Abschlussbericht zum DFG-Forschungsprojekt
„Dynamische Gebäudebestandsklassifizierung“

1 Allgemeine Angaben

1.1 DFG-Geschäftszeichen

BR 3513/1-1

1.2 Antragsteller

Dr. rer. nat. Patrick Erik Bradley

1.3 Institut/Lehrstuhl

Universität Karlsruhe (TH)

Institut für Industrielle Bauproduktion	01.01.2007 – 31.12.2008
Institut für Photogrammetrie und Fernerkundung	01.01.2009 – 28.02.2009

1.4 Aus DFG-Mitteln bezahlte wissenschaftliche Mitarbeiter mit Angabe des Beschäftigungszeitraums

Dr. rer. nat. Patrick Erik Bradley	01.01.2007 – 31.12.2008

1.5 Thema des Projekts

Analyse und Modellierung des raumzeitlichen Verhaltens des deutschen Gebäudebestandes unter Einsatz mathematischer und geoinformatischer Methoden

1.6 Berichtszeitraum, Förderungszeitraum insgesamt

01.01.2007 – 28.02.2009

1.7 Fachgebiet, Arbeitsrichtung

Architektur, Mathematik

Arbeitsrichtung: Gebäudebestandsmodellierung, Geoinformatik, nichtarchimedische Geometrie, Datamining

1.8 Verwertungsfelder

Gebäudebestands-, Liegenschaftsverwaltung, Maschinelles Lernen

1.9 Am Projekt beteiligte Kooperationspartner

Dr. Ing. Martin Behnisch (Inst. f. Denkmalpflege und Bauforschung, ETH Zürich)

Dr. Ing. Norbert Paul (Lehrstuhl für Computation in Engineering, TU München)

2 Zusammenfassung

2.1 Darstellung der wesentlichen Ergebnisse

Das Wissen über den deutschen Gebäudebestand ist trotz seiner enormen gesellschaftlichen Bedeutung erstaunlich gering. Um Kenntnisse über seine Zusammensetzung im Laufe der Zeit zu erhalten, wurden neuartige Klassifkationsmethoden entwickelt und mit deren Hilfe seine Dynamik untersucht. Daten zu Urbanisierung und Diversität von Gemeinden in ihrer zeitlichen Entwicklung wurden aus verschiedenartigen Quellen herangezogen. Zur Strukturerkennung in den Daten und zur Wissenskonversion wurden Emergente Selbst-Organisierende Merkmalskarten eingesetzt, Klassifikationen durchgeführt und Klassifikatoren entwickelt. Räumliche Informationen wurden zwischen verschiedenen, hinreichend ähnlichen Objekten übertragen. Dies führte zu einer erstmaligen gemeindescharfen Schätzung der Zusammensetzung und Größe des deutschen Gebäudebestandes, deren Genauigkeit nicht allein über die amtlichen Statistiken erreicht werden kann. Geschätzte Übergangswahrscheinlichkeiten zwischen den Nutzungen ermöglichte den Aufbau von Markovketten zur Simulation der Altersverteilung und des Schwundes im Nichtwohngebäudebestand Baden-Württembergs.

Konzeptionell wurde der Einsatz der nichtarchimedischen Geometrie in der hierarchischen Klassifikation bis hin zur Entwicklung von Algorithmen untersucht. Der Hauptvorteil der p-adischen Klassifikation ist ihre hohe Effizienz und die Eindeutigkeit des von Daten abgeleiteten Dendrograms. Beides steht im Gegensatz zur klassischen, archimedischen Situation. Dafür erforderlich ist eine binäre — oder allgemeiner: p-adische — Datenkodierung, wobei die Klassifikationsergebnisse von der Wahl einer Kodierung abhängen. Die neuen p-adischen Methoden wurden auf Projektdaten zum exemplarischen Vergleich der Dynamik urbaner Gebäudebestände angewandt.

Zur Modellierung raumzeitlicher Gebäudedaten in relationalen Datenbanken und ihrer geoinformatischen Verarbeitung wurden erstmals relationale Kettenkomplexe eingesetzt. Dieses dient als ein erstes Anwendungsbeispiel topologischer Datenbanken für den Aufbau eines 3D-/4D-Geoinformationssystems im Gebäudebestand.

2.2 Ausblick auf künftige Arbeiten und Beschreibung möglicher Anwendung

In der Zukunft sollten im Urbanen Data Mining charakteristische Kenngrößen für die Dynamik von Flurstücken (oder kleinerer Gebiete) zum effizienten Erfassungsaufwand von Longitudinaldaten längerer Zeiträume erarbeitet werden. Hinzu kommen sollte der detaillierte Vergleich der Dynamik des verschwundenen Teils der gebauten Umwelt mit der des bestehenden Teils, um so für Ersteren wenigstens eine indirekte Schätzung ermöglichen zu können. Dies gilt insebsondere im Hinblick auf die Frage, wieviel des früher Gebauten noch heute steht. Dabei sollten verschiedene Methoden der Datengewinnung und Wissenserzeugung Bedeutung erlangen, um auch die Projektergebnisse zu validieren.

