Gemeinden Ähnlichkeitsmuster, die sich mit Hilfe einer mehrdimensionalen Datenbasis berechnen lassen.[565]

Mit Blick auf die Komplexität von Schrumpfungs- und Wachstumsprozessen handelt es sich bei den Untersuchungen in diesem Kapitel um eine erste Annäherung, die sich auf eine subjektiv ausgewählte Anzahl von nur wenigen Untersuchungsvariablen stützt. Es handelt sich um ein sehr einfach aufgebautes Messmodell, das einen gewissen abstrakten Charakter aufweist und primär als Arbeitsgrundlage für das Funktionsprinzip eines Klassifikators im Abschnitt 7 dienen soll. Das hier dargestellte Messmodell muss zukünftig ggf. zusätzliche Ergänzungen erfahren. Es erfolgt eine Annäherung an die Komplexität des Phänomens mit Hilfe des Wirkprinzips des Klassifikators und damit verbundener Identifizierungsvariablen.

Im Jahr 2003 untersuchen GATZWEILER et al.[566] bereits die Städte in Deutschland unter dem Aspekt des Schrumpfungs- bzw. Wachstumsverhaltens. Dabei wurden sechs Variablen (Bevölkerungsentwicklung, Gesamtwanderungssaldo, Arbeitsplatzentwicklung, Arbeitslosenquote, Realsteuerkraft und Kaufkraft) verwendet. Das Ergebnis ist eine manuelle Klassenbildung, die auf einem Ansatz basiert, der die Häufigkeit von Werten der Untersuchungsvariablen im untersten bzw. obersten Quintil berücksichtigt. Im Jahr 2006 wurde auf Ebene der Stadtteile eine kleinräumige Untersuchung in ähnlicher Weise durchgeführt.[567]

SIEVERTS[568] sieht als Konsequenz aus dem Phänomen ‚Stadtwachstum und Stadtschrumpfung zur gleichen Zeit' die verstädterte Landschaft. Im Falle von Wachstum treten demnach ggf. Prozesse des ‚Urban Sprawl' auf und verändern das bisherige Landschaftsbild, so dass Stadt nicht mehr ohne Region zu denken ist. Im Falle von Schrumpfung wird eine Perforierung der Stadt erkennbar, die der Natur im bisherigen Stadtraum weiteren Platz anbietet. Ergänzend sei anzumerken, dass die Suche nach positiven Wertvorstellungen für die städtischen Schrumpfungsprozesse als große Herausforderung angesehen wird.[569]

---

565 Vgl. WEISKE et al. [2005, S.9] heben die Mehrdimensionalität der Problemstellung hervor.

566 Vgl. GATZWEILER et al. [2003]

567 Vgl. GATZWEILER et al. [2006, S. 12]

568 Vgl. SIEVERTS [1997, u.a. S. 18 ff.]: „In Ermangelung eines besseren Begriffs wollen wir diese Gebilde, die aus ‚Feldern' unterschiedlicher Nutzungen, Bebauungsformen und Topographien bestehen, Zwischenstädte nennen: Sie breiten sich in großen Feldern aus, sie haben sowohl städtische wie landschaftliche Eigenschaften. […] Die Ursachen, die zu dieser diffusen Gestalt führen, sind jeweils zwar unterschiedlich, gemeinsam ist ihnen aber auf der ganzen Welt der Tatbestand, dass in jedem Fall die historischen stadtbildenden Kräfte und die durch sie gesetzten Begrenzungen an ihr Ende gekommen waren."

569 Vgl. LÜTKE-DALDRUP [2001, S. 40 ff.], GANSER [2001], LARSEN [1999], ROMERO [2003].

6.2 Herleitung und Darstellung der Ergebnisse

Um das Arbeitsprinzip eines Klassifikators im urbanen Kontext anzudeuten bzw. erklären zu können, wird – wie bereits erwähnt – als kanonisches Beispiel eine Fragestellung untersucht, die sich auf die regionalen Entwicklungstendenzen der Gemeinden in Deutschland bezieht. Auf Grundlage einer ersten Auswahl von Variablen ist die Frage zu klären, ob eine Klassenbildung sich eignet, dynamische Prozesse bzw. die Bipolarität von Schrumpfungs- oder Wachstumstendenzen in der jüngeren Vergangenheit zwischen 1999 und 2003 abzubilden.[570]

Die Veränderungen der Beschäftigten, der Bevölkerung und der Zu- und Fortzüge werden als erste Messgrößen verstanden, um eine Tendenz zum Wachstum bzw. Schrumpfen einer Gemeinde einzuschätzen. Als bestimmende Komponente für die Bevölkerungsentwicklung sind die Wanderungen (Zu- und Fortzüge in einer Gemeinde) neben den Geburten und Sterbefällen anzusehen. Wanderungsgewinne und -verluste ermöglichen die Messung der unterschiedlichen Standortattraktivitäten von Gemeinden. Die Auspendlerquote dient zur Messung von Pendlervorgängen und ihren Veränderungen über die Zeit. Tabelle 6-1 enthält die vier Variablen, die für eine erste Klassenbildung auch in bereits existierenden Arbeiten zu dieser Thematik als relevant erachtet wurden.[571] Zusätzlich zur Messgröße ist die Variablennotation der Verteilungsuntersuchungen aus Tabelle 6-2 aufgeführt.

Messgröße	Einheit	Notation
(1) Relative Bevölkerungsentwicklung zwischen 1999 und 2003	[%]	‚Bevölkerung'
(2) Zuzugssaldo zwischen 1999 und 2003	[Personen]	‚Zuzugssaldo'
(3) Relative Beschäftigungsentwicklung zwischen 1999 und 2003	[%]	‚Beschäftigung'
(4) Veränderung der Auspendlerquote zwischen 1999 und 2003	[%]	‚Auspendlerquote'

Tabelle 6-1: Übersicht zu ausgewählten dynamischen Kenngrößen

Die Ergebnisse der Verteilungsuntersuchung belegen, dass es sich nicht um normalverteilte Daten handelt (siehe QQ-Plot). Durch die Transformation mit $S\log(x) = sign(x)\cdot\log(|x|+1)$ wurde ein dichtomes Verhalten in den Daten erkannt und eine Bimodalität bzw. Multimodalität (siehe Auspendlerquote) von Lognormalverteilungen festgestellt.

[570] Da eine Untersuchung von Gemeindedaten eine sehr hohe räumliche Auflösung betrifft, sind im Jahr 2006 Daten der Wanderungs-, Beschäftigungs- und Bevölkerungsstatistik lediglich bis zum Jahr 2003 in aufbereiteter Form einsetzbar. Für das Jahr 2004 liegen aus der Wanderungsstatistik lediglich Daten zu den Zuzügen bereits vor. Darüber hinaus ist im Abschnitt 3.4.1.7 auf die Problematik der Vergleichbarkeit langer Zeitreihen von Gemeindedaten aus der Beschäftigungsstatistik hingewiesen. Weiterhin liegen gemeindebezogene Beschäftigtendaten erst seit 1996 überhaupt in Ostdeutschland vor.

[571] Vgl. GATZWEILER et al. [2003], vgl. SIEDENTOP et al. [2003, S. 89 ff.]

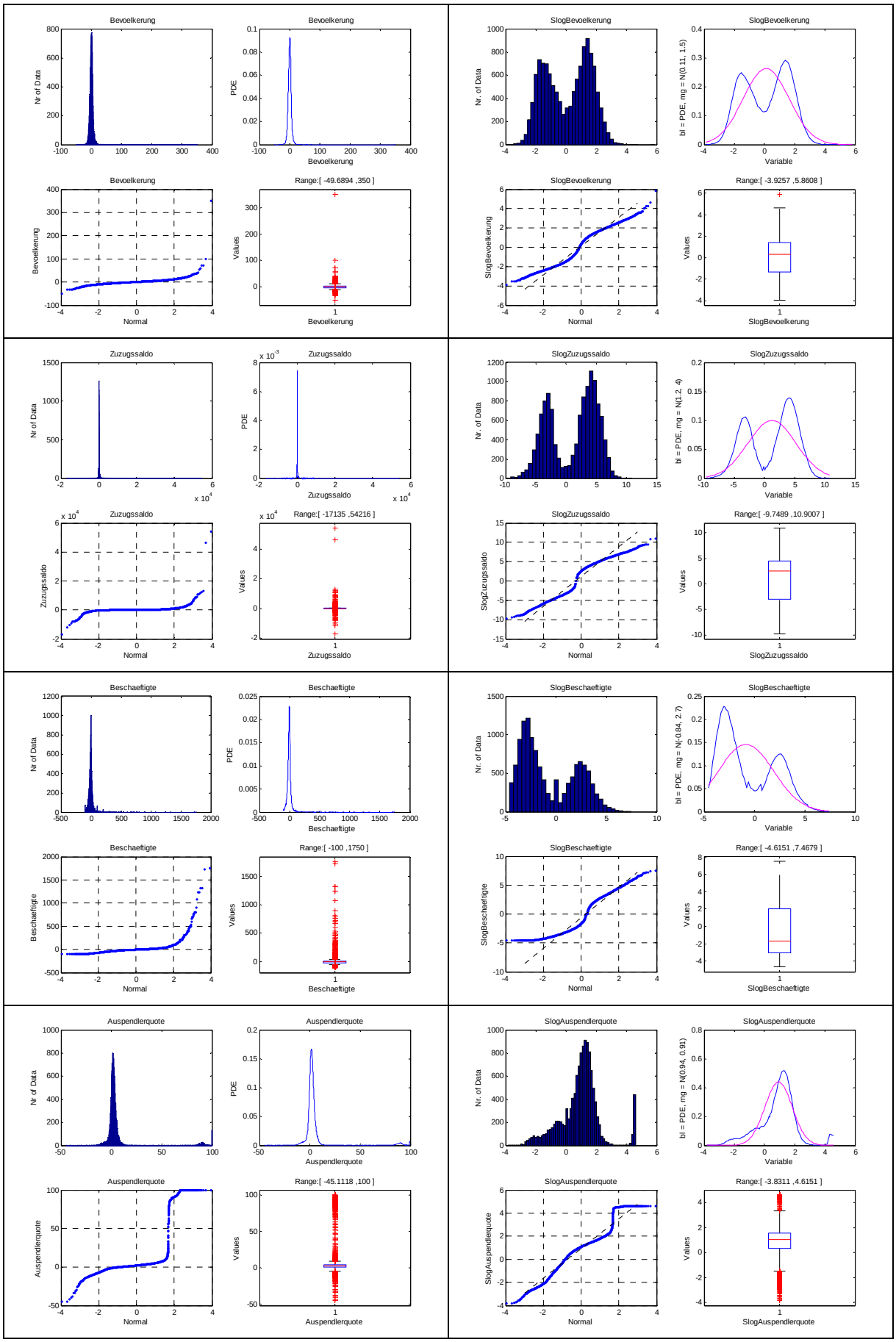


Tabelle 6-2: Übersicht zu den Verteilungsuntersuchungen der 4 dynamischen Kenngrößen

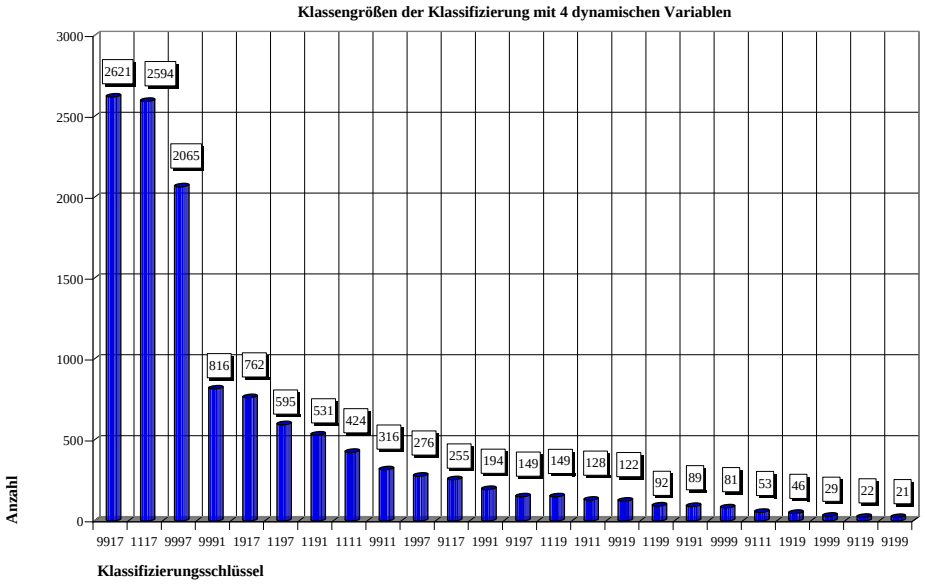
Die bereits ermittelten Ergebnisse der Verteilungsuntersuchung von vier dynamischen Variablen werden in Tabelle 6-3 zusammengefasst. Das dichotome Verhalten der Variablen wird im weiteren Verlauf für eine Klassenbildung innerhalb jeder einzelnen Variablen herangezogen, d.h. beobachtet werden Werteausprägungen im positiven bzw. negativen Bereich. Auf Basis der Entscheidungsregeln wurde die Klassenstärke zu jeder Variablen ermittelt.

Messgröße	Verteilungshypothese	Verteilung (grob inspiziert)	Entscheidungsregel	Klassenstärke
(1) Bevölkerung	**Bimodale Verteilung von Lognormalverteilungen, schiefe Verteilung im positiven wie im negativen Wertebereich (Data>0: linkssteil, Data<0 rechtssteil)** Große Mehrheit der Gemeinden mit geringen Veränderungen. Es werden sowohl im positiven wie im negativen Wertebereich um Null % sehr viele Gemeinden erwartet. Es gibt aber auch eine kleinere Menge an Gemeinden, die größere Bevölkerungsrückgänge und größere Bevölkerungszuwächse aufweisen. Es werden sowohl im positiven wie auch im negativen Wertebereich lognormalverteilte Daten erwartet, da es im positiven wie im negativen Wertebereich sich um links- bzw. rechtssteile Verteilungen handelt. Dadurch ist mit einer bimodalen Verteilung aus zwei lognormalverteilten Datensätzen zu rechnen.	**SLog(Data)** Bimodale Ausprägung a) Lognormalverteilte Daten bei Data>0 b) Lognormalverteilte Daten bei Data<0	Klasse 1: **Data <=0** Klasse 2: **Data >0**	Klasse 1: [5820], 46,82 % Klasse 2: [6610], 53,18 %
(2) Zuzugs saldo	**Bimodale Verteilung von Lognormalverteilungen, schiefe Verteilung im positiven wie im negativen Wertebereich (Data>0: linkssteil, Data<0 rechtssteil)** Es wird erwartet, dass die große Anzahl der Gemeinden ein relativ ausgeglichenes Zuzugssaldo aufweist, d.h. viele Gemeinden mit einem Saldo im Intervall [-10 10]. Darüber hinaus gibt es Gemeinden mit sehr großen Zuzugsmengen im Vergleich zu den Fortzügen bzw. sehr große Fortzugsmengen im Vergleich zu den Zuzügen. Es wird im positiven Wertebereich mit einer linkssteilen Verteilung gerechnet. Im negativen Wertebereich wird mit einer rechtssteilen Verteilung gerechnet. Dadurch ist mit einer bimodalen Verteilung zu rechnen, die durch zwei Lognormalverteilungen charakterisiert wird (log(data)>0 und log(data)<0).	**SLog(Data)** Bimodale Ausprägung a) Lognormalverteilte Daten bei Data>0 b) Lognormalverteilte Daten bei Data<0	Klasse 1: **Data <=0** Klasse 2: **Data >0**	Klasse 1: [4974], 40,02 % Klasse 2: [7456], 59,98 %
(3) Beschäftigung	**Bimodale Verteilung von Lognormalverteilungen, schiefe Verteilung im positiven wie im negativen Wertebereich (Data>0: linkssteil, Data<0 rechtssteil)** Große Mehrheit der Gemeinden mit geringen Veränderungen. Es werden sowohl im positiven wie im negativen Wertebereich um Null % viele Gemeinden erwartet. Dann gibt es auch eine kleinere Menge, die größere Beschäftigungsrückgänge und größere Beschäftigungszuwächse aufweisen. Es werden sowohl im positiven wie auch im negativen Wertebereich lognormalverteilte Daten erwartet, da es im positiven wie im negativen Wertebereich sich um linkssteile Verteilungen handelt. Dadurch ist mit einer bimodalen Verteilung aus zwei lognormalverteilten Datensätzen zu rechnen.	**SLog(Data)** Bimodale Ausprägung a) Lognormalverteilte Daten bei Data>0 b) Lognormalverteilte Daten bei Data<0	Klasse 1: **Data <0** Klasse 2: **Data >=0**	Klasse 1: [7492], 60,27 % Klasse 2: [4938], 39,73 %
(4) Auspendlerquote	**Bimodale Verteilung von Lognormalverteilungen, schiefe Verteilung im positiven wie im negativen Wertebereich (Data>0: linkssteil, Data<0 rechtssteil)** Die Veränderung der Auspendlerquote setzt sich aus dem Saldo der Auspendlerquote des Jahres 2003 und der Auspendlerquote des Jahres 1999 zusammen. Dadurch ergeben sich sowohl Rückgänge und Zuwächse und die Veränderung drückt sich in negativen und positiven Prozentwerten aus. Vermutlich gibt es einige Gemeinden, die sich nur um einige wenige Prozentpunkte über die gewählte Zeitspanne von 1999 bis 2003 verändern. Darüber hinaus gibt es wahrscheinlich Gemeinden mit größeren Zuwächsen und größeren Rückgängen. Im positiven Wertebereich ist eine linkssteile Verteilung denkbar bzw im negativen Wertebereich eine rechtssteile Verteilung.	**SLog(Data)** Multimodale Verteilung, Lognormalverteilungen a) Lognormalverteilte Daten bei Data>=0 b) Lognormalverteilte Daten bei Data<0	Klasse 1: **Data <=0** Klasse 2: **0>Data<50** Klasse 3: **Data >=50**	Klasse 1: [2551], 20,52 % Klasse 2: [9317], 74,96 % Klasse 3: [562], 4,52 %

Tabelle 6-3: Zusammenfassung der Datenaufbereitung der 4 dynamischen Kenngrößen

Unter Beachtung der Ausprägungen der vier Variablen ist die mehrdimensionale Klassifizierung der Gemeinden umsetzbar. Aufgrund theoretischer Kombinationsmöglichkeiten ist eine Struktur mit 72 Klassen zunächst mit Blick auf die Entscheidungsregeln denkbar. Die nachfolgende Tabelle zeigt die entwickelte Struktur mit 24 Klassen, die sich auf die tatsächlich vorhandenen dichotomen Eigenschaften jeder Gemeinde bezieht.

Dargestellt sind in Tabelle 6-4 die Klassen und die jeweilige Klassengröße sowie die Ausprägungen für jede der vier Variablen. Exemplarisch erklärt sei die objektstärkste Klasse mit 2621 Gemeinden (21,07 %), die einen Bevölkerungsanstieg mit enthaltenen Wanderungsgewinnen aufweist. Darüber hinaus wird bei dieser Klasse in den Gemeinden ein Beschäftigungsrückgang festgestellt sowie eine mittlere Zunahme der Auspendlerquote.



Ziffer 1	Ziffer 2	Ziffer 3	Ziffer 4
1 = Bevölkerungsrückgang, [%] 9 = Bevölkerungswachstum, [%]	1 = Fortzug, [Personen] 9 = Zuzug, [Personen]	1 = Beschäftigungsrückgang, [%] 9 = Beschäftigungszuwachs, [%]	1 = Auspendlerrückgang, [%] 7 = Mittlere Auspendlerzunahme, [%] 9 = Hohe Auspendlerzunahme, [%]

Tabelle 6-4: Klassengrößen der mehrdimensionalen Variablenbetrachtung

Im Hinblick auf die Bevölkerungs- und Beschäftigungsentwicklung als allgemein anerkannte Pressure-Faktoren von siedlungsstrukturellen Veränderungen wird eine weitere Aggregation der 24 Klassen vorgenommen. Es werden sechs Oberklassen erzeugt, die insbesondere Eigenschaften der Bevölkerungs- bzw. Beschäftigungsentwicklung berücksichtigen (Tabelle 6-5).

Bezeichnung	Bevölkerung	Zuzugssaldo	Beschäftigung	Auspendlerquote	Gemeinden
‚Wachstum der Bevölkerung und Beschäftigung'	Zunahme	Zunahme	Zunahme	nicht berücksichtigt	**2962**
‚Wachstum der Bevölkerung'	Zunahme	Zunahme	Abnahme	nicht berücksichtigt	**3059**
‚Wachstum der Beschäftigung'	Abnahme	Abnahme	Zunahme	nicht berücksichtigt	**1218**
‚Kompensation von Wanderungsverlusten'	Zunahme	Abnahme	Abnahme	nicht berücksichtigt	**330**
‚Kompensation von Wanderungsgewinnen'	Abnahme	Zunahme	nicht berücksichtigt	Zunahme	**1113**
‚Schrumpfung der Bevölkerung und Beschäftigung'	Abnahme	Abnahme	Abnahme	nicht berücksichtigt	**3167**
Sonderfälle					**581**

Tabelle 6-5: Aggregierte Klassen zu räumlichen Entwicklungstendenzen[572]

[572] Zusätzlich im Nebenteil B hinterlegt sind PDE-Plots zu den Variablen der 6 Oberklassen.

Abbildung 6-1 zeigt die Lokalisierung der Klassen. Es wird deutlich, dass die Entwicklung der Städte und Gemeinden in Deutschland sich durch eine Bipolarität von Wachstum und Schrumpfung kennzeichnen lässt. Im Osten konzentrieren sich Gemeinden mit Schrumpfungsansätzen und im Westen Gemeinden mit Wachstumsansätzen.

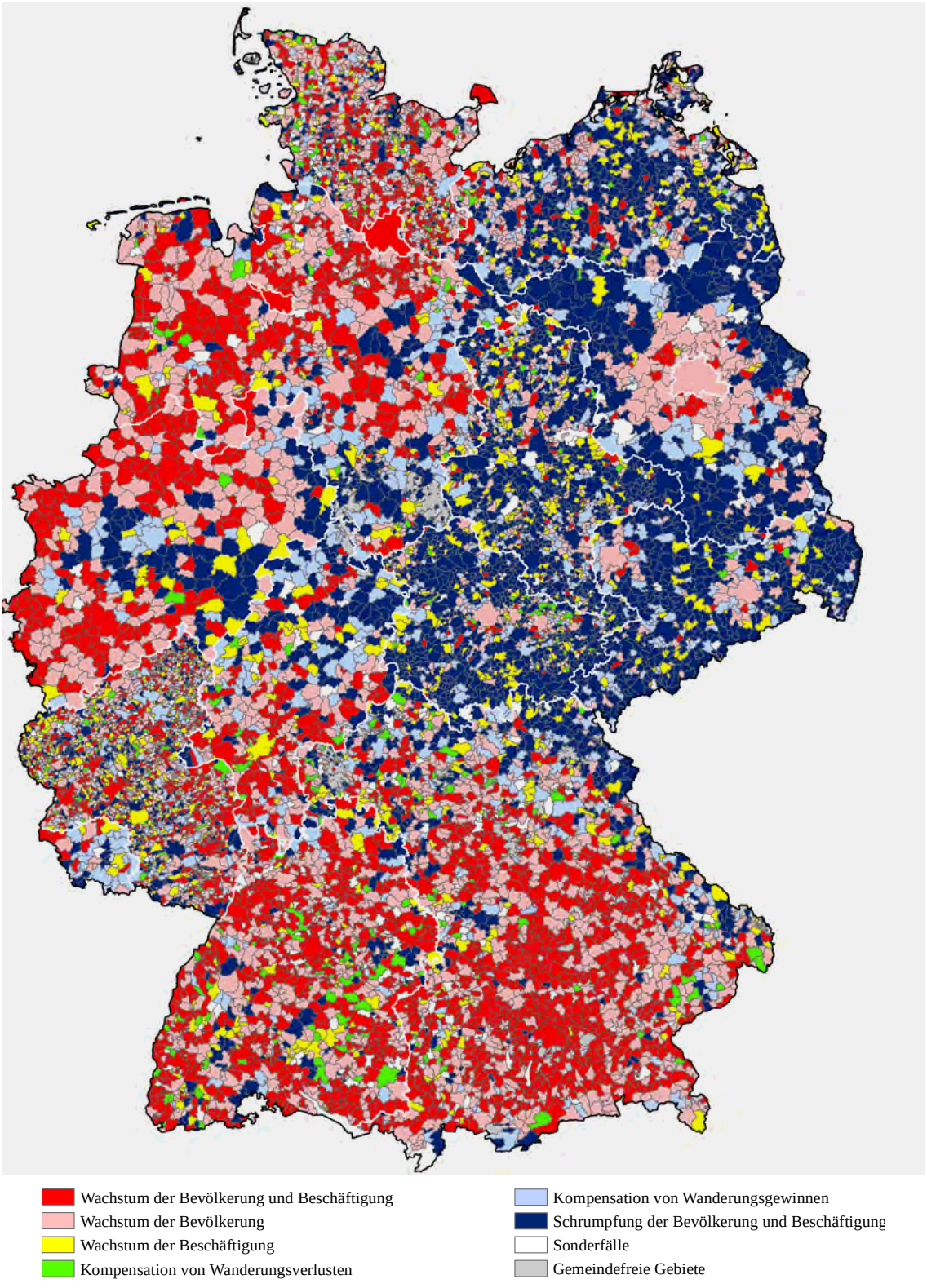


Abbildung 6-1: Verortung des Klassifizierungsergebnisses zu räumlichen Entwicklungstendenzen

Die Daten zu den räumlichen Entwicklungstendenzen werden im Folgenden genauer untersucht, indem zunächst Objekte ausgewählt werden, die über Eigenschaften verfügen, welche eine Klasse besonders charakterisieren (Kernelklassen). Dann werden diese einer zusätzlichen Strukturprüfung unterzogen. Die emergenten SOM werden hierzu eingesetzt, um sich zunächst einen Überblick von der hochdimensionalen Datenstruktur zu verschaffen und im Anschluss einen Clusteralgorithmus (U*C-Algorithmus) einzusetzen, der Cluster aus der U*C-Matrix ermittelt.

Klassische Algorithmen wie z.B. Ward oder andere hierarchische Clusteralgorithmen setzen voraus, dass die Daten unabhängig voneinander sind und einer gleichen Verteilung folgen. Es sei darauf verwiesen, dass im Forschungsfeld der Stadtplanung und der Regionalforschung die Daten oftmals mehrdimensional, räumlich autokorreliert und insbesondere heterogen vorliegen und damit gegen die Grundannahmen vieler Algorithmen verstoßen, falls vollständig auf eine Dateninspektion und Transformation verzichtet wird.[573]

Abbildung 6-2 zeigt das Ergebnis der Vorverarbeitung infolge von Standardisierung und Skalierung. Dargestellt sind die Zusammenhänge der vier Variablen in einem Streudiagramm (Scatter-Plot) und der Verteilungsverlauf jeder Variablen in einem Histogramm.

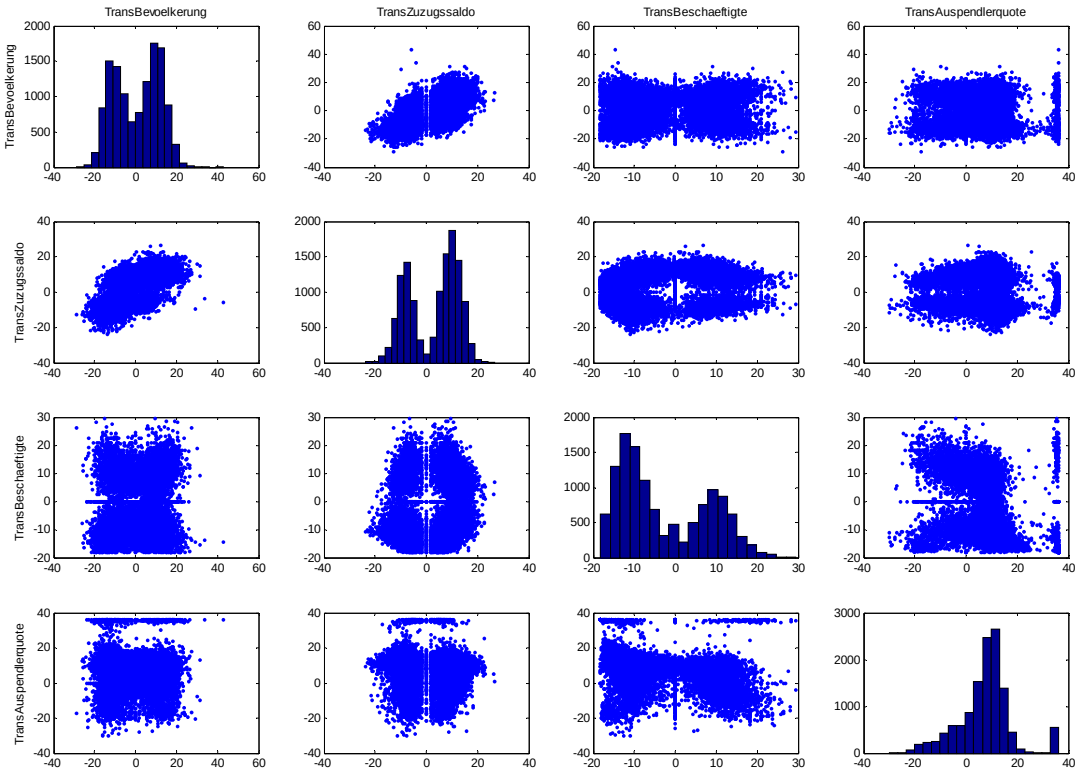


Abbildung 6-2: Scatter-Plot und Histogramme der vorverarbeiteten Daten

573 Vgl. DEMSAR [2006]

Die Abbildungen zeigen die Untersuchung der charakteristischen Gemeinden auf Basis einer Emergenten SOM, die mit 50x82 Neuronen trainiert wurde. Es handelt sich um Inseldarstellungen der U*-Map und der U*-Matrix. Die U*-Map entwickelt eine geographische Landschaft aus den vorverarbeiteten Daten (X und Y sind imaginäre Projektionsachsen). Die Darstellungen zeigen eine deutliche Struktur, wobei die Clustergrenzen durch Bergrücken (Z-Achse) ausgedrückt werden. Ein großer Höhenwert (U-Höhenwert) hebt die Unterschiedlichkeit zwischen Objekten hervor. Datenpunkte, die sich in kohärenten Regionen finden, lassen sich einem Cluster zuweisen. Mit Hilfe des U*C-Algorithmus wurden 11 Cluster ermittelt.

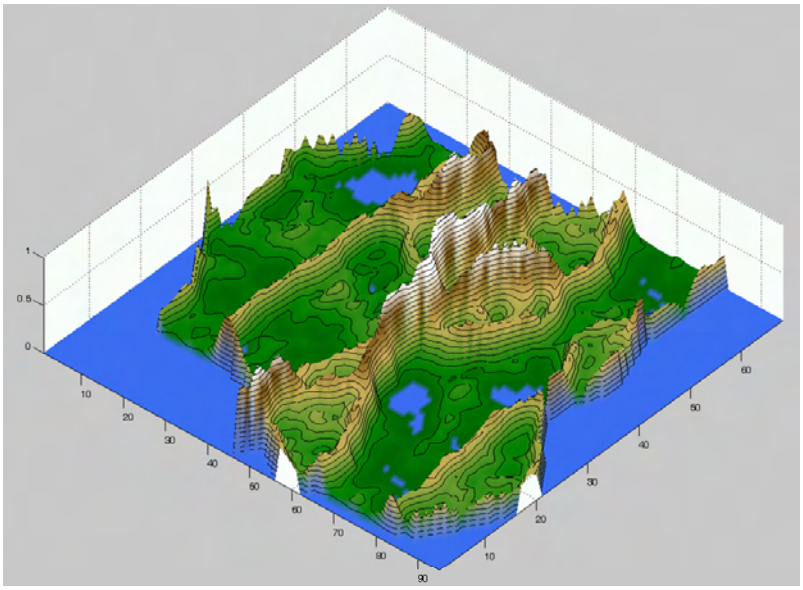


Abbildung 6-3: U*-Map (N=8113, D=4, 50x82 Neuronen, Inseldarstellung)[574]

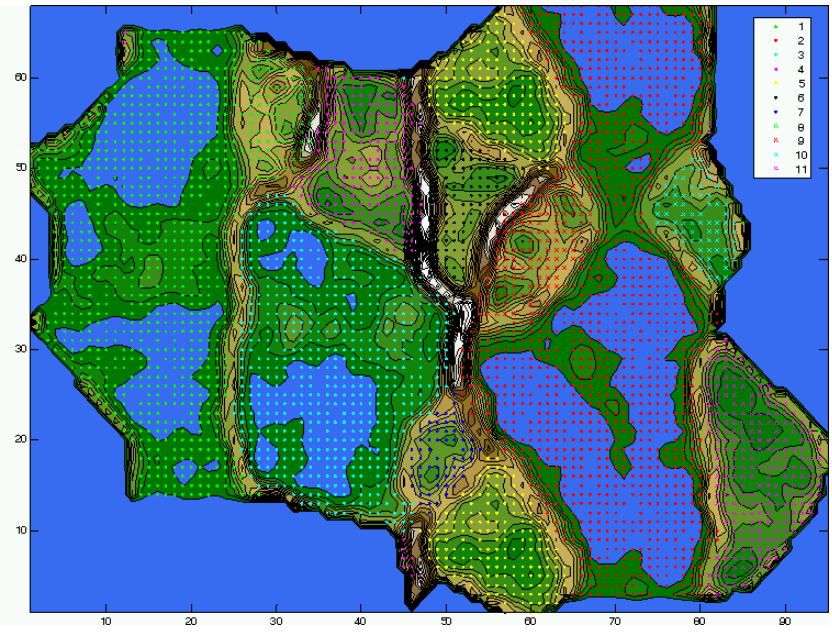


Abbildung 6-4: U*-Matrix mit der Clusterung des U*-C-Algorithmus (N=8113, D=4, 50x82 Neuronen)[574]

574 Vgl. BEHNISCH / ULTSCH [2007]

Es ist darauf hinzuweisen, dass sowohl der Ansatz der Entscheidungsregeln auf Basis dichotomer Variablenausprägungen als auch der Ansatz der Emergenten SOM mit dem U*-C-Algorithmus zu vergleichbaren Klassenstrukturen geführt haben. Die sechs Oberklassen lassen sich bei beiden Ansätzen durch Interpretation und Aggregation aufbauen. Abbildung 6-5 zeigt im Vergleich zu den vorherigen redundanzfreien Inseldarstellungen eine randlose gekachelte Projektion der Emergenten SOM (siehe Abschnitt 2.2.2). Abgebildet ist die U*-Matrix und weiterhin wird ein Vergleich zwischen der existierenden Struktur in den Daten und einer auf dichotomen Eigenschaften aufgebauten Klassifikation gezogen. Die sechs Oberklassen folgen deutlich den emergenten Strukturen.

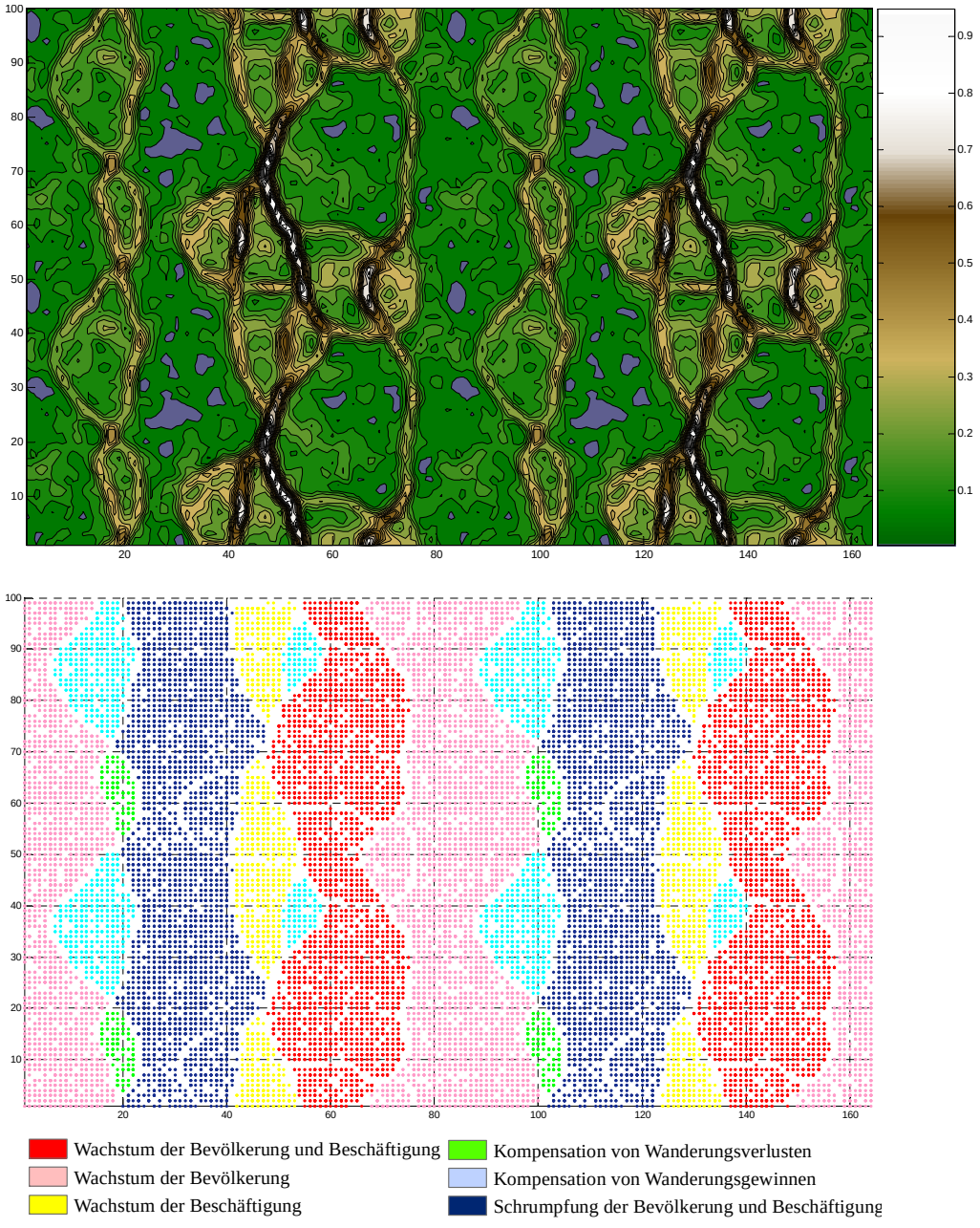


Abbildung 6-5: U*-Matrix (N=8113, D=4, 50x82 Neuronen) mit Vergleichsmöglichkeit zu Oberklassen

Abbildung 6-6 zeigt die Lokalisierung der Klassifizierungsresultate zu ausgewählten Gemeinden auf Basis charakteristischer Variablenausprägungen.

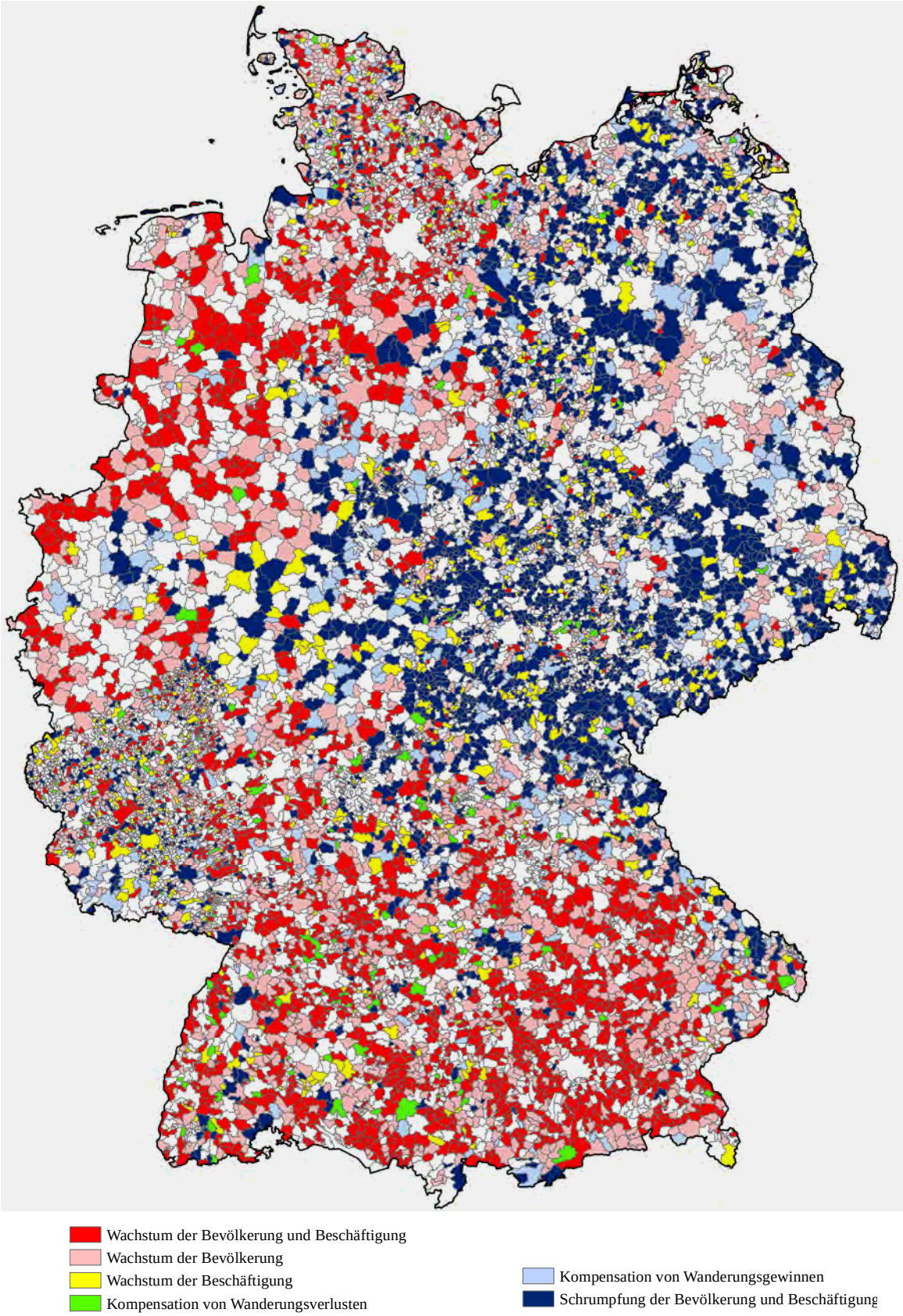


Abbildung 6-6: Verortung der Klassifizierung zu charakteristischen räumlichen Entwicklungstendenzen

Infolge eines relativ kurzen Untersuchungzeitraumes von 1999 bis 2003 werden keine durchschnittlichen prozentualen Veränderungen von einzelnen Gemeinden einer Klasse detailliert dargestellt, da es sich um eine Tendenzerfassung handelt. Es erfolgt dagegen eine Betrachtung von festgestellten Besonderheiten im Zusammenhang mit dem Klassifizierungsergebnis.[575]

Für den Untersuchungszeitraum zeigt sich ein deutliches flächenhaftes räumliches Bevölkerungswachstum in Westdeutschland und in weiten Teilen Ostdeutschlands ein anhaltender Rückgang der Bevölkerung. In den suburbanen Zonen um die Kernstädte ist in Ostdeutschland und insbesondere im Berliner Umland ein Bevölkerungszuwachs festzustellen. Vielfach ist darüber hinaus im Umland der ostdeutschen Oberzentren ein hoher Sterbefallüberschuss messbar, so dass eigentlich zusätzliche suburbane Wanderungsgewinne überdeckt werden (‚Kompensation von Wanderungsgewinnen'). Weitere Tendenzen zu fortschreitenden Schrumpfungsentwicklungen lassen sich in altindustrialisierten Regionen auch im westdeutschen Raum erkennen (Region Saar und Rhein-Ruhr) sowie Kompensationsvorgänge von Wanderungsgewinnen. Die natürliche Bevölkerungsentwicklung ist trotz festgestellter Wanderungsverluste in einigen Gemeinden sogar maßgebend für ihre Zunahme (‚Kompensation von Wanderungsverlusten').