Für die nichtarchimedischen Klassifikationsmethoden gilt es, in der Zukunft, die existierenden statistischen Klassifikationsalgorithmen um ihre p-adischen Entsprechungen zu erweitern, wobei die p-adische Analysis die entsprechenden Konzepte liefern sollte. Es wird dann zu prüfen sein, inwieweit ihr Einsatz im Data Mining der (urbanen) dynamischen Prozesse in einer Vielzahl von Situationen zu einer effizienteren Wissensgewinnung beiträgt. Zu einer vollständigen Klärung dieser Frage wird im allgemeinsten Fall vermutlich das p-adische Kodierungsproblem zu lösen sein.

Schließlich wird in der Zukunft die Überführung von Raumzeitdaten aus Fernerkundungsbeobachtungen in n-dimensionale geographische Informationssysteme von Bedeutung sein. Auf Grund der riesigen Datenmengen sollten zur Gebäudeerkennung erstens die oben

beschriebenen Methoden in der Segmentierung eine Rolle spielen, und zweitens sind die Informationssysteme konsequent auf eine topologische Basis zu stellen.

3 Arbeits- und Ergebnisbericht

3.1 Ausgangslage

Die Kenntnis über die Beschaffenheit und Dynamik des deutschen Gebäudebestandes ist erstaunlich gering, obwohl er schon länger Forschungsgegenstand ist. Die Daten existieren zumeist summarisch und beruhen auf wenigen Zählungen, während detaillierte Daten für größere Zeiträume in der Regel verstreut und nicht digital vorliegen.

Ausgangsfrage. Die eingangs des Projekts gestellte Frage lautete angesichts der schwierigen Datensituation:

Wie verhält sich der deutsche Gebäudebestand im Laufe des 20. Jahrhunderts?

Die Frage enthält implizit auch einen Vergleich der Dynamik von Gebäuden unterschiedlicher Nutzung. Eine Beantwortung der Frage würde sich beispielsweise direkt auf Investitionsentscheidungen im Bausektor auswirken können.

Zielsetzung. Als Projektziel wurde die Bestimmung der Größe und Beschaffenheit des Gebäudebestandes gesetzt. Dabei sollte seine Dynamik untersucht werden unter Verwendung von Methoden aus:

- Data Mining
- Geoinformatik
- Nichtarchimedische Geometrie

Dabei wurde auf die Entwicklung neuer Methoden Wert gelegt.

Arbeitshypothesen. Es wurde für das Projekt vorausgesetzt, dass eine Zuhilfenahme der Mathematik, insbesondere der Topologie und der Nichtarchimedischen Geometrie, erforderlich ist. Die Topologie liefert eine Methode zur Modellierung raumzeitlicher Gebäudedaten in relationalen Datenbanken, während die nichtarchimedische Geometrie der hierarchischen Klassifikation zu Grunde liegt und darüber hinaus neue Klassifikationsmethoden liefert.

3.2 Beschreibung der durchgeführten Arbeiten

Nach der Anfertigung einer Internetpräsenz und der Bereitstellung eines Projektservers wurden die Datenbanken eingerichtet und mit der Datenerfassung begonnen. Gebäudebestandsrelevante Daten wurden von den statistischen Ämtern sowie den Katasterämtern angefordert. Im Laufe des Projekts kamen die Übergabelisten der *toten Akten* der Gebäudeversicherung hinzu. Diese wurde digitalisiert und daraus eine zufällige Stichprobe gezogen. Aus mehreren Archiven wurden Versicherungsakten zusammengetragen und Daten extrahiert. In der zeitgleich bearbeiteten Dissertation [4] wurde der Begriff *Urban Data Mining* definiert und mit neuen Klassifiktionsmethoden eine detaillierte Untersuchung urbaner Prozesse durchgeführt, die über die reine Betrachung des Gebäudebestandes hinausgehen. Zu ihrem Studium wurden ESOM-Karten angefertigt, Schätzverfahren zur Informationsübertragung entwickelt und Klassifikatoren aufgebaut.

Für die Anwendung nichtarchimedischer Methoden wurde untersucht, wie eine p-adische Kodierung eines multivariaten Datensatzes effektiv bewerkstelligt werden kann. Zunächst

wurde vorausgesetzt, dass entweder eine Rangfolge der Variablen vorgegeben werden kann oder die Daten als Wörter über einem Alphabet repräsentierbar sind. Diese Voraussetzung konnte später auf verschiedene Weisen abgeschwächt werden. Darüber hinaus gehend, wurde der Frage nachgegangen, wie sich die klassischen Klassifikationsverfahren auf p-adische Daten übertragen lassen, um deren baumartige Struktur auszunutzen. Die Motivation ist, dass im p-adischen Fall Algorithmen wesentlich effizienter laufen können und Klassifikationsergebnisse, im Gegensatz zum klassischen Fall, eindeutig werden. Der Preis hierfür ist bei realen Datensätzen das Problem ihrer p-adischen Kodierung. Diese ist in der Regel nicht eindeutig.

Aus der Stichprobe der toten Akten wurden Grundrisse digitalisiert und zur Aufnahme in einer topologische Datenbank vektorisiert. Mit den extrahierten Ereignishistoriendaten wurden Überlebensanalysen durchgeführt sowie Markovketten zur Simulation aufgebaut. Ebenso wurden darauf die entwickelten p-adischen Methoden zur Klassifikation der Dynamik angewandt.

Erste Ergebnisse wurden schon auf der GfKl-Jahrestagung 2007 vorgestellt und veröffentlicht [7, 16]. Es folgten weitere Beiträge auf nationalen und internationalen Konferenzen [3, 6, 7, 8, 16, 17], wobei [17] auf Veranlassung der Organisatoren in der Zeitschrift *p-Adic Numbers, Ultrametric Analysis and Applications* veröffentlicht wurde. Ebenso wurden Aufsätze in Fachzeitschriften publiziert bzw. eingereicht [2, 12, 13, 14, 15, 17]. Die Veröffentlichung weiterer Ergebnisse ist geplant [5, 18, 22].

Abweichungen. Die Abweichungen in den untenstehenden Arbeitspaketen sind, bis auf eine Ausnahme, meist durch einen unterschätzten Arbeitsaufwand begründet.

DErf. Es wurde auf die Erarbeitung eines in GIS eingebetteten historischen Katasters verzichtet.

DSich. Die nichtarchimedische Klassifikationsdatenbank wurde nicht aufgebaut. Stattdessen wurde die Bedeutung der p-adischen Methoden vielmehr für die hierarchische Klassifikation selbst als für die Datenhaltung gesehen. Infolgedessen wurde mehr Wert auf die Entwicklung p-adischer Klassifikationsmethoden und ihrer Erprobung gelegt. Die Frage nach einem effektiven Einsatz des Modulraums $M_{0,n}$ bleibt bestehen, wenn auch in einem erweiterten Sinn als ursprünglich geplant.

DAuf. Auf die Geolokalisierung von einzelnen Gebäuden wurde verzichtet.

DAna. Auf die Schätzung des Stofflagers wurde verzichtet. Die Bestimmung der „interessant“ zu nennenden Flurstücke erwies sich als schwieriger als geplant. Daher wurde das Markovmodell mit der Stichprobe aus den *toten Akten* aufgebaut (s. Abschnitt 3.3). Ebenso wurde auf die Schätzung der historischen Entwicklung typischster Objekte verzichtet.

Besondere Probleme. Das Hauptproblem betraf den unterschätzten Arbeitsumfang. Dieses lässt sich durch Einsatz weiterer Ressourcen leicht lösen. Problematischer als ursprünglich gedacht erscheint die neu hinzugekommene Aufgabe des Erarbeitens von Kriterien für Flustücke mit einer hohen Langzeitdynamik anhand der Informationen über einen relativ kurzen Zeitraum. Hierzu ist weitere Forschungsarbeit von Nöten (s. auch Abschnitt 3.4. „Unerwartete Fragestellungen“).

3.3 Darstellung der erzielten Ergebnisse

Es wurden Ergebnisse in den projektrelevanten Teildisziplinen *Urban Data Mining, p-adische Klassifikation und Topologische Datenbanken* erzielt.

Urban Data Mining. Es wurde der Begriff *Urban Data Mining* definiert als ein methodischer Zugang zur Aufdeckung von logischen Beschreibungen urbaner Muster und Regelmäßigeiten in Daten [7]. Konkret wurden Daten herangezogen, die Aussagen über Schrumpfungs- und Wachstumsprozesse von Gemeinden treffen können. In der Ausführung gelang erstmals eine Schätzung der Gebäudezahl (ca. 38 Mio.) sowie eine gemeindescharfe Schätzung der Zusammensetzung des deutschen Gebäudebestandes in Wohn- und Nichtwohngebäude für das Jahr 2006 [4]. Für 6550 Gemeinden konnten die Anzahlen exakt bestimmt werden, während für die verbleibenden 5870 Gemeinden eine log-lineare Regression angewandt werden konnte. Es stellte sich dabei heraus, dass etwa 20% der Gemeinden etwa 80% des Bestandes ausmachen, was eine Informationsoptimierung ermöglicht [8]. Die Gemeinden wurden nach ähnlichen dynamischen Charakteristika klassifiziert und so die Urbanisierung und die regionale Diversität analysiert [3, 6].