Bezogen auf die in Deutschland geltende Zentrale-Orte-Kategorisierung ist festzustellen, dass im ostdeutschen Bundesgebiet nur vereinzelt Mittelzentren existieren, die über Bevölkerungs- und Beschäftigungsanstieg verfügen. Ähnliches lässt sich für die dortigen Oberzentren erkennen. Betrachtet man nur den Bevölkerungszuwachs in den Mittelzentren, so zeigt nur ein Zehntel eine positive Bevölkerungsentwicklung. Der deutlichste Bevölkerungsrückgang insgesamt entfällt bei den ostdeutschen Gemeinden sogar auf die Mittel- und Unterzentren. In Bezug auf das Klassifizierungsergebnis zeichnet sich im westdeutschen Bundesgebiet dagegen der überwiegende Teil der Oberzentren und Mittelzentren durch Bevölkerungs- und Beschäftigungswachstum aus, wobei zusätzlich im Umland weitere kombinierte Wachstumstendenzen aus Bevölkerung und Beschäftigung messbar werden.

Im Kontext des Raumstrukturtypenansatzes[576] ist festzustellen, dass sich ca. 67 % der Bevölkerung in Deutschland auf den inneren und äußeren Zentralraum sowie den Zwischenraum mit Verdichtungsansätzen verteilen. Dabei entfallen auf den westdeutschen Zentralraum ca. 45 % der Bevölkerung und 16 % auf den Zwischenraum mit Verdichtungsansätzen. Es deutet sich im Folgenden an, dass diese Konzentrationen sich weiter regional ausweiten.

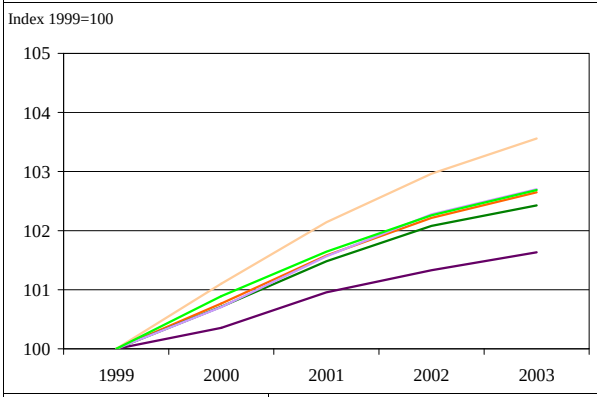
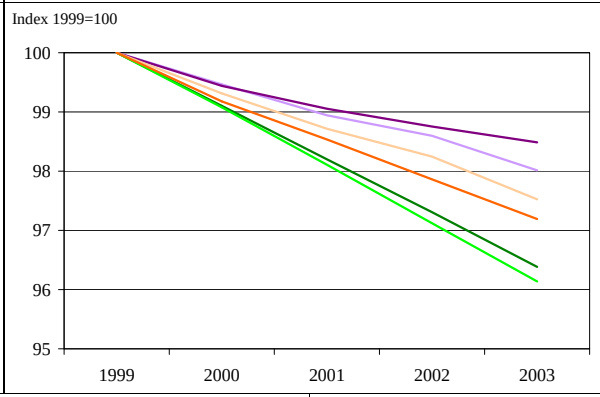
---

575 Die Verteilungen der Variablen zu jeder Klasse sind im Nebenteil B zusätzlich dargestellt (PDE).

576 Vgl. Raumstrukturtypen des Bundesamtes für Raumordnung (siehe Abschnitt 10.2.2). Die Klassifizierungsvariablen wurden mit Hilfe der PDE und den Raumstrukturtypen untersucht (siehe Nebenteil B).

In Tabelle 6-6 und Tabelle 6-7 wird die prozentuale Veränderung der Gesamtbevölkerung und Gesamtbeschäftigung nach Raumstrukturtypen gezeigt und damit eine räumliche Ausdehnung der Entwicklungsmuster vorstellbar. Es ist keine eindeutige Aussage über eine einzelne Gemeinde möglich, da es sich um die Entwicklung der Gesamtbevölkerung getrennt nach Wachstums- und Schrumpfungstendenzen handelt. In den peripheren Räumen sind Bevölkerungsverluste am deutlichsten. Die größten Zuwächse weist der Zwischenraum geringer Dichte auf. Der innere Zentralraum zeigt Bevölkerungsgewinne in den letzten Jahren.

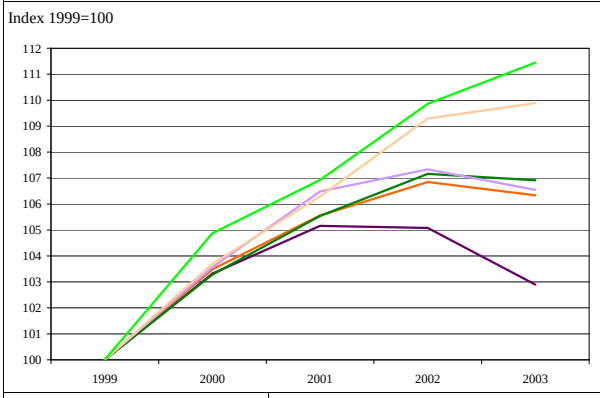
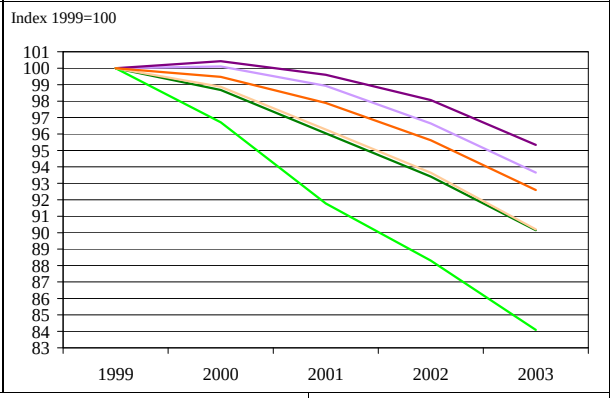
Prozentuale Bevölkerungsentwicklung insgesamt nach Raumstrukturtypen (1999 bis 2003)

Wachstumstendenz, Grundmenge: 6610 Gemeinden	Schrumpfungstendenz, Grundmenge: 5820 Gemeiden
	

enthaltene Gemeinden	Raumstrukturtyp des Bundesamtes für Bauwesen und Raumordnung	enthaltene Gemeinden
621 (Ost: 24, West: 597)	Äußerer Zentralraum	129 (Ost:14, West: 115)
178 (Ost: 10, West: 168)	Innerer Zentralraum	72 (Ost: 3, West: 69)
1425 (Ost: 137, West: 1197)	Zwischenraum mit Verdichtungsansätzen	592 (Ost: 199, West: 393)
1561 (Ost: 136, West: 1355)	Zwischenraum geringer Dichte	802 (Ost: 307, West: 495)
1491 (Ost: 234, West: 1191)	Peripherraum mit Verdichtungsansätzen	1736 (Ost: 981, West: 755)
1334 (Ost: 386, West: 1175)	Peripherraum sehr geringer Dichte	2489 (Ost: 1498, West: 991)

Tabelle 6-6: Bevölkerungsentwicklung in den Raumstrukturtypen

Beschäftigungsentwicklung in den Raumstrukturtypen (1999 bis 2003)

Wachstumstendenz, Grundmenge: 4938 Gemeinden	Schrumpfungstendenz, Grundmenge:7492 Gemeiden
	

enthaltene Gemeinden	Raumstrukturtyp des Bundesamtes für Bauwesen und Raumordnung	enthaltene Gemeinden
401 (Ost: 5, West: 396)	Äußerer Zentralraum	349 (Ost:33, West: 316)
122 (Ost: 3,West: 119)	Innerer Zentralraum	128 (Ost:10, West: 118)
930 (Ost: 74, West: 856)	Zwischenraum mit Verdichtungsansätzen	996 (Ost: 262, West: 734)
1057 (Ost: 103, West: 954)	Zwischenraum geringer Dichte	1236 (Ost: 340, West: 896)
1077 (Ost: 226, West: 851)	Peripherraum mit Verdichtungsansätzen	2084 (Ost: 989, West: 1095)
1351 (Ost: 400, West:951)	Peripherraum sehr geringer Dichte	2699 (Ost: 1484, West: 1215)

Tabelle 6-7: Beschäftigungsentwicklung in den Raumstrukturtypen

Die Verteilung der sozialversicherungspflichtig Beschäftigten auf die Raumstrukturtypen zeigt, dass ca. 56 % im Zentralraum einen Arbeitsplatz besitzen (Innerer Zentralraum: 42 %, Äußerer Zentralraum: 14 %). Im Zwischenraum mit Verdichtungsansätzen sind 17 % beschäftigt und im Peripherraum mit Verdichtungsansätzen weitere 17 %.

Mit Blick auf die Verortung der Gemeinden mit Beschäftigungswachstum werden in westdeutschen Regionen neben dem Zentralraum großräumliche Veränderungstendenzen sichtbar, und es zeigen sich Zuwächse im Zwischenraum geringer Dichte und in der Peripherie geringer Dichte (Verortung der Klasse ‚Wachstum der Beschäftigung'). Die Entwicklung in den Raumstrukturtypen insgesamt zeigt unabhängig von einer gemeindescharfen Betrachtung, dass die vom Zentralraum entfernten Gebiete (Zwischen- und Peripherraum) in der jüngeren Vergangenheit nicht nur Bevölkerungszuwächse, sondern auch Beschäftigungszuwächse aufweisen. In Gemeinden mit abnehmender Beschäftigung verzeichnet der Peripherraum zwischen 1999 und 2003 eine besonders deutliche negative Beschäftigungsentwicklung.

6.3 Diskussion der Ergebnisse

Im Rahmen der hier vorgestellten Untersuchung bestand das Ziel darin, die deutschen Gemeinden einem mehrdimensionalen Klassenbildungsprozess zu unterziehen, um auf diese Weise diejenigen Gemeinden zu einer Klasse zusammenzufassen, welche ähnliche Entwicklungstendenzen aufweisen.

Die Charakterisierung von räumlichen Entwicklungstendenzen wurde mit Hilfe von Verfahren des Visual Mining vorgenommen, wobei erstmalig zur Beschreibung des deutschen Gemeindesystems emergente selbstorganisierende Merkmalskarten eingesetzt wurden. Zur Strukturbildung wurden insgesamt zwei verschiedene Untersuchungsansätze verfolgt.

Zu Beginn wurde eine Klassenbildung mit Hilfe von Entscheidungsregeln umgesetzt. Dadurch wurde nochmals die Datenvorverarbeitung in ihrer Relevanz für eine zielführende Untersuchung betont, da ein dichotomes Verhalten in den Variablen aufgedeckt werden konnte, welches sich, wie im weiteren Verlauf nochmals diskutiert wird, für eine Klassenbildung anhand daraus definierter Entscheidungsregeln als sehr geeignet erwiesen hat. Mit Blick auf die vier Variablen und die definierten Entscheidungsregeln, die zu je zwei bzw. maximal drei Ausprägungen geführt haben, wäre zunächst eine weitaus größere Klassenzahl theoretisch denkbar gewesen, als sich später real durch Klassenbildung herauskristallisiert hat. Durch inhaltliche Interpretation und Aggregation dieser Klassen konnten sechs Oberklassen mit spezifischen Eigenschaften gebildet werden.

Weiterhin wurde die Datenstruktur mit Hilfe eines Projektionsverfahrens (ESOM) untersucht, um im hochdimensionalen Datenraum bestehende Strukturen sichtbar zu machen und mit einem dichtebasierten Clusteralgorithmus die darin enhaltenen Objekte zu identifizieren. Eine Interpretation der Strukturen, die mit dem U*C-Algorithmus aus der Emergenten SOM ermittelt wurde, hat gezeigt, dass die Eigenschaften der Cluster mit den zuvor durch Entscheidungsregeln definierten Klassen übereinstimmen. Es konnte dadurch die Struktur der Oberklassen merkmalsbasiert nachvollzogen werden, so dass sich die vorgenommene Aggregation mit beiden Ansätzen bestätigen lässt.

Es ist hervorzuheben, dass sowohl der Ansatz auf Basis von Entscheidungsregeln als auch der Ansatz mit Emergenten SOM zu den gleichen Ergebnissen geführt haben. Darüber hinaus wurde durch Überprüfung mit den emergenten SOM gezeigt, dass deutliche Strukturen in den Daten existieren und auf dieser Grundlage die Entwicklung des kanonischen Klassifikators möglich ist (siehe Abschnitt 7). Eine nachträgliche Zuordnung von Gemeinden zu den

Klassen ist damit nach wie vor operationalisierbar bzw. ergänzt eine zusätzliche Variablenauswahl den Untersuchungsansatz.

Neben der Verortung des Klassifizierungsergebnisses auf Gemeindeebene erfolgte durch Einbeziehung der bundesweit einheitlich definierten Raumstrukturtypen eine zusätzliche Ergebnisbewertung. Hierzu wurden die Gemeindedaten aufsummiert und für die sechs Raumstrukturtypen die prozentuale Bevölkerungs- und Beschäftigungsentwicklung insgesamt aufgeführt. Dadurch erhalten die Entwicklungsmuster der Gemeinden zusätzlich eine raumstrukturelle Ausdehnung, die ggf. eine strategische Planung von zukünftigen vertiefenden Untersuchungen einzelner Raumstrukturtypen unterstützen kann.

Es sei darauf verwiesen, dass aus Gründen der Datenverfügbarkeit zum jetzigen Zeitpunkt lediglich die jüngere Vergangenheit zwischen 1999 und 2003 einem Messansatz unterzogen werden konnte und so nur gewisse Entwicklungstendenzen ausdrückbar sind. Zukünftig ist bei Erhöhung der Betrachtungsschärfe durch weitere Variablen und insbesondere längerer Zeitreihen das komplexe Phänomen der Schrumpfung oder des Wachstums von Gemeinden genauer messbar. Erschwert wurde der Datenaufbau für eine mehrdimensionale Betrachtung besonders durch die hohe räumliche Auflösung (Gemeindeebene) und dem Anspruch einer flächendeckenden Erfassung der Daten der 12430 Gemeinden. Gerade mit Blick auf das Datenmaterial aus den neuen Bundesländern (gemeindescharfe Daten erstmals 1996) und aufgrund der aktualisierten Erfassungsmethodik bei der Beschäftigungsstatistik ist die zeitliche Betrachtungstiefe zum jetzigen Zeitpunkt eingeschränkt.

In Bezug auf die Wanderungsdynamik ist zu unterscheiden zwischen dem Wanderungssaldo und der Wanderungsintensität (Wanderungssaldo je 1000 Einwohner). Die Klassifikation basiert auf dem Wanderungssaldo und erfasst die Zu- und Fortzüge. Es ergibt sich dadurch mit Blick auf die Daten eine Abhängigkeit von der Gesamtbevölkerung in einer Gemeinde (siehe auch Abbildung 6-6), da in der Regel das Wanderungssaldo beispielsweise in Großstädten im positiven wie im negativen Wertebereich wesentlich größer ausfällt.

Darüber hinaus wurde beim Aufbau der Variablenbasis darauf Wert gelegt, dass sich alle Variablen auf den gleichen Zeitabschnitt beziehen und dies stellte eine zusätzliche Herausforderung dar, die aus Kenntnis anderer Arbeiten im urbanen Kontext nicht immer aufrecht zu erhalten ist. Zukünftig sind sicherlich weitere inhaltliche Aspekte mit in die Untersuchung einzubeziehen, um die Aussageperspektive weiter zu vergrößern.

In der hier dargestellten Untersuchung lag der Schwerpunkt im Wesentlichen auf der Darstellung und Anwendung einer methodisch orientierten Vorgehensweise. Mit Blick auf die große Objektanzahl (z.B. 12430 Gemeinden) und die damit verbundenen Variablen ist festzustellen, dass die emergenten SOM zur Untersuchung von derartig hochdimensionalen Daten besonders geeignet sind, da sich diese sowohl für eine erste Visualisierung bzw. Strukturerkennung eignen als auch eine spätere Strukturbildung auf Basis von dichtebasierten Clusteralgorithmen ermöglichen. Mit Hilfe von den in dieser Arbeit ebenfalls dargestellten hierarchischen Clusteralgorithmen wäre beispielsweise eine vollständige Untersuchung nur annähernd durch eine Stichprobe möglich gewesen, wobei dann nach wie vor der visuelle Überblick über die hochdimensionale Datenstruktur fehlen würde.

Von großer Relevanz ist der Aspekt, dass man die hier gezeigten Ergebnisse und gefundenen Klassen in einem Gesamtzusammenhang sehen muss, d.h. es ist zu berücksichtigenn, dass die auf vier Variablen noch relativ einfach gehaltene Messung von Entwicklungstendenzen im weiteren Verlauf dem Aufbau eines Klassifikators dient. Die Aussagekapazität kann ggf. infolge weiterer von der ersten Klassenbildung unabhängiger Untersuchungsvariablen erhöht werden. Es wird zusätzlich darauf verwiesen, dass Ansätze aus dem Bereich der Operationalisierung (siehe Abschnitt 2.5) sich zukünftig eignen werden, eine nachträgliche Identifizierung der Klassen auf Basis zusätzlicher Variablen zu verfolgen bzw. nach weiteren Variablen gesucht werden kann, die eine Klasse erlären. Die zu Beginn formulierte Untersuchungsaufgabe zu räumlichen Entwicklungstendenzen dient dementsprechend in dieser Arbeit zusätzlich einer Darstellung des Wirkprinzips eines Klassifikators.

6.4 Fazit

Welche Städte und Gemeinden in Deutschland zeigen eine Tendenz zum Wachsen bzw. zum Schrumpfen? Diese Frage wurde auf der Grundlage von vier gewählten Variablen für eine relativ kurze Zeitreihe von 1999 bis 2003 mit einem Klassifizierungsansatz für die 12430 Gemeinden beantwortet. Dabei wurde eine methodische Vorgehensweise zur Untersuchung hochdimensionaler Daten (ESOM) und darin enthaltener Strukturen vorgestellt.

Die Untersuchungsergebnisse belegen nachdrücklich, dass die Entwicklung der Gemeinden in Deutschland durch regionale Wachstums- und Schrumpfungseigenschaften geprägt ist. Dabei zeigt sich eine Vielfalt von unterschiedlichen Ausprägungen, so dass sehr differenzierte Gemeindeklassen bereits erkennbar werden. Als charakteristisches räumliches Bild ist ein anhaltender Bevölkerungsrückgang in weitgehend agglomerationsfernen Räumen Ostdeutschlands und ein nahezu flächenhaftes Bevölkerungswachstum in Westdeutschland feststellbar. Dort findet sich zusätzlich Beschäftigungs- und Bevölkerungswachstum, wobei regional sich ausdehnende Konzentrationsprozesse erkennbar werden.

Mit Blick auf zukünftige Veränderungen infolge des demographischen Wandels und darüber hinaus sich weiter vollziehender ökonomischer Umwandlungsprozesse wird in Deutschland mit einer Zunahme von Wachstums- und Schrumpfungsvorgängen gerechnet.[577] Vermutlich besteht der Handlungsbedarf für die Entscheider und Planer noch stärker darin auf die Bipolarität von Schrumpfung und Wachstum zu reagieren und ggf. in unterschiedlichen Klassen bzw. Typen von Gemeinden zu denken, so dass regionale Entwicklungskonzepte für verschieden aufgebaute Gemeindeklassen entwickelt werden können. Gerade durch Klassenbildung sind regionale Zusammhänge stärker erfahrbar, und unterstützt wird ein Prozess, dass Gemeinden voneinander lernen und gemeinsame regionsprägende Identitäten entwickeln bzw. stärken können. In Schrumpfungsregionen sei zusätzlich auf Möglichkeiten der gemeinsamen Infrastrukturversorgung durch Kooperationen verwiesen.

Es bedarf in der Zukunft wirksamer Instrumente, die zielgenau Entwicklungstendenzen aufdecken und identifizieren können, um spezifische Handlungsstrategien für die von positiven oder negativen Entwicklungen betroffenenen Gemeinden bzw. ganzen Regionen zu entwickeln. Im Rahmen regelmäßiger Erfolgskontrollen ist zu prüfen, inwieweit die in Strategien benannten Maßnahmen umgesetzt wurden und ob angestrebte Ziele erreicht werden. Schließlich müssen die Strategien selbst durchgängig weiterentwickelt werden.

---

[577] Vgl. „Aktion Demographischer Wandel“, siehe: http://www.aktion2050.de (30.01.2007)
Vgl. WINTERMANN [2005, S. 343-363]

7 Aufbau eines kanonischen Klassifikators zur nachträglichen Klassenidentifikation

7.1 Untersuchungsaufgabe: Wissensbasierte Systeme in Stadt- und Regionalplanung

Für die Stadtplanung sind zukünftig wissensbasierte Systeme von großer Bedeutung, da sich mit ihrer Hilfe auch komplexe, schlecht strukturierte bzw. strukturierbare Probleme mit der Computertechnik bearbeiten lassen. Dabei ist das für die Stadtplanung und den Städtebau zweckmäßige Methodenrepertoire auf der Grundlage von Wissensbasierten Systemen gegenwärtig noch im Aufbau begriffen.[578] Die wissensbasierten Systeme als Teilmenge des Gebietes der künstlichen Intelligenz (KI) umfassen im Allgemeinen fünf Komponenten (Wissensbasis, Problemlösungskomponente, Dialogkomponente, Wissensakquisitionskomponente und Erklärungskomponente). Für die Stadtplanung sind die folgenden Anwendungsmöglichkeiten von wissensbasierten Systemen denkbar:

- Die Regelbasierte Wissensrepräsentation, welche mittels Wenn-Dann-Konstruktionen einen Entscheidungsprozess begleitet (z.B. Assistenz- oder Expertensysteme).
- Die fallbasierte Wissensrepräsentation, deren Wissenskomponente durch Sammlung von Fällen mit ihren typischen Merkmalen beschrieben und erweitert wird. Bei Bedarf ist es später möglich, auf dieser Grundlage relevante Fälle zu recherchieren und auf den aktuellen Untersuchungsfall zu reagieren (z.B. städtebauliche Entwurfswerkstatt).
- Die objektorientierte Form der Wissensrepräsentation, welche die Einzelgegenstände (‚entities') der realen Welt und ihre spezifische Semantik in eine computerinterne Form bringt und über Beziehungen (Relationen) eine Kommunikation zwischen Objekten realisiert (z.B. Stadtmodellierung).
- Die semantischen Netze, die das Wissen auf der Grundlage von Netzwerkstrukturen abbilden und eine komplexe Verzahnung von Begriffen, Objekten, Ereignissen oder Konzepten erlauben (z.B. Semantische Netzwerkstruktur von Stadtgefügen).

Die Verfahren des Data Mining und der Knowledge Discovery unterstüzten eine empirische Erfassung und fördern den Aufbau städtebaulicher Orientierungswerte. Wie gezeigt, wird durch eine Klassifikation eine gegebene Datenmenge strukturiert. In diesem Kapitel soll in kanonischer Weise das Arbeitsprinzip eines Klassifikators erklärt werden. Die Klassenzuordnung wird dadurch operationalisierbar, d.h. es ist eine automatisierte Zuordnung möglich bzw. lassen sich auch je nach Klassifikator regelbasierte Zuordnungen vornehmen. Ein aufgebauter Klassifikator eignet sich zusätzlich zu einer Identifizierung von Klassen, falls Datenlücken bei den nicht klassifizierten Objekten in den Ausgangsvariablen bestehen.

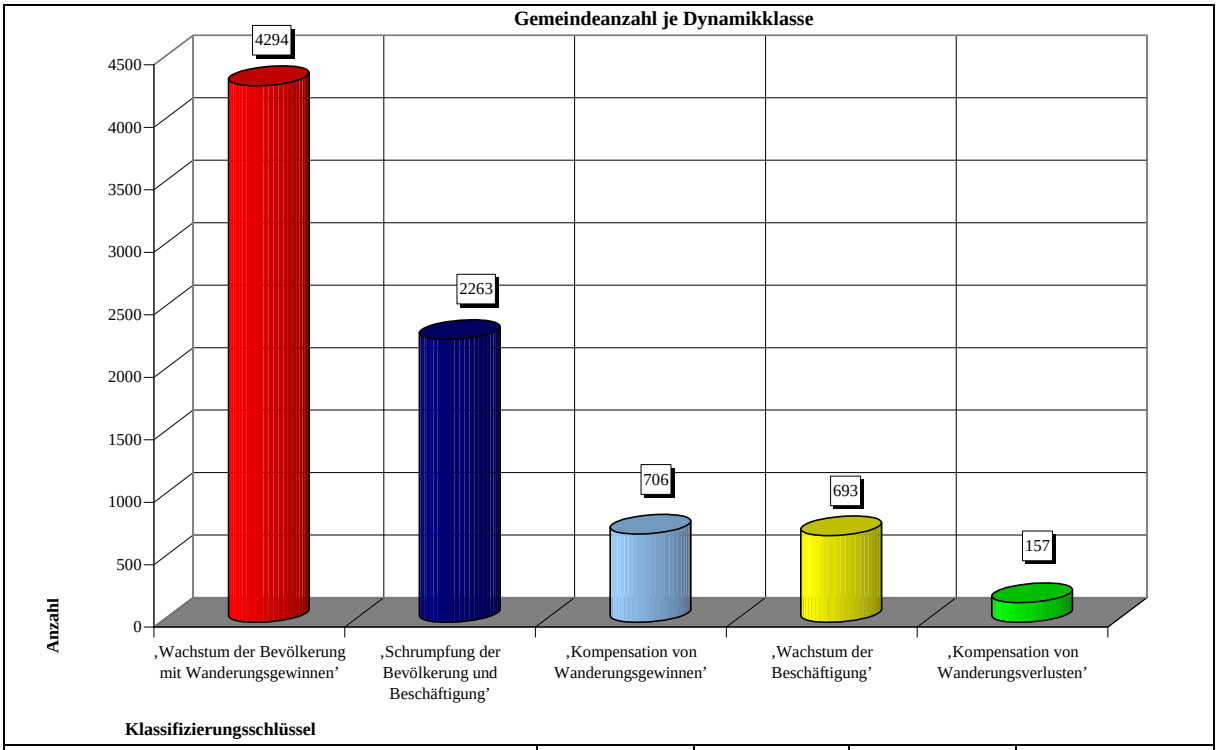
[578] Vgl. STREICH [2006, S. 194]

7.2 Herleitung und Darstellung der Ergebnisse

Gegeben sei eine Auswahl von Gemeinden (8113 Gemeinden in Deutschland). Diese sind einer definierten Klassenstruktur zugeordnet, wobei sich in diesem Fall die Klassen aus der Untersuchung zu räumlichen Entwicklungstendenzen (siehe Abschnitt 6) ableiten lassen. Es handelt sich um eine unterscheidbare Struktur mit fünf Klassen, die im Folgenden dazu dient, einen Klassifikator kanonisch aufzubauen (siehe Theorie aus Abschnitt 2.5).

Gesucht wird hierzu ein Algorithmus, der Datensätze in die Klassen einordnen kann. Gemessen wird die Genauigkeit, mit der dieser Klassifikator die Gemeinden einer zuvor gefundenen Klassenbildung zuordnen kann. Es stellt sich die Frage, inwieweit ausgewählte unabhängige Variablen sich überhaupt zur Identifikation einer bestehenden Klassenstruktur eignen und eine Zuordnung von Gemeinden unterstützen.

Tabelle 7-1 enthält die relevante Klassenstruktur, die im weiteren Verlauf als ‚Dynamikklassen' bezeichnet werden. Dargestellt ist die Anzahl der Gemeinden, die diesen Klassen zugeordnet sind und die Eigenschaft einer jeden Klasse.



Bezeichnung	Bevölkerung	Zuzugssaldo	Beschäftigung	Auspendlerquote
‚Wachstum der Bevölkerung mit Wanderungsgewinnen'	Zunahme	Zunahme	nicht berücksichtigt	
‚Schrumpfung der Bevölkerung und Beschäftigung'	Abnahme	Abnahme	Abnahme	nicht berücksichtigt
‚Kompensation von Wanderungsgewinnen'	Abnahme	Zunahme	nicht berücksichtigt	Zunahme
‚Wachstum der Beschäftigung'	Abnahme	Abnahme	Zunahme	nicht berücksichtigt
‚Kompensation von Wanderungsverlusten'	Zunahme	Abnahme	Abnahme	Zunahme

Tabelle 7-1: Bedeutung und Größe der Dynamikklassen

Abbildung 7-1 zeigt die Verortung der fünf Dynamikklassen. Es handelt sich um die bereits beschriebenen Gemeinden (siehe Abschnitt 6) mit spezifischen Entwicklungstendenzen.

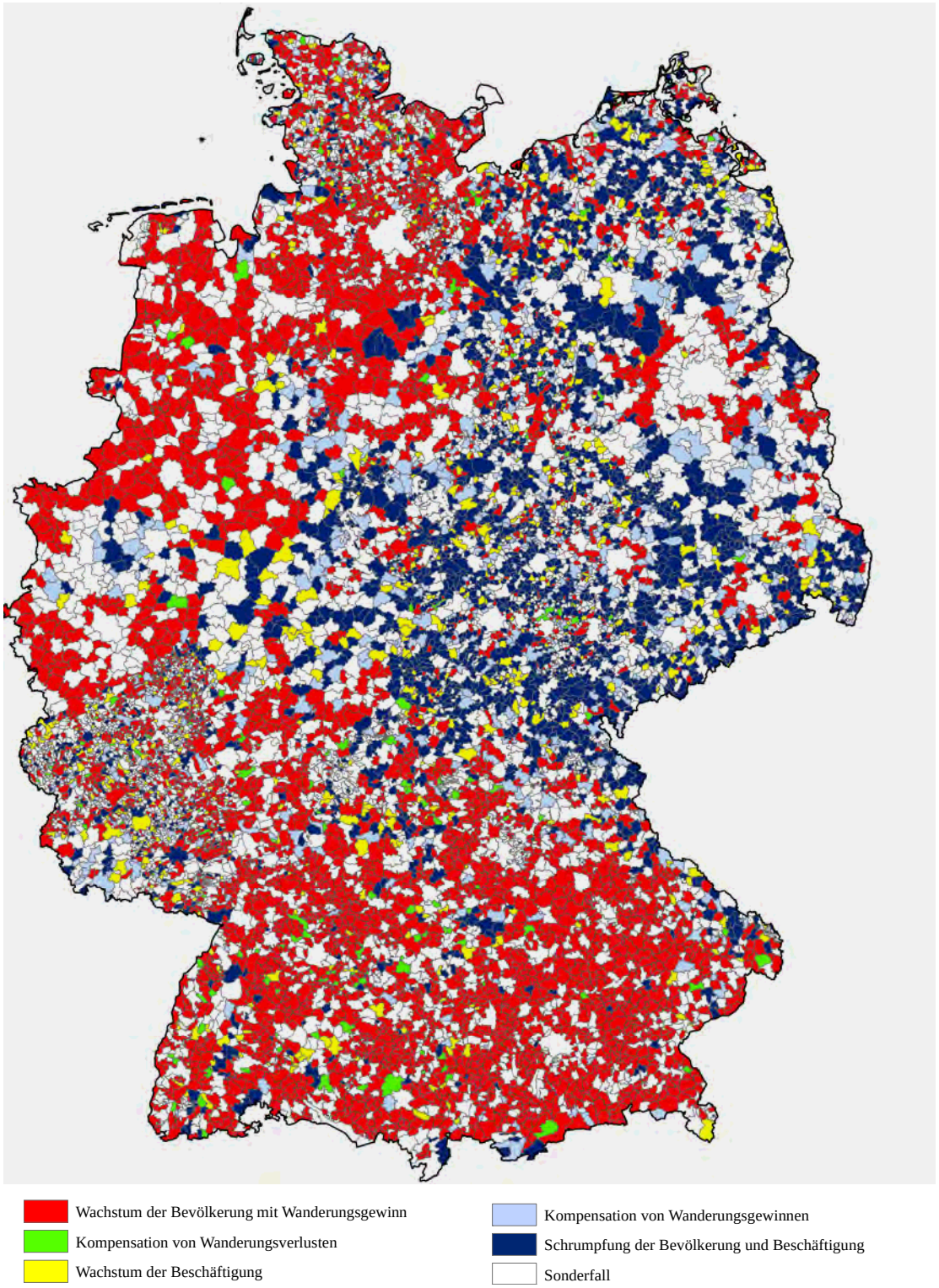


Abbildung 7-1: Gemeinden der Dynamikklassen

Um die Klassen nachträglich identifizieren zu können, wurden neun Variablen ausgewählt, die spezifische Eigenschaften tragen. Zusätzlich ist die Variablennotation aufgeführt.

Die Daten zur ‚Erreichbarkeit' eignen sich, um die Raumstruktur hinsichtlich der Lagegunst zu zentralen Orten als Träger wichtiger Raumfunktionen zu bewerten. Die Qualität der Verkehrsinfrastruktur stellt einen wichtigen Einflussfaktor für die Zugänglichkeit von Zentren und damit für die Vermittlung räumlicher Standortattraktivität dar.

Der Anteilswert der Landwirtschaftsfläche an der Gemeindefläche liefert Anhaltspunkte über die regionale Wirtschaftsstruktur, insbesondere die Bedeutung des primären Sektors. Die Vermutung besteht, dass je größer der Anteilswert in einer Gemeinde wird, umso eher mit Monostrukturen und dadurch bedingten Problemen zu rechnen ist (‚Landwirtschaftsfläche').

Das Bevölkerungspotential erfaßt die Möglichkeit räumlicher Interaktionen (‚Räumliche Interaktion'). Je mehr Bevölkerung in der Ortsumgebung erreichbar ist und je geringer der dabei zur Raumüberwindung benötigte Aufwand ist, desto höher ist dessen Kontaktpotenzial.

Als modifizierte Siedlungsdichte (‚Flächenkonzentration') wird die Auslastung der Gebäude- und Freifläche je km² durch Bevölkerung und Arbeitsplätze verstanden.[579] Mit zunehmender Dichte steigt oftmals die Versorgung mit Infrastruktureinrichtungen und anderen Gütern.

Der Anteilswert der Beschäftigten im Tertiären Sektor an den Beschäftigten insgesamt gibt Aufschluss über den Entwicklungsstand der Wirtschaft, da mit fortschreitender Entwicklung der Wirtschaft nach FOURASTIÉ dieser Sektor an Bedeutung gewinnt (‚Tertiärer Sektor'). Als dynamische Messgröße dient die Veränderung des Anteils der Beschäftigten im Primären und Tertiären Sektor in Bezug auf die Gesamtbeschäftigten (‚DynamikSektorenstruktur').

Das ‚Pendlersaldo' gibt an, ob mehr Arbeitskräfte regelmäßig von ihrem Wohnort zum Arbeiten kommen oder mehr in der Gemeinde Wohnhafte diese regelmäßig verlassen, da ihr Arbeitsplatz außerhalb der Gemeinde liegt. Ein Pendlerüberschuss (Einpendler- oder Auspendlerüberschuss) steht vielfach in engem Zusammenhang mit der Arbeitsplatzdichte.

Die Steuerkraft (‚Steuerkraft) ist als Messgröße für die wirtschaftliche bzw. finanzielle Lage einer Gemeinde zu verstehen. Das Istaufkommen der Grundsteuer A und B sowie der Gewerbesteuer ist durch den Einfluss der Hebesätze als Vergleichsmaßstab ungeeignet und wird durch Rückrechnung auf die Steuerbemessungsbasis neutralisiert.

Da der Wohnungsbau in räumlicher Hinsicht vielfach der Bevölkerungsentwicklung folgt und oftmals gerade Regionen mit hoher Beölkerungsdynamik auch eine intensive Wohnungs-neubautätigkeit verzeichnen, wird die Messgröße zur Dynamik des Wohnungsbestandes (‚DynamikWohnungsbestand) der letzten 10 Jahre als weitere unabhängige Variable genutzt.

[579] Vgl. STAACK [1995, S. 128], SIEDENTOP et al. [2005, S. 77]

Tabelle 7-2 enthält die unabhängigen Variablen, die der ersten Vermutung zufolge in einem Zusammenhang zu räumlichen Entwicklungstendenzen stehen könnten.

Messgröße	Messvorschrift	Einheit	Notation, Wertebereich[580]
(1) Erreichbarkeit	Messdaten aus dem BBR-Erreichbarkeitsmodell, 2003	[Minuten]	‚Erreichbarkeit', [0,∞]
(2) Landwirtschaftsfläche	Anteilswert der Landwirtschaftsfläche an der Gemeindefläche, 2004	[%]	‚Landwirtschaftsflaeche', [0,100]
(3) Räumliche Interaktion	Messdaten des Bundesamtes für Bauwesen und Raumordnung: Bevölkerungspotential. Es handelt sich um eine Messgröße zur Charakterisierung der Umgebung in Bezug auf die Umlandbevölkerung in einem Umkreis von 50 km, 2004	[Personen]	‚RaeumlicheInteraktion', [1,∞]
(4) Flächenkonzentration	Summe der Einwohner und Beschäftigten dividiert durch die Gebäude- und Freifläche. Es handelt sich um einen Dichtewert, 2004	[Personen pro km²]	‚FlaechenKonzentration', [1,∞]
(5) Dynamik Sektorenstruktur	Veränderung des Anteilswertes der Beschäftigten im Primären und Tertiären Sektor in Bezug auf die Gesamtbeschäftigten, Zeitintervall 1999 bis 2003	[%]	‚DynamikSektorenstruktur', [-100,100]
(6) Tertiärer Sektor	Anteilswert der Beschäftigten im Tertiären Sektor an den Beschäftigten insgesamt, 2003	[%]	‚TertiaererSektor', [-100,100]
(7) Pendlersaldo	Saldo der Ein- und Auspendler, 2004	[Personen]	‚Pendlersaldo', [-∞,∞]
(8) Steuerkraft	Berechnungsformel basiert auf der Statistik Öffentlicher Finanzen (Grundsteuer A und B, Gewerbesteuer, Einkommensteuer, Umsatzsteuer), 2003	[Euro]	‚Steuerkraft', [-∞,∞]
(9) Dynamik Wohnungsbestand	Veränderung des Wohnungsbestandes in Prozent zwischen 1994 und 2004	[%]	‚DynamikWohnungsbestand', [-100,∞]

Tabelle 7-2: Übersicht zu den unabhängigen Variablen für den Klassifikatoraufbau

Tabelle 7-3 fasst die Ergebnisse der Verteilungsuntersuchung zusammen und enthält die mit Hilfe des Gaussansatzes ermittelten Entscheidungsgrenzen und Klassenbeschreibungen.

Messgröße	Verteilungshypothese	Verteilung (inspiziert)	Klasse K_n: GMM (Modellierung)
(1) Erreichbarkeit	**Linkssteile Verteilung, Daten folgen einer Lognormalverteilung** Es gibt einige Gemeinden mit einer relativ kurzen Fahrzeit. Die überwiegende Zahl der Gemeinden weist aber eine hohe Fahrzeit auf. Die Oberzentren haben infolge der Modelldefinition zur Erreichbarkeit eine Fahrzeit von Null. Es handelt sich um ca. 160 Gemeinden.	**Log(Data+1)** folgt Normalverteilung	**Grenze:** 30 Minuten, Log (3.4) K1: Oberzentrum: **Data=0** K2: oberzentrennah: **0<Data≤30** K3: oberzentrenfern: **Data > 30**
(2) Landwirtschafts flaeche	**Gleichverteilung oder Normalverteilung** Es ist keine genaue Vorstellung zum Verteilungsverlauf vorhanden, bzw. werden weder eindeutig linkssteile oder rechssteile Verteilungen erwartet.	**Gleichverteilung**	**Grenze:** 40 % und 80 % K1: Kernstadt: **Data≤40 %** K2: Verstädtertes Umland: **40 %<Data≤80 %** K3: Ländlich: **Data>80 %**
(3) Räumliche Interaktion	**Linkssteile Verteilung, Daten folgen einer Lognormalverteilung** Die Großstädte erhöhen das Bevoelkerungspotential der Umlandgemeinden in einem Umkreis von 50 km sehr deutlich. Darüber hinaus wird aber eine große Menge an Gemeinden eher kleinere Ausprägungen für diese Variable aufweisen. Dadurch ist eine linkssteile Verteilung zu erwarten.	**Log(Data)** folgt Normalverteilung	**Grenze:** 100000 und 300000 Einwohner Log (11.4), Log(12.6) K1: Peripherie **Data≤100000** K2: Zwischenraum **100000<Data≤300000** K3: Zentralraum **Data>300000**

[580] Die Notation ist für die Verteilungsuntersuchung und -modellierung relevant (Nebenteil A).