Überraschend war die Verfügbarkeit der Addressenlisten zu den Versicherungsakten zu Adressen, deren Gebäude vor 1994 nicht mehr versichert waren, den sog. *toten Akten*. Eine zufällige Stichprobe von 100 Adressen daraus bildete die Datengrundlage für die Untersuchung der Dynamik mit neuartigen p-adischen Methoden. Dazu wurden die gebäuderelevanten Ereignisse extrahiert. Die untenstehenden p-adischen Methoden erlaubten via binärer Kodierung der Ereignisdaten eine erste Klassifikation der Dynamik von Gebäudebeständen [14]. Im Datensatz weisen die Großstädte eine ähnliche Dynamik auf, die sich von der kleinerer und mittlerer Städte unterscheidet. Hieraus konnten Übergangswahrscheinlichkeiten zwischen verschiedenen Nutzungsklassen geschätzt werden. Die erhaltenen Matrizen erlaubten erste Simulationen des Nichtwohngebäudebestandes in Baden-Württemberg mit Markovketten: der Stadtkreise, Landkreise sowie des Gesamtbestandes. In Letzterem ergab sich ein Schwund von etwa 16% des Bestandes von 1950, der etwa zu 12% im Jahr 2006 vertreten ist. Die Altersverteilungen sind in den drei Beständen ähnlich, wobei die Gebäude vor 1950 in den Stadtkreisen etwas stärker repräsentiert sind als in den Landkreisen [18].

Die Frage nach der Verschiedenheit der Überlebnswahrscheinlichkeiten der Wohngebäude im Vergleich zu den Nichtwohngebäuden lässt sich letztlich zwar nicht mit vollständiger Sicherheit beantworten. Dennoch ergibt sich für die Dynamik der beiden Gebäudenutzungskategorien folgende Präzisierung [18]:

Vermutung. *In urbanen Gebieten hoher Diversität oder mit hohem Wohnnutzungsanteil verhalten sich Wohn- und Nichtwohngebäude nicht wesentlich verschieden. In den anderen urbanen Gebieten ist die Dynamik der Nichtwohngebäude höher als die der Wohngebäude. In beiden Fällen ist die Überlebensfunktion unabhängig vom Baujahr des Gebäudes.*

Insbesondere wird im zweiten Fall die Überlebenswahrscheinlichkeit von Wohngebäuden höher als die der Nichtwohngebäude sein.

p-adische Klassifkation. Zunächst wird in [15] ein Dendrogramm konzeptionell als Punkt eines so-genannten *Modulraums* $M_{0,n}$ erklärt. Dies erlaubt die Beschreibung zeitlich veränderlicher Daten als eine Folge von Punkten in diesem Modulraum. Deren Dynamik kann dann z.B. durch Interpolation oder Regression geschätzt werden. Unscharfe Klassifikation entspricht einer Wahrscheinlichkeitsverteilung auf $M_{0,n}$, und Kollisionen von Punkten finden auf dessen Rand statt [16]. In [12] wird ein p-adisches agglomeratives hierarchisches Klassifikationsverfahren eingeführt. Ein Klassifikationsverfahren mit vorgegebener Höchstgrenze für die Klassenzahl, angelehnt an [26], findet sich in [13]. Hierbei wird eine Energiefunktion minimiert. Minimale Klassenzentren erlauben den Aufbau p-adischer Klassifikatoren, die beispielsweise im Maschinellen Lernen verwendet werden künen. Dabei zeigt sich, dass die optimalen Klassen und Zentren bis auf endlich viele Ausnahmen nicht von der Wahl der Primzahl p abhängen.

Die p-adische Kodierung von Daten wird in [12] ausfürlich beschrieben, falls entweder das zugehörige Dendrogramm bekannt ist oder diese durch Wörter über einem gegebenen

Alphabet repräsentiert sind. Im letzteren Fall ist das Dendrogramm D mit linearem Aufwand berechenbar. Entscheidend für die Kodierung ist die maximale Zahl c von Kindern eines Knotens in D. Für eine Kodierung über den *rationalen p-adischen Zahlen* $\mathbb{Q}_p$ muss gelten:

$$p \geq c. \tag{1}$$

Es wird in [12] gezeigt, wie für jede Wahl einer festen Primzahl mit *algebraischen p-adischen Zahlen* $K \supseteq \mathbb{Q}_p$ kodiert werden kann, ungeachtet der Bedingung (1). Die algebraischen p-adischen Zahlen sind eine nichtarchimedische Entsprechung der *komplexen Zahlen* $\mathbb{C}$. Damit wird ein Wechsel der Primzahl allein auf Grund einer Datenänderung unnötig.

Zur p-adischen Kodierung eines beliebigen multivariaten quantitativ-qualitativen Datensatzes gibt es verschiedene Möglichkeiten, wie in [14] aufgezeigt wird. Zum Einen gibt es die bereits in [10] verwendete diskrete Fouriertransformation, welche einen Vektor $x = (x_0, \ldots, x_{N-1})$ in den Bereich der linear geordneten Frequenzen k transformiert:

$$x \mapsto \hat{x}_k = \sum_{\ell=0}^{N-1} x_\ell \cdot e^{-2\pi i \frac{\ell \cdot k}{N}}, \qquad k = 0, \ldots, N-1,$$

und nach einer Quantisierung $q(|\hat{x}_k|)$ wird der transformierte Vektor $\hat{x}$ p-adisch kodiert:

$$\hat{x} \mapsto \sum_{k=0}^{N-1} q(|\hat{x}_k|) \cdot p^k.$$

Hierbei stehen die niedrigen Frequenzen höher in der Hierarchie als die höheren Frequenzanteile. Eine weitere Möglichkeit bietet eine Priorisierung der Attribute nebst Quantisierung: dann definiert das Attribut mit höchster Priorität den konstanten Term a_0, das nächsthöchste Attribut den Koeffizient a_1 usw. Damit wird die p-adische Zahl $a_0 + a_1 p + a_2 p^2 + \ldots$ definiert. Beide Methoden kommen in [14] zur Anwendung und liefern ähnliche Ergebnisse.

Topologische Datenbanken. Die digitalisierten Grundrisse der Einschätzungsverzeichnisse der toten Akten wurden erstmals als relationaler Kettenkomplex in einer temporalen $(2\frac{1}{2}+1)$D-Datenbank abgespeichert, wie in [22] beschrieben, und stehen somit dem Projekt [23] als erster Beispieldatensatz zur Verfügung.

Umfeld. *Urban Data Mining.* Nähe besteht zu den Arbeiten von H. Bergsdal et al. (Trondheim), welche auf physikalischen Modellen für die Massenströme aufbauen [11, 28], ebenso zu denen des Leibniz-Instituts für ökologische Raumentwicklung (IÖR), insbesondere [27], sowie der Arbeitsgruppe um P. Steadman (London). Der Bericht [24] enthält einen Überblick über Datenlage und dynamische Charakteristika zu den Gebäudebeständen ausgewählter europäischer Länder.

p-adische Klassifikation. Die inhärente hierarchische Struktur von Daten aufzudecken ist Ziel der Arbeiten von F. Murtagh (London, Dublin), während A. Khrennikov (Växjö) p-adische Lernalgorithmen entwickelt, die u.A. in der Verarbeitung bewegter Bilder angewendet werden [10, 9, 25].

Topologische Datenbanken. Im Bereich der Verwendung von Raumzeitdatenbanken besteht Nähe zur Entwicklung von 3D-/4D-GIS (M. Breunig, Osnabrück).

3.4 Ausblick auf zukünftige Arbeiten

Unerwartete Fragestellungen. Im Bereich der Analyse räumlicher Daten ergibt sich aus der prinzipiellen Verfügbarkeit der toten Akten die Aufteilung der Gebäudebestandsdynamik

in einen „sichtbaren“ und einen „unsichtbaren“ Anteil. Ersterer betrifft die Entwicklung aller zu einem Referenzzeitpunkt sichtbaren Objekte (Gebäude, Flurstücke, Quartiere, ...), während der andere Teil die Objekte betrifft, die etwa durch Abriss oder Verschmelzung mit anderen Objekten zum Referenzzeitpunkt nicht mehr vorhanden sind. Im vorliegenden Fall handelt es sich um die zu den toten Akten gehörenden Flurstücke. Diese weisen eine von den übrigen Flurstücken verschiedene Neubautätigkeit auf. Es ergibt sich die Frage nach weiteren Unterschieden zwischen den beiden Anteilen der Dynamik, auch wenn die Stichprobe eine ansonsten erwartungsgemäße Klassifikation der betroffenen Städte liefert [14]. Ebenso gilt es, weitere Massen- oder Flächenstrommodelle für den deutschen Gebäudebestand, angelehnt an [11, 28], aufzustellen und zu validieren. Eine weitere unerwartete Frage im urbanen Kontext ergab sich durch die Schwierigkeit, Flurstücke mit hoher Dynamik anhand der Informationen für kurze Zeiträume zu erkennen. Die Aufgabe besteht nun darin, charakterische Kriterien für derartige „interessante“ Flurstücke zu erarbeiten, um den Erfassungsaufwand beispielsweise in Versicherungsakten möglichst klein zu halten. Für die „uninteressanten“ Flurstücke sind dynamische Kenndaten noch zu schätzen, wenngleich die Resultate aus Ettlingen [19] in erster Näherung herangezogen werden können.