(4) **Flächen-konzentration**	**<u>Linkssteile Verteilung, Daten folgen einer Lognormalverteilung</u>** Es handelt sich um viele Gemeinden mit einer geringen Konzentration, d.h. geringe Auslastung der Gebäude- und Freifläche mit Personen. Zusätzlich existieren wenige Gemeinden mit einer sehr hohen Flächenauslastung.	**Log(Data)** folgt Normalverteilung	**Grenze:** 2500 und 4000 Personen/km² Log (7.8), Log(8.3) K1: <u>geringe Konzentration</u> **Data≤2500** K2: <u>mittlere Konzentration</u> **2500<Data≤4000** K3: <u>hohe Konzentration</u> **Data>4000**
(5) **Dynamik Sektoren-struktur**	**<u>Bimodale Verteilung von Lognormalverteilungen, schiefe Verteilung im positiven wie im negativen Wertebereich (Data>0: linkssteil, Data<0 rechtssteil)</u>** Die Variable charakterisiert gleichzeitig zwei Sektoren, da vielfach Arbeitsplätze des Primären Sektors durch Arbeitsplätze im Tertiären Sektor ersetzt werden. Es wird mit einer gewissen Menge an Gemeinden zu rechnen sein, die deutliche Zuwächse in den letzten Jahren in Bezug auf die Variable aufweisen. Dies sind vermutlich Gemeinden, die verstärkt Beschäftigte im Tertiären Bereich hinzugewinnen konnten. Der weitaus größte Teil der Gemeinden zeigt aber eher geringere Zuwächse bzw. aufgrund wirtschaftlicher Veränderungen geringfügige Abnahmen der Beschäftigten. Dadurch kann ggf. im positiven Wertebereich eine linkssteile Verteilung der Daten vorliegen und im negativen Bereich infolge von einigen Ausreißern im rücklaufigen Sinne eine rechtssteile Verteilung.	**SLog(Data)** Bimodale Verteilung, Lognormalverteilungen	**Grenze:** 0 Dichotomie K1: <u>Arbeitsplatzabbau</u> **Data ≤ 0** K2: <u>Arbeitsplatzaufbau</u> **Data >0**
(6) **Tertiärer Sektor**	**<u>Linkssteile Verteilung, Daten folgen einer Lognormalverteilung</u>** Viele Gemeinden zeigen einen geringen Anteilswert bei den Beschäftigten im tertiären Sektor. Einige Beschäftigungszentren weisen dagegen sehr große Anteilswerte auf.	**Gleichverteilung**	**Grenze:** 25 % K1: <u>gering tertiärisiert</u> **Data ≤ 25 %** K2: <u>hoch tertiärisiert</u> **Data > 25 %**
(7) **Pendlersaldo**	**<u>Bimodale Verteilung von Lognormalverteilungen, schiefe Verteilung im positiven wie im negativen Wertebereich (Data>0: linkssteil, Data<0 rechtssteil)</u>** Mehrzahl der Gemeinden hat ein nahezu ausgeglichenes Pendlersaldo, d.h. einen gering positiven bzw. gering negativen Wert. Die Beschäftigungszentren haben eine sehr große positive Einpendlerzahl. Des Weiteren ist mit Gemeinden zu rechnen, die deutliche negative Salden besitzen.	**SLog(Data)** Bimodale Verteilung von Lognormal-verteilungen	**Grenze:** 0 Dichotomie K1: <u>Auspendlergemeinde</u> **Data ≤ 0** K2: <u>Einpendlergemeinde</u> **Data >0**
(8) **Steuerkraft**	**<u>Linkssteile Verteilung, Daten folgen einer Lognormalverteilung</u>** Die Mehrzahl der Gemeinden verfügt wahrscheinlich über eine eher geringere Steuerkraft. Die Beschäftigungszentren zeigen dagegen aufgrund der Bedeutung der Gewerbesteuer, die in die Berechnungsformel eingeht, sehr große Werte	**SLog(Data)** Bimodale Verteilung von Lognormal-verteilungen	**Grenze:** 200 und 900 Euro/Einwohner Slog (5.2), Slog(6.8) K1: <u>niedrige Steuerkraft</u> **Data≤180** K2: <u>mittlere Steuerkraft</u> **180<Data≤900** K3: <u>hohe Steuerkraft</u> **Data>900**
(9) **Dynamik Wohnungs-bestand**	**<u>Bimodale Verteilung von Lognormalverteilungen, (Data>0: linkssteil, Data<0 rechtssteil)</u>** Der Wohnungsbestand nimmt bei der Mehrzahl der Gemeinden zu. Dennoch gibt es auch Gemeinden, deren Bestand zurückgeht. Dies ist aber eine eher kleinere Menge. Bei den Gemeinden mit zunehmendem Wohnungsbestand ist zwischen Gemeinden mit sehr großer Zunahme und Gemeinden mit einer eher geringen Zunahme zu unterscheiden. Es bildet sich im positiven Wertebereich wahrscheinlich eine linkssteile Verteilung aus. Im negativen Bereich könnte eine rechtssteile Verteilung existieren, da auch mit Gemeinden mit großem Rückgang des Bestandes und einer größeren Anzahl von Gemeinden mit eher kleinerem Rückgang zu rechnen ist.	**SLog(Data)** Bimodale Verteilung von Lognormal-verteilungen	**Grenze:** 0 Dichotomie K1: <u>B</u>estandszuwachs **Data ≤ 0** K2: <u>Bestandsrückgang</u> **Data >0**

Tabelle 7-3: Zusammenfassung der Datenaufbereitung von unabhängigen Variablen

Auf Grundlage der zuvor ausgewählten 9 Variablen wird nach einer Zuordnungsvorschrift gesucht, welche sich eignet, eine Gemeinde in die bestehenden fünf Dynamikklassen einzuordnen. Abbildung 7-2 zeigt in allgemeingültiger Form das Wirkprinzip des Klassifikators unter Berücksichtigung von Trainings- und Testdaten. Das Ziel der Klassifikatorkonstruktion ist, die Klassifikation nachvollziehbar zu gestalten, d.h. eine Zuordnungsvorschrift zu finden, welche die gewonnene Klassifikation nachvollziehen kann. Für diesen Zweck wird in dieser Arbeit ein k-Nearest-Neighbour-Klassifikator realisiert (siehe Methode im Abschnitt 2.5.1.1), der subsymbolisch ohne ein explizites Verständnis der zu leistenden Klassifikation arbeitet. Als Abstandsmaß dient die quadrierte Euklidische Distanz. Falls eine ausreichend große Zuordnungsgenauigkeit auf Basis der 9 unabhängigen Variablen erzielt wird, besteht zukünftig die Möglichkeit, weitere bislang nicht klassifizierte Gemeinden nachträglich in die Dynamikklassen einzuordnen.

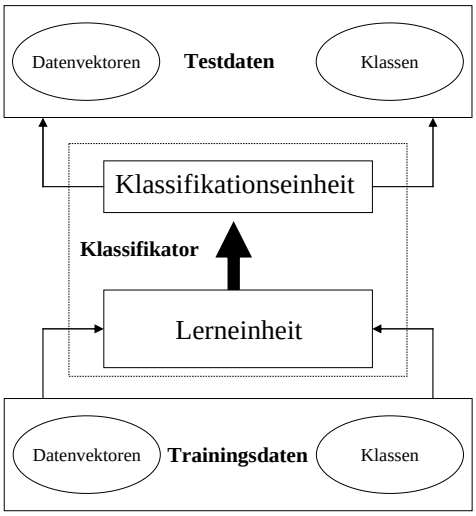


Abbildung 7-2: Klassifikatoraufbau mit Trainings- und Testdatensatz[581]

Zunächst wird der k-Nearest-Neighbour-Klassifikator an der Ausgangsklassifikation getestet, um überhaupt beurteilen zu können, inwieweit auch die bestehende Klassifikation wiedererkannt wird. Hierzu wird die Anzahl der nächsten Nachbarn $k = 1$ gesetzt und die bereits verwendeten vier Klassifikationsvariablen (Bevoelkerung, Zuzugssaldo, Beschaeftigte und Auspendlerquote) der 8113 Gemeinden werden benutzt, um die Zuordnungsgenauigkeit (OverallAccuracy) zu den Dynamikklassen zu bestimmen. Man erhält ein Beurteilungsmaß für die Genauigkeit, mit der der Klassifikator die bereits aufgebaute Klassifikation anhand der dazugehörigen Gemeindeobjekte wiedererkennen kann. Durch diesen Arbeitsschritt konnte eine Zuordnungsgenauigkeit von 100 % ermittelt werden.

581 Eigene Bearbeitung (Entwurf: BEICHEL [2002])

Im zweiten Arbeitsschritt werden die neun zusätzlich ausgewählten Variablen eingesetzt, um mit diesen neuen Informationen über die Gemeinden die Dynamikklassen zu verfolgen. Es wird die Anzahl der nächsten Nachbarn ebenfalls $k = 1$ gesetzt und eine Zuordnungsgenauigkeit von 98,6 % gemessen. Dieser Wert liegt vermutlich aufgrund einiger nicht eindeutig definierter Objekte im Merkmalsraum etwas niedriger.

Im Folgenden beinhaltet ein Referenzdatensatz die bestehende Klassifikation und dient als Grundlage für den Vergleich mit sogenannten unbekannten Daten. Eine Klassenzuordnung erfolgt unter Berücksichtigung der k-nächsten Nachbarn. Für ein zu klein gewähltes k besteht im Allgemeinen die Gefahr, dass Rauschen in den Trainingsdaten die Klassifikationsergebnisse verschlechtert. Wählt man k zu groß, so besteht die Gefahr, Punkte mit großem Abstand zu einem Klassifikationsobjekt bei der Klassifikationsentscheidung zu berücksichtigen. In dieser Arbeit wurde $k = 7$ als geeignete Nachbarschaftszahl ermittelt.

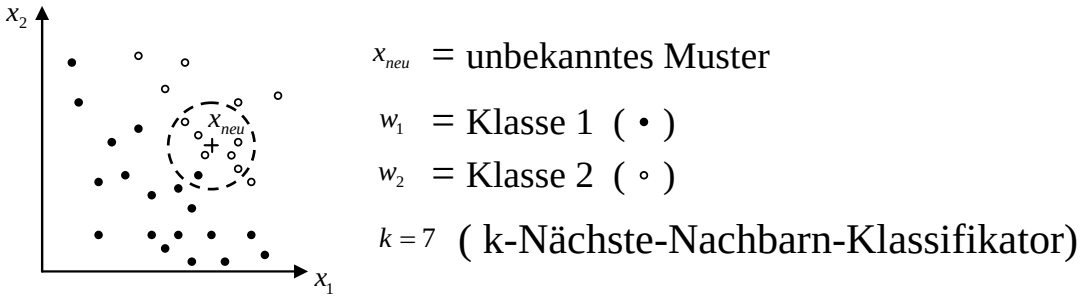


Abbildung 7-3: Prinzipschema für eine gewählte 2-Klassen-Problematik mit zweidimensionalen Mustern

Der k-Nearest-Neighbour-Klassifikator verwendet das Abstandsmaß zwischen einer neu zu klassifizierenden Gemeinde und den klassifizierten Gemeinden der Dynamikklassen. Zu der unklassifizierten Gemeinde werden diejenigen Datensätze gesucht, die dieser am nächsten liegen und eine Klassenzuweisung ermöglichen (siehe Abbildung 7-3). Die Klassifikationsleistung wird im Vergleich zu der gegebenen Klassifikation untersucht, d.h. es werden somit alle 8113 Gemeinden mit einer bekannten Klassenidentität erneut mit Hilfe des k-NN-Klassifikators klassifiziert, und es wird geprüft, inwieweit ein übereinstimmendes Ergebnis vorliegt. Insgesamt ist festzustellen, dass eine Zuordnungsgenauigkeit von **67 %** erreicht wird. Der Wert beschreibt, mit welcher Genauigkeit der Klassifikator anhand der neun unabhängigen Variablen und einer Nachbarschaftszahl von sieben eine Zuordnung auf die Dynamikklassen vornehmen kann. Es stellt sich heraus, dass mehr als zwei Drittel aller Zuordnungsfälle richtig auf die bereits gefundenen Klassen zugeordnet werden. Im Vergleich zu einer zufallsbedingten Zuordnung auf die fünf Dynamikklassen führt der Klassifikatoransatz bereits ohne weitere Optimierungsschritte zu deutlich besseren Ergebnissen.

7.3 Diskussion der Ergebnisse

Der Aufbau des Klassifikators wurde kanonisch anhand der bestehenden 5 Dynamikklassen durch Verwendung eines k-NN-Algorithmus umgesetzt. Hierzu wurden von der Ausgangsklassifikation unabhängige Variablen ausgewählt und auf Grundlage der Ausprägungen von 9 selektierten Variablen ist bereits eine Zuordnungsgenauigkeit von **67 %** erzielt werden. Diese lag weit über einer zufallsbasierten Zuordnung.

Um die Güte des Klassifikators zu prüfen, wird im Folgenden nicht nur die Klassifikationsleistung im Vergleich zu der gegebenen Klassifikation untersucht, sondern auch die Fähigkeit zur Generalisierung, d.h. wie gut reagiert der Klassifikator auf neue Daten, die nicht zur Konstruktion des Klassifikators verwendet werden. Das vorhandene Datenmaterial wird hierzu in zufällige Trainings- und Testdatensätze aufgeteilt. Der Trainingsdatensatz dient der Klassifikatorkonstruktion. Der Testdatensatz ist der Datensatz, mit dem die Generalisierungsfähigkeit des konstruierten Klassifikators gemessen wird.

Abbildung 7-4 zeigt die Zuordnungsgenauigkeit des Nearest-Neighbour-Klassifikators bei 20 zu Grunde liegenden Versuchsreihen mit Trainings- und Testdatensätzen. Ablesbar sind der Mittelwert der Zuordnungsgenauigkeit und die berechnete Standardabweichung, welche sich auf die 20 Versuchsreihen beziehen. Es ist festzustellen, dass die Zuordnungsgenauigkeit im Durchschnitt bei 66 % liegt und sich das zuvor errechnete Ergebnis bestätigt.

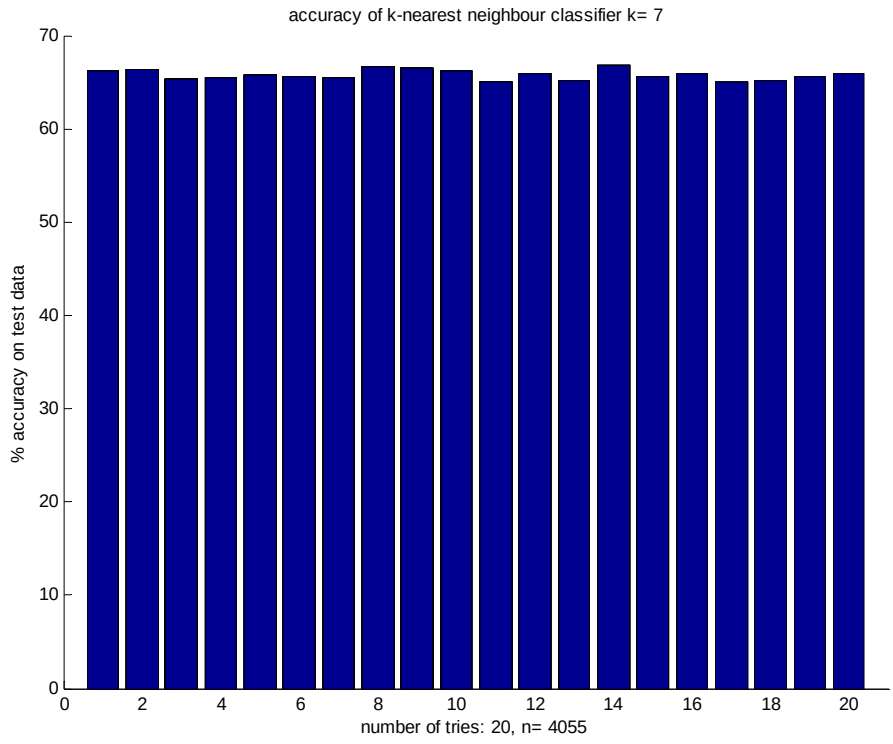


Abbildung 7-4: Zuordnungsgenauigkeit zu den 5 Dynamikklassen bei 20 zufälligen Testdatensätzen

Um den Klassifikator als Werkzeug in der Regionalforschung bzw. Regionalplanung längerfristig einsetzen zu können, besteht mit Blick auf die bisher erreichte Zuordnungsgenauigkeit weiterer Optimierungsbedarf. Durch den Aufbau des hier gezeigten Klassifikators wurde das Wirkprinzip eines Klassifikators zunächst genauer charakterisiert und die Möglichkeit angedeutet, zukünftig ein komplexes Messsystem zu entwickeln, welches aufbauend auf einer bestehenden Klassenstruktur eine Zuordnung weiterer Gemeinden vornehmen kann.

Im Sinne der Optimierung der Zuordnungsgenauigkeit ist der Zusammenhang zwischen den für die Klassifikation verwendeten Variablen und den für den Aufbau des Klassifikators eingesetzen Variablen genauer zu untersuchen. Dies kann einerseits zunächst mit Hilfe von grafischen strukturerkennenden Verfahren erfolgen. Beispielsweise ist durch Einsatz von Projektionsverfahren zu prüfen, inwieweit die unabhängigen Variablen überhaupt auch die Klassenunterschiedlichkeit abbilden. Andererseits ist mit Hilfe von zusätzlichen Korrelationsbetrachtungen oder insbesondere Ansätzen der Diskriminanzanalyse (siehe Abschnitt 2.4.2) eine weitere Optimierung der Variablenauswahl denkbar.

Zusätzlich sei auf die in Abschnitt 2.5 bereits hingewiesenen unterschiedlichen Klassifikatortypen einzugehen. Während der hier eingesetzte subsymbolische Klassifikatortyp (Nearest Neighbour-Klassifikator) ohne ein explizites Verständnis der zu leistenden Klassifikation arbeitet, kann z.B. ein symbolischer Klassifikator aufgrund seiner Beschreibungseigenschaften in einer eher natürlichsprachigen Form helfen, das Verständnis für jede einzelne Klasse zu präzisieren und damit auch die Bedeutung einer jeden Variablen für das Zuordnungsergebnis klären. Auf diese Weise können Variablen sehr genau in ihren Eigenschaften und ihrer Bedeutung für einen Klassifikator charakterisiert werden. Der sig*-Algorithmus (siehe Abschnitt 2.5.2.1) misst beispielsweise die Signifikanz einer merkmalsbasierten Klassenbeschreibung und generiert Regeln, die oftmals über eine sehr hohe Sensitivität verfügen, um vorgelegte bisher nicht zugeordnete Fälle nachträglich korrekt einer Klasse zuzuordnen. Nach Meinung der Entwickler[582] ist der sig*-Algorithmus besonders in Fachbereichen einsetzbar, bei denen es auf die Kommunizierung von neuen Erkenntnissen ankommt.

Diskutiert sei abschließend die Problematik einer Fehldimensionierung von Klassifikatoren, d.h. die Klassifikatorgüte hängt maßgeblich von der Festlegung der Anzahl von frei wählbaren Parametern ab. Für den k-NN-Algorithmus als Vertreter der subsymbolischen Klassifikatoren gilt: Je größer die Stichprobenmenge und Anzahl k, desto besser wird das Klassifikationsergebnis sein. Der frei wählbare Parameter ist bei diesem Ansatz durch k gegeben.

[582] Vgl. ULTSCH [1991, S. 169 ff]

7.4 Fazit

In diesem Kapitel wurde basierend auf einer bestehenden Klassenstrukur des deutschen Gemeindesystems der Aufbau eines Klassifikators kanonisch beschrieben. Dadurch ist gerade die Operationalisierung als wichtiger Teilschritt des Data Mining in ihrer Wirkungsweise charakterisiert worden, d.h. es wurde ein Algorithmus angewendet, der beliebige Daten bzw. in diesem Fall Gemeinden mit klassifikationsunabhängigen Variablen auf Klassen aufteilen kann. Es wurde gezeigt, dass es zur Bestimmung der Güte von Klassifikatoren nicht ausreicht, nur die Klassifikationsleistung im Vergleich zu einer gegebenen Klassifikation zu untersuchen. Vielmehr ist auch die Generalisierungsfähigkeit zu klären, d.h. wie gut reagiert ein Klassifikator auf neue Daten, also neue Datensätze, die nicht zur Konstruktion des Klassifikators verwendet wurden. Dargestellt wurden für diesen Zweck die unterschiedlichen Formen von Datensätzen (Trainings-, Test- und Validierungsdatensatz), die sich für einen Klassifikatoraufbau eignen.

Mit Hilfe der Clusteranalyse wurde in vielen Arbeiten der Stadtklassifizierung der Ansatz verfolgt, einen Ausschnitt der Realität standpunkts- bzw. zweckabhängig mehrdimensional abzubilden. In einigen Fällen wurde die Diskriminanzanalyse zusätzlich eingesetzt, um die Eigenschaft von Variablen zur Trennfähigkeit zu untersuchen. Es ist jedoch festzustellen, dass die Möglichkeit zur Integration von Klassifikatoren in einen Planungs- und Steuerungsprozess auf dem Gebeit der Stadt- und Regionalplanung bisher kaum umfassend untersucht worden ist. Gerade durch Nutzung von Klassifikatoreigenschaften bieten sich Ergänzungen zu bestehenden Urban Monitoring Systemen an, um eine Diagnose zu einer räumlichen Fragestellung zu bieten. Zukünftig ist verstärkt an der Entwicklung von symbolischen Klassifikatoren zu arbeiten, um ein Verständnis für die Klassen und die Klassenzuordnung anhand der unabhängen Variablen zu erhalten.

Im Zusammenhang mit wissensbasierten Systemen müssen Klassifikatoren in der Lage sein, nicht nur eine Diagnose zu treffen, sondern sie auch zu begründen, wobei eine sprachliche Darstellung des Wissens benötigt wird. Auf der Grundlage von wissensbasierten Systemen und den hier vorgestellten Ansätzen der Klassifikatorentwicklung sind zukünftig weitere raumwissenschaftliche Diagnosen oder individuell entwickelte Stadtpolitiken vorstellbar. Im Kontext einer „Lernenden Stadtregion“, einer „Wissensstadt“ oder auch einer „Knowledge Based City“[583] ist zukünftig nach weiteren methodischen Anwendungsgebieten zu suchen.

[583] SIMMIE / LEVER [2002]

8 Entwicklung eines Schätzverfahrens zur Übertragung räumlicher Information

8.1 Untersuchungsaufgabe: Deutscher Gebäudebestand

In diesem Kapitel wird eine Methode zur Schätzung des deutschen Gebäudebestandes dargestellt. Grundsätzlich besteht die Schwierigkeit darin, den Gebäudebestand in seiner ganzen Komplexität zu erfassen und dokumentieren zu können. Gezeigt wird an dieser Stelle, dass es möglich ist, anhand von bekannten Untersuchungsobjekten auf weitere bisher nicht charakterisierbare Objekte zu schließen. Als bekannt wird dabei ein Untersuchungsobjekt bezeichnet, wenn für dieses statistisches Datenmaterial oder andere Information vorliegt. Abbildung 8-1 zeigt Gemeinden mit erhobenen Daten sowie Gemeinden mit Datenlücken.

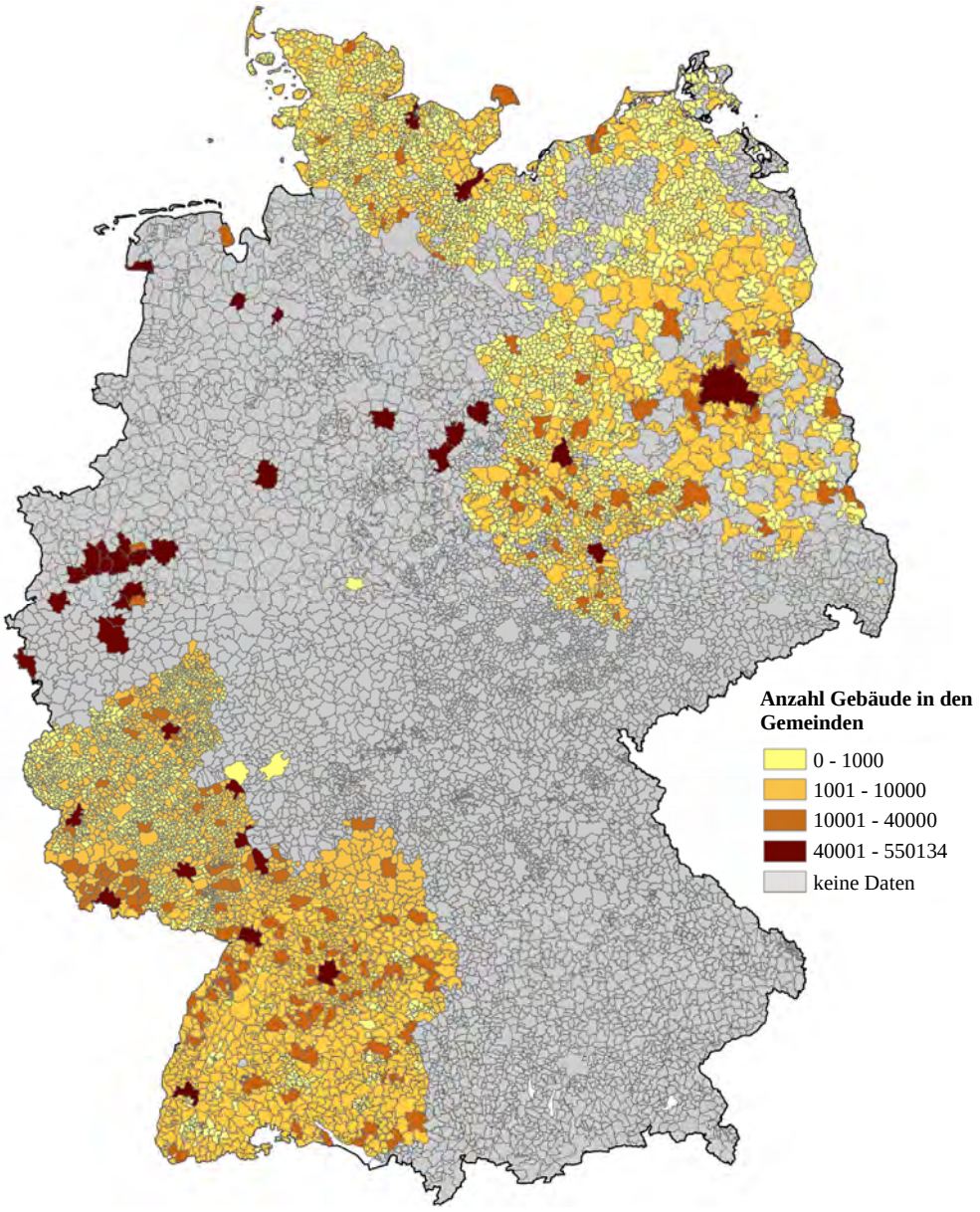
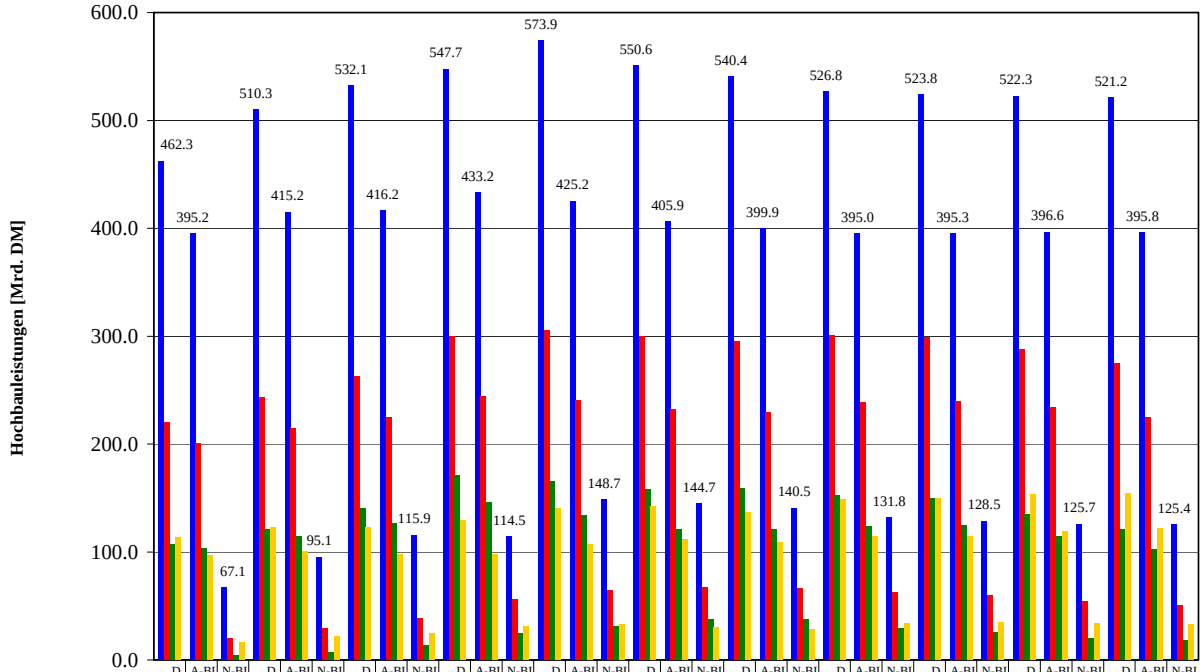


Abbildung 8-1: Status quo der Erfassung (6560 Gemeinden mit Daten; 5870 Gemeinden ohne Daten)[584]

[584] Es wurde versucht, die ALK-Daten möglichst flächendeckend und ausreichend gut verteilt zu beschaffen. Es handelt sich bei den Daten um den maximal noch ohne größere Aufwandssteigerung zugänglichen Bestand.

In der ENQUETE-KOMMISSION[585] wird festgestellt, dass nur knapp 1 % des vorhandenen Gebäudebestandes jährlich neu entsteht und vom Baubestand, der im Jahr 2020 genutzt werden wird, schon rund 80 % gebaut ist. Es besteht die Problematik, den Gesamtgebäudebestand in Deutschland gemeindescharf zu beziffern. In den vergangenen Jahren verzeichnet der Anteil der Baumaßnahmen an bestehenden Gebäuden eine starke Zunahme, und diese prägen die Tätigkeiten im gesamten Baubereich (Abbildung 8-2).[586]



	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL	D	A-BL	N-BL
	1991			1992			1993			1994			1995			1996			1997			1998			1999			2000			2001		
Gesamthochbauvolumen	462.3	395.2	67.1	510.3	415.2	95.1	532.1	416.2	115.9	547.7	433.2	114.5	573.9	425.2	148.7	550.6	405.9	144.7	540.4	399.9	140.5	526.8	395.0	131.8	523.8	395.3	128.5	522.3	396.6	125.7	521.2	395.8	125.4
davon im Wohnungsbau	220.5	200.4	20.1	243.6	214.9	28.7	262.6	224.6	38.0	300.0	244.2	55.8	305.1	240.8	64.3	299.5	232.3	67.2	295.6	229.7	65.9	301.0	238.2	62.8	298.9	239.3	59.6	287.7	234.0	53.7	275.0	224.5	50.5
Neubau	107.3	103.3	4.0	121.1	114.4	6.7	140.3	126.7	13.6	170.9	146.0	24.9	165.0	133.8	31.2	157.7	120.7	37.1	158.7	120.7	37.9	152.7	123.4	29.3	149.7	124.7	25.0	134.9	114.9	20.0	120.5	102.6	17.9
Bauleistungen im Bestand	113.2	97.1	16.1	122.5	100.5	22.0	122.3	97.9	24.4	129.1	98.2	30.9	140.1	107.0	33.1	141.7	111.6	30.1	136.9	109.0	27.9	148.3	114.8	33.5	149.2	114.6	34.6	152.8	119.1	33.7	154.5	121.9	32.6

Abbildung 8-2: Hochbauleistungen in Mrd. DM in Deutschland in Preisen von 1999[587]

Das Brutto-Bau-Anlagevermögen in Deutschland betrug in Wiederbeschaffungspreisen im Jahr 2000 insgesamt 15,81 Bill. DM (~8,08 Bill. €), wovon 8,75 Bill. DM (~4,47 Bill. €) auf Wohngebäude und 7,06 Bill. DM (~3,61 Bill. €) auf Nichtwohngebäude entfielen.[588] Im Zuge der Produktion von Gütern und Dienstleistungen entstehen 32 % der deutschen Nettowertschöpfung, die mittelbar über Vorleistungen oder unmittelbar dem Aktivitätsfeld ‚Wohnen und Bauen' dienen.[589] Der Gebäudebestand bildet einen bedeutenden Wirtschaftsfaktor, und daher ist eine bessere Kenntnis über seine Nutzungsstrukturen, Bauzustände, Abrissraten und Energiekosten für die Volkswirtschaft von großem Interesse.

585 Vgl. HASSLER, U. / KOHLER, N. / PASCHEN, H. [1999, S. 1]
586 Vgl. DIW (Hg.) [1998] und BfLR [1996]
587 Vgl. BUNDESMINISTERIUM FÜR VERKEHR, BAU- UND WOHNUNGSWESEN (Hrsg.) [2003, S.19]
588 Vgl. BUNDESMINISTERIUM FÜR VERKEHR, BAU- UND WOHNUNGSWESEN (Hrsg.) [2003, S. 15]
589 Vgl. STATISTISCHES BUNDESAMT [1997, S.97 ff.]

In ausgeprägter Abhängigkeit von der Bebauungsstruktur stehen zusätzlich die Kosten für die Bereitstellung, den Betrieb und den Unterhalt von Infrastrukturanlagen und -einrichtungen.[590]

Mit dem Bedarfsfeld ‚Wohnen und Bauen' sind rund zwei Fünftel der weltweiten Stoffflüsse verbunden.[591] In Deutschland entfallen auf das Bauwesen etwa 25 % der jährlich anthropogen verursachten mineralischen Stoffflüsse. Es ist mit einem jährlichen Stoffinput von 660 Mio. Tonnen der materialintensivste Wirtschaftszweig Deutschlands. Die in der Technosphäre gespeicherten Massen betragen etwa 60 Mrd. Tonnen.[592] Der Bausektor nimmt mit einem Anteil von 75 % am Gesamtaufkommen in der Abfallentsorgung eine wichtige Rolle ein. Die baubedingten Abfälle umfassen den Erdaushub, Straßenaufbruch, Bauschutt und Baustellenabfälle, wobei pro Jahr ungefähr 50 bis 60 Mio t Bauschutt entstehen. Ca. 70 % werden davon recycelt, größtenteils für niederwertige Nutzungen (z.B. Lärmschutzwände, Hinterfüllungen und ungebundene Tragschichten im Straßen-/Wegebau).[593]

Der Anteil am Primärenergieverbrauch insgesamt beträgt für den Bereich ‚Wohnen und Bauen' ca. 38 %.[594] Der Wohnungsneubau (Baustoffherstellung und Baugewerbe) verursacht pro Jahr 5 % des Primärenergieverbrauchs. Deutlich höher ist der Energieverbrauch während der Nutzungsphase der Gebäude, insbesondere durch die Bereitstellung von Heizenergie aber auch durch den Energieverbrauch für Warmwasserbereitung und Stromverbrauch. Circa 75 % des Hochbaubestandes stammen aus einer Zeit, in der keine Anforderungen durch eine Wärmeschutzverordnung gestellt wurden (vor Inkrafttreten der 1. WSVO in 1977). Etwa 96 % der gesamten benötigten Heizenergie verbrauchen die Gebäude, die vor Inkrafttreten der 2. WSVO im Jahr 1983 errichtet wurden (ca. 82 % des Gebäudebestandes).[595] Die WSVO und die Energieeinsparverordnung berücksichtigen, dass eine erhebliche Verminderung des Heizenergieaufwandes nur zu erreichen ist, wenn der Energieverbrauch im Bestand reduziert wird. AIBAU[596] beklagt den niedrigen Anteil nachträglicher Wärmeschutzmaßnahmen.

Der Gebäudebestand ist als Zwischenlager von Baumaterialien und Bauteilen aufzufassen.[597] Das Zwischenlager kann dabei als Ressource für die Baustoffe und Bauteile der Zukunft interpretiert werden. Eine Schließung von Stoffkreisläufen ist derzeit nicht möglich, da der stoffliche Input an Baumaterialien weitaus größer ist als sein Output an Bauabfällen.

---

590 Vgl. SIEDENTOP [2003]
591 Vgl. ROODMAN/LENSSEN [1995]: Es handelt sich nur um die direkten Materialflüsse.
592 Vgl. HOLZKAMP [1999, S. 1/59]
593 Vgl. COENEN/GRUNWALD (Hrsg.) [2003, S.166]
594 Vgl. COENEN/GRUNWALD (Hrsg.) [2003, S.160]
595 Vgl. REISS et al. [1999, S. 98-105]
596 Vgl. BUNDESMINISTERIUM FÜR VERKEHR, BAU- UND WOHNUNGSWESEN (Hrsg.) [2003, S.63]
597 GLENCK [1996, S. 85-94]

Die jährlich in Neubau-, Umbau- und Sanierungsmaßnahmen verwendeten Baumaterialien und Bauteile (Input) beschreiben die Zugangskurve, und die Abgangskurve wird durch die Summe an verwertbaren und nicht verwertbaren Bauabfällen (Output) gebildet. Die Differenz von Zugang zu Abgang entspricht der jährlichen Veränderung des „Lagerbestandes", wobei je nach verwendetem Rechenmodell[598] die Relation von Input- zu Outputströmen erheblich schwankt. Die Materialmenge, die 1993 im „Bauwerk Deutschland" (Hoch- und Tiefbau) gebunden war, beträgt 18 Mrd. Tonnen. Dabei wird der jährliche Input für Neubauten, Renovierungen und Sanierungen für die alten Bundesländer auf 575 Mio. Tonnen beziffert.[599] Abbildung 8-3 zeigt die Relation zwischen Stoffinput- und Stoffoutputströmen.

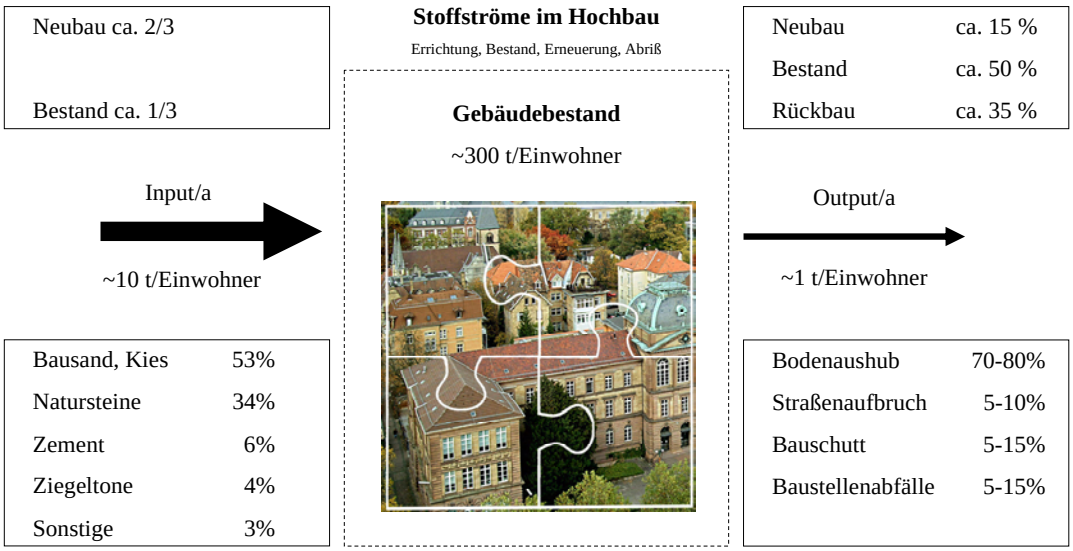


Abbildung 8-3: Stoffströme im Hochbau (makroökonomische Ergebnisse)[600]

A priori ist die Kenntnis über mehrdimensionale Ähnlichkeiten von Gemeinden in Bezug auf den in der Regel über Jahrhunderte gewachsenen Gebäudebestand sehr gering. Eine Untersuchung von ausgewählten Teilbeständen auf regionaler Ebene kann zukünftig dazu beitragen, den Gebäudebestand in seiner Struktur und Dynamik genauer zu charakterisieren. Hier fördern gerade methodische Ansätze zur Übertragung von räumlicher Information die Verallgemeinerung von gewonnenen Erkenntnissen.

---

598 Die ENQUETE-KOMMISSION [1997, S.93] gibt ein Verhältnis in der Spanne 2:1 bis 10:1 an. SCHMIDT-BLEEK et al. [1996, S.II./2] errechnen ein Verhältnis von Stoff-Input zu Stoff-Output von 2,6:1 mit dem Bezugsjahr 1991, GRIEßHAMMER/BUCHERT [1996, S.46] nennen ein Verhältnis von 9:1 mit dem Bezugsjahr 1991. Für die Schweiz kommen vergleichbare Untersuchungen zu Verhältnissen von 13:1 am BUNDESAMT FÜR KONJUNKTURFRAGEN [1991], 2:1 durch BACCINI [1994] und 1,8:1 durch KOHLER [1994], für Österreich zu 10:1 durch GLENCK [1996]; Differenzen sind auf die mangelnde Datenlage, auf unterschiedliche Bezugsjahre und Abgrenzungen bzw. Systemgrenzen sowie auf die Berücksichtigung bzw. Nicht-Berücksichtigung des Erdaushubs als Bestandteil des Outputs zurückzuführen.

599 Vgl. Grießhammer/Buchert [1996]

600 Vgl. ENGELBACH [2000] und HOLZKAMP [1999] sowie Öko-Institut [1996] (Input-Zusammensetzung); Statistisches Bundesamt [1997] (Output-Zusammensetzung, Kohler u.a. [1994] (Neubau-Bestand-Rückbau-Verhältnis); Baccini, Brunner [1996] (Input-Bestand-Output-Verhältnis); Enquete [1997]

8.2 Herleitung und Darstellung der Ergebnisse

PARETO[601] definiert allgemein: „In einer beliebigen Menge von Elementen, die etwas bewirken sollen, bewirkt immer eine zahlenmäßig kleine Menge von Elementen den größten Effekt“. Konkret kann zu Beginn des Schätzansatzes eine Optimierung des Informationsgehaltes in Bezug auf das potentiell einzusetzende Datenmaterial erfolgen. Hierzu wird ein graphischer Ansatz basierend auf Lorenzkurven[602] verfolgt, der eine theoretische Begründung der Pareto-80 / 20 Regel mit einbezieht.[603] Die rote Markierung definiert dabei den kürzesten Abstand *a* von einer konkreten Datensituation (Verlauf der Lorenzkurve, blaue Linie) zu einem theoretischen Optimum (linke obere Ecke bzw. abstrakt ausgedrückt 0 % Aufwand, 100 % Nutzen). Tabelle 8-1 vermittelt eine Vorstellung von der Verteilung spezifischer Gebäudebestände (Gesamtbestand und Teilbestände) auf die 6050 Gemeinden. Es wird eine ungleichmäßige Verteilung erkannt: Ca. 20 % der Gemeinden enthalten 80 % der Gebäude.

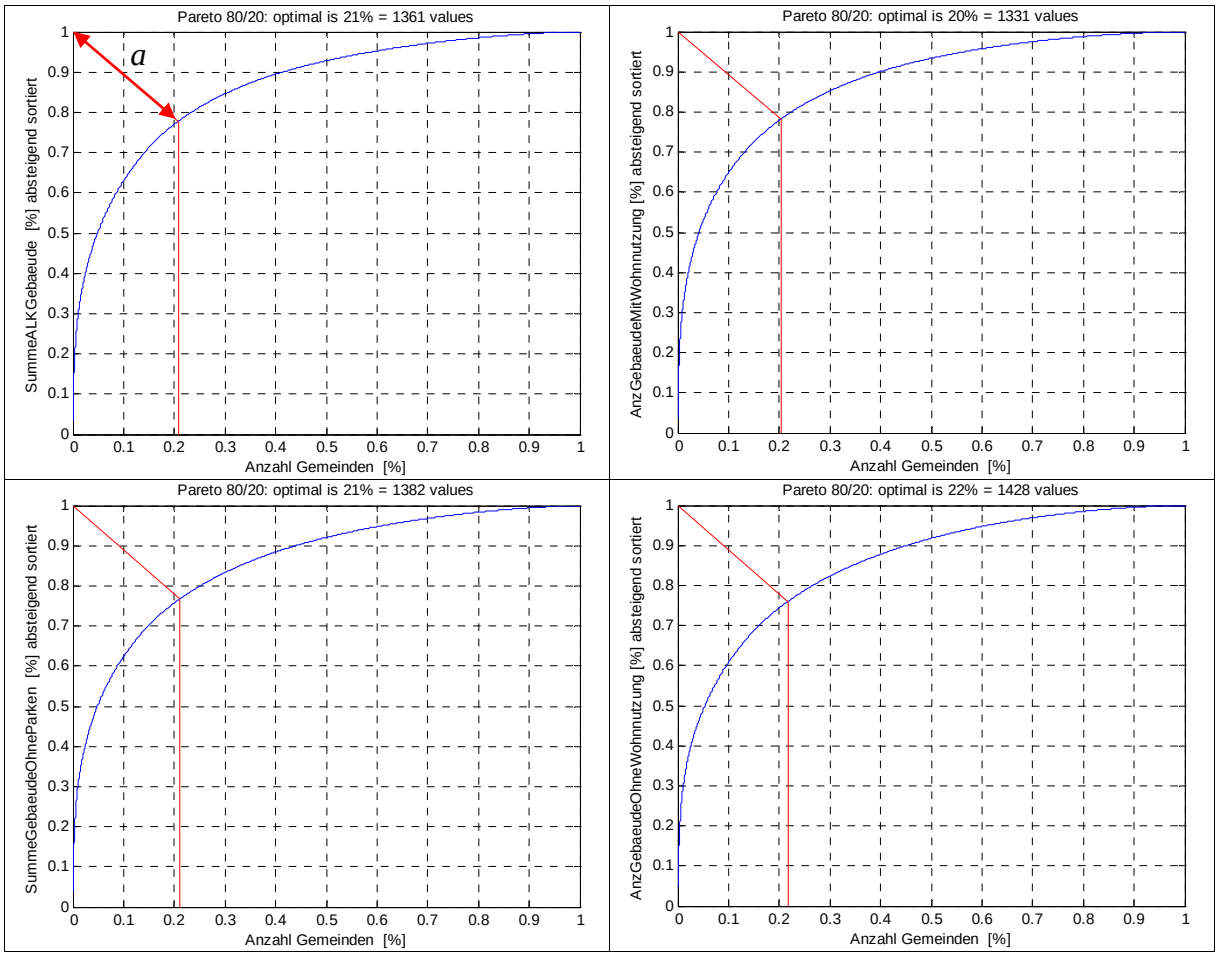


Tabelle 8-1: Optimierung des Informationsgehaltes (6050 Gemeinden, Gesamtbestand: 16.264.344)[604]

601 Vgl. SOMBART [1967], Der italienische Volkswirtschaftler Wilfredo Pareto (1848-1923) erkannte, dass im Italien des 19. Jahrhunderts 20 % der Bevölkerung 80 % des Landes besaßen. Siehe auch LAUKAT [1999]

602 Vgl. Abschnitt 2.1.1.2: Lorenzkurven zur graphischen Veranschaulichung der Datenkonzentration.

603 Vgl. ULTSCH [2001]: Theoretische Begründung der Pareto-80 / 20 Regel (siehe ‚unrealisiertes Potential’)

604 Gesamtbestand in einer Gemeinde = Summe Gebäude aus der ALK (siehe Nebenteil B).

Zusätzlich geprüft wird der Zusammenhang zwischen der Größe des Gebäudebestandes in einer Gemeinde und den gültigen Zentrale-Orte-Kategorien sowie den Raumstrukturtypen des BBR. Es handelt sich in Tabelle 8-2 um die Untersuchung von transformierten Daten mit LOG(Data). Auf der Y-Achse sind die Wahrscheinlichkeitsdichten (PDE)[605] für die Gebäudebestandsgröße ablesbar. In der Regel lassen sich die Kategorien anhand ihrer Ausprägungen deutlich unterscheiden (z.B. Oberzentren und Gemeinden ohne zentralörtliche Einstufung). Die PDE der Unter- und Kleinzentren sind hingegen nur schwer unterscheidbar.

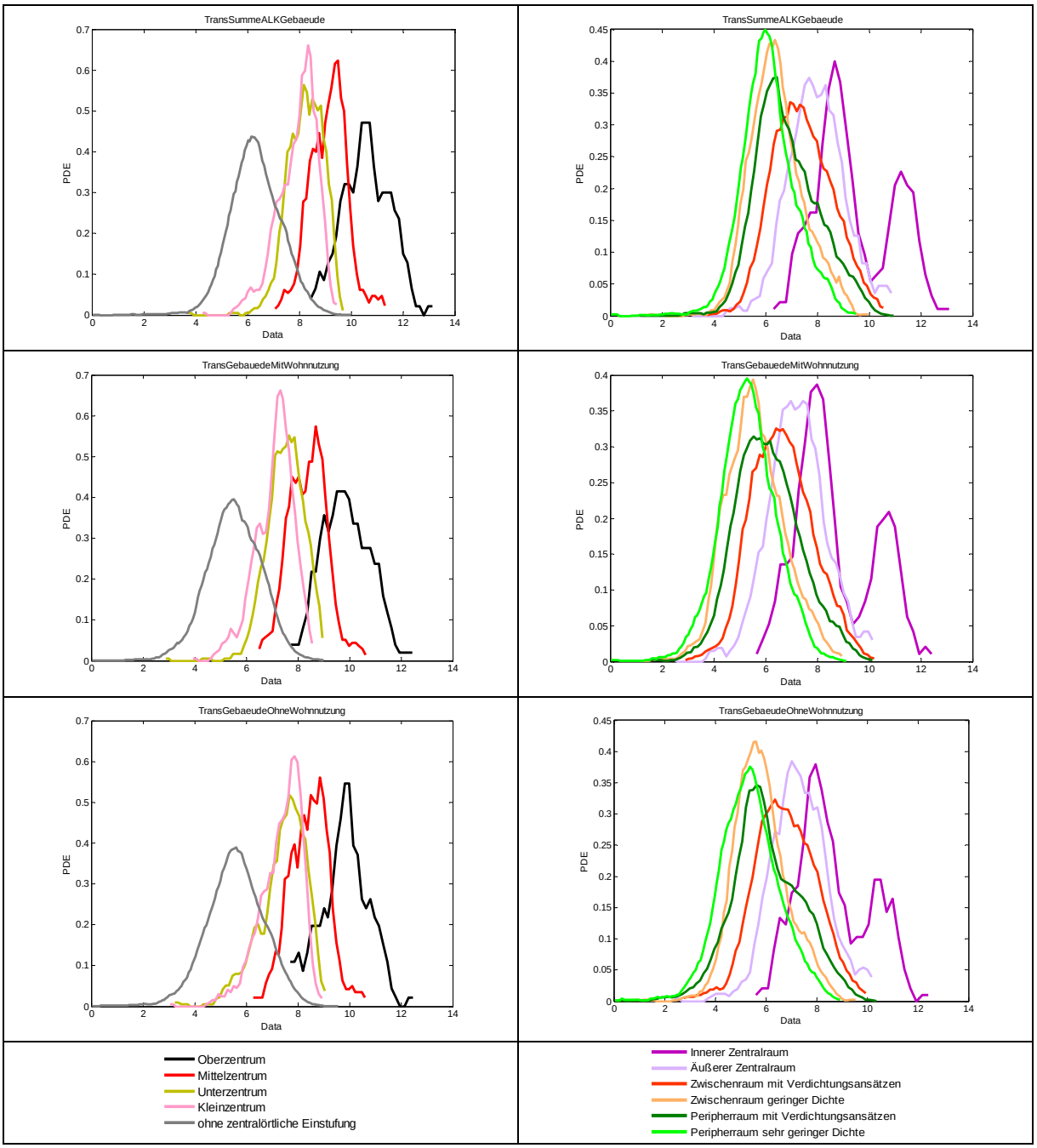


Tabelle 8-2: Gebäudebestände nach Zentrale-Orte-Kategorie (links) und Raumstrukturtyp (rechts)

605 Vgl. Abschnitt 2.1.1.2 Pareto Density Estimation, X=Variablenwert, Y=Wahrscheinlichkeitsdichte (PDE)

Das Ziel besteht darin, auf Basis von relativ leicht zugänglichen Gemeindestrukturdaten im weiteren Verlauf mit großer Präzision die Gesamtsumme der Gebäude in Deutschland mit Hilfe eines Algorithmus zu schätzen. Zunächst sind zu diesem Zweck aus einer Datenbasis mit insgesamt ca. 130 Variablen potentielle Schätzgrößen zu extrahieren, da vermutet wird, dass es sich um ein mehrdimensionales Schätzproblem handelt.

Die Auswahl erfolgt einerseits aus rein theoretischer Überlegung, wobei sich die Variablen im Sinne einer vergrößerten Aussageperspektive auf möglichst verschiedenene Dimensionen (z.B. Fläche, Personen, Mobilität oder Finanzen u.a.) beziehen sollen. Andererseits werden Scatter-Plots und Korrelationswerte eingesetzt, um nach Zusammenhängen zwischen potentiellen Schätzgrößen und den Gebäudebestandsdaten (‚SummeALKGebaeude') der 6560 Gemeinden zu suchen bzw. die bereits aus inhaltlicher Überlegung getroffene Vorauswahl zu prüfen und ggf. danach zu optimieren. Tabelle 8-3 vermittelt einen Eindruck von den Variablenuntersuchungen. Gezeigt sind die Eigenschaften von sechs Variablen, die in einem individuell spezifischen Zusammenhang zur Gesamtsumme der Gebäude stehen.

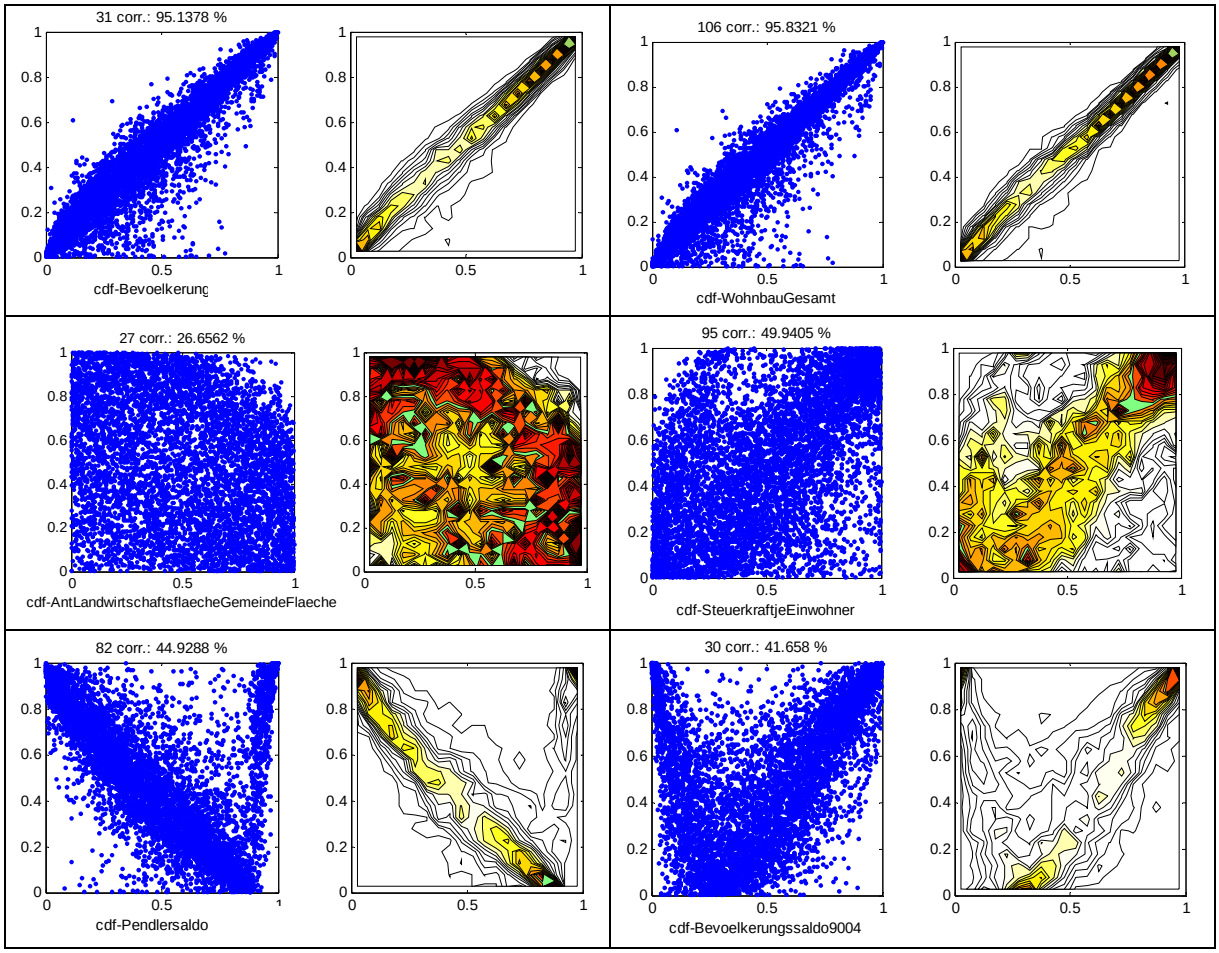


Tabelle 8-3: Scatter-Plots und PDEscatter von potentiellen Schätzgrößen und der Gebäudesumme [606]

606 Eigene Bearbeitung, weitere Variablenuntersuchungen finden sich im Nebenteil B.

Es sind zwei Arten von Scatter-Plots zu unterscheiden. Der PDEscatter ermöglicht eine Abbildung von Dichtestrukturen, so dass im Gegensatz zum üblichen Scatter-Plot zwei Modi sichtbar werden. Die zuvor dargestellten Variablenuntersuchungen basieren auf Daten, die CDF-transformiert[607] und auf das Intervall 0 bis 1 normiert wurden. Die Gesamtsumme der Gebäude ist auf der Y-Achse aufgetragen. Ein Wert bei 1 charakterisiert eine große Ausprägung und ein Wert nahe 0 eine kleine Ausprägung der Variablen. Zusätzlich ausgewiesen ist der Korrelationswert zwischen der betreffenden Variable und der Gesamtsumme der Gebäude.

Die Scatter-Plots und Korrelationswerte verweisen auf einen linearen Zusammenhang zwischen der Gesamtsumme der Gebäude und der Variable ‚Bevoelkerung'. Ein etwas stärkerer Zusammenhang ist mit der Variable ‚WohnbauGesamt' festzustellen, aufgrund eines höheren Korrelationswertes und der deutlicheren Ausprägung großer Gebäudemengen im Scatter-Dichteplot. Beide Variablen eignen sich, um mit Hilfe eines einfachen linearen Schätzers den Gebäudebestand in den deutschen Gemeinden zu prognostizieren.[608]

Die darüber hinaus gezeigten Scatter-Plots zu den Variablen ‚Steuerkraft je Einwohner', ‚AntLandwirtschaftflaecheGemeindefläche', ‚Pendlersaldo' und ‚Bevoelkerungssaldo9004' sind exemplarisch herausgegriffen, um anhand dieser weitere charakteristische Abhängigkeiten zum Gesamtgebäudebestand (‚SummeALKGebaeude') herauszustellen. Durch Interpretation der PDEscatter lassen sich spezielle Strukturzusammenhänge, wie z.B. bei der Variable zum Anteil der Landwirtschaftsfläche an der Gemeindefläche aufdecken. Für eine mehrdimensionale Schätzaufgabe lassen sich auf diese Weise geeignete Variablen finden.

Infolge der Variablenuntersuchung wurde die Entscheidung getroffen, den Gebäudebestand zunächst mit Hilfe eines linearen Schätzerverfahrens zu prognostizieren. Erst wenn die Genauigkeit einer derartigen Schätzung nicht zu zufriedenstellenden Ergebnissen führt, werden weitere Optimierungsansätze verfolgt (z.B. mehrdimensionale Schätzverfahren, multipler Regressionsansatz).

---

607 Die kumulative Verteilungsfunktion (cumulative distribution function, CDF) wird dazu eingesetzt, um die Wahrscheinlichkeit bzw. die Häufigkeit für das Auftreten eines Wertes x_i einer Messgröße zu beschreiben. Die Berechnungsvorschrift für die Verteilungsfunktionen der Messgröße lautet:

$$P(x < X) = \sum_{x_i \le X} P(x = x_i)$$

Dabei stellt $P(x = x_i)$ die Häufigkeit bzw. die Wahrscheinlichkeit des Auftretens des Wertes x_i dar.

608 Die Datenquellen werden im Abschnitt 3 beschrieben. Es besteht ein vom Autor erkanntes Schätzrisiko, da aktuell statistische Werte zu Gebäuden auf Gemeindebene nur schwer mit vollständig einheitlichem jährlichem Zeitbezug bundesweit zu beschaffen sind.

Auf Grund der Datenvorverarbeitung wurde erkannt, dass die Variablen ‚WohnbauGesamt' und ‚SummeALKGebaeude' (Summe der Gebäude aus der ALK) einer Lognormalverteilung folgen. Es konnte ein linearer Zusammenhang aufgedeckt werden, der zwischen diesen logarithmierten Variablen besteht und für die lineare Regressionsrechnung (siehe Abschnitt 2.4.3) die elementare Basis bildet. Abbildung 8-4 zeigt die gegeneinander aufgetragenen Variablen sowie die Regressionsgerade. Der bereits erkannte Zusammenhang zwischen den logarithmierten Variablen ‚Bevoelkerung' und ‚SummeALKGebaeude' ist in Abbildung 8-5 ergänzend aufgeführt, um durch Gegenüberstellung die Modellierungseigenschaften der im Folgenden verwendeten Schätzgröße ‚WohnbauGesamt' zu betonen.

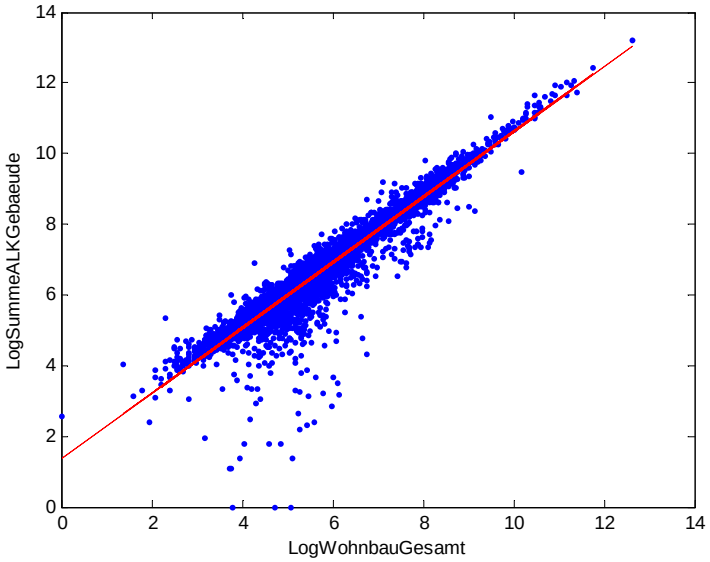


Abbildung 8-4: Regressionsgerade (‚LogWohnbauGesamt' und ‚LogSummeALKGebaeude') [609]

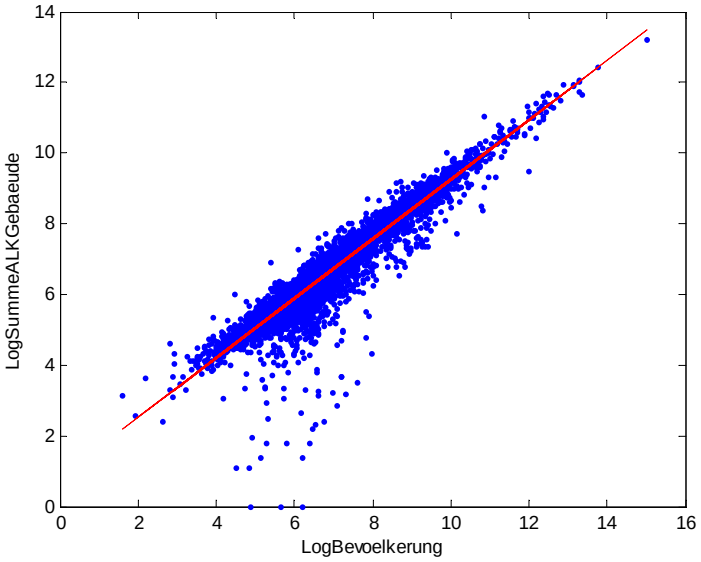


Abbildung 8-5: Regressionsgerade (‚LogBevoelkerung' und ‚LogSummeALKGebaeude') [610]

[609] Eigene Bearbeitung, Gleichung der Regressionsgerade: $p(t) = 1{,}34 + 0{,}929 \cdot t$

[610] Eigene Bearbeitung, Gleichung der Regressionsgerade: $p(t) = 0{,}83 + 0{,}84 \cdot t$

Auf Basis der Regressionsgleichung und der bereits bekannten Ausprägung der Variablen ‚WohnbauGesamt' wird für alle 12430 Gemeinden der Gesamtgebäudebestand in einer Gemeinde prognostiziert. Abbildung 8-6 zeigt das verortete Ergebnis zur Schätzung des deutschen Gebäudebestandes und die bereits vorhandenen Daten der 6560 Gemeinden.

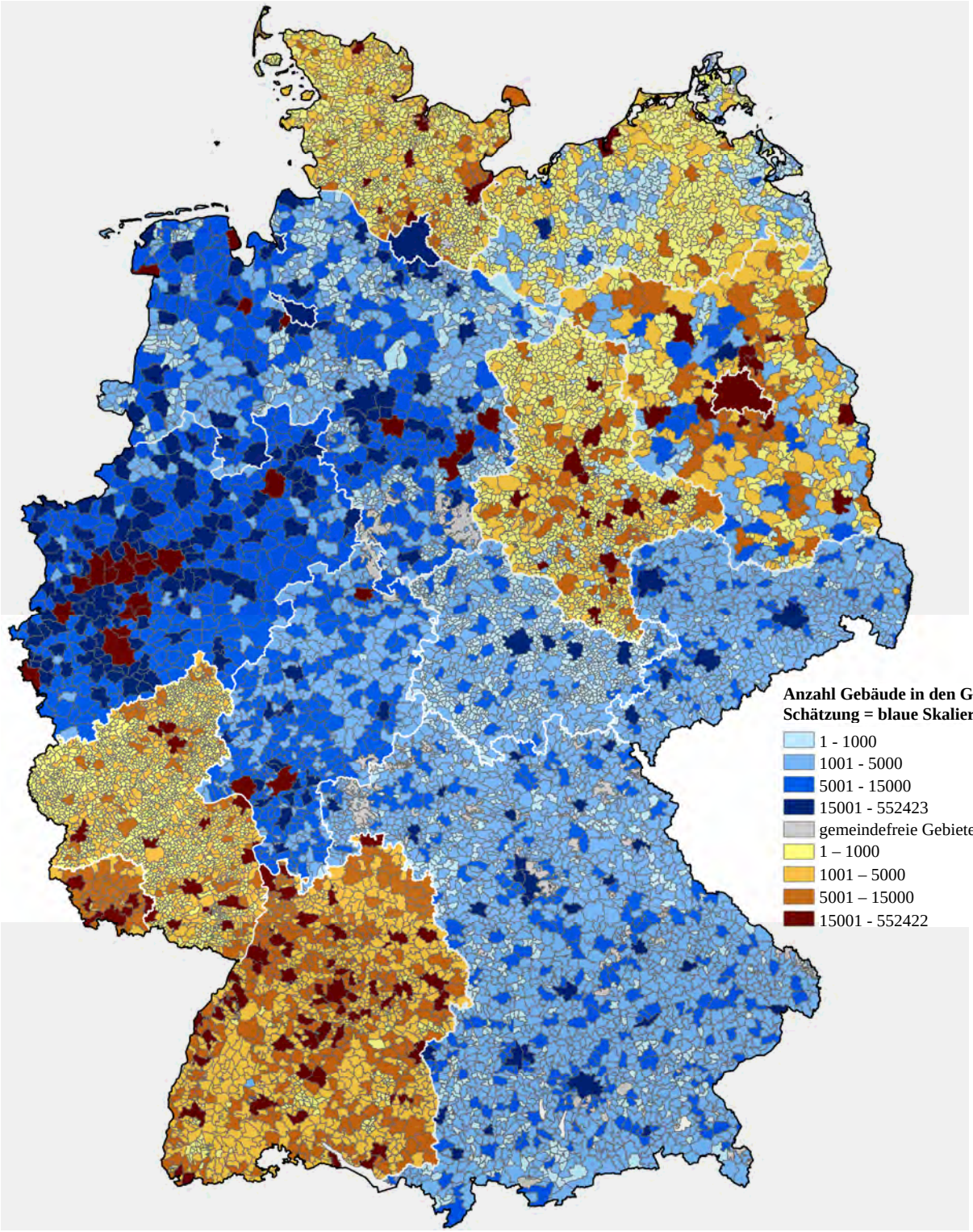


Abbildung 8-6: Verortung der Schätzung des deutschen Gebäudebestandes[611]

[611] Die in blau dargestellten 5870 Gemeinden zeigen die Schätzwerte des Gesamtbestandes in einer Gemeinde.

Infolge der Aufsummierung der zuvor verortet dargestellten Bestandszahlen auf Gemeindeebene lässt sich der Gesamtgebäudebestand in Deutschland auf insgesamt **38 Millionen Gebäude** beziffern.[612]

Tabelle 8-4 enthält die Ergebnisse des Ansatzes zur Schätzung des Deutschen Gebäudebestandes, gegliedert nach den bestehenden Zentrale-Orte-Kategorien.[613] Es wird der geschätzte Gesamtbestand je Kategorie bestimmt, wobei zusätzlich als erster Orientierungswert der Median sowie der Mittelwert aus den zu jeder Kategorie zugewiesenen Gemeinden berechnet wurde. Unter Berücksichtigung der Anzahl der Gemeinden und den jeweils berechneten Teilbeständen wird erkennbar, dass je nach Zentrale-Orte-Kategorie die Größe des Gebäudebestandes in einer Gemeinde unterschiedlich ausfällt. Zusätzlich enthalten ist in Tabelle 8-4 eine Angabe aus der amtlichen Statitistik zum bereits bekannten Wohnbaubestand der Gemeinden in Deutschland.[615]

Bezeichnung [Anzahl Gemeinden]		**Deutschland [12430]**	**Oberzentrum [157]**	**Mittelzentrum [930]**	**Unterzentrum [1484]**	**Kleinzentrum [1263]**	**ohne Einstufung [8670]**
Gesamtbestand[614]		37.997.958	7.574.334	10.989.500	7.408.938	3.727.835	8.297.351
	Anteilswert		19,93 %	28,92 %	19,50 %	9,81 %	21,84 %
Gebäudebestand je Gemeinde	Median		25.365	9.555	4.446	2.574	676
	Mittelwert		48.244	11.817	4.992	2.952	965
Wohnbau[615]		17.458.934	3.871.434	5.240.573	3.383.516	1.592.910	3.370.501
	Anteilswert		22,17 %	30,02 %	19,38 %	9,12 %	19,31 %
Gebäudebestand je Gemeinde	Median		12.516	4.406	2.002	1.106	262
	Mittelwert		24.659	5.635	2.280	1.261	392

Tabelle 8-4: Schätzergebnisse des deutschen Gebäudebestandes nach Zentrale-Orte-Kategorien

Die nachfolgende Tabelle 8-5 und Abbildung 8-7 enthalten die Ergebnisse des Schätzungsansatzes, gegliedert nach Gemeindegrößenklassen des Statistischen Bundesamtes. Bezogen auf einzelne Gemeindegrößenklassen ist der Tabelle die Gesamtanzahl der dazugehörigen Gemeinden zu entnehmen. Weiterhin sind für eine Gemeinde der jeweiligen Größenklasse der durchschnittliche Gesamtbestand und der berechnete Median ausgewiesen.

612 Dieses hier berechnete Ergebnis hat einen gewissen Pioniercharakter, da nach Kenntnis des Autors dieser Arbeit bisher keine weitere Arbeit existiert, die methodisch begründet eine Abschätzung des deutschen Gebäudebestandes nach den letzen Gebäude- und Wohnungszählungen der 1950er Jahre ermöglicht.

613 Die zu Grunde gelegten Zentralen-Orte-Kategorien wurden vom Autor teilweise eigenständig erfasst, so dass Fehler bei der Datenerhebung aus den Informationen der Landesentwicklungs- und Regionalplänen nicht vollständig auszuschließen sind.

614 Der Fehler der Schätzung des Gesamtbestandes beeinflusst maßgeblich die Mengenangaben.

615 Die Bestandsgröße im Wohnbau (‚WohnbauGesamt') basiert auf der Gebäude- und Wohnungsstatistik des Statistischen Bundesamtes und charakterisiert den Wohngebäudebestand in einer Gemeinde. Es besteht bei Vergleich der Angaben zum Wohnbau und den Ergebnissen der Schätzung des Gesamtbestandes die Problematik der teilweise unterschiedlich verwendeten Erfassungsdefinitionen und Erhebungsgenauigkeiten für ein Wohngebäude in der Automatisierten Liegenschaftskarte und der Amtlichen Statistik.

Größenklasse [Anzahl Gemeinden]	Deutschland [12430]	500.000 und mehr [12]	200.000 bis 500.000 [25]	100.000 bis 200.000 [45]	50.000 bis 100.000 [107]	20.000 bis 50.000 [511]
Gebäudebestand	37.997.958	2.540.393	2.166.559	1.989.879	2.809.766	7.002.448
	Anteilswert	6,69 %	5,70 %	5,24 %	7,39 %	18,43 %
Gebäudebestand je Gemeinde	Median	168.145	83.600	42.214	25.115	12.998
	Mittelwert	211.699	86.662	44.219	26.259	13.703
Zwischensumme der 5 Anteilswerte am Gesamtbestand: **43,45 %**, (= 16.509.045 Gebäude), 700 Gemeinden						

Größenklasse [Anzahl Gemeinden]		10.000 bis 20.000 [875]	5000 bis 10.000 [1306]	4000 bis 5000 [514]	3000 bis 4000 [772]	2000 bis 3000 [1163]
Gebäudebestand		6.353.002	5.457.872	1.471.649	1.789.910	1.994.632
	Anteilswert	16,72 %	14,36 %	3,87 %	4,71 %	5,25 %
Gebäudebestand je Gemeinde	Median	7.085	4.025	2.813	2.277	1.687
	Mittelwert	7.261	4.179	2.863	2.318	1.715
Zwischensumme der 5 Anteilswerte am Gesamtbestand: **44,92 %** (= 17.067.065 Gebäude), 4630 Gemeinden						

Größenklasse [Anzahl Gemeinden]		1000 bis 2000 [2192]	500 bis 1000 [2283]	200 bis 500 [1897]	100 bis 200 [530]	unter 100 [198]
Gebäudebestand		2.338.269	1.372.352	602.643	90.282	18.302
	Anteilswert	6,15 %	3,61 %	1,59 %	0,24 %	0,05 %
Gebäudebestand je Gemeinde	Median	1.042	589	310	165	90
	Mittelwert	1.067	601	317	170	92
Zwischensumme der 5 Anteilswerte am Gesamtbestand: **11,64 %** (= 4.421.848 Gebäude), 7100 Gemeinden						

Tabelle 8-5: Schätzergebnisse des deutschen Gebäudebestandes nach Gemeindegrößenklassen (StaBu)[616]

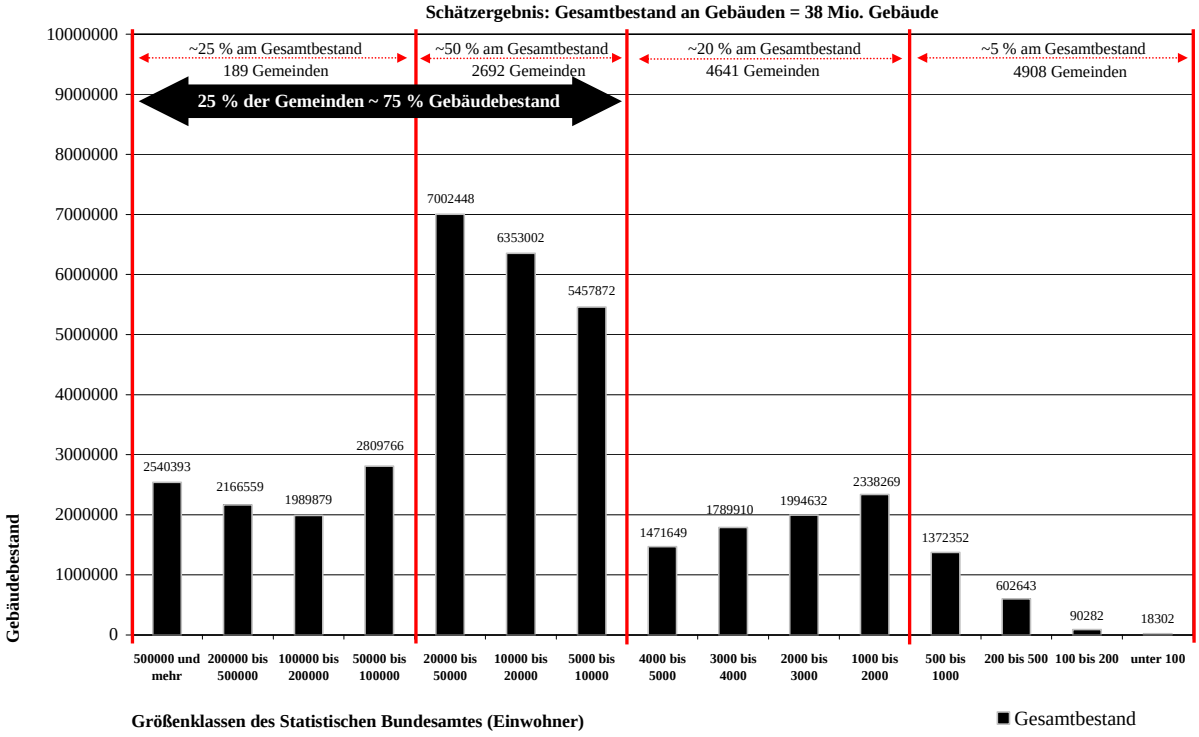


Abbildung 8-7: Schätzung des Gesamtbestandes unterteilt in Teilbestände nach Gemeindegrößenklassen

616 Die Tabelle berücksichtigt die 15 Gemeindegrößenklassen des Statistischen Bundesamtes, die eine Kategorisierung nach der Einwohnerzahl der deutschen Gemeinden repräsentieren.

8.3 Diskussion der Ergebnisse

Das vorhandene Datenmaterial mit Gesamtwohnbaubestand und bekanntem Gesamtgebäudebestand (‚SummeALKGebaeude', ‚Wohnbau Gesamt') wurde zuvor dazu verwendet, um den Deutschen Gebäudebestand zu schätzen. Im Folgenden findet eine kritische Auseinandersetzung mit den Schätzergebnissen statt, wobei insbesondere die Genauigkeit diskutiert wird.

Es werden hierzu aus dem vorhandenen Datenmaterial zufällige Trainings- und Testdatensätze[617] gebildet und anhand dieser der Fehler der prognostizierten Daten ermittelt. Durchgeführt wird die Schätzung für die Datensätze der Stichproben in gleicher Weise wie zuvor der dargelegte Schätzansatz für die unbekannten Objekte. Zunächst wird die Regressionsgerade aus den Trainingsdaten ermittelt. Danach wird eine Prognose für die fehlenden Objekte gestellt, die anhand der erzeugten Testdatensätze abgeglichen wird und die Ermittlung des prozentualen Fehlers zulässt. Abbildung 8-8 zeigt den auf Grundlage der gebildeten Trainings- und Testdatensätzen bestimmten Generalisierungsfehler. Ablesbar sind einerseits der Mittelwert und andererseits die postive bzw. negative einfache Standardabweichung des Generalisierungsfehlers.

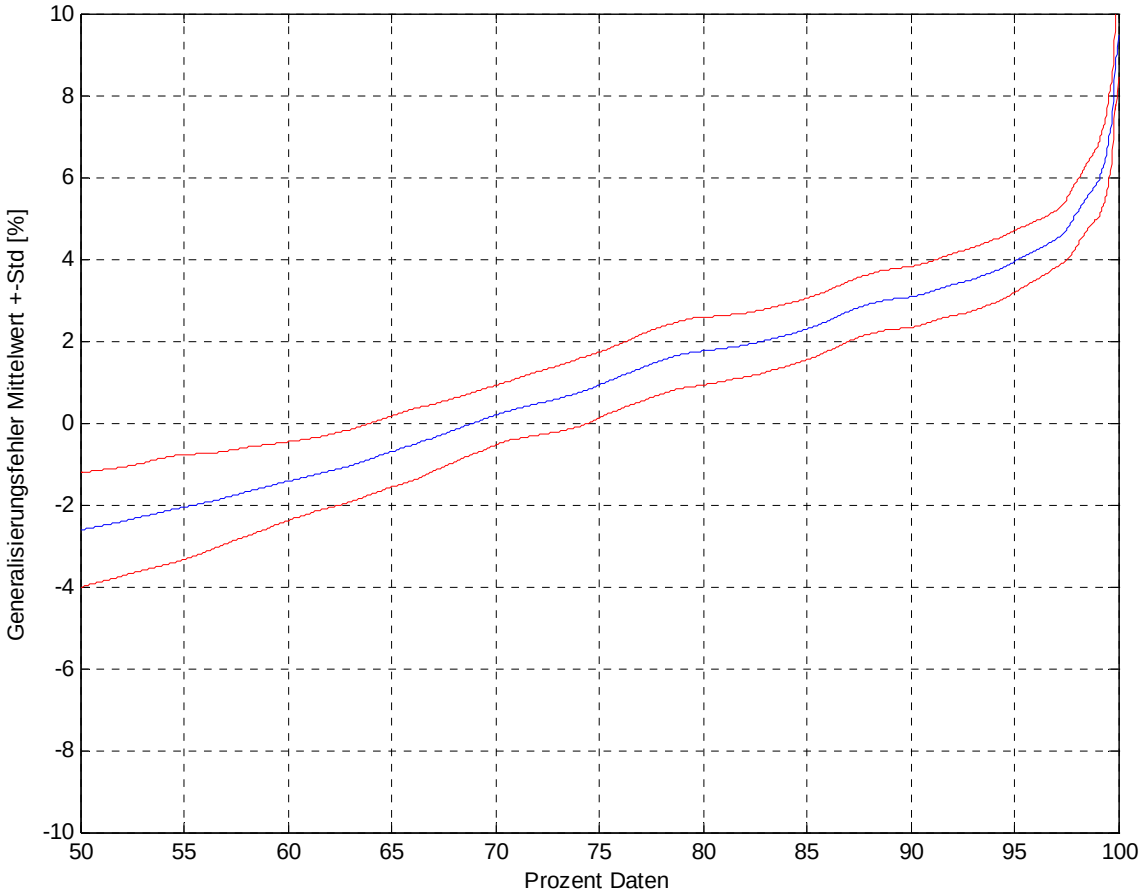


Abbildung 8-8: Generalisierungsfehler der Schätzung des deutschen Gebäudebestandes

[617] Im Nebenteil B sind die Ergebnisse der Testdatensätze detailliert aufgeführt und einsehbar.

Abbildung 8-9 enthält zur Ergänzung exemplarisch herausgegriffen die Ergebnisse der Überprüfung des Schätzansatzes für einen Trainings- und Testdatensatz. Gegeneinander aufgetragen sind die Größe ‚LogWohnbauGesamt' und ‚LogSummeALKGebaeude'. Eingezeichnet ist die aus dem Trainingsdatensatz ermittelte Regressionsgerade (Abbildung links), die für die Schätzung der weiteren Objekte dient und mit den Testdatensätzen verglichen wird.

Die Differenz der Schätzergebnisse von den bekannten Gesamtsummen der Gebäude in den Gemeinden ist hierzu grafisch aufgetragen (Abbildung rechts). Für ca. 80 % der Daten ist der prozentuale Schätzfehler als relativ gering (< 2 %) einzuschätzen.

Die Untersuchung der Abweichung in Bezug auf den vorhandenen Wohnbaubestand deutet bei diesem Testdatensatz darauf hin, dass die Genaugkeit der Schätzung in großem Maß von den Gemeinden mit großen Wohnbaubeständen abhängt. Ersichtlich wird dieses aus der mittleren Abbildung, bei der die Abweichung (Y-Achse) und der logarithmierte Wohnbaubestand (X-Achse) gegeneinander aufgetragen sind.

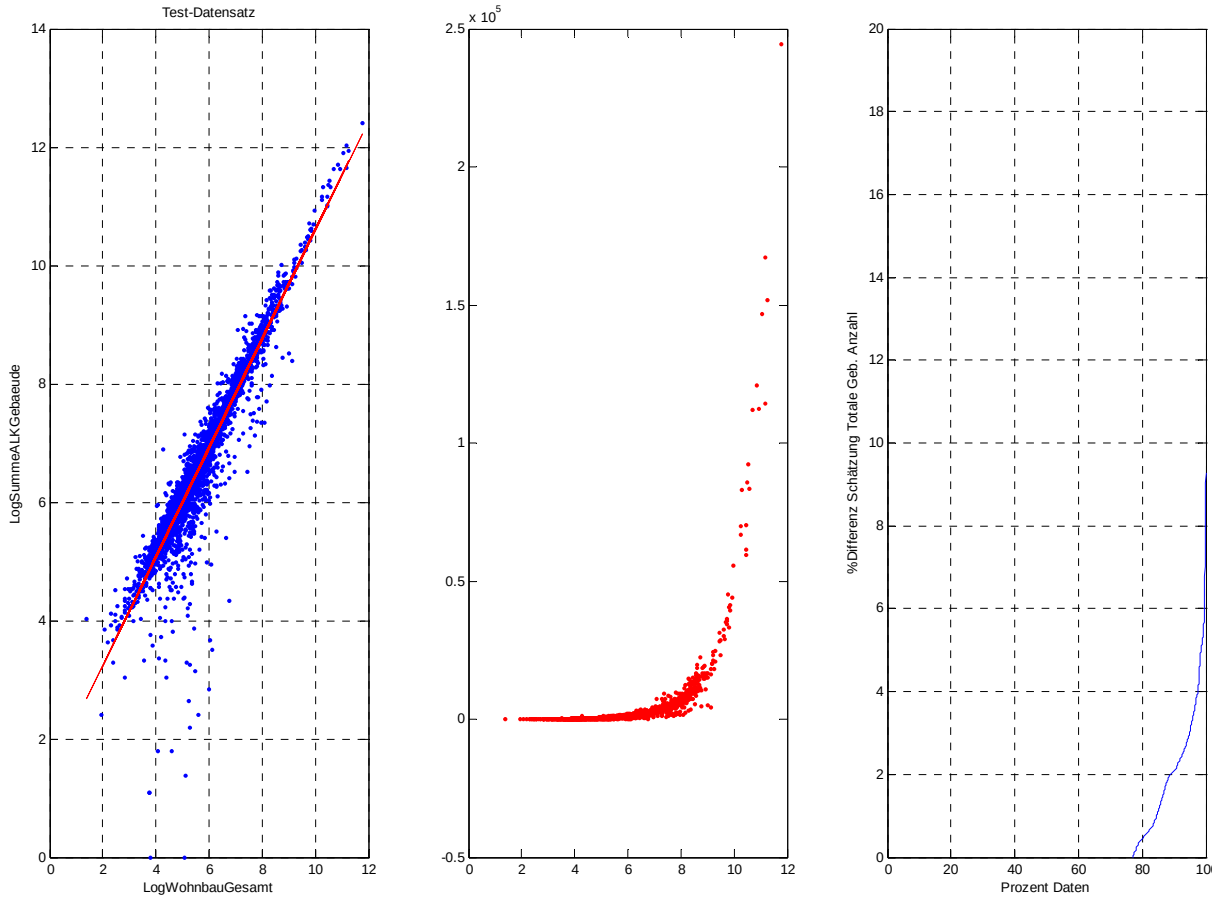


Abbildung 8-9: Qualitätskontrolle der Schätzergebnisse anhand von zufälligen Testdatensätzen[618]

618 In Nebtenteil B sind zu jedem Lern- und Testdatensätzen die Überprüfungsergebnisse hinterlegt.

Die Qualitätskontrolle der Schätzergebnisse durch Testdatensätze hat bereits angedeutet, dass die Prognose des vorhandenen Gesamtgebäudebestandes auf Gemeindeebene gerade für kleinere Gemeinden sehr genau ist und damit der Gebäudebestand für die Mehrzahl der deutschen Gemeinden relativ zuverlässig prognostiziert werden kann. Erst mit zunehmender Größe der Gebäudebestände treten größere Abweichungen auf, und es bieten sich für zukünftige Optimierungsansätze an dieser Stelle weitere Verbesserungsmöglichkeiten.

Abbildung 8-10 verdeutlicht diesen Aspekt anhand des Schätzungsergebnisses für die 6560 bekannten Gemeinden, und es wird auch hier besonders deutlich, dass die Abweichung in großem Maße von der Größe des Gesamtbestandes im Wohnbau abhängt. Will man die Genauigkeit der Schätzung also erhöhen, so ist es wichtig, gerade diese Gemeinden zu identifizieren, die einen großen Wohnbaubestand aufweisen. An dieser Stelle ist dann zur Verbesserung des Schätzergebnisses nach einer Optimierungsmöglichkeit zu suchen. In der Rückschau auf die Untersuchungen des Informationsgehaltes anhand der Lorenzkurven sind wahrscheinlich gerade die 20 % der Gemeinden relevant, die 80 % des Gebäudebestandes umfassen. Es handelt sich infolge der PDE-Dichteschätzungen überwiegend um die Ober- und Mittelzentren (Tabelle 8-2), die über große Wohnbaubestände verfügen.