Die ursprüngliche Motivation für die Verwendung p-adischer Methoden war die Datenkodierung zur Speicherung von Dendrogrammen in einer Datenbank. Dabei sollte der Modulraum $M_{0,n}$ als abstraktes Datenbankmodell dienen. Konkret wird dabei jedes Dendrogramm einfach als n-tupel natürlicher Zahlen abgelegt — entweder in p-adischer oder in irgendeiner anderen Darstellung. Im Fall natürlicher Zahlen rekonstruiert der euklidische Algorithmus das Dendrogramm [12]. Durch die in [12, 13] begonnene Übertragung von Klassifikationsalgorithmen in die (effizientere) p-adische Welt und ihre erste Anwendung auf die Projektdaten ergibt sich allgemein die Frage nach ihrer Einsetzbarkeit im (Urban) Data Mining bzw. der Vergleich mit klassischen Methoden in der Untersuchung dynamischer Prozesse. Eng verbunden damit ist die Frage nach einer geeigneten Kodierung von Daten, mit der inhärente hierarchische Strukturen sicher aufgedeckt werden. Dabei sind auch die Resultate in [14] zu validieren. Schließlich steht die Frage nach einer Weiterentwicklung der p-adischen Klassifikationsalgorithmen bis hin zu einer *effektiven* Einbeziehung des Modulraums $M_{0,n}$, der in diesem Zusammenhang bisher lediglich *konzeptionell* von Bedeutung ist.

Die Überführung von Daten in Informationssysteme lässt sich allgemeiner beschreiben als Datengewinnung, Kodierung und Bereitstellung des Informationssystems. Der Fokus des Projekts lag mehr auf den ersten beiden Teilen. Die Einbindung an das IPF erweitert die Datengewinnung zusätzlich um verschiedenartige Messdaten von Satelliten oder Überfliegungen, die in Raumzeitinformationssysteme zu überführen sind. Dadurch gewinnt die Entwicklung von Klassifikations- und Segmentierungsverfahren zur Gebäudeerkennung an Bedeutung, ebenso die Bereitstellung geeigneter Informationssysteme.

Beteiligung anderer Partner oder Fachdisziplinen. Die Kopplung der verschiedenen Methoden der Gebäudedatengewinnung erfordern insbesondere die Expertise fernerkundlicher Methoden und sind am IPF vorhanden. Die Entwicklung von 3D-/4D-Informationssystemen geschieht bereits durch Kooperation mit M. Breunig (Osnabrück) und N. Paul (München). Für die Wissenskonversion durch Zusammenführung urbaner Daten aus der Vielzahl verschiedenartiger Datenquellen bieten sich eine Fortführung der Zusammenarbeit mit M. Behnisch (ETH Zürich) und Kooperation mit A. Ultsch (Marburg), dem IÖR und weiteren europäischen Institutionen an.

Forschungsinitiative. Eine größer angelegte Forschungsinitiative unter Einbeziehung der oben vorgeschlagenen Partner würde neben einem Erkenntnisgewinn um den deutschen Gebäudebestand im europäischen Kontext die Weiterentwicklung von wissensgenerierenden Methoden zu diesem Zwecke erwarten lassen. Eine interdisziplinär angelegte Initiative

würde dabei eine Effizienzsteigerung im methodischen Teil bei gleichzeitiger Erhöhung des Detaillierungsgrades erwarten lassen (vgl. auch nachfolgenden Abschnitt 3.5).

3.5 Interdisziplinäre Weiterentwicklung

Architektur. Die gewonnenen Erkenntnisse über den Gebäudebestand und vielmehr die Methoden ihrer Gewinnung sind für die Architektur von Interesse. Der Begriff des Urban Data Mining [4] schafft den theoretischen Rahmen für die Anwendung moderner explorativer Datenanalyseverfahren auf regionalplanerische Fragestellungen. Die Langzeitdynamik von Gebäudebeständen wird noch länger relevant sein, wie auch deren Vergleich. Dazu wird es einer Weiterentwicklung der Methoden bedürfen, schon allein um die Frage zu beantworten, wieviel der früher gebauten Umwelt heute noch vorhanden ist. Darüber hinaus vertiefen die entwickelten Methoden die Verbindungen zu den untenstehenden Disziplinen, insbesondere auf dem Niveau der Datenbehandlung und Modellierung.