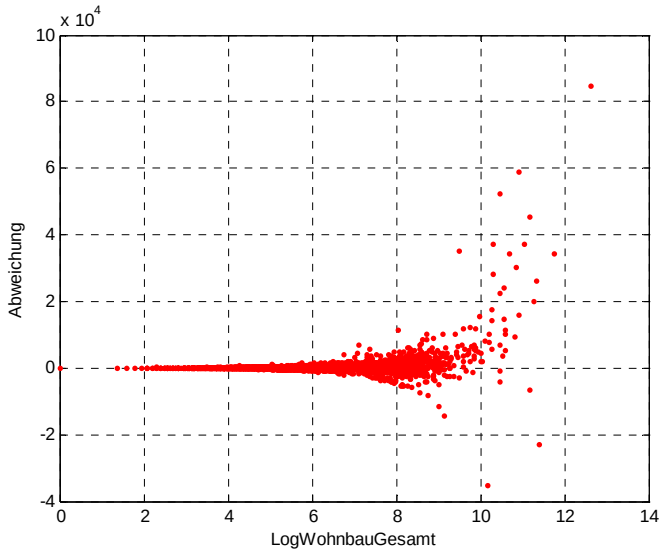


Abbildung 8-10: Abweichung des Schätzergebnisses in Abhängigkeit von ‚LogWohnbauGesamt'

Der hier vorgestellte Schätzansatz basiert auf Daten aus der automatisierten Liegenschaftskarte, die momentan leider noch nicht in allen Bundesländern zeitgleich und vollständig inhaltlich vergleichbar umgesetzt ist. Zukünftig lässt sich dieser Schätzansatz bei besserer Datenlage hinsichtlich seiner Genauigkeit weiter überprüfen, indem eine Stichprobe mit zufälliger Gemeindeauswahl gezogen werden kann und sich die Beschaffung der Ausgangsdaten nicht mehr nur an der tatsächlichen Datenverfügbarkeit orientieren muss.

8.4 Fazit

Die gesellschaftliche Bedeutung des Gebäudebestandes kann nicht genug betont werden. Der Gebäudebestand ist ein bedeutender Wirtschaftsfaktor, und eine bessere Kenntnis über seine Nutzungsstrukturen, Bauzustände, Abrissraten und Energiekosten sind für die Volkswirtschaft von Interesse. Festzustellen ist, dass mit Ausnahme einiger Basisdaten zum Wohnungsbau, den öffentlichen Bauten und Denkmälern der Gesamtbestand an Hochbauten bisher nicht detailliert in amtlichen Statistiken erfasst ist. Dies ist nicht nur ein nationales, sondern ebenso ein europäisches Problem, dessen Lösung noch aussteht.

Vor diesem Hintergrund ist in diesem Kapitel ein gemeindescharfer Schätzansatz für den deutschen Gebäudebestand erarbeitet worden. Für 6560 der 12430 Gemeinden in Deutschland konnten Daten zum Gesamtbestand aus der Automatisierten Liegenschaftskarte erhoben werden. Grundvoraussetzung zum Aufbau der Eingangsdaten für die Schätzung war eine langfristige Planung und sehr zeitintensive Aufbereitung der erfassten Daten. Darüber hinaus wurden 130 Variablen aus den allgemein zugänglichen Daten der amtlichen Statistik erzeugt.

Infolge der Verteilungsuntersuchung und Datentransformation (Logarithmierung) konnte mit Hilfe von Korrelationsrechnungen und Scatter-Dichte-Plots ein linearer Zusammenhang zwischen dem Wohnbaubestand und dem Gesamtbestand an Gebäuden erkannt werden. Mit Hilfe einer Regressionsrechnung wurde daraufhin eine gemeindescharfe Prognose möglich. Mittels Lern- und Testdatensätzen wurde das Ergebnis in seiner Qualität überprüft. Um den Gesamtbestand nochmals zu kontrollieren, wurde der gleichermaßen bestehende log-lineare Zusammenhang zwischen Bevölkerung und Gesamtbestand an Gebäuden herangezogen. Auch diese Regressionsrechnung bestätigte den zuvor prognostizierten Gesamtbestand.

Aus diesem Schätzansatz lassen sich drei wesentliche Erkenntnisse gewinnen:

1. In Deutschland existieren ungefähr 38 Millionen Gebäude.
2. 20 % der Gemeinden enthalten ca. 80 % des gesamten Gebäudebestandes.
3. Gerade diejenigen Gemeinden mit einem großen Bestand an Wohngebäuden führen zu den größten Abweichungen bei den prognostizierten Bestandsgrößen.

Zukünftig ist darüber zu reflektieren, inwieweit zum Schließen von Datenlücken ggf. Schätzansätze dieser Art einsetzbar sein könnten. Beispielsweise lassen sich sicherlich Gebäude aus besonders relevanten Nutzungsklassen oder insbesondere historische Gebäudebestände in ähnlicher Weise prognostizieren.

9 Resümee

Durch den Fortschritt in der Informationstechnologie und das immer rapidere Anwachsen von Datenmengen sind im letzten Jahrzehnt die Anforderungen an Systeme gestiegen, die Wissen aus Daten extrahieren und abbilden. In der Zukunft werden Daten und Informationen zwar erwartungsgemäß im Überfluss verfügbar sein, jedoch wird die Einbindung in Erfahrungszusammenhänge, durch welche erst Wissen geschaffen wird, schwieriger.

Der Begriff des ‚Urban Data Mining' wurde definiert, um für den urbanen Kontext eine Methodik zur Problemlösung vorzustellen, die dazu geeignet ist, logische oder mathematische und zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen zu entdecken. Als Forschungsgegenstand bietet die Raum- und Stadtstruktur für verschiedene wissenschaftliche Disziplinen interessante Arbeitsfelder, und der Erkenntnisgewinn erfolgt zunehmend durch interdisziplinäre Ansätze. Um einen strukturierten und methodenorientierten Handlungsrahmen für die Stadt- und Regionalforschung zu erarbeiten, war ein umfangreiches Studium der Verfahren des Data Mining erforderlich. Dabei wurde im Vergleich zu anderen fachspezifischen Arbeiten besonders Wert darauf gelegt, auch die in jüngster Zeit entwickelten Verfahren sowie Arbeitstechniken der Data Miner zu integrieren.

Als grundlegende Verfahrensschritte des ‚Urban Data Mining' sind die Dateninspektion, die Strukturerkennung, die Strukturbildung, die Strukturprüfung, die Operationalisierung und die Wissenskonversion anzusehen. Es wurde verstärkt auf die Bedeutung der Datenaufbereitung als Teilschritt der Dateninspektion in dieser Arbeit hingewiesen.

Dargestellt wurden Methoden, die Ähnlichkeiten zwischen Untersuchungsobjekten feststellen und solche Methoden, die geeignet sind, daraus Erkenntnisse abzuleiten und zu bewerten. Die Klassifikation ist dabei ein wichtiges Instrument des Data Mining, das sich zur Entwicklung von Maßstäben und Bewertungsskalen in Bezug auf urbane Phänomene eignet. Es werden induktive Verallgemeinerungen über die Untersuchungsobjekte vorgenommen, indem ein gemeinsamer Begriff (Semantik) gefunden wird. Klassifikatoren eignen sich dazu, Klassifikationen nachzuvollziehen und die Zuordnung bislang nicht klassifizierter Objekte zu ermöglichen.

Data Mining beschreibt hier einen zyklischen Prozess, wobei die in jedem Schritt gewonnenen Erkenntnisse wieder validiert und als Eingangsstufe eines nachfolgenden Schrittes zu verstehen sind.

Die Regionalisierung dient der gedanklichen Ordnung der Vielfalt der Realität und ermöglicht die Erfassung und Charakterisierung der räumlichen Struktur- und Interaktionsmuster. Als traditionelles Anliegen der Geografie kann sie zur räumlichen Theoriebildung herangezogen werden aber auch zur Verwerfung bereits existierender Hypothesen oder Theorien einen entscheidenden Beitrag leisten. Die Verfahren des Data Mining tragen zum automatischen Erzeugen und Prüfen von Hypothesen und Modellen bei. Der Begriff der Wissensentdeckung (Knowledge Discovery) charakterisiert die Aufdeckung bzw. Extraktion von neuen bzw. verborgenen Zusammenhängen in Daten, aber insbesondere die Wissensgenerierung und (natürlich-) sprachliche Darstellung von Wissen aus Datensammlungen. Zukünftig ist verstärkt die maschinelle Verarbeitbarkeit des gewonnenen Wissens zu berücksichtigen, die in Form von wissensbasierten Systemen umgesetzt wird. Gegenwärtig ist das Methodenrepertoire für die Stadt- und Regionalplanung jedoch erst noch im Aufbau begriffen.

Der konzeptionelle Ansatz des ‚Urban Data Mining' sah neben einer umfangreichen Recherche nach einsetzbaren Methoden auch eine zeitintensive Auseinandersetzung mit bestehenden Datenquellen vor. Es wurde das Ziel verfolgt, die Möglichkeiten der Datenbeschaffung auf Gemeinde- und Kreisebene weitgehend transparent darzustellen, um dadurch zukünftigen Forschungsansätzen im urbanen Kontext die bestehenden Möglichkeiten der Datenbeschaffung zu erleichtern und einen schnelleren Einstieg zu ermöglichen. Neben den sogenannten üblichen Daten aus der amtlichen Statistik werden Daten aus der Automatisierten Liegenschaftskarte und Maßzahlen der Geocomputation in ihren Besonderheiten dargestellt. Zum Aufbau einer zuverlässigen digitalen Datenbasis war dabei der Einsatz von GIS zu Kontrollzwecken unverzichtbar.

Es besteht aktuell noch die Herausforderung, das erwünschte Datenangebot bei zunehmender räumlicher Auflösung in großem Umfang zu beschaffen. Zusätzlich ist für die zeitlich statistische Vergleichbarkeit eine möglichst konstante Gebietsstandsregelung wünschenswert. Aus diesem Grunde wurden Strategiekonzepte entwickelt, die auch eine spätere Beurteilung von Untersuchungsobjekten ermöglichen und in der Lage sind, Ähnlichkeitsmuster genauer zu definieren oder wiederzuerkennen. Ein strategischer Orientierungspfad zeigt, wie hochdimensionale Untersuchungsobjekte schrittweise beurteilt werden können und eine Strukturerkennung und Strukturbildung von Ähnlichkeitsmustern operationalisierbar wird. Mit Hilfe der Operationalisierung lassen sich bereits gewonnene Erkenntnisse zukünftig über untersuchte räumliche Objekte in eine Untersuchung von tiefer oder höher aggregierten Objekten einbeziehen.

Anhand von vier sogenannten Untersuchungsaufgaben wurden die Anwendungsmöglichkeiten von Verfahren des ‚Urban Data Mining' gezeigt. Es handelte sich um regionsvergleichende Untersuchungen nach raumstrukturellen Ausprägungen. Das Ziel bestand darin, auf möglichst hoher räumlicher Auflösung das Gemeindesystem zu untersuchen. Betrachtet wurden einerseits statische Eigenschaften zur Raumstruktur und andererseits jene, die Aussagen über regionale Enwicklungstendenzen ermöglichten. Darüber hinaus wurden methodische Möglichkeiten gesucht, um räumliche Informationen zwischen verschiedenen Objekten zu übertragen. Die gemeindescharfe Schätzung des deutschen Gebäudebestandes wurde dabei als besondere Herausforderung angesehen, da bisher keine amtlichen Statistiken in Deutschland vorliegen, die den Gesamtbestand oder den Nichtwohnbaubestand auf Gemeindeebene ausweisen.

Die erste Untersuchungsaufgabe ‚Polyzentralität und räumliche Vielfalt' enthielt eine deskriptive Erfassung von räumlichen Struktureigenschaften unter Einsatz von sogenannten Gauß-Mixtur-Modellen. Ermittelt wurden die Beschäftigungsschwerpunkte bzw. die Diversität der 12430 Gemeinden in Deutschland. Abbildung 9-1 vermittelt einen Eindruck von den Ergebnissen dieses Ansatzes. Dargestellt sind die Beschäftigungsschwerpunkte von Gemeinden nach Wirtschaftssektoren und die Diversität der Gemeinden in Bezug auf die Anzahl der Beschäftigungsschwerpunkte nach Wirtschaftszweigen. Auch der Zwischen- und der Peripherraum zeigen Beschäftigungsschwerpunkte im Tertiären Sektor.

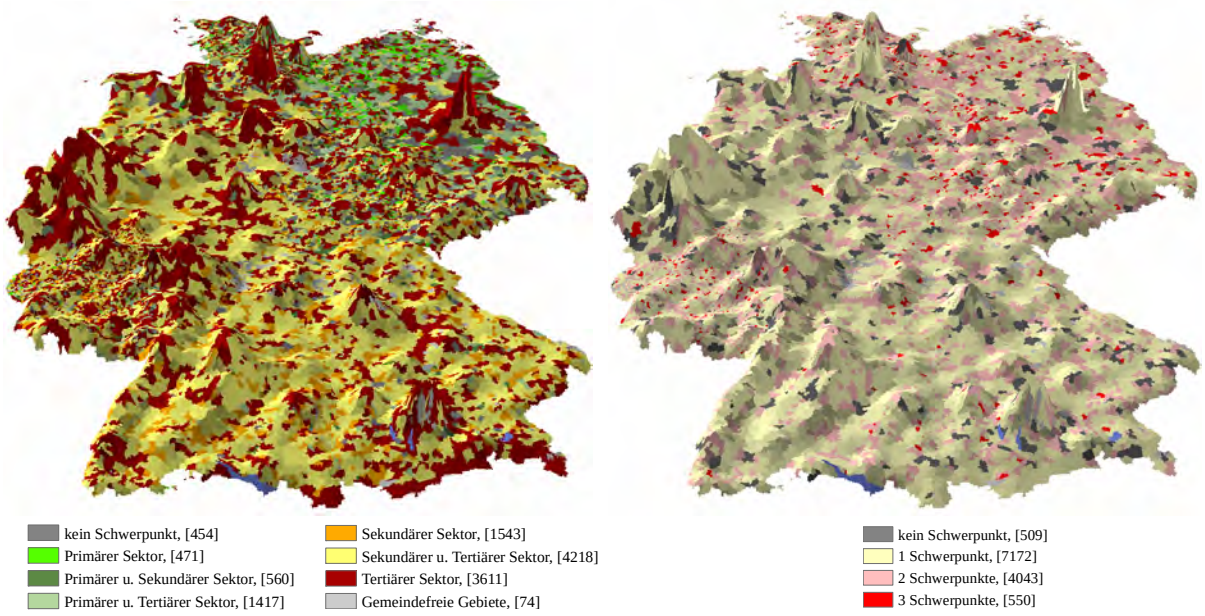


Abbildung 9-1: Überlagerung von Beschäftigungsschwerpunkten und der Gebäudedichte[619]

619 Siehe ausführliche Ergebnisbeschreibung im Abschnitt 5.2.1. Bei diesen Darstellungen ist zusätzlich das gemeindescharfe Schätzergebnis zum deutschen Gebäudebestand unterlegt. Große Höhenwerte sind Ausdruck einer größeren Gebäudedichte mit Bezug auf die Katasterfläche der Gemeinde.

Vor dem Hintergrund einer sachlich-räumlichen Differenzierung des deutschen Gemeindesystems wurden im Spannungsfeld von ‚Polyzentralität und räumlicher Vielfalt' zusätzlich ausgewählte raumstrukturelle Kenngrößen einer Einzeluntersuchung unterzogen. Geprüft wurde die Möglichkeit der Klassenbildung mit sechs Messgrößen. Gerade aufgrund der Verteilungsuntersuchung und einer Modellierung von Verteilungsverläufen mit Hilfe von Gauß-Mixtur-Modellen konnte gezeigt werden, dass es möglich ist, mit drei der sechs Variablen eine eindeutige Klassifizierung aufzubauen. Die Anwendung eines Clusteralgorithmus hätte nicht zu einer klar unterscheidbaren Gruppenstruktur geführt. Abbildung 9-2 skizziert das Ergebnis in anderer Darstellung, welches sich eignet, Agglomerations- und Verdichtungseigenschaften der Gemeinden zu erfassen und zu strukturieren. Die Suburbanisierung hat dazu geführt, dass sich in Teilbereichen flächenhaft bebaute Stadtregionen entwickelt haben, die sich weit in das Umland ausdehnen. Es konnte gezeigt werden, dass sich die aktuelle Debatte über das Konzept der Zentralen Orte gerade mit den polyzentrischen Kooperationsräumen (‚zentralörtliche Kooperationsräume') auseinandersetzt, die nicht von klassischen Zentrale-Orte-Kategorien repräsentiert werden.

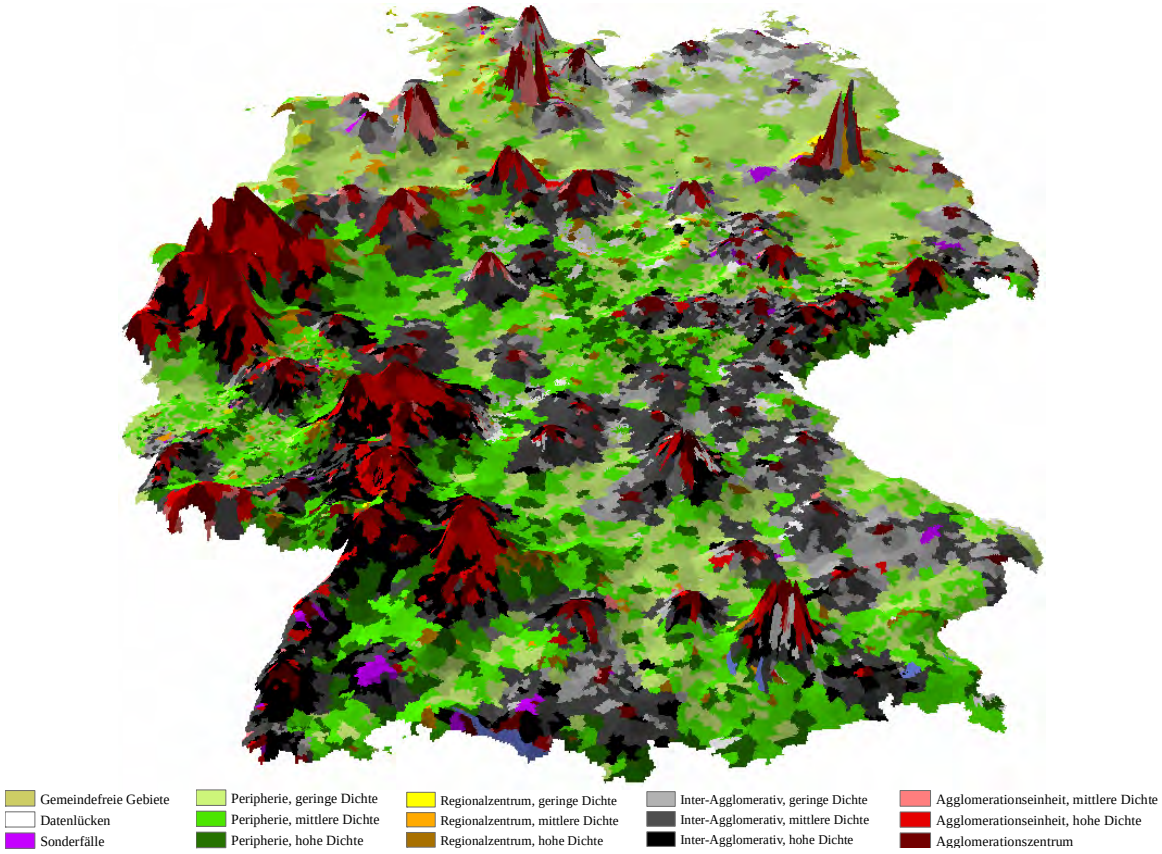


Abbildung 9-2: Überlagerung von Verstädterungseigenschaften und der Gebäudedichte[620]

620 Siehe ausführliche Ergebnisbeschreibung im Abschnitt 5.2.3. Bei diesen Darstellungen ist zusätzlich das gemeindescharfe Schätzergebnis zum deutschen Gebäudebestand unterlegt. Große Höhenwerte sind Ausdruck einer größeren Gebäudedichte mit Bezug auf die Katasterfläche der Gemeinde.

Die zweite Untersuchungsaufgabe ‚Schrumpfung, Wachstum oder Stagnation' diente als Ausgangspunkt um die Techniken des Visual Mining vorzustellen und die räumlichen Entwicklungstendenzen zu erfassen. Geklärt wurde die Frage, ob eine Klassenbildung sich eignet, dynamische Prozesse bzw. die Bipolarität von Schrumpfungs- oder Wachstumstendenzen in der jüngeren Vergangenheit abzubilden. Zur Strukturbildung wurden zwei verschiedene Untersuchungsansätze verfolgt, die sich sogar im Ergebnis bestätigt haben. Einerseits wurde zur Beschreibung des deutschen Gemeindesystems ein auf Entscheidungsregeln basierter Ansatz dargestellt. Andererseits sind sogenannte emergente selbstorganisierende Merkmalskarten in den urbanen Kontext überführt und nach Kenntnis des Verfassers erstmalig für eine Forschungsfrage in der Raumordnung angewendet worden. Das relativ einfach auf vier Variablen aufgebaute Messmodell diente zusätzlich als Arbeitsgrundlage, um in der dritten Untersuchungsaufgabe ‚Wissensbasierte Systeme in Stadt und Regionalplanung' das Funktionsprinzip eines Klassifikators kanonisch darzustellen.

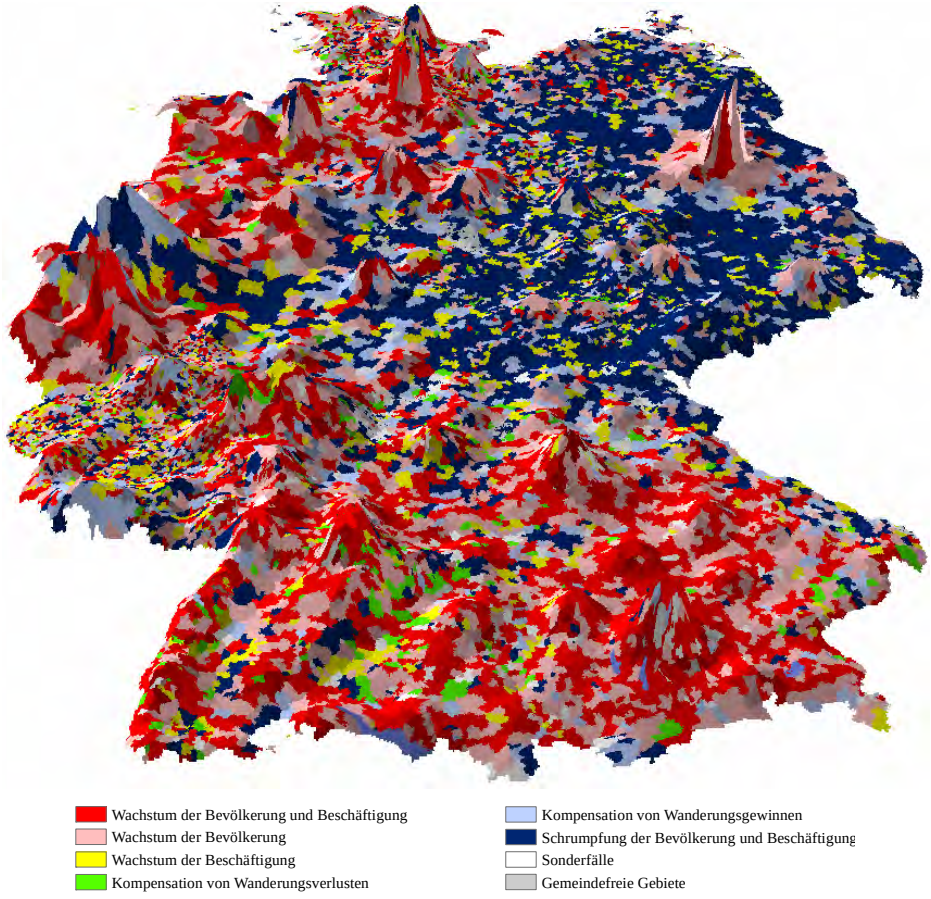


Abbildung 9-3: Überlagerung von Wachstums- und Schrumpfungstendenzen und der Gebäudedichte[621]

621 Siehe ausführliche Ergebnisbeschreibung im Abschnitt 6. Bei diesen Darstellungen ist zusätzlich das gemeindescharfe Schätzergebnis zum deutschen Gebäudebestand unterlegt. Große Höhenwerte sind Ausdruck einer größeren Gebäudedichte mit Bezug auf die Katasterfläche der Gemeinde.

Es wurde auf der Grundlage von zusätzlich ausgewählten Variablen nach einer Zuordnungsvorschrift gesucht, welche sich eignet, eine Gemeinde in ein bestehendes Klassensystem nachträglich einzuordnen. Der bisher aus rein kanonischen Zwecken aufgebaute k-Nearest-Neighbour-Klassifikator zeigte bereits eine Zuordnungsgenauigkeit, die weit über einer zufälligen Einordnung liegt.

Die vierte Untersuchungsaufgabe bezog sich auf den ‚Deutschen Gebäudebestand'. Es sollte ein Schätzverfahren zur Übertragung räumlicher Information erarbeitet werden, welches sich eignet den Deutschen Gebäudebestand erstmalig in seiner Gesamtgebäudezahl gemeindescharf zu prognostizieren. Gezeigt wurde, dass es methodisch möglich ist, anhand von Untersuchungsobjekten mit einem bekannten Gebäudebestand auf weitere Objekte ohne bekannten Gebäudebestand zu schließen. Zunächst wurde das Datenmaterial der 6560 Gemeinden mit bekanntem Gesamtbestand mit Hilfe eines graphischen Ansatzes untersucht. Dieser basiert auf der theoretischen Begründung der Pareto-80 / 20 Regel, und es konnte darauf verwiesen werden, dass nahezu 20 % der Gemeinden 80 % des Gesamtbestandes enthalten. Um für die 5870 Gemeinden ohne Gesamtbestand eine zuverlässige Prognose über den Gesamtbestand gemeindescharf zu stellen, wurden die als relevant erachteten Schätzgrößen aus einer im Rahmen dieser Arbeit erarbeiteten Datenbasis extrahiert (Gesamtumfang ca 130 Variablen). Gerade durch die Datenvorverarbeitung ist es möglich gewesen einen deutlichen Zusammenhang zwischen den durch Logarithmierung transformierten ausgewählten Schätzgrößen (Wohnbaubestand sowie Bevölkerung) und dem Gesamtbestand einer Gemeinde zu erkennen. Die lineare Regressionsrechnung wurde verwendet, um den Gesamtbestand auf ungefähr 38 Millionen Gebäude in Deutschland beziffern zu können. Das Schätzergebnis wurde in seiner Qualität geprüft und für die Mehrzahl der Gemeinden ist ein relativ geringer prozentualer Schätzfehler zu bennenen. Zusätzlich wurde erkannt, dass gerade diejenigen Gemeinden genauer untersucht werden sollten, die über eine große Anzahl an Wohngebäuden verfügen.

Abschließend sei darauf verwiesen, dass das Ziel dieser Arbeit darin bestand, auf der Grundlage von bestehenden Methoden des Data Mining und der Knowledge Discovery ein für die Stadt- und Regionalforschung strukturiertes methodisches Arbeitskonzept zu entwickeln. Dieses wurde letztendlich durch den Begriff des ‚Urban Data Mining' charakterisiert. Im theoretischen Teil dieser Arbeit wurden hierzu die als besonders relevant erachteten Verfahren vorgestellt und im empirischen Teil an Untersuchungsaufgaben angewendet.

10 Ausblick

10.1 Zukünftige Bearbeitungsmöglichkeiten

Die geschilderten Ergebnisse belegen neben der kritischen Bestandsaufnahme und Diskussion von räumlichen Eigenschaften und Entwicklungstendenzen insbesondere eine Ausseinandersetzung mit methodischen Vorgehensweisen des DATA MINING und des KNOWLEDGE DISCOVERY. Auf der Grundlage dieses explizit für den urbanen Kontext ausgewählten Methodenrepertoires kann in Zukunft ggf. in einem größeren Arbeitskreis ein umfassendes Regelwerk aufgebaut werden, welches zur dauerhaften Bewertung von räumlichen Strukturen geeignet wäre. Die Entwicklung von Erklärungs- und Messmodellen wird auch eine grundlegende Interpretation von Gebäude- und Infrastrukturen verstärkt fördern. Den in der Stadt- und Regionalforschung erst im Aufbau begriffenen wissensbasierten Systemen ist zukünftig eine wesentlich größere Bedeutung beizumessen.

Verwiesen sei im methodischen Sinne in diesem Zusammenhang nochmals auf Techniken der Klassenerklärung (Operationalisierung), die eine Klassifizierung mit weiteren Eigenschaften belegen und so das Verständnis für eine Klasse insgesamt vergrößern. Genannt sei an dieser Stelle der erst kürzlich entwickelte Algorithmus U-KNOW (ULTSCH, 2007), der eine Ergänzung zu dem in dieser Arbeit vorgestellten SIG*-Algorithmus bietet. Darüber hinaus ist gerade der Aufbau von sogenannten symbolischen Klassifikatoren anzustreben. Diese berücksichtigen z.B. Entscheidungsbäume, Entscheidungsregeln oder statistische Wahrscheinlichkeiten, wie sie die häufig verwendeten Bayes'schen Klassifikatoren berechnen. Das den Bayes'schen Klassifikatoren zugrunde gelegte Modell geht von einer Datengenerierung durch einen Zufallsprozess aus und wird spezifiziert als eine Verteilung mit einer zentralen Lage und gewissen Streuparametern.

Unter dem Aspekt einer zunehmend feststellbaren Verfügbarkeit von raumbezogenen digitalen Daten und leistungsstarker GI-Systeme ist damit zu rechnen, dass sich weitere umfassende Lösungsmöglichkeiten für die räumliche Analyse in den nächsten Jahren ergeben werden. Das Datenmaterial muss nicht mehr auf nur administrative Einheiten aggregiert werden, so dass im Sinne der geographischen Abbildung eine genauere Wiedergabe der tatsächlich vorhandenen räumlichen Situation vorstellbar ist. Zukünftig ist bei ausreichend verfügbarem Datenmaterial auch eine Einbeziehung von grenznahen Gemeinden im benachbarten Ausland als sinnvoll zu erachten, um gegenseitige Grenzbeziehungen bzw. regional spezifische Einflüsse und Wirkungsbeziehungen in Europa aufzudecken.

In naher Zukunft ist damit zu rechnen, dass durch die ALKIS-Einführung (siehe Abschnitt 3.6.1) auch eine zeitliche und räumliche Vergleichbarkeit von Gebäudenutzungsdaten auf Gemeindeebene nach ersten denkbaren Umstellungsschwierigkeiten in Deutschland möglich sein wird. Dadurch kann das Wissen auch über mehrdimensionale Ähnlichkeiten von Gemeinden in Bezug auf den in der Regel über Jahrhunderte gewachsenen Gebäudebestand deutlich vergrößert werden. Die in dieser Arbeit bereits geschilderten Prognoseansätze bzw. Schätzverfahren können zukünftig die Verallgemeinerung von weiteren Erkenntnissen über den Deutschen Gebäudebestand unterstützen.

Abschließend werden die städtebauliche Strukturtypeneinteilung des IÖR und der Raumstrukturtypenansatz des BBR zusätzlich in verkürzter Form aufgeführt, um für einige Verfahren dieser Arbeit mögliche Anknüpfungspunkte zu nennen.

10.2 Integrationsmöglichkeiten in bestehende Stadt- und Raumstrukturtypenansätze

10.2.1 Städtebauliche Strukturtypeneinteilung des IÖR

Die räumliche Struktur wird oftmals durch vorhandene Bestände an Arbeitskräften, Kapital, verfügbaren Flächen sowie Gebäuden oder vorhandenen Rohstoffen beeinflusst und unterliegt sich überlagernden Prozessen. Im Zeitablauf können sich räumliche Strukturen verändern und fordern Anpassungsreaktionen, so dass jede gegenwärtige Entscheidung auch von den in der Vergangenheit geschaffenen Voraussetzungen beeinflusst und begrenzt wird. Eine einmal entstandene Struktur ist als Ausgangssituation für die darauffolgende Periode anzusehen. In einem grundlegenden Zusammenhang zu Phänomenen städtischer und ländlicher Raumstrukturen stehen die Stadtstrukturen, wobei eine bessere Kenntnis ihrer Eigenschaften das Gesamtverständnis von Wirkungszusammenhängen vergrößert.

Um homogene städtebauliche Flächeneinheiten abzugrenzen und damit in Zusammenhang stehende Kenngrößen zu quantifizieren, wurde am IÖR, Dresden[622] ein Strukturtypenansatz entwickelt. Dieser ermöglicht nach HEBER / LEHMANN[623] eine räumliche Stadtgliederung in Strukturtypenflächen und ist von ARLT[624] im Jahr 2001 qualifiziert worden. Auf Basis von topografischen Karten werden zunächst strukturbestimmende Merkmale gemäß der folgenden Tabelle 10-1 ausgewählt. Spezifische Gebäudeformen, Gebäudeabstände und Gebäudeanordnungen lassen zusätzlich Rückschlüsse auf die Geschosszahl und Bauzeit zu.

[622] Leibniz-Institut für ökologische Raumentwicklung e.V. (IÖR), http://www.ioer.de (Stand: 07.01.2007)
[623] Vgl. HEBER / LEHMANN [1996]
[624] Vgl. ARLT et al. [2001]

Die Städte und Stadtregionen werden anhand der Merkmale in neun städtebauliche Strukturtypflächen mit weitgehend homogener Ausprägung unterteilt: Verdichtet geschlossen bebaute Flächen (1), Geschlossen bebaute Flächen (2), Offen bebaute Flächen (3), Aufgelockert offen bebaute Flächen (4), Unbebaute Flächen im Siedlungsraum (5), Freiraumfläche (6), Fließgewässer (7), Stehende Gewässer (8) und Meeresgewässer (9).

Strukturbestimmendes Merkmal	**Darstellungsart in topografischen Karten**
Bauform, Bebauungsstruktur	Einzelhaus, Gebäudegruppe
Bebauungsdichte	Gebäudeabstand, Grundstücksgröße
Lage in der Stadt	Rand- oder innerstädtisch
Erschließung	Straßen, Gleistrassen, Straßenklassifizierung
Topographie	Höhenlinie, Höhenpunkt, Böschung
Nutzung	Erläuterung in der Kartenlegende
Geschosszahl und Bauzeit bzw. Baualter	Typische Gebäudeanordnung

Tabelle 10-1: Bestimmende Merkmale der städtebaulichen Strukturtypeneinteilung des IÖR[625]

THINH[626] formuliert den Strukturtypenansatz in allgemeiner Form wie folgt: „Untersucht man eine Eigenschaft eines Objektes als ein System aus vielen Elementen, so kann man zunächst die Elemente anhand von ausgewählten Strukturmerkmalen so in Klassen aufteilen, dass der Unterschied zwischen zwei beliebigen Klassen und die Ähnlichkeit zwischen den Elementen jeder Klasse möglichst groß ausfallen (Clusteranalyse). Danach wird die Eigenschaft in den einzelnen Klassen untersucht und auf das Gesamtsystem übertragen."

Der Strukturtypenansatz in allgemeiner Form entspricht der Vorgehensweise dieser Arbeit um Ähnlichkeitsmuster von Gemeinden zu ermitteln, wobei nicht immer das Verfahren der Clusteranalyse in dieser Arbeit zur Strukturbildung (siehe Abschnitt 0) verwendet wird. Es besteht in Zukunft die Möglichkeit, die Einsatzfelder der in dieser Arbeit enthaltenen Klassifikationen zu prüfen. Durch Einsatz von Geodaten (z.B. Automatisierte Liegenschaftskarte oder Topographischen Karten) ist der Frage nachzugehen, ob andere Gemeinden der gleichen Klasse die gewonnenen Ergebnisse einer Einzelfalluntersuchung auf Grundlage z.B. der ALK in ähnlicher Weise bestätigen können. Die Zielsetzung könnte darin bestehen, ohne umfassende Einsicht in die Topografischen Karten diejenigen Gemeinden zu suchen, welche annähernd über die gleichen städtebaulichen Strukturtypen verfügen.

Durch Ziehung einer Stichprobe von Elementen aus einer bereits bekannten Klasse wird dabei sicherlich der Untersuchungsumfang reduziert.

---

625 Eigene Bearbeitung, vgl. THINH [2004 a, S. 13]

626 Vgl. THINH [2004 a, S. 12 / 13]: Elemente sind mit der in dieser Arbeit verwendeten Definition des Objektes gleichzusetzen und der von THINH verwendete Begriff ‚Objekt' entspricht einem Untersuchungsgebiet, wie z.B. eines Bundeslandes oder eines Kreisgebietes mit einer Anzahl von Untersuchungsgemeinden.

10.2.2 Raumstrukturtypenansatz des BBR

Wie bereits dargestellt, ermöglichen Raumgliederungen den Aufbau eines Untersuchungsrasters, um die Raum- und Siedlungsstruktur bewerten und beschreiben zu können. Als politikberatende Forschungseinrichtung des Bundes befasst sich das Bundesamt für Bauwesen und Raumordnung (BBR)[627] mit vielfältigen Fragestellungen der Raumentwicklung und verfolgt dabei in der Regel flächendeckende empirische Analysen innerhalb des Bundesgebietes. Der europäische Kontext erfährt zunehmend eine größere Integration.

Von Bedeutung für eine qualifizierte Raumanalyse ist die Berücksichtigung unterschiedlicher Raumstrukturen, um raumordnungspolitische Wertungen und Schlussfolgerungen räumlich differenziert zu treffen. Üblicherweise basieren Gebietstypisierungen in ihrer Abgrenzung auf administrativen Gebietseinheiten (Regionen, Kreise, Gemeinden) und anhand von regionalstatistischen Daten werden siedlungsstrukturell differenzierte Aussagen getroffen (z. B. ‚Agglomerationsraum', ‚Verstädterter Raum' und ‚Ländlicher Raum').

Im Jahr 2005 hat das BBR eine Raumstrukturgliederung entwickelt, welche eine Zusammenfassung von vergleichbaren Raumstrukturtypen auf Basis der Bevölkerungsdichte[628] und Zentrenerreichbarkeit[629] ermöglicht. Mit Hilfe von Geodaten zur Ausdehnung von Ortslagen aus dem Basis-Landschaftsmodell erfolgt eine Umschätzung regionalstatistischer Daten. Die Raumstrukturgliederung des BBR findet eine von Verwaltungsgrenzen weitgehend unabhängige Einteilung, die zunächst nach dem Kriterium der Zentrenerreichbarkeit zwischen ‚Zentralraum', ‚Zwischenraum' und ‚Peripherraum' unterscheidet. Danach erfolgt eine Untergliederung der gefundenen Grundtypen nach spezifischen Bevölkerungsdichten in weitere sechs Typen.

Durch Überlagerung von Oberflächendarstellungen zur Zentrenerreichbarkeit und Bevölkerungsdichte in einem GIS wird eine zonale Raumabgrenzung erzeugt. Es handelt sich um einen problemorientierten Ansatz zur Raumstrukturtypisierung, der die Anforderung erfüllt, die Abstraktion von regional unterschiedlich strukturierten administrativen Einheiten sicherzustellen und insbesondere die aktuelle Raumstruktur möglichst geographisch genau abzubilden (Verwaltungsgrenzenabhängigkeit).

---

627 Bundesamt für Bauwesen und Raumordnung (BBR): http://www.bbr.bund.de (Stand 06.01.2007)

628 Die Bevölkerungsdaten werden unter Annahme einer Gleichverteilung der Bevölkerung in den Ortslagen innerhalb der Gemeinden bzw. Stadtteile/-bezirke auf die Ortslagenflächen umgeschätzt. Im Vergleich zum gängigen Nachweis der Bevölkerungsdichte bezogen auf das jeweilig gesamte Gemeindegebiet wird auf diese Weise eine Annäherung an die tatsächliche Verteilung erreicht.

629 Die Basis für die Ermittlung eines Indexwertes der Zentrenerreichbarkeit bildet die PKW-Fahrzeit zu hochrangigen Zentren. Den Berechnungen liegt ein gleichmäßig verteiltes Raster von ca. 20.000 Messpunkten zu Grunde. Das BBR verfügt über eine modellhafte Abbildung des gesamten europaweiten Straßennetzes.

Abbildung 10-1 gibt einen Überblick zu den Raumstrukturtypen. Extreme Strukturen von sehr dichten zentralen und sehr dünn besiedelten, peripheren Räumen werden durch Überlagerung betont und vermischte Strukturen in den Zwischenräumen stärker differenziert.

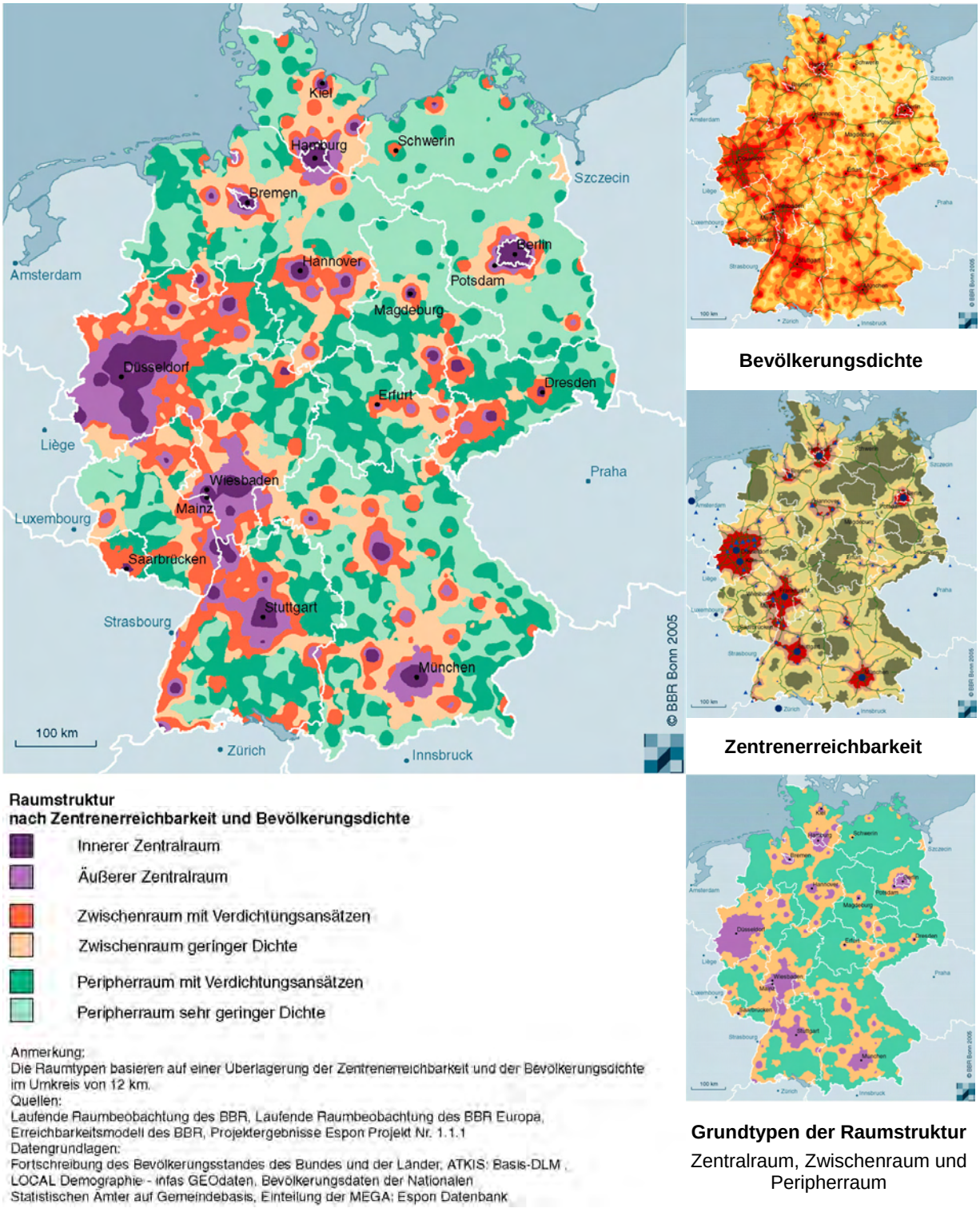


Abbildung 10-1: Raumstrukturtypen nach Zentrenerreichbarkeit und Bevölkerungsdichte (Genese)[630]

[630] Eigene Anordnung, Quelle: Bundesamt für Bauwesen und Raumordnung (BBR) – Einflussgrößen sind die Bevölkerungs- und Siedlungsentwicklung, zentralörtliche Strukturen sowie Erreichbarkeitsveränderungen.