Mathematik. Die neuen p-adischen Klassifikationsmethoden gehören in die angewandte Mathematik, wobei auch Teile der so-genannten „reinen“ Mathematik einen Anwendungsbezug bekommen. Daher ergibt sich die multidisziplinäre Zeitschrift *p-Adic Numbers, Ultrametric Analysis, and Applications* als Kommunikationsmedium für [13, 17]. Die archimedischen Klassifikationstechniken sollten nichtarchimedische Entsprechungen finden lassen. Dazu bieten sich Anleihen aus der p-adischen Analysis an, wie auch der mathematischen Physik. Der Mehrwert wäre eine Effizienzsteigerung unter der Vorgabe einer p-adischen Datenkodierung.

Geoinformatik. Das Projekt gab den Ergebnissen des vorangegangenen Projekts KO 1488/8-1, 8-2 *Architektonische Komplexe* [21, 20] eine erste Möglichkeit zur praktischen Anwendung topologischer Datenbanken [22]. Die Projektdaten stehen dem gemeinsamen Projekt BR 2128/12-1 und BR 3513/3-1 *Modellierung und Verwaltung der Topologie für Gebäudeinformationsmodelle unter besonderer Berücksichtigung von Planungsalternativen und Versionen* [23] mit M. Breunig (Osnabrück) zur Verfügung.

Fernerkundung. Durch die abschließende Überführung des Projekts an das Institut für Photogrammetrie und Fernerkundung ergibt sich eine umfassende Zusammenarbeit mit der Geodäsie. So bieten sich, über das anschließende Projekt [23] hinausgehend, nichtarchimedische Klassifikationsmethoden in der Fernerkundung zur Segmentierung von Laserscandaten, insbesondere zur detaillierten Erkennung von Gebäuden und Integration der gewonnenen Daten in Raumzeitinformationssysteme oder zur 3D-/4D-Gebäudemodellierung, an.

3.6 Verwertungspotenzial

Bewertung. Ein Verwertungspotenzial der erreichten Ergebnisse ist vorhanden, wenn auch mit weiterem Forschungs- und Implementierungsaufwand.

Zukünftige Verwertungsmöglichkeiten Eine Produktrealisierung der p-adischen Klassifikationsalgorithmen für maschinelles Lernen ist in der Form

$$\text{DFT} \rightarrow \text{Quantisierung} \rightarrow p\text{-adische Kodierung} \rightarrow \text{Klassifikation}$$

nach Erbringung der erforderlichen Implementierungsarbeiten möglich. Für die Verwertungsfelder Gebäudebestands- und Liegenschaftsverwaltung können nach weiterer Forschungsarbeit urbane Benchmark- und Monitorsysteme entstehen.

Verwertungsmaßnahmen Es wurden weder Verwertungsmaßnahmen eingeleitet noch geplant.

Patente, Industriekooperationen o.Ä. Es wurden keine Patente angemeldet oder Industriekooperationen o.Ä. geplant.

3.7 Beteiligte Wissenschaftler

Patrick Erik Bradley. Antragsteller und Projektbearbeiter. Beratung bei der Dissertation [4], p-adische Methoden im Data Mining, Datenerhebung, Datenbankentwürfe und Abfragen, Datenauswertung, Überlebens- und Ereignishistorienanalyse, Bestandssimulation, Raumzeitdatenbanken.

Publikationen: [12, 13, 14, 15, 16, 17, 18, 22].

Martin Behnisch. Projektmitarbeiter und Kooperationspartner. Dissertation [4], Methodik des Urban Data Mining, Datenerhebung, Klassifikatorenaufbau, Emergente SOM, Entwicklung von Schätzverfahren zur Informationsübertragung, Gemeindescharfe Schätzung des deutschen Gebäudebestandes.

Publikationen: [2, 3, 4, 6, 8, 7, 18, 22].

Norbert Paul. Kooperationspartner. Beratung bei Implementierung Topologischer Datenbanken.