Die Konzeption der Raumstrukturtypen umfasst eine Beschränkung auf wenige Basisindikatoren (Einfachheit), eine Erfassung der Raumstruktur in ihrer Entwicklung (Dynamik) und eine variable Anpassungsmöglichkeit je nach Fragestellung und Maßstabsebene für die der Typisierung zugrunde liegenden Schwellenwerte (Skalierbarkeit). Abbildung 10-2 zeigt eine analoge Umsetzung der Raumstrukturtypisierung auf europäischer Ebene.

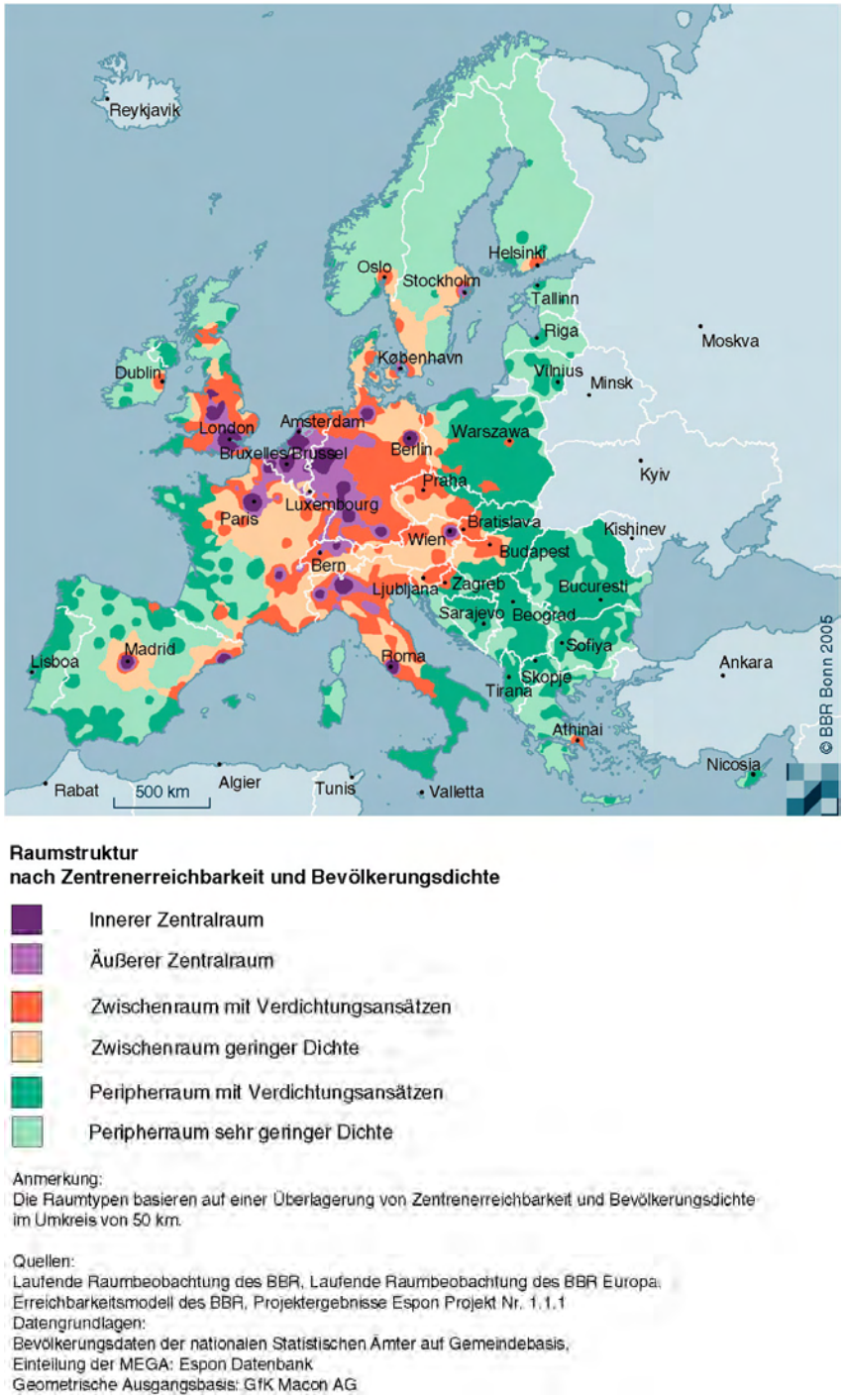


Abbildung 10-2: Raumstruktur Europa nach Zentrenerreichbarkeit und Bevölkerungsdichte[631]

631 Quelle: BUNDESAMT FÜR BAUWESEN UND RAUMORDNUNG (BBR) – Einflussgrößen sind die Bevölkerungs- und Siedlungsentwicklung, zentralörtliche Strukturen und Erreichbarkeitsveränderungen.

SCHÜRT[632] hebt hervor, dass die Raumstrukturtypen über die Zeit hinweg relativ beständig sind. Deshalb wird die Möglichkeit diskutiert, ausgehend von den Raumstrukturtypen eine Überlagerung mit zusätzlichem statistischem Datenmaterial zu verfolgen. Gezeigt wird diese Möglichkeit am Beispiel von Beschäftigungsdaten der Jahre 1997 und 2003 (siehe Abbildung 10-3). Es kann geprüft werden, inwieweit beobachtete dynamische Prozesse sich auf bestimmte Raumstrukturtypen konzentrieren oder ggf. auf andere Typen übergreifen.

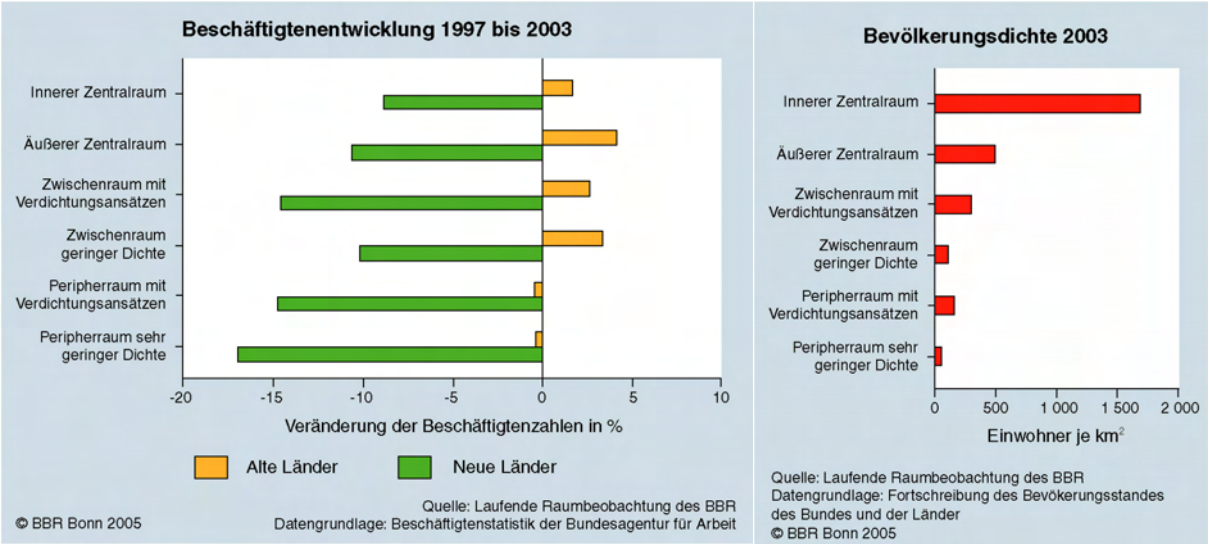


Abbildung 10-3: Überlagerung von statistischem Datenmaterial mit den Raumstrukturtypen des BBR[633]

Die hier kurz dargestellte Raumtypisierung des BBR basiert primär auf der Kopplung von Geo-Daten des Basis-Landschaftsmodells (Abschnitt 3.8) mit üblichem statistischem Datenmaterial und zusätzlichem Datenmaterial aus dem Erreichbarkeitsmodell des BBR.

Im Rahmen des Untersuchungsansatzes dieser Arbeit lagen aus Kostengründen gebietsscharfe Siedlungsdaten in einer vergleichbaren Genauigkeit nicht vor. Denkbar ist eine zukünftige Erweiterung des in dieser Arbeit dargelegten Forschungsansatzes durch verstärkte analytische Überlagerung von hier gefundenen Untersuchungsergebnissen mit den bereits existierenden Raumstrukturtypen. Die Methodik des ‚Urban Data Mining' und insbesondere die Verfahren zur Beschreibung von gefundenen Klassifikationen (Abschnitt 2.5) können ggf. weitere Eigenschaften den bereits bekannten Raumstrukturtypen mit einer ausdrückbaren Zuverlässigkeit zuweisen. In naher Zukunft ist damit zu rechnen, dass eine Umschätzung anderer statistischer Daten auf derartige Geodaten auch zu universitären Forschungszwecken flächendeckend möglich wird und eine bessere Anlehnung an die realen Siedlungs- und dazugehörigen Gebäudestrukturen stattfinden kann.

632 SCHÜRT et al. [2005]
633 Quelle: BUNDESAMT FÜR BAUWESEN UND RAUMORDNUNG (BBR)

Im Kontext der Raumstrukturtypen des Bundesamtes für Bauwesen und Raumordnung sei abschließend an dieser Stelle eine erste Möglichkeit der Ergebnisüberlagerung gezeigt. Die dreidimensionale Abbildung 10-4 entstand aus den in dieser Arbeit ermittelten Schätzwerten zum Deutschen Gebäudebestand (siehe Abschnitt 8). Große Höhenwerte sind Ausdruck einer größeren Gebäudedichte in einer Gemeinde, wobei als Bezugsgröße in diesem GIS-basierten Darstellungsansatz (siehe Abschnitt 3.1) die Gemeindefläche insgesamt eingesetzt wird. Die farbliche Skalierung repräsentiert die sechs zuvor beschriebenen Raumstrukturtypen. Aus der Visualisierung wird ein erster Hinweis darauf gegeben, dass die Gemeinden des Inneren und Äußeren Zentralraums nicht nur eine hohe Bevölkerungsdichte und gute Zentrenerreichbarkeit aufweisen, sondern zusätzlich über eine große Gebäudedichte verfügen. In den Peripherräumen mit geringer Dichte sind erwartungsgemäß auch geringere Gebäudedichten vorhanden. Auf Grundlage der vorgestellten Raumstrukturtypen, bietet sich in Zukunft die Möglichkeit an, einzelne Teilbestände des deutschen Gebäudebestandes genauer zu untersuchen und das Wissen über die Struktur und Dynamik des Bestandes zu vergößern.

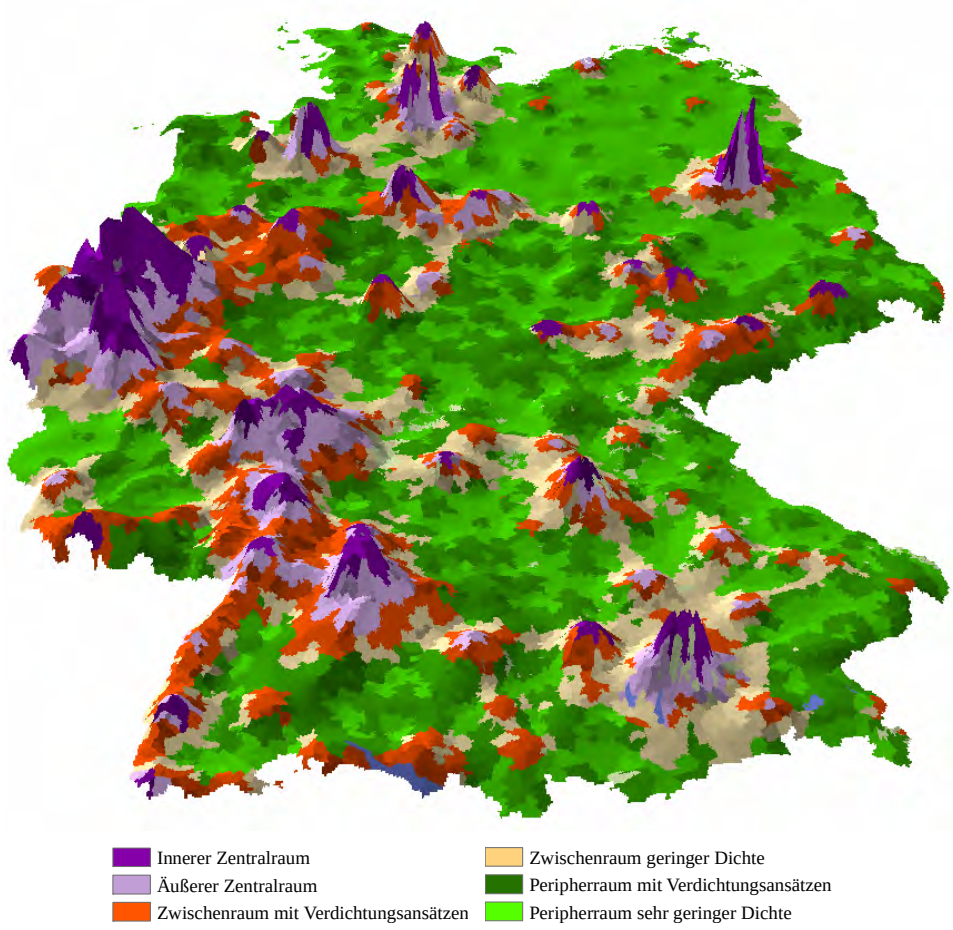


Abbildung 10-4: Überlagerung der bestehenden Raumstrukturtypen des BBR und der Gebäudedichte[634]

[634] Bei dieser Darstellung ist zu berücksichtigen, dass es sich um Schätzdaten des deutschen Gebäudebestandes handelt (siehe Verfahrensansatz im Abschnitt 8).

Thesen

1. **Die Klassifizierung ist ohne eine grundlegende Verteilungsuntersuchung der Meßgrößen und ohne die Verwendung von strukturerkennenden Verfahren kaum vorstellbar.**

2. **Durch Einsatz von Gauß-Mixtur-Modellen und den Aufbau von Entscheidungsgrenzen erfährt die deskriptive Raumbeobachtung eine zusätzliche Betrachtungsperspektive zur Messung der Diversität.**

3. **Die Erfassung und Strukturierung des Agglomerations- und Verdichtungsprozesses deutet an, dass in einigen Gebieten in Deutschland polyzentrische Regionalstrukturen entstanden sind, die gemäß der öffentlichen Debatte über das Zentrale-Orte-Konzept nicht von den klassischen ausgewiesenen Zentrale-Orte-Kategorien repräsentiert werden.**

4. **Die Emergenten SOM sind ein sehr leistungsfähiges Verfahren, um hochdimensionale Daten im urbanen Forschungsfeld zu untersuchen, da sowohl die reine Strukturerkennung als auch die spätere Strukturbildung auf Basis von dichtebasierten Clusteralgorithmen unterstützt wird.**

5. **Es ist festzustellen, dass die Möglichkeit zur Integration von Klassifikatoren in einen Planungs- und Steuerungsprozess auf dem Gebiet der Stadt- und Regionalplanung bisher kaum umfassend untersucht worden ist.**

6. **20 % der Gemeinden in Deutschland enthalten 80 % des Gesamtbestandes.**

7. **Will man das Schätzergebnis des deutschen Gebäudebestandes zusätzlich optimieren, so sollte man gerade diejenigen Gemeinden genauer untersuchen, die über eine große Anzahl an Wohngebäuden bzw. Einwohnern verfügen.**

Literaturverzeichnis

AACHENER INSTITUT FÜR BAUSCHADENSFORSCHUNG UND ANGEWANDTE BAUPHYSIK (HRSG.): HOFMAN, FRANZ GEORG [2001]: Urban heritage – building maintenance. Final report. COST Action C5. European Commission.

ALBERS, G. [2000]: Die kompakte Stadt im Wandel der Leitbilder. In: WENTZ, M. (Hrsg.): Die kompakte Stadt. Campus Verlag, Frankfurt/New York, 23-29.

ALDENDERFER, M.S.; BLASHFIELD, R.K. [1984]: Cluster Analysis. Beverly Hills-London-New York.

ALLISON, P.D. [2002]: Missing Data. Sage, Thousand Oaks, CA.

ALONSO, W. [1964]: Location and Land Use – Toward A General Theory Of Land Rent, Harvard University Press, Cambridge, Massachusetts.

ALPAR, P. [2005]: Anwendungsorientierte Wirtschaftsinformatik. Strategische Planung, Entwicklung und Nutzung von Informations- und Kommunikationssystemen. 4. Auflage, Vieweg Verlag, Braunschweig.

ALTENKRÜGER, DORIS; BÜTTNER, Winfried [1992]: Wissensbasierte Systeme – Architektur, Entwicklung, Echtzeit-Anwendungen. Vieweg Verlag, Braunschweig.

AMBROSI, K [1980]: Aggregation und Identifikation in der numerischen Taxonomie. Königstein: Verlag Anton Hain.

ANDERBERG, M.R. [1973]: Cluster Analysis for Applications. John Wiley & Sons, New York.

ANSELIN, LUC [1988]: Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht.

ARING, J. [1999]: Suburbia – Postsuburbia – Zwischenstadt. Die jüngere Wohnsiedlungsentwicklung im Umland der großen Städte Westdeutschlands und Folgerungen für die Regionale Planung und Steuerung. ARL Arbeitsmaterial Bd. 262, Hannover.

ARING, J. [2005]: Leitbilder und Handlungsstrategien für die Raumentwicklung in Deutschland – Diskussionspapier. Auftraggeber: BBR, Bundesamt für Bauwesen und Raumordnung, Berlin (im Auftrag des BMVBW).

ARL (AKADEMIE FÜR RAUMFORSCHUNG UND LANDESPLANUNG) [1987]: Flächenhaushaltspolitik – Ein Beitrag zum Bodenschutz. Forschungs- und Sitzungsberichte der Akademie für Raumforschung und Landesplanung, Band 173, Hannover.

ARL (AKADEMIE FÜR RAUMFORSCHUNG UND LANDESPLANUNG) [1995]: Handwörterbuch der Raumordnung, ARL, Hannover.

ARL (AKADEMIE FÜR RAUMFORSCHUNG UND LANDESPLANUNG) [1999]: Flächenhaushaltspolitik – Feststellungen und Empfehlungen für eine zukunftsfähige Raum- und Siedlungsentwicklung. Forschungs- und Sitzungsberichte der Akademie für Raumforschung und Landesplanung, Band 208, Hannover.

ARL (AKADEMIE FÜR RAUMFORSCHUNG UND LANDESPLANUNG) [2002]: Fortentwicklung des Zentrale-OrteSystems. BLOTEVOGEL (Hrsg.): Bericht des Ad-hoc-Arbeitskreises „Fortentwicklung des Zentrale-Orte-Systems“. ARL, Hannover.

ARL (AKADEMIE FÜR RAUMFORSCHUNG UND LANDESPLANUNG) [2004]: Flächenhaushaltspolitik – Ein Beitrag zur nachhaltigen Raumentwicklung. Positionspapier Nr. 58 der Akademie für Raumforschung und Landesplanung. Hannover.

ARLT, GÜNTHER; GÖSSEL, JÖRG; HEBER, BERND; HENNERSDORF, JÖRG; LEHMANN, IRIS UND THINH, NGUYEN XUAN [2001]: Auswirkungen städtischer Nutzungsstrukturen auf Bodenversiegelung und Bodenpreis. IÖR-Schriften, Band 34, Leibniz-Institut für ökologische Raumentwicklung, Dresden.

ARLT, GÜNTHER; HENNERSDORF, JÖRG; LEHMANN, IRIS; THINH, NGUYEN XUAN [2005]: Auswirkungen städtischer Nutzungsstrukturen auf Grünflächen und Grünvolumen. IÖR-Schriften, Band 47, Leibniz-Institut für ökologische Raumentwicklung, Dresden.

ATKINSON, P.M. / MARTIN, D. (Eds.) [2000]: INNOVATIONS IN GIS 7 – GIS and Geocomputation. Taylor & Francis, London.

BACCINI, PETER [1991]: Metabolism of the Anthroposphere / Peter BACCINI ; Paul H. BRUNNER. – Berlin ; Heidelberg : Springer – XII, 157 S. : Ill., zahlr. graph. Darst.; (dt.).

BACCINI, PETER [1994]: Stoffwechsel der Anthroposphäre (Vorlesungsskript). ETH Zürich: Lehrstuhl für Stoffhaushalt und Entsorgungstechnik.

BACCINI, PETER [1996]: Regionaler Stoffhaushalt: Erfassung, Bewertung und Steuerung / Peter BACCINI und H.-Peter BADER. – Heidelberg : Spektrum Akad. Verl. - XII, 420 S. : Ill., graph. Darst.; (dt.).

BACCINI, PETER; KYTZIA SUSANNE; OSWALD, FRANZ [2002]: Restructuring Urban Systems. In: Future Cities: Dynamics and Sustainability (MOAVENZADEH, HANAKI, BACCINI, eds.), pp. 17-43. Kluwer Academic Publishers, Dordrecht.

BACCINI, PETER; OSWALD, FRANZ [1999]: Netzstadt. Transdisziplinäre Methoden zum Umbau urbaner Systeme: Ergebnisse aus dem Forschungsprojekt SYNOIKOS — Nachhaltigkeit und urbane Gestaltung im Raum Kreuzung Schweizer Mittelland. Hochschulverlag AG an der ETH Zürich.

BACHER, JOHANN [1994]: Clusteranalyse – Anwendungsorientierte Einführung. Oldenbourg Verlag München/Wien.

BÄCHTHOLD, H.-G. [1998]: Nachhaltigkeit. Herkunft und Definitionen eines komplexen Begriffs. In: Schweizer Ingenieur und Architekt, Nr. 13.

BACKHAUS; ERICHSON; PLINKE; WEIBER [2000]: Multivariate Analysemethoden – Eine anwendungsorientierte Einführung, 9. Auflage. Springer Verlag Berlin.

BACKHAUS; ERICHSON; PLINKE; WEIBER [2006]: Multivariate Analysemethoden – Eine anwendungsorientierte Einführung. 11. Auflage. Springer Verlag, Berlin.

BAHRENBERG, G.; GIESE, E.; NIPPER, J. [2003]: Statistische Methoden in der Geographie. Gebrüder Borntraeger, Berlin.

BALL, G.H.; HALL, B.J. [1965]: ISODATA, A novel method of data analysis and pattern classification, Technical Report, Stanford Research Institute, Melno Park.

BARTELS, D. [1978]: Erläuterungen zur Ermittlung der Entwicklungsverlaufklassen. In: BLOTEVOGEL, H.H. und SCHÖLLER, P.: Erläuterung zur Karte ’Bevölkerungsentwicklung in den Gemeinden 1837-1970 nach Entwicklungsverlaufsklassen’. Veröffentlichung der Akademie für Raumforschung und Landesplanung, Deutscher Planungsatlas, Band 1. 13. Lieferung, Berlin.

BATTY, M.; CHIN, N.; BESUSSI, E. [2002]: The Scatter Project - Sprawling Cities and Transport: From Evaluation to Recommendations. Project funded by the European Commission (Deliverable 1: Work package 1: State of the Art Review of Urban Sprawl Impacts and Measurement Techniques), London.

BAUMARD, PHILIPPE [1999]: Tacit Knowledge in Organizations. Sage, Thousand Oaks, CA.

BBR - BUNDESAMT FÜR BAUWESEN UND RAUMORDNUNG (HRSG.) [2002]: Endbericht zum Forschungsprojekt „Dialog Bauqualität“ - Az: Z6 – 4.4 – 01 – 110, http://www.bbr.bund.de/bauwesen/download/endbericht_dialog_bauqualitaet.pdf, (Stand: 08.06.2005).

BBR - BUNDESAMT FÜR BAUWESEN UND RAUMORDNUNG (HRSG.) [2003]: Siedlungsstrukturelle Veränderungen im Umland der Agglomerationsräume. Bonn.

BUNDESAMT FÜR BAUWESEN UND RAUMORDNUNG (HRSG.) [2004]: Flächenerhebung 2004. http://www.bbr.bund.de/index.html?/raumordnung/siedlung/flaechenerhebung_2004.htm, (Stand: 29.07.2005).

BBR - BUNDESAMT FÜR BAUWESEN UND RAUMORDNUNG (HRSG.) [2006]: Referat I6 Raum- und Stadtbeobachtung – Betreuung und Weiterentwicklung des raumbezogenen Informationssystems ‚Laufende Raum- und Stadtbeobachtung'. http://www.bbr.bund.de/nn_21190/DE/Bundesamt/Organigramm/I/ReferatI6_node.html__nnn=true (Stand: 30.10.2006).

BEATLEY, T. [2000]: Green Urbanism Learning from European Cities. Island Press Washington.

BECHER, STEPHAN [1995]: Klassifikation der regionalen Immobilienmärkte der Bundesrepublik Deutschland (Dissertation). Universität Mainz.

BEHNISCH, MARTIN [2004]: Klassifizierung deutscher Gemeinden nach bestandsorientierten Merkmalen. Diplomarbeit des Fachbereichs Architektur der Technischen Hochschule Karlsruhe, 2004.

BEHNISCH, MARTIN [2005]: Bestandsorientiertes Klassifikatormodell — Ein Informations- und Analysewerkzeug zur Untersuchung von Gebäuden und Stadt. In: Workshop Simulation in den Geowissenschaften und Umweltwissenschaften, Leibnitz Institut für ökologische Raumentwicklung.

BEHNISCH, MARTIN / VIEJO GARCIA, PABLO [2005]: Comparing German building stocks in space and time from city to region. Computers in Urban Planning and Urban Management, Conference – CUPUM, London.

BEHNISCH, MARTIN [2006]: Bestandsorientiertes Klassifikatormodell – Fuzzy-Pattern Klassifikation und GIS als Informations- und Analysewerkzeuge zur Untersuchung von Stadt- und Gebäudestrukturen. In: Doctoral Workshop im Rahmen der 30. Jahrestagung der Gesellschaft für Klassifikation, Berlin.

BEHNISCH, MARTIN / ULTSCH, ALFRED [2007]: Urban Data Mining. In: BOCK, H.H., GAUL, W., VICHI, M. (Hrsg.): Studies in Classification, Data Analysis, and Knowledge Organization. Eingereichter Beitrag zu den Proceedings der 31. Jahrestagung der Gesellschaft für Klassifikation. Springer, Freiburg.

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 Baden-Württemberg, S. 165-186, Statistisches Landesamt Baden-Württemberg, Stuttgart.

BEICHEL, REINHARD [2002]: Klassifikatorprinzip, gefunden in: http://www-gs.informatik.tu-cottbus.de/~wwwgs/bia2_v09.pdf (Stand: 05.04.2007).

BENKE, CARSTEN [2004]: Historische Umbrüche. Schrumpfungen und städtische Krisen in Mitteleuropa seit dem Mittelalter. In: Städte im Umbruch – Online Magazin für Stadtentwicklung, Stadtschrumpfung, Stadtumbau & Regenerierung, Heft 1, 7-14.

BELL, DANIEL [1976]: The coming of post-industrial society a venture of social forecasting. Basic Books, New York.

BELLMANN, R.; KALABA, R.; ZADEH, L.A. [1966]: Abstraction and Pattern Classification. In: Journal of Mathematical Analysis and Applications, Vol. 13, S.1-7.

BERGS, SIEGFRIED [1981]: Optimalität bei Clusteranalysen – Experimente zur Bewertung numerischer Klassifikationsverfahren, Münster.

BERRY, BRIAN J.L. / F.E. HORTON [1970]: Geographic Perspectives on Urban Systems with Integrated Readings. New Jersey.

BERRY, BRIAN J.L. [1972]: City Classification Handbook - Methods and Applications. John Wiley & Sons, New York.

BERRY, BRIAN J.L. [1996]: Technology-sensitive urban typology. Urban Geography 17, pp. 674-689, Bellwether Publishing, Columbia.

BERRY, BRIAN J.L.; KASARDA, J.O. [1977]: Contemporary Urban Ecology. Macmillan, New York.

BERRY; LINOFF [1997]: Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons, New York.

BEZDEK, J.C. [1974]: Cluster Validity with Fuzzy Sets. In: Journal of Cybernetics, Vol. 3, S. 58-73.

BEZDEK, J.C. [1981]: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York 1981.

BEZDEK, J.C. [1993]: A Review of Probalistic, Fuzzy and Neural Models for Pattern Recognition. Journal of intelligent and Fuzzy Systems, Vol. 1 (1), S. 1-25.

BIEWER, BENNO [1997]: Fuzzy-Methode – Praxisrelevante Rechenmodelle und Fuzzy-Programmiersprachen. Springer Verlag, Berlin.

BILL, R. [1999]: Grundlagen der Geo-Informationssysteme, Band 2 – Analysen, Anwendungen und neue Entwicklungen. 2. Auflage. Wichmann Verlag, Heidelberg.

BILMES, J. (1997): A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of Berkeley, ICSI-TR-97-021. http://crow.ee.washington.edu/people/bulyko/papers/em.pdf.

BLASHFIELD, R.K. [1976]: Mixture Model Tests of Clusteranalysis, Accuracy of four Agglomerative Hierarchical Methods. In: Psychological Bulletin, Vol. 83, No. 3, S.377-388.

BLOTEVOGEL, H. [1996]: Zentrale Orte: Zur Karriere und Krise eines Konzeptes in Geographie und Raumplanung – In: Erdkunde 50, 9-25, Geographisches Institut, Bonn.

BLOTEVOGEL, H. [2002]: Evaluierung des Zentrale-Orte-Systems und ihre Auswirkungen auf die Regionalplanung. Empfehlung zur Weiterentwicklung des Zentrale-Orte-Konzepts. In: ARL (Hrsg.): Regionalplanung in Baden-Württemberg: Weiterentwicklung der 12 Regionen und ausgewählte Handlungsfelder. Arbeitsmaterial, Nr. 290, Akademie für Raumforschung und Landesplanung, Hannover.

BLOTEVOGEL, H. [2005]: Neuformulierung des Zentrale-Orte-Konzepts. Kurzfassung eines Vortrages im Rahmen einer Fachtagung des Ministeriums des Innern und für Sport, Oberste Landesplanungsbehörde, am 14.07.2005 in Budenheim bei Mainz.

BMBau - Bundesministerium für Raumordnung, Bauwesen und Städtebau (Hrsg.) [1996]: Siedlungsentwicklung und Siedlungspolitik. Nationalbericht Deutschland zur Konferenz HABITAT II. Bonn.

BMVBW - BUNDESMINISTERIUM FÜR VERKEHR, BAU- UND WOHNUNGSWESEN (Hrsg.) [2003]: Endbericht T 3010, Systematische Instandsetzung und Modernisierung im Wohnungsbestand. Fraunhofer IRB Verlag, Stuttgart.

BOBEK, H. [1928]: Innsbruck, eine Gebirgsstadt. Ihr Lebensraum und ihre Erscheinung. Forschungen zur deutschen Landes- und Volkskunde, 25(3):220–372, Stuttgart.

BOCK, HANS HERMANN [1974]: Automatische Klassifikation. Studia Mathematica/Mathematische Lehrbücher, Band XXIV. Vandenhoeck & Ruprecht in Göttingen.

BOCK, HANS HERMANN [1979]: Clusteranalyse mit unscharfen Partitionen. In: BOCK Hans Hermann (Hrsg.): Klassifikation und Erkenntnis III, Frankfurt, S. 137-163.

BOCKLISCH, S. F. [1987]: Prozeßanalyse mit unscharfen Verfahren, VEB Verlag Technik, Berlin.

BONTJE, M. [2004]: From suburbia to postsuburbia in the Netherlands: Potentials and threats for sustainable regional development. In: Journal of Housing and the Built Environment, Vol. 19, S. 25–47, Springer Netherlands.

BORCHERDT, CHR. [1983]: Geografische Landeskunde von Baden-Württemberg. Schriften zur politischen Landeskunde Baden Württembergs 8. Stuttgart.

BORGELT, C.; KLAWONN, F.; KRUSE, R.; NAUCK, D. [2003]: Neuro-Fuzzy-Systeme – Von den Grundlagen künstlicher Neuronaler Netze zur Kopplung mit Fuzzy-Systemen. 3. Auflage. Vieweg Verlag, Braunschweig.

BORGELT, C.; TIMM, H. [2000]: Advanced Fuzzy Clustering and Decision Tree. Plug-Ins for DataEngine™. Intelligent Systems and Soft Computing, 188-212. B. AZVINE, N. AZARMI, and D. NAUCK, eds. LNCS 1804. Springer, London, United Kingdom. http://fuzzy.cs.uni-magdeburg.de/~borgelt/papers/bt.pdf, (Stand 20.05.2006).

BOSSARD, M.; FERANEC, J.; OTAHEL, J. [2000]: CORINE Land Cover Technical Guide – Addendum 2000. European Environmental Agency, Technical Report No. 40.

BOSSEL, HARTMUT [1987]: Systemdynamik – Grundwissen, Methoden und BASIC-Programme zur Simulation dynamischer Systeme. Vieweg Verlag, Braunschweig.

BOUSTEDT, OLAF [1953]: Die Stadtregion. Ein Beitrag zur Abgrenzung städtischer Agglomerationen. Allgemeines statistisches Archiv 37, S.13-26.

BOUSTEDT, OLAF [1970]: Stadtregionen. In: Akademie für Raumforschung und Landesplanung (Hrsg.). Handwörterbuch der Raumforschung und Raumordnung. 2. Auflage. S. 3207-3237, Hannover.

BOUSTEDT, OLAF [1975 a]: Grundriß der empirischen Raumforschung – Teil 3 (= Taschenbücher zur Raumplanung Band 6): Siedlungsstrukturen. Veröffentlichung Akademie für Raumforschung und Landesplanung, Hannover.

BOUSTEDT, OLAF [1975 b]: Grundriß der empirischen Raumforschung – Teil 4: Regionalstatistik. (= Taschenbücher zur Raumplanung Band 7), Veröffentlichung Akademie für Raumforschung und Landesplanung, Hannover.

BRADLEY, PATRICK ERIK; FERRARA, CLAUDIO [2004]: Abschlussbericht des DFG-Projekts: Validierung eines integrierten, dynamischen Modells des deutschen Gebäudebestandes. (BEVAL) http://ww1.mathematik.uni-karlsruhe.de/»bradley/Arch/KO1488-4-2.pdf, (Stand 17.01.2006).

BRADLEY, PATRICK ERIK; FERRARA, CLAUDIO; KOHLER, NIKLAUS; PAUL, NORBERT [2005]: Sustainable Management of building stocks. Abschlussbericht zum EIFER-Projekt. http://ww1.mathematik.uni-karlsruhe.de/»bradley/Arch/sub/SUB-rapfin-2005.pdf., (Stand 10.01.2006).

BRAKE, KLAUS; EINACKER, INGO; MÄDING, HEINRICH [2005]: Kräfte, Prozesse, Akteure – Zur Empirie der Zwischenstadt. Müller+Busmann, Wuppertal.

BRAUN, CHRISTOPH [1990]: Regional- und Standortgrobplanung in der Europäischen Gemeinschaft – Neue Möglichkeiten der Auswertung von Regionaldaten, Münster.

BRAUSE, RÜDIGER [1995]: Neuronale Netze. 2. Auflage, Teubner, Stuttgart.

BREHENY, M. J. [1992]: The Contradictions of the Compact City: A Review. In: BREHENY, M.J. (Ed.) Sustainable Development and Urban Form, European research in regional science, London, 138-159.

BREHENY, M.J. [1996]: Centrists, Decentrists and Compromisers: Views on the Future of Urban Form. In: JENKS, M.; BURTON, E.; WILLIAMS, K. (Eds.): The Compact City – A Sustainable Urban Form, Spon Press, London.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. [1984]: Classification and Regression Trees. Wadsworth, California.

BRESSLER, CHRISTIAN [2001]: Das Bevölkerungspotential – Messgröße für Interaktionschancen. In: Nationalatlas der Bundesrepublik Deutschland. Band: Bevölkerung. Institut für Länderkunde, Leipzig (Hrsg.). Specktrum Akademischer Verlag, Heidelberg, Berlin.

BRINGEZU, S. [2000]: Ressourcennutzung in Wirtschaftsräumen. Stoffstromanalysen für eine nachhaltige Raumentwicklung. Berlin.

BRÖCKER, J. [1984]: Räumliche Querschnittsregressionen mit potentialisierten Variablen. Seminarberichte der Gesellschaft für Regionalforschung. Band 21, Heidelberg.

BRÖCKER, J. [1989]: Determinanten des regionalen Wachstums im sekundären und tertiären Sektor der Bundesrepublik Deutschland 1970 bis 1982. Schriften des Instituts für Regionalforschung der Universität Kiel, München.

BRÖCKER, J., SCHÖLER, K. HIRSCHFELD, M., NIESE, M. UND RICHTER, F. [1998]: Langfristige Entwicklungsmuster deutscher Stadtregionen unter wechselnden Wirtschaftssystemen. Forschungsprojekt im Rahmen des DFG-Schwerpunktprogramms „Wirtschaftliche Strukturveränderungen, Innovationen und regionaler Wandel in Deutschland nach 1945“. Abschlussbericht (2 Bände). Dresden / Kiel.

BROCKHAUS (Hrsg.) [1885]: Conversations-Lexikon, 13. Auflage, Band 10.

BROCKHAUS (Hrsg.) [1979]: Der große Brockhaus. Brockhaus, Wiesbaden.

BROSKA, E. [2000]: Die Beteiligung der (Privat-)Wirtschaft an Lokalen Agenda-Prozessen. Eine Übersicht über bisherige Erfahrungen einschließlich eines konkreten Fallbeispiels (Düsseldorf) aus der kommunalen Sicht. Ruhr-Universität Bochum. Geographisches Institut, Diplomarbeit.

BÜHL; ZÖFEL [2002]: SPSS 11 – Einführung in die moderne Datenanalyse. Addison-Wesley, Boston.

BULMER, M. G. [2003]: Francis Galton: Pioneer of heredity and biometry. John Hopkins University Press, Baltimore, Maryland.

BUND/MISEREOR (Hrsg.) [1996]: Zukunftsfähiges Deutschland - Ein Beitrag zu einer global nachhaltigen Entwicklung. Studie des Wuppertal Instituts für Klima. Umwelt, Energie. Birkhäuser Verlag, Basel, Boston.

BUNDESAMT FÜR KONJUNKTURFRAGEN; IP-BAU; INFRAS [1991]: Recycling, Verwertung und Behandlung von Bauabfällen. Bern.

BUNDESFORSCHUNGSANSTALT FÜR LANDESKUNDE UND RAUMORDNUNG [1996]: Raumordnungsprognose 2010. Bonn.

BURGESS, E.W. [1925]: The Growth of the City: An Introduction to a Research Project. In: PARK, R.E.; BURGESS, E.W.; McKenzie R.D.: The City. Suggestions for Investigation of Human Behaviour in the Urban Environment. Chicago, London: The University of Chicago Press. Reprint 1967.

CAIN, A.J. [1962]: Zoological Classification, aslib-proceedings, 14, S. 226-230.

CANTOR, G. [1895]: Beiträge zur Begründung der transfiniten Mengenlehre; Halle.

CASETTI [1964 a]: Multiple discriminant functions. Technical Report No. 11, Computer Applications in the Earth Sciences Project, Department of Geography, Northwestern University, Evanston.

CASETTI [1964 b]: Classificatory and regional analysis by discriminant iterations. Computer Applications in the Earth Sciences Project, Dissertation (Technical Report No. 11), Department of Geography, Northwestern University, Evanston.

CASTELLS, MANUEL [2003]: Das Informationszeitalter. Band I: Der Aufstieg der Netzwerkgesellschaft, Band II: Die Macht der Identität, Band III: Jahrtausendwende. Leske + Budrich (UTB), Opladen.

CERVERO, R. [2001]: Efficient Urbanisation: Economic Performance and the Shape of the Metropolis. In: Urban Studies, Vol. 38, No. 10, S. 1651–1671, Taylor Francis, Abingdon Oxfordshire.

CEST – Center for Science and Technology Studies, Hrsg: BERWERT, Adrian; VOCK, Patrick; TIRI, Marc (Editors) [2004]: Cluster in der schweizerischen Volkswirtschaft und im Espace Mittelland – Identifikation, Analyse und Diskussion aufgrund von Input-Output Daten. CEST Edition, Bern
http://www.cest.ch/Publikationen/2004/CEST_2004_8b.pdf, (Stand: 20.05.2006).

CHALMERS, ALAN F. [2001]: Wege der Wissenschaft. Einführung in die Wissenschaftstheorie. Springer, Berlin.

CHRISTALLER, WALTER [1980]: Die zentralen Orte in Süddeutschland. Eine ökonomischgeographische Untersuchung über die Gesetzmäßigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen. Jena, 1933. Nachdruck. Wissenschaftliche Buchgesellschaft Darmstadt.

CLAUS, ROGER [2003]: Entwurf und Implementation von Fuzzy Pattern Klassifikator-Netzen. Diplomarbeit, Technische Universität Chemnitz, Fakultät für Elektrotechnik / Informationstechnik, Lehrstuhl für Systemtheorie.

CLIFF, ANDREW D.; ORD, JOHN K. [1973]: Spatial Autocorrelation. Pion Limited, London.

COENEN, REINHARD; GRUNWALD, ARMIN [2003] : Nachhaltigkeitsprobleme in Deutschland – Analyse und Lösungsstrategien. Edition sigma, Berlin.

CORMACK, R.M. [1971]: A Review of Classification. Journal of the Royal Statistical Society, (Series A), Vol. 134, S. 321-367.

CRAIG, J. [1985]: A 1981 Socio-economic Classification of Local and Health Authorities of Great Britain Studies on Medical and Population Subjects, HSMO, London.

DAGNELIE, P. [1966]: A propos des différentes méthodes de classification numérique. Rev. Stat. Appl., 14, S. 55-75.

DECKER; FOCARDI [1995]: Technology overview: a report on data mining. Technical Report CSCS TR-95-02, CSCS-ETH, Swiss Scientific Computing Center.

DEGGAU, M.; STRALLA, H.; WIRTHMANN, A. [1998]: Klassifizierung von Satellitendaten (CORINE Land Cover), Endbericht zum Forschungsprojekt UFOPLAN 291 91 055/00, Statistisches Bundesamt,Wiesbaden.

DEICHSEL; TRAMPISCH [1985]: Clusteranalyse und Diskriminanzanalyse. Gustav Fischer Verlag, Stuttgart.

DEIMER [1986]: Unscharfe Clusteranalysemethoden – Eine problemorientierte Darstellung zur unscharfen Klassifikation gemischter Daten. Dissertation. Johann Wolfgang Goethe-Universität, Frankfurt am Main.

DEIMER, J. [1998]: Leitstern kompakte Stadt. In: Der Städtetag 1/1998, 1-2.

DEITERS, J. [1978]: Zur empirischen Überprüfbarkeit der Theorie zentraler Orte. Fallstudie Westerwald. Dümmler, Bonn.

DEMANT, BERND [1993]: Fuzzy-Theorie oder die Faszination des Vagen - Grundlagen einer präzisen Theorie des Unpräzisen. Vieweg Verlag, Braunschweig.

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. [1977]: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, B, 39, 1-38.

DER RAT DER SACHVERSTÄNDIGEN FÜR UMWELTFRAGEN [2000]: Schritte ins nächste Jahrhundert. Umweltgutachten 2000. Stuttgart.

DESTATIS (Hrsg.) [2005]: Qualitätsbericht – Statistik der Baufertigstellungen (Stand: November 2005). Statistisches Bundesamt, Wiesbaden.

DESTATIS (Hrsg.) [2005]: Qualitätsbericht – Statistik der Baugenehmigungen (Stand: November 2005). Statistisches Bundesamt, Wiesbaden.

DESTATIS (Hrsg.) [2005]: Qualitätsbericht – Statistik des Bauabgangs (Stand: November 2005). Statistisches Bundesamt, Wiesbaden.

DESTATIS (Hrsg.) [2005]: Qualitätsbericht – Statistik des Bauüberhangs (Stand: November 2005). Statistisches Bundesamt, Wiesbaden.

DEUTSCHER BUNDESTAG [1996], 13. Wahlperiode: Dritter Bericht über Schäden an Gebäuden, Unterrichtung durch die Bundesregierung, Drucksache 13/3593 vom 25.01.1996.

DEUTSCHES INSTITUT FÜR WIRTSCHAFTSFORSCHUNG (Hrsg.) [1998]: Zur Entwicklung der Bauwirtschaft in West- und Ostdeutschland. Wochenbericht 42/97.
http://www.diw-berlin.de/diwwbd97-42-1.html, (Stand: 03.04.1998).

DEUTSCHES ZENTRUM FÜR LUFT- UND RAUMFAHRT [2000]: CORINE LANDCOVER 2000 – Bodenbedeckungsdaten für Deutschland.
http://www.corine.dfd.dlr.de, (Stand: 20.02.2006).

DIEHL; STAUFENBIEHL [2001]: Statistik mit SPSS. Dietmar Klotz Verlag.

DOLD, G. [1996]: Computerunterstützung der produktbezogenen Ökobilanzierung, Wiesbaden.

DOSCH, F. [2001]: Flächenverbrauch in Deutschland und Mitteleuropa - Struktur, Trends und Steuerungsoptionen durch das Boden-Bündnis. Beitrag zur 1. Internationalen Jahrestagung des Boden-Bündnis europäischer Städte und Gemeinden. 12-13.11.2001, Osnabrück.

DOSCH, F. [2002]: Auf dem Weg zu einer nachhaltigen Flächennutzung? In: Informationen zur Raumentwicklung 1/2, S. 31-43.

DROTH, W. / FISCHER, M.M. [1980]: Zur Theoriebildung und Theorietestung: Eine Diskussion von Grundlagenproblemen am Beispiel der Sozialraumanalyse. In: OSTHEIDER, M. und STEINER, D. (eds.): Theorie und quantitative Methodik in der Geographie. Geographische Schriften 1. Geographisches Institut der ETH Zürich, Zürich.

DUBOIS, DIDIER; PRADE, HENRI [1980]: Fuzzy Sets and Sytems, Academic Press, New York.

DUBOIS, DIDIER; PRADE, HENRI [1988]: Possibility Theory. Plenum Press, New York.

DUBRAL, CHRISTOPH [1986]: Bautätigkeit im Nichtwohnbau, in Wirtschaft und Statistik, Heft 7, S. 523-528.

DYLLICK, T.; BELZ, F.; SCHNEIDEWIND, U. [1997]: Ökologie und Wettbewerbsfähigkeit, München, Wien.

ECKEY, HANS-FRIEDRICH; KOSFELD, REINHOLD; TÜRCK, MATTHIAS: Ökonometrische Eingleichungsmodelle mit SPSS. PDF-Manuskript.
http://www.wirtschaft.uni-kassel.de/Eckey/Lehre/Oekonometrie/SPSS/SPSS_Skript.pdf,
(Stand: 16.06.2006) als ergänzendes Hilfsmittel zum Lehrbuch von ECKEY, Hans-Friedrich; KOSFELD, Reinhold; DREGER, Christian [2001]: Ökonometrie. Grundlagen – Methoden – Beispiele, 2. Aufl., Wiesbaden.

ECOINVENT [2004]: Ökoinventare. Schweiz. Zentrum für Ökoinventare.
http://www.ecoinvent.ch/de/index.htm, (Stand 18.12.2004).

EEA [1997]: Technical and Methodological Guide for Updating CORINE Land Cover Data Base. European Environmental Agency.
http://www.ec-gis.org/docs/F27057/CORINE.PDF (Stand: 31.10.2006).

EGENHOFER, M. / GOLLEDGE, R. [1998]: Spatial and Temporal Reasoning in Geographic Information Systems. Oxford University Press.

ELIAS, BIRGIT [2006]: Extraktion von Landmarken für die Navigation. Fakultät für Bauingenieurwesen und Geodäsie, Universität Hannover.

ELSEN, INGO [2000]: Ansichtenbasierte 3D-Objekterkennung mit erweiterten Selbstorganisierenden Merkmalskarten. VDI Verlag, Düsseldorf.

ENGELBACH, W.D. [2000]: Bestandsorientiertes Stoffstrommanagement als Mittel einer nachhaltigen Bewirtschaftung von Wohngebäuden. Dissertation, RWTH Aachen.

ENQUETE-KOMISSION: „Schutz des Menschen und der Umwelt“ (Hrsg.) [1994]: Die Industriegesellschaft gestalten. Perspektiven für einen nachhaltigen Umgang mit Stoff- und Energieströmen. Bonn.

ENQUETE-KOMISSION: „Schutz des Menschen und der Umwelt“ (Hrsg.) [1997]: Konzept Nachhaltigkeit. Fundamente für die Gesellschaft von morgen. Zwischenbericht. Bonn.

ENQUETE-KOMISSION: „Schutz des Menschen und der Umwelt“ (Hrsg.) [1998]: Konzept Nachhaltigkeit. Vom Leitbild zur Umsetzung. Abschlußbericht. Drucksache 13/11200, Bonn.

ENQUETE-KOMISSION: „Schutz des Menschen und der Umwelt“ (Hrsg.), HASSLER, U. / KOHLER, N. / PASCHEN, H. [1999]: Stoffströme und Kosten in den Bereichen Bauen und Wohnen. Springer, Berlin.

EPPING, GÜNTER [1977]: Bodenmarkt und Bodenpolitik in der Bundesrepublik Deutschland, Berlin.

ERB, WOLF-DIETER [1990]: Anwendungsmöglichkeiten der linearen Diskriminanzanalyse in Geographie und Regionalwissenschaft. Schriften des Zentrums für regionale Entwicklungsforschung der Justus-Liebig-Universität Gießen, Bd. 39, Weltarchiv, Hamburg.

EURAC – Europäische Akademie Bozen, Fachbereich Alpine Umwelt, Hrsg.: TAPPEINER, Ulrike; TAPPEINER, Gottfried; HILBERT, Andreas; MATTANOVICH, Ernst (Editors) [2003]: The EU Agricultural Policy and the Environment – Evaluation of the Alpine Region. Blackwell Verlag, Berlin Vienna:

EVERITT, B.S. [1980]: Cluster Analysis. Quality and Quantity, New York.

FAHRMEIER, LUDWIG; KÜNSTLER, RITA; PIGEOT, IRIS; TUTZ, GERHARD [2004]: Statistik – Der Weg zur Datenanalyse. Fünfte Auflage, Springer, Berlin.

FAHRMEIER; HAMERLE [1984]: Multivariate statistische Verfahren. De Gruyter Berlin/New York.

FAIST, MIREILLE [2000]: Akteurbezogene Stoffflussanalyse (Dissertation), EAWAG Dübendorf.

FALK; BECKER; MAROHN [1995]: Angewandte Statistik mit SAS. Springer Verlag, Berlin.

FAYYAD; PIATESKY-SHAPIRO; SMYTH [1996]: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM, Vol. 39, No 11, November 1996, ACM Press, New York.

FERBER, REGINALD [2003]: Information Retrieval – Suchmodelle und Data Mining-Verfahren für Textsammlungen und das Web. 1. Auflage, Dpunkt Verlag, Heidelberg.

FERNAU, ALF [1997]: Werkzeuge zur Analyse und Beurteilung der internationalen Wettbewerbsfähigkeit von Regionen (Dissertation). Universität St. Gallen.

FERRARA, CLAUDIO [2004]: Validierung eines integrierten, dynamischen Modells des deutschen Gebäudebestandes: Allgemeiner Teil. Technischer Bericht 2004-04, Institut für industrielle Bauproduktion, Universität Karlsruhe.

FEYNMAN, Richard P. [1988]: What do you care, What Other People Think? Norton & Company, New York.

FISCHER, M. M. [1982]: Eine Methodologie der Regionaltaxonomie: Probleme und Verfahren der Klassifikation und Regionalisierung in der Geographie und Regionalforschung. Heft 3 der Bremer Beiträge zur Geographie und Raumplanung (Hrsg.: BAHRENBERG, Gerhard; BARTH, Hans-Karl; LEUZE, Eva; TAUBMANN, Wolfgang), Presse- und Informationsdienst, Universität Bremen.

FISCHER, M. M. [2003]: Arbeitsunterlagen zur Lehrveranstaltung: Methoden und Techniken der Raumwirtschaft. Institut für Wirtschaftsgeographie, Regionalentwicklung und Umweltwirtschaft. Abteilung für Wirtschaftgeographie & Geoinformatik, Wirtschaftsuniversität Wien.

FISCHER, M. M. [2006]: Spatial Analysis and GeoComputation. Springer Verlag, Berlin.

FIX, E. / HODGES, J. [1951]: Discriminatory analysis. Nonparametric discrimination: Consistency Properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

FIX, E. / HODGES, J. [1952]: Discriminatory analysis: small sample performance. Technical Report Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

FLACKE, JOHANNES [2003]: Mehr Stadt – weniger Fläche. Deutsche Akademie für Landeskunde, Selbstverlag, Flensburg.

FLAKE, R.H.; TURNER, B.L. [1968]: Numerical Classification for Taxonomic Problems, In: Journal of Theoretical Biology, Vol. 20, 1968, S. 260-270.

FLOODGATE, G.D. [1962]: Some remarks on the theoretical aspects of bacterial Taxonomy. In: Bact. Rev. 26, S. 277-291.

FLOREK, K; LUKASZEWICZ, J.; PARKAL, J.; STEINHAUS, H.; ZUBRZYCKI, S. [1951]: Sur la liason et la division despoints d'un ensemble fini. In: Colleg. Math., Vol.2, S.282-285.

FORESTALL, R. [1967]: Economic classification of places over 10000, 1960-1963. In: Municipal Yearbook, pp. 30-65, International City Managers'Association, Chicago, IL.

FORST, HANS THEO [1974]: Zur Klassifizierung von Städten nach wirtschafts- und sozialstatistischen Strukturmerkmalen. Physica Verlag, Würzburg.

FOUCAULT, MICHEL [1987]: Von der Subversion des Wissens. Walter SEITTER (Hrsg.): Übertragen aus dem Französischen und Italienischen, Fischer-Taschenbuch-Verlag, Frankfurt am Main.

FOURASTIE, JEAN [1952]: Le Grand Espoir du XXe siècle, progrès technique – progrès social, 3ème ed., Paris.

FRANKE, J.; HERR, D. [1987]: Klassifikation von Wohngebieten durch Laien. Deutscher Studien Verlag, Weinheim.

FRENKEL, AMMON [2004]: Land-use patterns in the classification of cities: the Israeli case. In: Environmental and Planning B: Planning and Design, Volume 31, pp. 711-730. Pion Publication, London.

FRIEDMAN, H.P.; RUBIN, J. [1967]: On some invariant criteria for grouping data. In: JASA (Journal of the Acoustic Society of America), Vol. 62, No. 320, S. 1159-1178, American Institute of Physics, New York.

FRÜHWALD, W. [1997]: Neue Perspektiven in der Wissenschaft. In: ZEIT Punkte 6/97, S. 62-64, Franz Steiner Verlag, Stuttgart.

GALE, S. / ATKINSON, M. [1979]: On the set theoretic foundation of the regionalization problem. In: GALE, S. und OLSSON, G. (eds.): Philosophy in geography, pp. 65-107, Reidel Verlag, Dordrecht.

GALSTER, G. [2000]: Wrestling Sprawl to the Ground: Defining and Measuring an Elusive Concept. In: Housing Policy Debate, Vol. 12, Issue 4, S. 681–717, Fannie Mae Foundation, Washington, DC.

GANSER, K. [2001]: Hände weg, liegen lassen. In: Der Architekt 4/2001. Bund Deutscher Architekten (Hrsg.), Berlin (Abgedruckt in: Schrumpfende Städte fordern neue Strategien für die Stadtentwicklung. Aus dem Leerstand in neue Qualitäten? Deutsche Akademie für Städtebau und Landesplanung (Hrsg.) [2002], Berlin.).

GATH, I. [1989]: Unsupervised Optimal Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, pp. 773 - 781, IEEE Computer Society, Washington.

GATZWEILER, H.-P. [1996]: Siedlungsentwicklung und Siedlungspolitik in Deutschland. In: Raumforschung und Raumordnung 57, 2/3, S. 129-136. Carl Heymanns Verlag, Köln.

GATZWEILER, HANS-PETER; MEYER, KATRIN; MILBERT, ANTONIA [2003]: Schrumpfende Städte in Deutschland? – Fakten und Trends. In: BBR (Hrsg.), Informationen zur Raumentwicklung, Heft 10/11, 557-574, Selbstverlag, Bonn.

GATZWEILER, HANS-PETER; KUHLMANN, PETRA; MEYER, KATRIN; MILBERT, ANTONIA; PÜTZ, THOMAS; SCHLÖMER, CLAUS; SCHÜRT, ALEXANDER [2006]: Herausforderungen deutscher Städte und Stadtregionen - Ergebnisse aus der laufenden Raum- und Stadtbeobachtung des BBR zur Entwicklung der Städte und Stadtregionen in Deutschland. In: BBR (Hrsg.), BBR-Online-Publikation, Nr. 8/2006, Bonn.

GAUL, W. G.; SÄUBERLICH, F. [1998]: Classification and Positioning of Data Mining Tools. Herausforderungen der Informationsgesellschaft an Datenanalyse und Wissensverarbeitung. 22. Jahrestagung der Gesellschaft für Klassifikation, Proceedings. Springer Verlag, Berlin.

GAUSS [1887]: Abhandlungen zur Methode der kleinsten Quadrate. Hrsg.: BÖRSCH, A. [1964], Neudruck der ursprünglichen Ausgabe. Physica-Verlag, Würzburg.

GEER, STEN [1923]: Greater Stockholm: A Geographical Interpretation. Geographical Review, Vol. 13, pp. 487–506, Graduate School of Geography, Clark University, Worcester.

GEISSLER, K. [1998]: „Diplomarbeit – Entwurf eines Klassifikator-Netzes“; Fakultät für Elektrotechnik / Informationstechnik, Lehrstuhl für Systemtheorie, TU Chemnitz.

GHOLAMREZA, NAKHAEIZADEH [1987]: Data Mining – Theoretische Aspekte und Anwendungen. Physica-Verlag, Heidelberg.

GIESE, E. [1980]: Entwicklung und Forschungsstand der ‚Quantitativen Geographie’ im deutschsprachigen Bereich. In: Geographische Zeitschrift, 68, Franz Steiner Verlag, Stuttgart.

GIFFINGER, RUDOLF; KALASEK, ROBERT; WONKA, ERICH [2006]: Ein neuer Ansatz zur Abgrenzung von Stadtregionen: methodische Grundlagen und Perspektiven zur Anwendung. In: Proceedings CORP 2006 & Geomultimedia06, Wien.

GITMAN, I.; LEVINE, M.D. [1970]: An Algorithm for Detecting Unimodal Fuzzy Sets and its Applications as a Clustering Technique, In: IEEE Transactions on Computers, Vol. C-19, No.7, pp.583 - 593, IEEE Computer Society, Washington.

GLENCK, E. [1996]: Güter- und Stoffbilanzen im Bauwesen. In: Wissenschaft und Umwelt: Dokumentation des Hearings des Graduiertenkollegs „Interdisziplinäre Strategien zum Schutz der Umwelt“ der RWTH Aachen, Heft 1-2/1996, S. 85-94.

GÖB, RÜDIGER [1977]: Die schrumpfende Stadt. In: Archiv für Kommunalwissenschaften, 16 II, S. 149-177, Kohlhammer / Deutscher Gemeindeverlag, Stuttgart.

GOODCHILD, M.F.; STEYART, L.T.; PARKS, B.O.; JOHNSTON, C.; MAIDMENT, D., CRANE, M.; GLENDINNING, S. (Eds.) [1996]: GIS and Environmental Modelling: Progress and Research Issues. GIS World Books, Fort. Collins.

GORDON. A.D. [1981]: Classification. New York: Chapman and Hall.

GOWER, J.C. [1969]: A survey of numerical methods useful in taxonomy. In: Acaralogia, 11, S. 357-375.

GRAUL, ADOLF [1992]: Neuronale Netze, Grundlagen und mathematische Modellierung. Wissenschaftsverlag, Mannheim.

GREEN, P.E.; CARMONE, JR. F.J.; SMITH, S.M. [1989]: Multidimensional Scaling, Concepts and Applications, Boston.

GREEN, P.E.; FRANK, R.E.; ROBINSON, P.J. [1967]: Cluster Analysis in Test Market Selection Management. Science 13, pp. B-387-B 400.

GRIEßHAMMER, R.; BUCHERT, M. [1996]: Nachhaltige Entwicklung und Stoffstrommanagement am Beispiel Bau. Freiburg.

GROB, H. L.; BENSBERG, F. [1999]: Das Data-Mining-Konzept, Arbeitsbericht Nr. 8, Münster.

GRONAU, N.; WEBER, E. [2005]: Analyse wissensintensiver Verwaltungsprozesse mit der Beschreibungssprache KMDL. In: KLISCHEWSKI, R.; WIMMER, M. (Hrsg.): Wissensbasiertes Prozessmanagement im E-Government, LIT (Münster), S. 171-183.

GRUHLER, K.; BÖHM, R.; DEILMANN, C.; SCHILLER, G. [2002]: Stofflich-energetische Gebäudesteckbriefe – Gebäudevergleiche und Hochrechnungen für Bebauungsstrukturen. IÖR Schriften, Band 38, Dresden.

GRÜNEWALD, WERNER [1982]: Klassifikation von Ländern nach ihrem demographischen Entwicklungsstand (Dissertation). Universität Bamberg.

GUNDRY, JOHN [2006]: Knowledge Ability.
http://www.knowab.co.uk/index.html, (Stand: 20.06.2006).

GÜßEFELDT, J. [1988]: Kausalmodelle in Geographie, Ökonomie und Soziologie. Eine Einführung mit Übungen und einem Computerprogramm. Springer, Berlin.

GUSTAFSON, E.E.; KESSEL, W.C. [1979]: Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE CDC, San Diego, Californien, pp. 761-766.

HABER, W. [1996]: Die ökologischen Grenzen menschlichen Handelns. In: Deutsche Bundesstiftung Umwelt (Hrsg.): Nachhaltigkeit 2000 – tragfähiges Leitbild für die Zukunft?. 1. Internationale Sommerakademie St. Marienthal. Bramsche, Rasch.

HADDEN, JEFFREY K.; BORGATTA, EDGAR F. [1965]: American cities: their social characteristics. McNally, Chicago.

HAFNER/WALDL [2001]: Statistik für Sozial- und Wirtschaftswissenschaftler Band 2 – Arbeitsbuch für SPSS und Microsoft Excel. Springer Wien/New York.

HÄGERSTRAND, TORSTEN [1976]: Innovation as a Spatial Process. Chicago University Press, Chicago, 1976.

HAHN, E. [1892, S. 8-12]: Die Wirtschaftsreformen der Erde. Geographische Mitteilungen 38, Petermann, Gotha.

HAND, DAVID; MANNILA, HEIKKI; SMYTH, PADHRAIC [2001]: Principles of Data Mining. MIT Press, Cambridge, Massachusetts.HÄNDEL, LARS [2003]: Clusterverfahren zur datenbasierten Generierung interpretierbarer Regeln unter Verwendung lokaler Entscheidungskriterien. Dissertation, Fakultät für Elektrotechnik und Informationstechnik, Universität Dortmund.

HANDL [2002]: Multivariate Analysemethoden (S-Plus). Springer Berlin.

HANNEMANN, CHRISTINE [2002]: Schrumpfende Städte: Überlegungen zur Konjunktur einer vernachlässigten Entwicklungsoption für Städte. In: Newsletter „Stadt 2030“.
http://www.newsletter.stadt2030.de/index6.htm (Stand: 05.01.2007).

HANNEMANN, CHRISTINE; KABISCH, SIGRUN; WEISKE, CHRISTINE (Hrsg.) [2002]: Neue Länder – Neue Sitten? Transformationsprozesse in Städten und Regionen Ostdeutschlands. Schelzky & Jeep, Berlin.

HANNEMANN, CHRISTINE [2004]: Marginalisierte Städte – Probleme, Differenzierungen und Chancen ostdeutscher Kleinstädte im Schrumpfungsprozeß. Wissenschafts-Verlag, Berlin.

HARRIS, CHAUNCY D. [1943]: A Functional Classification of Cities in the United States. Geographical Review, Vol. 33, No. 1, pp. 86-99. Graduate School of Geography, Clark University, Worcester.

HARRIS, CHAUNCY, D. [1972]: Cities of the Soviet Union. Monograph Series Number 5 of the Association of American Geographers, Second Printing, Rand McNally, Skokie, IL.

HART, J. [1955]: Functions and occupational structures of cities in the American south. In: Annals of the Association of American Geographers 45, pp.269-286, Blackwell Publishing, Oxford.

HARTIGAN, J.A. [1975]: Clustering Algorithms, John Wiley & Sons, New York.

HARTUNG, JOACHIM [2005]: Statistik – Lehr- und Handbuch der angewandten Statistik. 14. unwesentlich veränderte Auflage, Oldenbourg, München.

HASSINGER, H. [1916]: Kunsthistorischer Atlas der k. k. Reichshaupt- und Residenzstadt Wien, Band 15 der Reihe Österreichische Kunsttopographie. Schroll, Wien, 1916.

HASSLER, U. /KOHLER, N. [2004]: Das Verschwinden der Bauten des Industriezeitalters. – Lebenszyklen industrieller Baubestände und Methoden transdisziplinärer Forschung. Wasmuth Verlag, Tübingen.

HASSLER, U.; KOHLER, N; WANG, W. (edit) [1999]: Umbau – Über die Zukunft des Baubestandes. Wasmuth Verlag, Tübingen.

HÄUßERMANN; SIEBEL [1983]: Die Chancen des Schrumpfens. Plädoyer für eine andere Großstadtpolitik. In: Die Zeit Nr. 13/1983. Zeitverlag, Hamburg.

HÄUßERMANN; SIEBEL [1987]: Neue Urbanität. Suhrkamp Verlag, Frankfurt am Main.

HÄUßERMANN; SIEBEL [1988]: Die schrumpfende Stadt und die Stadtsoziologie. In: FRIEDRICHS, J. (Hrsg.), Soziologische Stadtforschung. S. 78-94. Westdeutscher Verlag, Opladen.

HÄUßERMANN; SIEBEL [1999]: Neue Urbanität. In: Raum Journal I, (S. 19-21), Schrift der Kulturregion Stuttgart e.V. (Hrsg.), Stuttgart.

HEBB, D.O. [1949]: The Organization of Behaviour: A Neuropsychological Theory. John Wiley & Sons, New York.

HEBER, B. / LEHMANN, I. [1996]: Beschreibung und Bewertung der Bodenversiegelung in Städten. IÖR Schriften 15, Leibniz-Institut für ökologische Raumentwicklung, Dresden.

HECKING, GEORG; MIKULICZ, STEFAN; SÄTTELE, ANDREAS [1988]: Bevölkerungsentwicklung und Siedlungsflächenexpansion. Schriftenreihe 15 des städtebaulichen Instituts der Universität Stuttgart, Karl Krämer Verlag.

HEDLUND, GUNNAR / NONAKA, IKUJIRO [1993, pp. 117-144]: Models of Knowledge Management in the West and Japan. In: LORANGE, PETER; CHAKRAVARTY, B.G.; ROOS, J. (editors.): Implementing Strategic Process: Change, Learning and Cooperation. Basil Blackwell, London.

HEINE, G.W. [1986]: A Controlled Study of Some Two-Dimensional Interpolation Methods. COGS Computer Contributions (Vol 2 - #2): 60-72. Computer Oriented Geological Society (Hrsg.).

HEMPEL, ARNE-JENS [2005]: Aggregation und Identifikation von Fuzzy-Objekten mit unterschiedlichen elementaren Unschärfen. Diplomarbeit, Technische Universität Chemnitz.

HEMPEL, C.G. / OPPENHEIM, P. [1948]: Studies in the Logic of Explanation. In: Readings in the Philosophy of Science, Vol. 15. University of Chicago.

HERBERTSON, A.J. [1905, P.300-310]: The major natural regions: An essay in systematic geography, geographical journal 25. Blackwell Publishing, Oxford.

HERZ; SCHLICHTER; SIEGENER [1992]: Angewandte Statistik für Verkehrs- und Regionalplaner. Werner Verlag Düsseldorf.

HEUER, JÜRGEN; KÜHNE-BÜNING, LIDWINA; NORDALM, VOLKER; DREVERMANN, MARLIS [1979]: Lehrbuch der Wohnungswirtschaft. Frankfurt am Main.

HILL, E.W.; BRENNAN, J.F.; WOLMAN, H.L. [1998]: What is a central city in the United States? Applying a statistical technique for developing taxonomies. In: Urban Studies 35, pp. 1935-1969. Taylor&Francis, Abingdon.

HILLMANN, M. [1996]: In Favour of the Compact City. In: JENKS, M.; BURTON, E.; WILLIAMS, K. [Eds.]: The Compact City – A Sustainable Urban Form. Spon Press, London.

HOBERG, RICHARD [2002]: Clusteranalyse, Klassifikation und Datentiefe (Dissertation). Universität Köln. Josef Eul Verlag Köln.

HOFMEISTER, B. [1996]: Die Stadtstruktur - ihre Ausprägung in den verschiedenen Kulturräumen der Erde. Wissenschaftliche Buchgesellschaft, Darmstadt.

HOFSTETTER, P. [1998]: Perspectives in Life Cycle Impact Assessment. A structured approach to combine models of technosphere, ecosphere and valuesphere. Kluwer, Boston.

HOLZKAMP, JOCHEN [1999]: Stoffstrommanagement „Bauen und Wohnen“ dargestellt am Beispiel der Wiederverwendung von Bauteilen. Dissertation. RWTH Aachen.

HÖPPNER, F.; KLAWONN, F.; KRUSE, R.; RUNKLER, T. [1999]: Fuzzy cluster Analysis. Aktualisierte Version der deutschen Ausgabe: HÖPPNER, F. [1997]. John Wiley & Sons, New York.

HÖPPNER, F.; KLAWONN, F.; KRUSE,R. [1997]: Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifikation und Datenanalyse. Vieweg, Braunschweig.

HUBER, J. [1995]: Nachhaltige Entwicklung - Strategien für eine ökologische und soziale Erdpolitik. Edition Sigma, Berlin.

HUBER, J. [1998]: Die Konsistenz-Strategie. Effizienz und Suffizienz alleine können Nachhaltigkeit nicht sichern. In: Politische Ökologie, 16 Jg., Sonderheft 11, S. 26-29.

HÜBLER, KARL-HERMANN; KAETHER, JOHANN (Hrsg.) [1999]: Nachhaltige Raum- und Regionalentwicklung – wo bleibt sie? Befunde, Perspektiven und Vorschläge. 1. Auflage. Verlag für Wissenschaft und Forschung, Berlin.

HYTTINEN, LAURA [2004]: Knowledge conversions in knowledge work - a descriptive case study. Licentiate Thesis. Espoo: Helsinki University of Technology. http://www.tuta.hut.fi/library/raportit/tps_report/pdf/raporttiLauraHyttinen1.pdf, (Stand: 16.06.2006).

INTERDEPARTEMENTALER AUSSCHUSS RIO IDARIO [1996]: Nachhaltige Entwicklung in der Schweiz – Bericht. Bern: Dokumentationsdienst Bundesamt für Umwelt, Wald und Landschaft (BUWAL).

INTERDEPARTEMENTALER AUSSCHUSS RIO IDARIO [1997]: Nachhaltige Entwicklung in der Schweiz – Stand der Realisierung. Bern: Dokumentationsdienst Bundesamt für Umwelt, Wald und Landschaft (BUWAL).

ISARD, WALTER [1960]: Methods of regional analysis: An introduction to regional science. Cambridge. Technology Press of the Massachusetts Institute of Technology.

ISENBERG, G. [1957]: Die Ballungsgebiete in der Bundesrepublik. Institut für Raumforschung, Vorträge 6, Bad Godesberg.

JANSSEN, J; LAATZ W. [2005]: Statistische Datenanalyse mit SPSS für Windows. 5. Auflage, Springer, Berlin.

JENKIS, H.W. (Hrsg.) [1996]: Raumordnung und Raumordnungspolitik. Oldenbourg Verlag, München.

JENKS, M.; BURTON, E.; WILLIAMS, K. (Eds.) [1996]: The Compact City – A Sustainable Urban Form. Spon Press, London.

JESSEN, J. [2005]: Städtebauliche Leitbilder – Entwicklungstendenzen. In: Stadtverkehrsplanung. Grundlagen, Methoden, Ziele. Hrsg. KÜNNE, G. / STEIERWALD, H.D. / VOGT, W., 2. Auflage, Springer Berlin / Heidelberg / New York.

JOHNSON, S.G. [1967]: Hierarchical Clustering Schemes, In: Psychometrika, Vol. 32, S.241-254.

JONES V.; FORESTALL, R. [1963]: Economic and social classification of metropolitan areas. In: The Municipal Yearbook (International City Managers Association, Chicago, IL), pp. 31-44.

KÁDAS, S. [1981]: Anwendung der Clusteranalyse zur Untersuchung regionaler Unterschiede der Industriestruktur und des sektoralen Strukturwandels in der Bundesrepublik Deutschland (=Arbeitspapier Nr. 33 Teilprojekt M1 des Sonderforschungsbereichs 26 Raumordnung und Raumwirtschaft Münster), Münster.

KAISER, H. F. [1958]: The Varimax Criterion for Analytic Rotation in Factor Analysis. In: Psychometrika 23 (ADKINS, Dorothy C., HORST, Paul, editors.), pp. 187–200, Springer, Heidelberg.

KANNING, HELGA [2000]: Umweltbilanzen: Instrumente einer zukunftsfähigen Regionalplanung? Dissertation, Universität Tübingen.

KAU, WIENAND [1970]: Theorie und Anwendung raumwirtschaftlicher Potentialmodelle. Dissertation, Universität Tübingen.

KEELER, E.; RODGERS, W. [1973]: A Classification of Large American Urban Areas. Rand Corporation. Santa Monica, CA.

KEIL, M.; MOHAUPT-JAHR, B.; KIEFL, R.; STRUNZ, G. [2002]: Das Projekt CORINE Land Cover 2000 in Deutschland. In: Tagungsband 19. DFD Nutzerseminar, 15.-16. Oktober 2002, Oberpfaffenhofen.

KEIM, Daniel, A. [2002, S.31]: Datenvisualisierung und Data Mining. In: Datenbank-Spektrum – Zeitschrift für Datenbanktechnologie. 2. Jahrgang, Heft 2, Februar 2002, dpunkt.verlag, Heidelberg.

KIERS; RASSON; GROENEN; SCHADER [2000]: Data Analysis, Classification, and related Methods. Springer Berlin/Heidelberg.

KILCHEMANN, A. [1968]: Untersuchungen mit quantitativen Methoden über die fremdenverkehrs- und wirtschaftsgeographische Struktur der Gemeinden im Kanton Graubünden (Schweiz). Dissertation, Zürich.

KILCHEMANN, A. [1970]: Statistisch/analytische Arbeitsmethoden in der regionalgeografischen Forschung. Arbor, Selbstverlag, University of Michigan.

KING, LESLIE J. [1966]: Cross Sectional Analysis of Canadian Urban Dimensions, 1951 and 1961, Canadian Geographer, Vol. 10, pp. 20-224, Blackwell Publishing, Malden.

KING, LESLIE J. [1967]: Discriminatory analysis of urban growth patterns in Ontario and Quebec, 1951-1961. Annals of the Association of American Geographers, Vol. 57, S. 566-578, Blackwell Publishing, Oxford.

KING, LESLIE J. [1969]: Statistical analysis in geography. Englewood Cliffs, N.J. (Prentice Hall).

KING, LESLIE J. [1970]: Discriminatory analysis: a review of recent theoretical contributions and applications. Economic Geography, Vol. 46, No. 2 (supplement).

KLANN, U.; SCHULZ, V. [2001]: Die Aktivitätsfeldanalyse auf Basis von Input-Output-Tabellen. In: GRUNDWALD, A.; COENEN, R; NITSCH, J.; SYDOW; A.; WIEDEMANN, O. (Hrsg.): Forschungswerkstatt Nachhaltigkeit. Auf dem Weg von der Diagnose zur Therapie. Global zukunftsfähige Entwicklung – Perspektiven für Deutschland. Berlin, S. 141-169.

KLAUA, D. [1966]: Grundbegriffe einer mehrwertigen Mengenlehre: In: Monatsberichte der Deutschen Akademie der Wissenschaften, Berlin-Ost, Nr. 8, S.782-802.

KLAWONN, FRANK; KRUSE, RUDOLF [1995]: Automatic Generation of Fuzzy Controllers by Fuzzy Clustering. Proc. 1995 IEEE International Conference on Systems, Man and Cybernetics, Vancouver (1995), 2040-2045.

KLAWONN, FRANK; KRUSE, RUDOLF [1997]: Constructing a fuzzy controller from data. Fuzzy Sets and Systems, 85:177-193.

KLEIN, KONRAD [1981]: Hierarchische Klassifikation einer Objektmenge (Dissertation). Universität Karlsruhe.

KNOP, WERNER [1989]: Bestand an Gebäuden und Wohnungen 1987 – Ergebnis der Gebäude- und Wohnungszählung. In: Wirtschaft und Statistik, Heft 8, S. 483-489.

KOBAYASHI, TAKESHI; SHINJI, IKARUGA; MAHITO, NAKAZONO [2005]: Study on Forecasting Urbanization by applying the area division system to the suburb in local cities. Computers in urban planning and urban management (CUPUM 05), 9th International Conference, UCL London.

KODRATOFF, Y.[1994]: Guest Editor's Introduction. In: AI Communications 7, pp. 83-85, IOS Press, Amsterdam.

KOHLER, N. [1994]: Energie- und Stofflussbilanzen von Gebäuden während ihrer Lebensdauer. Koordinationsgruppe des Bundes für Energie- und Ökobilanzen, Schlußbericht. Lausanne.

KOHLER, N. [2003] Cultural issues for a sustainable built environment. In: LORCH, R, and COLE, R. (edit.) Buildings, Culture and the Environment.

KOHLER, N.; HASSLER, U.; PASCHEN, H. (Hrsg.) [1999]: Enquete Kommission zum Schutz von Mensch und Umwelt des deutschen Bundestages. Konzept Nachhaltigkeit. Fundamente für die Gesellschaft von morgen. Springer-Verlag, Berlin.

KOHLER, N.; LÜTZKENDORF, TH. [2002]: Integrated Life Cycle Analysis. Building Research & Information 30(5), 338–348.

KOHONEN, T. [1982]: Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics Vol. 43, pp. 59-69.

KOHONEN, T. [1995]: Self-Organizing Maps. Springer, Berlin.

KOHONEN, T. [2001]: Self-organizing Maps. Bd. 30 von Information Sciences, 3. Aufl., Springer, Berlin.

KRESS, CELINA [2006]: Schrumpfungsprozesse versus Wachstumsparadigma in der DDR. Beitrag zum 36. Kolloquium des Instituts für vergleichende Städtegeschichte an der Universität Münster (IStG). Münster: Schrumpfende Städte in historischer Perspektive, 27. bis 29. März 2006.

KRESSLER, F.; STEINHOCKER, K [2001]: Monitoring urban development using satellite images. In: JÜRGENS, C. (Ed.): Remote Sensing of Urban Areas. Regensburger Geografische Schriften, Heft 35. Universtät Regensburg, Institut für Geographie.

KRISHNAPURAM, R.; KELLER, J. [1993]: A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems, pp. 98-100, IEEE Computer Society, Washington.

KROTZ, FRIEDRICH [1990]: Lebenswelten in der BRD. Leske+Budrich, Opladen.

KRUSKAL, J.B. [1964 a]: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesi.: In: Psychometrika (ADKINS, Dorothy C., HORST, Paul, editors.), Vol 29, März 1964, S. 1-27, Springer, Heidelberg.

KRUSKAL, J.B. [1964 b]: Nonmetric Multidimensional Scaling: A Numerical Method. In: Psychometrika (ADKINS, Dorothy C., HORST, Paul, editors.), Vol 29, Juni 1964, S. 115-129, Springer, Heidelberg.

KRUSKAL, J.B.; CARMONE, F.J. [1973]: How to use MDSCAL, a program to do Multidimensional Scaling and Multidimensional Unfolding (Version 5M). Bell Laboratories, Murray Hill, New York (vervielfältigtes Manual).

KUIPER, F.; FISHER, L.: A Monte Carlo Comparison of six Clustering Procedures. In: Biometrics, Vol. 31, 1975, S.777-783.

KULTURREGION STUTTGART E.V. (Hrsg.): Raum Journal I, 1999.

KUTTER, E. / HOLZ-RAU, C. [1994]: Städtebauliche Maßnahmen zur Verminderung von Verkehrserfordernis und Verkehrsindividualisierung. Forschungsvorhaben im Rahmen des Experimentellen Wohnungs- und Städtebaus. Endbericht. Berlin.

KÜBLER, HANS-DIETER [2005]: Mythos Wissensgesellschaft: Gesellschaftlicher Wandel zwischen Information, Medien und Wissen - eine Einführung. 1. Auflage, VS-Verlag für Sozialwissenschaften, Wiesbaden.

KÜNTZEL, THOMAS [2006]: Was ist eigentlich eine Stadtwüstung? Beitrag zum 36. Kolloquium des Instituts für vergleichende Städtegeschichte an der Universität Münster (IStG). Münster: Schrumpfende Städte in historischer Perspektive, 27. bis 29. März 2006.

KYTZIA, S. [1998]: Wie kann man Stoffhaushaltssysteme mit ökonomischen Daten verknüpfen? Ressourcen im Bau. T. LICHTENSTEIGER. Zürich, vdf: 69-79.

L.A. Zadeh [1973]: Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Trans. Systems, Man and Cybernetics, 3:28-44. IEEE Computer Society, Washington.

LANCE, G.N.; WILLIAMS, W.T. [1966]: A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems. The Computer Journal 9, S. 373-380.

LANE, ROBERT E. [1966]: The decline of politics and ideology in a knowledgeable society. In: American Sociological Review, 31, S. 649-662, Official Journal of the American Sociological Association, Washington.

LANGRAN, G. [1992]: Time in geographic information systems. Technical Issues in Geographical Information Systems. Taylor & Francis, New York.

LANG, THILO; TENZ, ERIC [2003]: Von der schrumpfenden Stadt zur Lean City. Vertrieb für Bau- und Planungsliteratur, Dortmund.

LARSEN, Kurt [1999]: Learning Cities: The new Recipe in regional development. In: OECD Observer, August, www.oecdobserver.org (Stand: 05.01.2007). Published by Organisation for Economic Co-operation and Development, Paris.

LAUKAT, A. [1999]: Vilfredo Pareto: Trattato di Sociologia Generale, In: Die Zeit Nr. 36/1999: “Friedhof der Eliten”. Zeitverlag, Hamburg.

LAURITZEN, S. [1996]: Graphical Models. Oxford University Press.

LAUSCHMANN, ELISABETH [1973]: Grundlagen einer Theorie der Regionalpolitik. Jänicke, Hannover.

LEUSCHNER, D. [1974]: Einführung in die numerische Taxonomie. Fischer Verlag, Jena.

LICHTENBERGER, ELISABETH [1998]: Stadtgeographie. B.G. Teubner, Stuttgart.

LICHTENSTIEGER, THOMAS [2006]: Bauwerke als Ressourcennutzer und Ressourcenspender in der langfristigen Entwicklung urbaner Systeme. vdf Hochschulverlag an der ETH Zürich.

LIEBMANN, HEIKE; ROBISCHON, TOBIAS (Hrsg.) [2003]: Städtische Kreativität. Potenzial für den Stadtumbau. IRS und Schader-Stiftung, Erkner/Darmstadt.

LINNÉ, CARL VON [1770]: Caroli Linnaei systema naturae, per regna tria naturae, secundum classes, ordines, genera, species. Vindobonae: Thoma de Trattern.

LIPPE, PETER v.d. [1990]: Wirtschaftsstatistik. 4. Auflage, Fischer Verlag, Stuttgart.

LITTLE, R.J.A.; RUBIN, D.A. [1987]: Statistical Analysis with missing data. John Wiley & Sons, New York.

LONGLEY, PAUL A.; BROOKS, SUE M.; MCDONNELL, RACHAEL; MACMILLAN, BILL (Hrsg.) [1998]: Geocomputation. John Wiley & Sons, Chichester, New York.

LONGLEY, PAUL A.; GOODCHILD M. F.; MAGUIRE D. J.; RHIND, D.W. [2001]: Geographic Information, Systems and Science. Chichester, New York.

LOSCH, S. [1997]: Der große Hunger. Landschaftsverbrauch in Deutschland – Anspruch und Wirklichkeit. Politische Ökologie, 15. Jg., Sonderheft 10, S. 27-32.

LO, C.P / YEUNG, Albert K. W. [2002]: Concepts and Techniques of Geographic Information Systems. Prentice Hall, New Jersey.

LUC, V. / SOILLE, P. [1991]: Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations, IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol. 13 (6), 583-598. Communication and Signal Processing Lab. at the University of Maryland.

LÜTKE-DALDRUP, E. [1989]: 88,9% des Bundesgebiets sind noch nicht bebaut! Ist das Ende des expansiven Stadtentwicklungsmodells absehbar? RaumPlanung, H.45, S. 85-95.

LÜTKE-DALDRUP, E. [2001]: Die perforierte Stadt. Eine Versuchsanordnung. In: Stadtbauwelt 150, Heft 24, S. 40-45, Bertelsmann Verlag, Frankfurt am Main.

LÜTZKENDORF, Th. [2002]: Nachhaltiges Planen, Bauen und Bewirtschaften von Bauwerken – Bewertungsmethoden und Hilfsmittel. Studie für das BMVBW im Auftrag des BBR.

LYNCH, Kevin [2001]: Das Bild der Stadt, unveränderter Nachdruck der Originalausgabe von 1965. Basel, Birkhäuser Verlag.

LYOTARD, JEAN-FRANCOIS [2005]: Das postmoderne Wissen. (Hrsg. von Peter Engelmann), 5. unveränderte Auflage, Passagen Verlag, Wien.

MAGGI, R. [1983]: Entwicklungsmöglichkeiten von Wintersportorten. St. Galler Beiträge zum Fremdenverkehr und zur Verkehrswirtschaft, Reihe Fremdenverkehr, 15. Dissertation, ETH Zürich.

MANAGEMENT INTELLIGENTER TECHNOLOGIEN GMBH (Hrsg.) [1998]: Plugin: Advanced Clustering – Benutzerhandbuch.

MANAGEMENT INTELLIGENTER TECHNOLOGIEN GMBH (Hrsg.) [2000]: Dokumentation DataEngine® Version 4.01.

MANHART, MICHAEL [1977]: Die Abgrenzung homogener städtischer Teilgebiete – Eine Clusteranalyse der Baublöcke Hamburgs. Christians, Hamburg.

MANN, S. [1983]: Ein Lernverfahren zur Modellierung zeitvarianter Systeme mittels unscharfer Klassifikation. Dissertationsschrift, Fakultät für Elektrotechnik / Informationstechnik, Technische Hochschule Karl-Marx-Stadt.

MARGRAF, OTTI [2005]: Vorlesungsunterlagen zur Multivariaten Statistik (Statistik II). Geografisches Institut, Abteilung Geoinformatik/Kartografie, Humbold-Universität Berlin.

MCLACHLAN, G.; BASFORD, K.E. [1989]: Mixture Models. Inference and Applications to Clustering. Marcel Dekker, New York.

MCLACHLAN, G.; KRISHNAN, T. [1997]: The EM Algorithm and Extensions. John Wiley & Sons, New York.

MCLACHLAN, G.; PEEL, D. [2000]: Finite Mixture Models. Wiley series in probability and statistics. John Wiley & Sons, New York.

MCQUITTY, L.L. [1957]: Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevances. In: Educational and Psychological Measurement, Vol. 17, S. 207-229.

MEYER, D. R. [1972, S. 61 ff.]: Classification of U.S. Metropolitan Areas by Characteristics of their Nonwhite Populations. In: BERRY Brian J.L.: City Classification Handbook - Methods and Applications. John Wiley & Sons, New York.

MILLIGAN, G.W. [1979]: Ultrametric Hierarchical Clustering Algorithms. In: Psychometrika, Vol. 44, S. 343-346.

MISNER, CHARLES W.; THORNE, KIP S.; WHEELER, JOHN ARCHIBALD [2000]: Gravitation, 23. Auflage, Freeman, New York.

MKRO [2006]: Ministerkonferenz für Raumordnung, http://www.bmvbw.de (Stand: 27.12.2006).

MOJENA, R. [1977]: Hierarchical Grouping Methods and Stopping-Rules: An Evaluation. In: The Computer Journal, Vol. 20, No. 4, S.359-363.

MÖLLERS, H. [1977]: Infrastrukturausstattung und Entwicklung von Städten – Methoden der multivariaten Analyse. Beiträge zum Siedlungs- und Wohnungswesen und zur Raumplanung, Band 42, Sonderdruck des Instituts für Siedlungs- und Wohnungswesen, Münster.

MOCK, J. [1993]: Ausgleich von Eingriffen in den Wasserhaushalt. In: Wasser und Boden. Heft 3 (1993), S. 148-151, Parey Buchverlag, Berlin.

MOON, TODD K. [1996]: The Expectation-Maximization Algorithm, IEEE Signal Processing Magazine, Vol. 11/96, Communication and Signal Processing Lab. at the University of Maryland.

MORAN, P.A.P. [1948]: The interpretation of statistical maps. In: Journal of the Royal Statistical Society B, Nr. 10, pp. 243-251, Blackwell Publishing, Oxford.

MOSER, C.A.; SCOTT, W. [1961]: British towns: A Statistical Study of Their Social and Economic Differences. Oliver and Boyd, Edinburgh.

MOTZKUS, A. [2000]: Zur Bedeutung der höherwertigen unternehmensorientierten Dienstleistungen für die Entwicklung der Metropolregionen Westdeutschlands. In: Raumforschung und Raumordnung, Heft 4, S. 265-275.

MÜLLER, A. v. [1997]: Denkwerkzeuge für Global Player. In: U. KRYSTEK, E. (Hrsg.): Internationalisierung. Eine Herausforderung für die Unternehmensführung. S. 465-473, Springer Verlag, Berlin.

MÜLLER, BERNHARD; SIEDENTOP, STEFAN (Hrsg.) [2003]: Schrumpfung – Neue Herausforderungen für die Regionalentwicklung in Sachsen/Sachsen-Anhalt und Thüringen. ARL-Arbeitsmaterial Nr. 303. Hannover.

MÜLLER, BERNHARD; SIEDENTOP, STEFAN (Hrsg.) [2004]: Wachstum und Schrumpfung in Deutschland – Trends, Perspektiven und Herausforderungen für die räumliche Planung und Entwicklung. In: Deutsche Zeitschrift für Kommunalwissenschaften (DfK), Band I, Deutsches Institut für Urbanistik (Difu), Berlin.

MUMFORD, LEWIS. [1970]: The Culture of Cities (Originalausgabe 1938). Thomson Learning, New York.

NAGLER, HEINZ; RAMBOW, RIKLEF; STURM, ULRIKE (Hrsg.) [2004]: Der öffentliche Raum in Zeiten der Schrumpfung. Bd. 8, edition stadt und region, Leue, Berlin.

NARR, W.D. [1971]: Theoriebegriffe und Systemtheorie. 2. Auflage, Kohlhammer Verlag, Stuttgart.

NELSON, H. [1955]: A Service classification of American cities. Economic Geography 31 - pp 189-210, Clark University, Worcester.

NETZBAND, M. [1998]: Möglichkeiten und Grenzen der Fernerkundung zur Versiegelungskartierung in Siedlungsräumen. Leibniz Institut für ökologische Raumentwicklung. IÖR-Schriften, Heft 28, Dresden.

NIEBUHR, ANNEKATRIN [2000]: Räumliche Wachstumszusammenhänge – Empirische Befunde für Deutschland. Hamburgisches Welt-Wirtschafts-Archiv (HWWA), Referatsbeitrag zum HWWA-Workshop: „Agglomerationen, Zentren und die Peripherie“ (02.12.1999).

NIPPER, J.; STREIT, U. [1977]: Zum Problem der räumlichen Enthaltensneigung in räumlichen Strukturen und raumvarianten Prozessen. In: Geographische Zeitschrift 65, S. 241-263, Franz Steiner Verlag, Stuttgart.

NONAKA, I.; TOYAMA R.; BYOSIÈRE, P. [2001]: A Theory of Organizational Knowledge Creation: Understanding the Dynamic Process of Creating Knowledge. In: DIERKES, M.; BERTHOIN ANTAL, A.; CHILD, J.; NONAKA, I. (Editors): Handbook of Organizational Learning and Knowledge. S. 491-517, Oxford University Press, New York.

NONAKA, I.; TAKEUCHI, H. [1995]: The Knowledge-Creating Company. How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, New York.

NONAKA, I.; TAKEUCHI, H. [1997]: Die Organisation des Wissens. Wie japanische Unternehmen eine brachliegende Ressource nutzbar machen. Campus Verlag, Frankfurt am Main.

NOYELLE, T. J.; STANBACK, T.M. [1983]: The Economic Transformation of American Cities. Rowan & Allanheld, Totowa, New Jersey.

ODUM, EUGENE P.[1999]: Ökologie: Grundlagen, Standorte, Anwendung. 3. Auflage. Thieme, - Stuttgart.

ÖKO-INSTITUT [1998]: Stoffflußbezogene Bausteine für ein nationales Konzept der nachhaltigen Entwicklung. Schlußbericht zum Vorhaben FKZ-Nr. 295 92 148 für das Umweltbundesamt; Darmstadt, Freiburg, Berlin.

OLIVER, M.A. [1990]: Kriging: A Method of Interpolation for Geographical Information Systems. International Journal of Geographic Information Systems 4. 313-332, Taylor & Francis, New York.

OPENSHAW, STAN; ABRAHART, R. J. [2000]: Geocomputation Putting. Taylor & Francis, London.

OSWALD, FRANZ; BACCINI, PETER [2003]: Netzstadt: Einführung in das Stadtentwerfen. Birkhäuser, Basel; Boston; Berlin.

OSWALT, PHILIPP (Hrsg.) [2004]: Schrumpfende Städte, Bd.1: Internationale Untersuchung. Hatje Cantz, Ostfildern-Ruit.

OSWALT, PHILIPP (Hrsg.) [2005]: Schrumpfende Städte, Bd. 2: Handlungskonzepte. Hatje Cantz, Ostfildern-Ruit.

OSWALT, PHILIPP / TIM RIENIETS (Hrsg.) [2006]: Atlas der schrumpfenden Städte. Hatje Cantz, Ostfildern-Ruit.

OTT, KONRAD. [2002]: Nachhaltigkeit des Wissens – was könnte das sein? Beitrag zum Kongress "Gut zu Wissen", aus: Heinrich-Böll-Stiftung(Hrsg.): Gut zu Wissen, Westfälisches Dampfboot. PDF-Manuskript: http://www.wissensgesellschaft.org/themen/wissensoekonomie/nachhaltigkeit.html, (Stand: 22.06.2006).

PÄßLER, M. [1998]: Mehrdimensionale Zeitreihenmodellierung und Prognose mittels Fuzzy Pattern Modellen. Diplomarbeit, Fakultät für Mathematik, TU Chemnitz.

PÄßLER, M. [1999]: Zeitreihenanalyse und –prognose mit Fuzzy Pattern Modellen, Professur für Systemtheorie, TU Chemnitz.

PEISER, R. [2001]: Decomposing Urban Sprawl. In: Town Planning Review, Vol. 72 (3), S. 275–298. University Press, Liverpool.

PESCHEL, GERALD [1991]: Klassifizierung geowissenschaftlicher Information. Sven von Loga Verlag.

PETRAK, J. [1997]: Data Mining - Methoden und Anwendungen. Technischer Report OEFAI-TR-97-15, Österreichisches Forschungsinstitut für Artificial Intelligence.

PEUQUET, D. / QUIAN, L. [1995]: An integrated database design for temporal Gis. In: KRAAK, M. - J.; MOLENAR, M.; FENDEL, E. M. (Eds.): Advances in GIS Research II: Proceedings of the Seventh International Symposium on Spatial Data Handling.

PIRKTL, LENNART [1983]: Probleme und Algorithmen der Clusteranalyse unter Berücksichtigung auf die landwirtschaftliche Typisierung (Dissertation). ETH Zürich.

POHL, CHRISTIAN [1997]: Unsicherheit und Ungenauigkeit in ökologischen Bewertungen : Nachbearbeitung des 3. Diskussionsforums Ökobilanzen vom 30. Oktober 1996 an der ETH Zürich.

POLANYI, MICHAEL [1966]: The Tacit Dimension, Doubleday, New York.

PREECE, J.; ROGERS, Y.; SHARP, H.; BENYON, D.; HOLLAND, S.; CAREY, T. [1994]: Human-Computer Interaction. Addision Wesley, Boston.

QU, WEIDONG [2000]: Zur Anwendung der Fuzzy-Clusteranalyse in der Grundstückswertermittlung (Dissertation). Hannover.

QUECKBÖRNER, SABINE [2004]: Was ist Data Mining? Seminarbeit zu Business Intelligence II Data Mining & Knowledge Discovery, Fakultät für Informatik, Universität Kaiserslautern.

QUINLAN, J. R. [1986]: Induction of Decision Trees. Machine Learning Vol. 1, Springer Netherlands, Dordrecht.

QUINLAN, J. R. [1997]. C5.0 and See 5: Illustrative examples. RuleQuest Research, http://ww.ruleguest.com. (Stand: 19.06.2006).

QUINLAN, J.R. [1993]: C4.5 - Programs for Machine Learning. Morgan Kaufmann, San Mateo, Californien.

RAO, C.R. [1952]: Advanced Statistical Methods in Biometric Research, New York.

Raumordnungsgesetz (ROG) vom 18. August 1997 (BGBl. I S. 2081, 2102), geändert durch Artikel 2 des Gesetzes zur Anpassung des Baugesetzbuches an EU-Richtlinien (Europarechtsanpassungsgesetz Bau – EAGBau) vom 24. Juni 2004 (BGBl. I S. 1359, 1379).

RAY, D. M.; MURDIE, R.A. [1972, S. 181 ff.]: Canadian and American Urban Dimensions. In: BERRY, Brian J.L.: City Classification Handbook - Methods and Applications. John Wiley & Sons, New York.

REIMER, U.[1991], Einführung in die Wissensrepräsentation. Teubner, Stuttgart.

REINBORN, DIETMAR [1996]: Städtebau im 19. und 20. Jahrhundert. Kohlhammer, Stuttgart.

REIß, J.; ERHORN, H.; KLUTTIG, H. [1999]: Hemmnisse bei der energetischen Altbaumodernisierung. Kann die Forschung Impulse geben? In: Bauphysik 21, Heft 3, S. 98-105.

REULECKE, JÜRGEN [1985]: Geschichte der Urbanisierung in Deutschland. 1. Auflage, Suhrkamp, Frankfurt am Main.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. (2000): Speaker verification using adapted gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41.

RHODENBURG, THOMAS [2003]: Klassifikation von Audio-Signalen. Fachbereich Physik/Elektrotechnik, Arbeitsbereich Nachrichtentechnik, Universität Bremen.

RIDD, M. / LIU, J. [1998]: A comparision of four algorithms for change detection in an urban environment. Remote Sensing of Environment, 63. Elsevier Science, Frankfurt am Main.

RIECHERS, RAINER [1977]: Analyse von Konjunkturzyklen – Eine Anwendung der Cluster-Analyse. Dissertation, Fachbereich Informatik, Technische Universität, Berlin.

RIFKIN, J. [2000]: Das Verschwinden des Eigentums. Campus Verlag, Frankfurt am Main (amerikanisches Original: The Age of Access, New York).

ROBINSON, G.M. [1998]: Methods and Techniques in Human Geography. John Wiley & Sons, New York.

RODOMAN, B.B. [1967]: Mathematical aspects of the formalization of regional characteristics. Soviet Geography: Review and Translation 8, pp. 687-708, University of Texas, Dallas.

ROG [2004]: Raumordnungsgesetz (ROG) vom 18. August 1997 (BGBl. I S. 2081, 2102), geändert durch Artikel 2 des Gesetzes zur Anpassung des Baugesetzbuches an EU-Richtlinien (Europarechtsanpassungsgesetz Bau – EAGBau) vom 24. Juni 2004 (BGBl. I S. 1359, 1379).

ROHLF, F.J. [1975]: Generalization of the Gap TEST for the Detection of Multivariate Outliers. Biometrics 31, S. 92-101.

ROIGER, R; GEATZ, M. [2003]: Data Mining – A Tutorial-Based Primer. Addison Wesley, Boston.

ROLLETT; BARTRAM (HRSG.) [1976]: Einführung in die hierarchische Clusteranalyse, Ernst Klett.

ROMMELFANGER, HEINRICH [1988]: Entscheiden bei Unschärfe – Fuzzy Decision Support-Systeme, Berlin.

ROMERO, A [2003]: Wissen und Kreativität. Zukunft München 2030. Schlussbericht des Teilprojekts „Wissen und Kreativität“ im Forschungsprojekt „Zukunft München 2030 – Visionen und Strategien für Stadt und Region“, München.
www.muenchen2030.de (Stand: 05.01.2007).

ROODMAN, D.M.; LENSEN, N. [1995]: A building revolution. How ecology and health concerns are transforming construction. Worldwatch Paper 124, Worldwatch Institute. Washington.

RUBIN, J. [1967]: Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem. In: Journal of Theoretical Biology, Vol. 15, S. 103-144.

RUSPINI, E.H. [1969]: A new Approach to Clustering. In: Information and Control, Vol. 15, S. 22-32.

RUSPINI, E.H. [1970]: Numerical Methods for Fuzzy Clustering. In: Informations Sciences, Vol. 2, S. 319 – 352.

SACHS, LOTHAR [2004, S. 551]: Angewandte Statistik – Anwendung statistischer Methoden. Elfte Auflage. Springer, Berlin.

SALLANDT [1987]: Determinanten des Mietniveaus auf regionalen Wohnungsmärkten. Beiträge zum Siedlungs- und Wohnungswesen und zur Raumplanung Band 119, Münster.

SCHAFER, J.L. [1997]: Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

SCHÄFFER, K.-A. [1969]: Klassifizierung landwirtschaftlicher Betriebe mit Hilfe multivariater statistischer Verfahren. S. 3 ff. Statistisches Amt der Europäischen Gemeinschaften – Agrarstatistik.

SCHMIDT-BLEEK; LIEDTKE, C.; BRINGEZU, S. (Hrsg.) [1996]: Bauen und Wohnen. Bausteine zum Schließen einer ökologischen Innovationslücke. Wuppertal.

SCHREIBER, G.; AKKERMANS, H.; ANJEWIERDEN, A.; DE HOOG, R.; SHADBOLT, N.; VAN DE VELDE, W.; WIELINGA, B. [2000]: Knowledge Engineering and Management - The CommonKADS Methodology. MIT Press, Cambridge, Massachusetts.

SCHUCHARD-FICHER; BACKHAUS; HUMME; LOHRBERG; PLINKE; SCHREINER [1985]: Multivariate Analysemethoden. Springer Verlag, Berlin.

SCHULZE, PETER M. [1980]: Region und Informationssystem. München.

SCHÜRT, ALEXANDER; SPANGENBERG, MARTIN; PÜTZ, THOMAS [2005]: Raumstruktur-typen Konzept – Ergebnisse – Anwendungsmöglichkeiten – Perspektiven. BBR Arbeitspapier, Bonn.

SCHWAIGER, BÄRBEL [2002]: Strukturelle und dynamische Modellierung von Gebäudebeständen. Universität Karlsruhe, Dissertation.

SCHWANINGER, MARKUS [1998]: Self-Organization and Self-Reference in the Cognition of Organizations. In: BRAITENBERG, VALENTINO; RADERMACHER, FRANZ-JOSEF (Editors): Interdisciplinary Approaches to a New Understanding of Cognition and Consciousness (Vigoni Conference). Universitäts-Verlag, Ulm.

SCOTT, D. W.; KEATING, J. P. [1999]: A Primer on Density Estimation for the Great Home Run Race of ‘98’. STATS 25 (Statistical Assessment Service), George Mason University, Washington. http://www.stats.org/about.htm, (Stand: 03.06.2006).

SCOTT, D.W. [1992]: Multivariate Density Estimation. John Wiley & Sons, New York.

SERUGENDO, G.; KARAGEORGOS, A.; RANA, O. F.; ZAMBONELLI, F. [2004]: Engineering Self-Organising Systems. Nature-Inspired Approaches to Software Engineering. Springer Verlag, Berlin.

SESTER, MONIKA [1995]: Lernen struktureller Modelle für die Bildanalyse. Dissertation, Fakultät für Bauingenieur- und Vermessungswesen, Universität Stuttgart.

SHEVKY, E.; BELL W. [1955]: Social area analysis: Theory, Illustrative Application, and Computational Procedures. Stanford University Press, Stanford.

SIBSON, R. [1981]: A Brief Description of Natural Neighbor Interpolation. Chapter 2 (21-36). In: Interpolation multivariate data, John Wiley & Sons, New York.

SIEDENTOP, STEFAN [2003]: Siedlungsentwicklung und Infrastrukturfolgekosten – Bilanzierung und Strategieentwicklung. Forschungsvorhaben im Auftrag des Bundesamtes für Bauwesen und Raumordnung sowie des Bundesministeriums für Verkehr, Bau- und Wohnungswesen. Zweiter Zwischenbericht. Dresden: Institut für ökologische Raumentwicklung e.V.

SIEDENTOP, STEFAN; KAUSCH, STEFFEN; EINIG, KLAUS UND GÖSSEL, JÖRG [2003]: Siedlungsstrukturelle Veränderungen im Umland der Agglomerationsräume. Bundesamt für Bauwesen und Raumordnung (Hrsg.), Heft 114, Bonn.

SIEDENTOP, STEFAN [2005 a, S. 23-35]: Urban Sprawl – verstehen, messen, steuern – Ansatzpunkte für ein empirisches Mess- und Evaluationskonzept der urbanen Siedlungsentwicklung. In: Disp 160 Netzwerk Stadt und Landschaft, ETH Zürich (Hrsg.), urban sprawl – Strategien und Instrumente einer nachhaltigen Flächenhaushaltspolitik. Zürich.

SIEDENTOP, STEFAN; SCHILLER, GEORG [2005 b, S. 83-93]: Infrastrukturfolgekosten der Siedlungsentwicklung unter Schrumpfungsbedingungen. In: Disp 160 Netzwerk Stadt und Landschaft, ETH Zürich (Hrsg.), urban sprawl – Strategien und Instrumente einer nachhaltigen Flächenhaushaltspolitik. Zürich.

SIEDENTOP, STEFAN; KAUSCH, STEFFEN; GUTH, DENNIS; STEIN, AXEL; WOLF, ULRIKE; LANZENDORF, MARTIN; HARBICH, RONNY; HESSE, MARKUS [2005 c]: Mobilität im suburbanen Raum. Neue verkehrliche und raumordnerische Implikationen des räumlichen Strukturwandels. Abschlussbericht zum Forschungsvorhaben 70.716 im Auftrag des Bundesministeriums für Verkehr, Bau- und Wohnungswesen (BMVBW).

SIEVERTS, THOMAS [1997]: Zwischenstadt zwischen Ort und Welt, Raum und Zeit, Stadt und Land. Birkhäuser Verlag, Basel, 1997, 1. Auflage.

SILVERMAN, B.W. [1986]: Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

SIMMIE, JAMES; LEVER, WILLIAM F. [2002]: Urban Studies – The knowledge based City. Taylor & Francis, London.

SITTERBERG, GEORG [1977]: Multivariate Analyse der Struktur und Entwicklung von Städten. (Dissertation). Universität Münster.

SMITH, R.H.T. [1965]: The Functions of Australian Towns. Tiejdschrift voor Economische en Soziale Geografie 56, pp. 81-92.

SNEATH, P.H.A. [1957]: The Application of Computers to Taxonomy. In: Journal of General Microbiology, Vol. 17, S. 201-226.

SOKAL, R.R.; MICHENER, C.D. [1958]: A Statistical Method for Evaluating Systematic Relationship. In: The University of Kansas Science Bulletin, Vol 38, S. 1409-1438.

SOKAL, R.R.; MICHENER, C.D. [1967]: The effects of different numerical technicals on the phenetic classification of bees of the hoplitis complex (megachilidae). Proc. Linn. Soc. London, 178, S. 59-74.

SOKAL, R.R.; SNEATH, P.H.A [1963]: Principles of Numerical Taxonomy. Freeman, W.H., San Francisco.

SOMBART, W. [1967]: Die drei Nationalökonomien. Geschichte und System der Lehre von der Wirtschaft. Zeller Verlag, Osnabrück.

SOMMER, E.; EMDE, W.; KIETZ, J.-U.; MORIK, K.; WROBEL, S. [1993]: Mobal 2.2 User Guide. German National Research Center for Computer Science (GMD), St. Augustin.

SPANGENBERG, MARTIN [2001]: Regionales Bevölkerungspotential. In: BBR (Hrsg.) [2003]: Informationen aus der Forschung des BBR – Nr.6/Dezember 2006, Selbstverlag, Bonn.

SPÄTH, H. [1975]: Cluster-Analyse-Algorithmen zur Objektklassifikation und Datenreduktion. Oldenbourg Verlag, München, Wien.

SPÄTH, H. [1977]: Fallstudien zur Clusteranalyse. Oldenbourg Verlag, München, Wien.

SPEHL, HARALD: Nachhaltige Entwicklung als Herausforderung für Raumordnung, Landes- und Regionalplanung. In: ARL (Hrsg.): Nachhaltige Raumentwicklung. Szenarien und Perspektiven für Berlin/Brandenburg, Hannover 1998, S. 19-33.

SPINNER, HELMUT F. [1998]: Die Architektur der Informationsgesellschaft: Entwurf eines wissensorientierten Gesamtkonzepts. Philo-Verlag, Bodenheim.

STAACK, JÖRG [1995]: Die Klassifikation deutscher Städte nach ihrer regionalen Zentralität. Dissertation, Universität Hamburg. Europäische Hochschulschriften, Peter Lang Verlag, Frankfurt am Main.

STADEL, WERNER [2006]: Lineare Regression. Unterlagen zum Block Rg1 des Kurses in Angewandter Statistik an der ETH Zürich.
http://stat.ethz.ch/~stahel/courses/regression/reg1-script.pdf (Stand: 16.06.2006).

STADTUMBAU OST [2002]: Internetseite der Bundestransferstelle des Stadtumbau Ost, http://www.stadtumbau-ost.info/ (Stand: 01.01.2007).

STADTUMBAU WEST [2004]: Internetseite des ExWoSt-Forschungsfeldes Stadtumbau West, http://www.stadtumbauwest.info/ (Stand: 01.01.2007).

STATISTISCHES BUNDESAMT (Hrsg.) [1997]: Volkswirtschaftliche Gesamtrechnung, Fachserie 18, Reihe 2: Input-Output-Tabellen 1993. Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [1950]: Statistik von Baden-Württemberg - Band 6, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [1961]: Statistik von Baden-Württemberg - Band 88, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [1968]: Statistik von Baden-Württemberg - Band 161, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [1987]: Statistik von Baden-Württemberg - Band 403 ff., Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [2001]: Statistik von Baden-Württemberg - Band 570 Heft 3, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [2002]: Statistische Berichte 20.08.2002 – Preise, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [2002]: Mikrozensus-Zusatzerhebung 2002, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [2003]: Statistisches Monatsheft 8/2003, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [2003]: Statistische Berichte 12.06.2003 – Bevölkerung und Erwerbstätigkeit, Stuttgart.

STATISTISCHES LANDESAMT Baden-Württemberg (Hrsg.) [2003]: Landesinformationssystem Baden-Württemberg (LIS), www.statistik-bw.de, Stuttgart.

STEINBUCH, KARL [1968]: Falsch programmiert: Über das Versagen unserer Gesellschaft in der Gegenwart und vor der Zukunft und was eigentlich geschehen müsste. DVA, Stuttgart.

STEINER, D. [1975]: Geographische Raumgliederung und Mustererkennung. Publikationen des Geographischen Instituts der ETH-Zürich, Vol. 55, S. 19-45, Zürich.

STEINHAUSEN; LANGER [1977]: Clusteranalyse – Einführung in Methoden und Verfahren der automatischen Klassifikation. De Gruyter Berlin.

STEPHAN, ACHIM [1997]: Emergenz – Historisch-systematische Studien zu einem zentralen Begriff der Metaphysik und Wissenschaftsphilosophie. Habilitationsschrift der Fakultät für Geistes- und Sozialwissenschaften der Univeristät Fridericiana zu Karlsruhe (TH).

STEPHAN, ACHIM [1999]: Emergenz – Von der Unvorhersagbarkeit zur Selbstorganisation. University Press, Dresden.

STEWART, J.Q. [1947]: Empirical Mathematical Rules Concerning the Distribution of Equilibrium of Population. Geographical Review 37, Graduate School of Geography, Clark University, Worcester.

STEWART, T.A. [1998]: Der vierte Produktionsfaktor – Wachstum und Wettbewerbsvorteile durch Wissensmanagement. Hanser Verlag, München.

STÖLZEL, M [1999]: Clusteranalyse zur Erkennung kohärenter Strukturen in Umweltdaten. (Diplomarbeit) GKSS Forschungszentrum GKSS99/E/47 (Hrsg.).

STRASSERT, G. [1975]: Regionale Kennziffern. Veröffentlichungen der Akademie für Raumforschung und Landesplanung, Forschungs- und Sitzungsberichte 105.

STREICH, BERND [2005]: Stadtplanung in der Wissensgesellschaft : ein Handbuch. 1. Auflage, VS Verlag für Sozialwissenschaften, Wiesbaden.

STREICH, B. [1987]: Der Einfluß neuer Technologien auf Flächenbedarf und Flächeninanspruchnahme. In: Forschungs- und Sitzungsberichte der Akademie für Raumforschung und Landesplanung (ARL), Band 173, zum Thema: Flächenhaushaltspolitik – Ein Beitrag zum Bodenschutz. Hannover.

STREIT, U. [2000]: Interoperable, Offene Geowissenschaftliche Informationssysteme, GIS Geo-Informations-Systeme. Zeitschrift für raumbezogene Information und Entscheidung 2/2000, Wichmann Verlag, Heidelberg.

STREIT, U. / WIESMANN, K. [1996]: Problems of integrating GIS and hydrological models. In: FISCHER, M.; SCHOLTEN, H.J.; UNWIN, D. (Eds.): Spatial Analytical Perspectives on GIS. Taylor & Francis, London.

STROBEL, MARKUS [2000]: Systematisches Flussmanagement (Dissertation). Ziel Verlag, Augsburg.

STROEKER, ELISABETH [1992]: Einführung in die Wissenschaftstheorie. 4. Auflage, Wissenschaftliche Buchgesellschaft, Darmstadt.

TAT INGENIEURBÜRO DR. JÖRN BURMEISTER (Hrsg.) [1997]: Fuzzy Pattern Classification Plug-In für Data-Engine®. 4. überarbeitete Auflage, Berlin.

THEUSS, MARTIN [2004] – Vorlesungsunterlagen zu Multivariaten Statistischen Verfahren, Universität Augsburg, Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse.

THIEME, KARL [1984]: Wohnungsbestand und Stadtentwicklung – Verwendung der Clusteranalyse zur Beurteilung der Wohnsituation in Augsburg (Diplomarbeit). Universität Augsburg.

THINH, N.X.; ARLT, G.; HEBER, B.; HENNERSDORF, J.; LEHMANN, I. [2000]: GIS-basierte Ableitung von funktionsräumlichen Kenngrößen für kreisfreie Städte in Deutschland. In: GNAUCK, A. (Hrsg.): Umweltforschung und Umweltinformatik. Cottbus: BTU, Aktuelle Reihe 7/2000, S. 61-72.

THINH, NGUYEN XUAN [2002 a]: Entwicklung von AML-Programmen zur räumlichen Analyse der Flächennutzungsmuster von 116 kreisfreien Städten in Deutschland. In: Photogrammetrie Fernerkundung Geoinformation 6/2004. E. Schweitzerbartsche Verlagsbuchhandlung (Nägele und Obermiller), Stuttgart.

THINH, NGUYEN XUAN [2002 b]: Evaluation of urban land-use structures with a view to sustainable development. In: Environmental Impact Assessment Review 22. Elsevier Science, Frankfurt.

THINH, NGUYEN XUAN [2004 a]: Entwicklung von mathematisch-geoinformatischen Methoden und Modellen zur Analyse, Bewertung, Simulation und Entscheidungsunterstützung in Städtebau und Stadtökologie. Habilitationsschrift der Universität Rostock, Fakultät der Agrar- und Umweltwissenschaften.

THINH, XUAN NGUYEN [2004 b]: Entwicklung von Maßen zur Charakterisierung und Bewertung der physischen und funktionalen Kompaktheit von Stadtregionen. In: Photogrammetrie Fernerkundung Geoinformation 6/2004. E. Schweitzerbartsche Verlagsbuchhandlung (Nägele und Obermiller), Stuttgart.

THINH, NGUYEN XUAN; BEHNISCH, MARTIN; ULTSCH, ALFRED [2006]: Examination of several results of different cluster analysis with a separate view to balancing the economic and ecological performance potential of towns and cities. In: BOCK, H.H., GAUL, W., VICHI, M. (Hrsg.): Studies in Classification, Data Analysis, and Knowledge Organization. Beitrag zu den Proceedings der 30. Jahrestagung der Gesellschaft für Klassifikation. Springer, Freiburg.

THINH, NGUYEN XUAN; SCHUMACHER ULRICH; GEIER, KATRIN [2007]: Modellierung und Bewertung der Ökoeffizienz von Siedlungsstrukturen – Ein pragmatischer Ansatz. Shaker Verlag, Aachen.

THORNWAITE, C.W. [1933]: The climates on the earth. Geographical Review, Vol. 23, pp. 433-440, Graduate School of Geography, Clark University, Worcester.

THÜNEN, J.H. [1826]: Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie oder Untersuchung über den Einfluß, den die Getreidepreise, der Reichthum des Bodens und die Abgaben auf den Ackerbau ausüben. Perthes, Hamburg.

THURSTONE, L.L [1931]: Multiple factor analysis. Psychological Review, 38, pp. 406-427, published by American Psychological Association, Washington.

THUVANDER, LIANE [2002]: Towards Environmental Informatics for Building Stocks. Chalmers University of Technology, Göteborg, 2002. ISBN 91-7291-242-1.

TIMM, HEIKO [2002]: Fuzzy-Clusteranalyse: Methoden zur Exploration von Daten mit fehlenden Werten sowie klassifizierten Daten (Dissertation). Otto-von-Guericke-Universität Magdeburg.

TOBLER, W. [1979]: Cellular Geography. In: S. Gale and G. Olsson, Philosophy in Geography, pp. 379-86, Reidel, Dordrecht.

UBA, Umweltbundesamt (Hrsg.) [2004]: Workshop CORINE Land Cover 2000 in Germany and Europe and its use for environmental applications, 20-21 January 2004, Berlin. UBA Texte 04/04, Berlin.

ÜBERLA, KARL [1971]: Faktorenanalyse - eine systematische Einführung für Psychologen, Mediziner, Wirtschafts- und Sozialwissenschaftler. 2. Auflage, Springer, Berlin.

ULLMANN, E.L.; DACEY, M.F. [1960]: The Minimum Requirements Approach to the Urban Economic Base. In: Proceedings of the IGU Symposium in UrbanGeography, Lund. Studies in Geography, B, 24 [1962, S. 121-143].

ULTSCH, A. [1991]: Konnektionistische Modelle und ihre Integration mit wissensbasierten Systemen. Habilitationsschrift der Universität Dortmund.

ULTSCH, A. [1993]: Self-organizing Neural Networks for Visualization and Classification, Information and Classification. Berlin, Springer-Verlag, pp. 307-313.

ULTSCH, A. [1999]: Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series, Kohonen Maps, pp. 33-46.

ULTSCH, A. [2000]: The Neuronal Data Mine. In: Proceedings 2nd Int. ICSC Symposium on Neural Computation NC, Berlin.

ULTSCH, A. [2001]: Eine Begründung der Pareto 80/20 Regel und Grenzwerte für die ABC Analyse. Technical Reports No. 30, Dept. of Mathematics and Computer Science, University of Marburg, Germany.

ULTSCH, A. [2003 a]: Pareto Density Estimation: A Density Estimation for Knowledge Discovery, BAIER D., WERNECKE K.D. (Eds), In: Innovations in Classification, Data Science, and Information Systems - Proceedings 27th Annual Conference of the German Classification Society (GfKL) 2003. Berlin, Heidelberg, Springer, pp. 91-100.

ULTSCH, A. [2003 b]: Maps for the Visualization of high-dimensional Data Spaces, In: Proceedings Workshop on Self-Organizing Maps (WSOM 2003), Kyushu, Japan, pp. 225-230.

ULTSCH, A. [2003 c]: U*-Matrix: a Tool to visualize Clusters in high dimensional Data: Technical Report No. 36, Dept. of Mathematics and Computer Science, University of Marburg, Germany.

ULTSCH, A. [2005 a]: Clustering with SOM U*C. In: Proceedings Workshop on Self-Organizing Maps (WSOM 2005), Paris, France, pp. 75-82.

ULTSCH, A. [2005 b]: U*C: Self-organized Clustering with Emergent Feature Map. In: Proceedings Lernen, Wissenentdeckung und Adaptivität (LWA / FGML / 2005), Saarbrücken, pp.240-246.

ULTSCH, A. [2006 a]: Datenbionik: Erklärung zu Emergenten Selbst-Organisierenden Merkmalskarten (ESOM), Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, http://www.mathematik.uni-marburg.de/~databionics/de//?q=esom, (Stand 03.05.2006).

ULTSCH, A. [2006 b]: Datenbionik: Software zur Berechnung der Emergenten Selbst-Organisierenden Merkmalskarten (ESOM), Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, http://www.mathematik.uni-marburg.de/~databionics/de//?q=software, (Stand 03.05.2006).

ULTSCH, A. [2006 c]: Vorlesungsunterlagen: Knowledge Discovery, Fachbereich Mathematik und Informatik, Philipps-Universität Marburg.

UMWELTBUNDESAMT (Hrsg.) [1999]: Betriebliche Umweltauswirkungen. Ihre Erfassung und Bewertung im Rahmen des Umweltmanagements. Berlin.

VOGEL, FRIEDRICH [1975]: Probleme und Verfahren der numerischen Klassifikation. Vandenhoeck & Ruprecht, Göttingen.

VOGEL, FRIEDRICH [1977]: Subjektivität bei der Klassifikation von Einheiten. In: Proceedings in Operations Research, Vol. 7, S. 105-129. Physica Verlag, Würzburg.

WALKER, P. / MOORE, D. [1988]: SIMPLE: An inductive modelling and mapping tool for spatially-oriented data. In: International Journal of Geographical Information Systems 2, pp. 347–363. Taylor&Francis, Abingdon.

WALLACE, D. L. [1968]: Clustering. International Encyclopedia of the Social Sciences, Elsevier, Oxford.

WARD, J.H. [1963]: Hierarchical Grouping to optimize an Objective Function. In: JASA (Journal of the Acoustic Society of America), Vol. 58, No. 301, 1963, S.236-244. American Institute of Physics, New York.

WASSMER, R.W. [2002]: Defining Excessive Decentralization in Californial and Other Western States. An Economist's Perspective on Urban Sprawl, Part 1. Sacramento, CA: California Senate Office of Research.

WATSON, D.F. / PHILIP, G.M. [1985]: A Refinement of Inverse Distance Weighted Interpolation. Geoprocessing, Vol. 2, 315-327. ACM Press, New York.

WEBER, A. [1909]: Über den Standort der Industrien. Teil 1: Reine Theorie des Standortes. Mohr, Tübingen.

WEBER, R.; CRAIG, J. [1978]: Socio-economic Classification of Local Authorities Areas Studies on Medical and Population Subjects 35, Office of Population Censuses and Surveys, London.

WEE, W.G. [1967]: On Generalization of Adaptive Algorithms and Applications of Fuzzy Sets - Concepts to Pattern Classification. Ph.D. Thesis, Purdue University Lafayette.

WEIGER, H. [1997]: Land in Sicht! Schutzkonzepte, die dem Flächenverbrauch den Boden entziehen. Politische Ökologie, 15 Jg., Sonderheft 10, S. 101-104, oekom Verlag, München.

WEISKOPF, LEONHARD [1984]: Die Beschreibung regionaler Strukturen durch Kennzahlen und deren Benutzbarkeit in der Clusteranalyse (Dissertation). Universität Augsburg.

WEISKE, CHRISTINE; KABISCH, SIGRUN; HANNEMANN, CHRISTINE (Hrsg.) [2005]: Kommunikative Steuerung des Stadtumbaus. Verlag für Sozialwissenschaften, Wiesbaden.

WEIZÄCKER, E.U. VON; LOVINS, A.B.; LOVINS, L.H. [1995]: Faktor vier. Doppelter Wohlstand – halbierter Naturverbrauch. Der neue Bericht an den Club of Rome. München.

WERSIG, GERNOT [1985]: Die kommunikative Revolution: Strategien zur Bewältigung der Krise der Moderne. Westdeutscher Verlag, Opladen.

WETHERILL, G. [1986]: Regression Analysis with Applications. Number 27 in Monographs on Statistics and Applied Probability, Chapmann and Hall, London.

WIEGANDT, C.-C. [1999]: Nachhaltige Siedlungspolitik in Deutschland – Beitrag des Bundesamtes für Bauwesen und Raumordnung. In: BERGMANN, A.; EINIG, K.; HUTTER, G. u.a. (Hrsg.): Siedlungspolitik auf neuen Wegen. Steuerungsinstrumente für eine ressourcenschonende Flächennutzung, S. 83-99. Berlin.

WIENER, NORBERT [1948]: Cybernetics or Control and Communication in the Animal and the Machine. MIT Press, Cambridge, Massachusetts. (deutsche Ausgabe von WIENER, NORBERT [1963]: Kybernetik. Regelung und Nachrichtenübertragung im Lebewesen und in der Maschine. 1. Auflage, Econ Verlag, Düsseldorf.).

WILKE, HELMUT [1998]: Organisierte Wissensarbeit. In: Zeitschrift für Soziologie 27, 3, S. 161-177, Lucius & Lucius Verlag, Stuttgart.

WILLIAMS, K.; BURTON, E.; JENKS, M. [2001]: Achieving Sustainable Form. Spon Press, London.

WISHART, D.: Clustan [1978] – User Manual, Edinburgh (= Inter University/Research Councils Series, Report No. 47).

WONG, DAVID WING SHUN / LEE, DAVID WONG JAY [2005]: Statistical Analysis of Geographical Information with ArcView GIS® and ArcGIS®. John Wiley & Sons, New Jersey.

WOODS, E.; KYRAL, E. [1997]: Ovum Evaluates: Data Mining. OVUM, London.

WORBOYS, M.F. [1998]: A generic model for Spatio-Temporal Modelling. In: Spatial and temporal reasoning in geographic information systems. Oxford University Press.

WORLD COMMISSION ON ENVIRONMENT AND DEVELOPMENT [WCED]; HAUFF, VOLKER (Hrsg.): Unsere gemeinsame Zukunft. Der Brundland-Bericht der Weltkommission für Umwelt und Entwicklung. Greven 1987.

WUTTKE, MICHAEL [2003]: Stadtumbau in Ostdeutschland – Eine Untersuchung zur Stadt Schwedt/Oder. Diplomarbeit an der Humboldt-Universität zu Berlin, Philosophische Fakultät III, Institut für Sozialwissenschaften (Prof. Dr. H. Häußermann / Prof. Dr. Gläßner).

WÜEST, HANNES; GABATHULER, CHRISTIAN [1989]: Bauwerk Schweiz, Eigenverlag Zürich.

YUAN, M. [1996]: Temporal GIS and Spato-Temporal Modelling. In: Proceedings of the 3rd International Conference / Workshop on Integrating GIS and Environmental Modelling, Santa Fe, USA.

ZADEH, L.A. [1965]: Fuzzy Sets. In: Information and Control, Vol. 8, S.338-353.

ZADEH, L.A. [1975]: Vorwort zu A. KAUFMANN: Introduction to the Theory of Fuzzy Subsets. Academic Press New York, San Francisco, London.

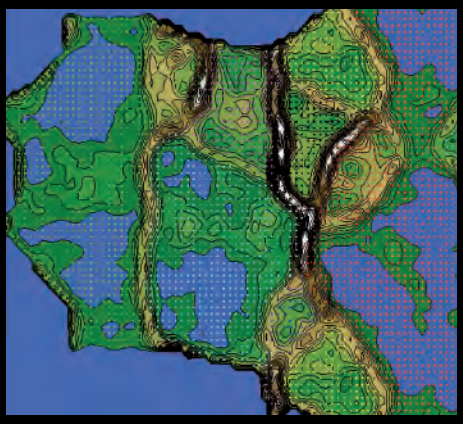
ZECK, HILDEGARD [2003]: Zentrale Orte als räumliches Konzept für Anpassungsstrategien. In: Information zur Raumentwicklung, Heft 12 / 03, S. 725 ff. Selbstverlag des Bundesamtes für Bauwesen und Raumordnung (BBR), Bonn.

ZIMMERMANN (Hrsg.) [1993]: FUZZY Technologien – Prinzipien, Werkzeuge, Potentiale. VDI-Verlag, Düsseldorf.

ZIMMERMANN (Hrsg.) [1995a]: Datenanalyse – Anwendung von DataEngine mit Fuzzy Technologien und Neuronalen Netzen. VDI Verlag, Düsseldorf.

ZIMMERMANN (Hrsg.) [1995b]: Neuro+Fuzzy – Technologien – Anwendungen. VDI-Verlag, Düsseldorf.

Durch den schnellen Fortschritt in der Informationstechnologie und das rapide Anwachsen raumbezogener Daten steigen die Anforderungen an Systeme, die Wissen aus diesen Daten extrahieren und darstellen. ‚Urban Data Mining' wird als Methodik zur Problemlösung verstanden, um logische oder mathematische, zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen zu entdecken. Auf der Grundlage von bestehenden Methoden des Data Mining und der Knowledge Discovery wird ein für die Stadt- und Regionalforschung strukturiertes methodisches Arbeitskonzept erarbeitet und am deutschen Gemeindesystem empirisch-analytisch vorgestellt. Neben Methoden, die eine kritische Bestandsaufnahme und Auseinandersetzung mit vorhandenen räumlichen Eigenschaften und Entwicklungstendenzen ermöglichen, werden Vorgehensweisen gesucht, die sich eignen, bereits vorhandene Informationen oder Erkenntnisse auf weitere Objekte zu übertragen.



ISBN: 978-3-86644-249-8

www.uvka.de