Publikationen: [22]

4 Publikationen

4.1 Publikationen in Fachzeitschriften

1. Martin Behnisch. *Urban Data Mining—Spatiotemporal exploration of multidimensional data.* Eingeladene Veröffentl. in Building Research & Information. Eingereicht. [2]

2. Martin Behnisch und Nguyen Xuan Thinh. *Spatial Data Mining.* In Arbeit. [5]

3. Patrick Erik Bradley. *Mumford dendrograms.* The Computer Journal. Im Druck. DOI: 10.1093/comjnl/bxm088. [12]

4. Patrick Erik Bradley. *On p-adic classification.* Eingereicht bei *p-Adic Numbers, Ultrametric Analysis and Applications.* [13]

5. Patrick Erik Bradley. *An ultrametric interpretation of building related event data.* Eingereicht bei *Construction Management and Economics.* [14]

6. Patrick Erik Bradley. *Degenerating families of dendrograms.* Journal of Classification, Vol. 25, Nr. 1 (2008), 27-42. DOI: 10.1007/s00357-008-9009-5. [15]

7. Patrick Erik Bradley. *Mumford dendrograms and discrete p-adic symmetries.* p-Adic Numbers, Ultrametric Analysis and Applications, Vol. 1, Nr. 2 (2009), 118–127. Im Druck. DOI: 10.1134/S2070046609020010. [17]

8. Patrick Erik Bradley und Martin Behnisch. *Building stocks in perspective of states and transition probabilies.* In Arbeit. [18]

9. Patrick Erik Bradley, Norbert Paul und Martin Behnisch. *A topological database for urban space-time data.* In Arbeit. [22]

4.2 Kongressbeiträge

1. Martin Behnisch. *Spatial similarities in urbanisation and regional diversity.* 22nd Int. Conf. on Informatics for Environmental Protection. Leuphana-Universität Lüneburg. Sept. 10-12, 2008. [3]

   Abstract. Most of the large databases currently available have a strong geospatial component and contain potentially useful information that might be of value. Data mining is defined as the inspection of data with the aim of discovering knowledge. Mining implies a laborious process of searching for hidden information in a large amount of data. Knowledge discovery is defined as the discovery and formal representation of knowledge from data collections. Data mining in connection with knowledge discovery techniques will be of increasing importance for the urban research and planning processes. "Urban Data Mining" is a methodological approach to discover logical, mathematical and partly complex descriptions of patterns and regularities inside a set of data. The main theme of this contribution is the definition of an urbanized area. 12430 German communes are structured by a classification approach. The issue of these grouping processes are urbanisation and regional diversity. The set of data is examined and it will be shown that the investigation of distributions leads to a better understanding of each attribute. Gaussian Mixture Models are presented as an appropriate method for regionalisation. Spatial Analysis (GIS) is part of the process of knowledge conversion and communication. Results suggest a general typology and can lead to the development of prediction models using subgroups instead of the total population. The procedures on the basis of knowledge-based systems are currently not sufficiently developed for a direct integration into the regional and urban planning and development processes. These approaches could lead to a benchmark system for regional policy or to other strategic instruments such as fully automated urban monitoring systems.

2. Martin Behnisch and Alfred Ultsch. *Are there groups (cluster) of communities with the same dynamic behaviour?* International Federation of Classification Societies 2009 Conference, Dresden University of Technology. March 13-18, 2009. Eingereichter Beitrag. [6]

   Abstract. More than half of the worlds population will be living in urbanized areas by the end of this decade. Urbanized Areas are therefore a major component of the modern environment. They are subject to environmental change. Demographic and economic changes lead to the phenomena of growing and shrinking cities. The issue of this article is to find groups (cluster) of communes with the same dynamic characteristics in Germany. Most of the large databases currently available have a strong geospatial and in particular an urban component containing potentially useful information that might be of value. Urban Data Mining represents a methodological approach that discovers logical, mathematical and partly complex descriptions of urban patterns and regularities inside statistical data. The approach relies on 12430 communes and refers to data from well known and easily accessible institutions. The use of Emergent SOMs is presented as an appropriate and powerful method for clustering and classification. The application of graphical methods especially U*-Matrix shows that it is of high value, first, to visualize the structure of highdimensional data and second, to detect meaningful classes. In addition the knowledge generating algorithm U*C is applied to find significant attributes for the description and recognition of a given set of cluster. The results suggests a general typology and lead to the development of prediction models using subgroups instead of the total population. In the future it might be possible to establish an instrument that defines objective criteria for the benchmark process.

3. Martin Behnisch. *Urban Data Mining using Emergent SOM.* 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007. [7]

   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 mining 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. In this study ESOM is used to examine communities in Germany. The data deals with the question of dynamic processes (e.g. shrinking and growing of cities). 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.

4. Martin Behnisch und Alfred Ultsch. *Estimating the number of buildings in Germany.* 32nd Annual Conference of the Gesellschaft für Klassifikation e.V., Helmut-Schmidt-Universität Hamburg, March 7–9, 2008. [8]

   Abstract. The debate on sustainable development has lead to the view of buildings as flows (mass, energy, money and information) or capitals. In this context buildings are considered as the largest physical, economical, social and cultural capital of a society. In Germany many institutions record different kind of data about buildings. Unfortunately there are just a few basic statistics about the amount of buildings. Collection of data is very complicated, often expensive and the handling of missing data is one of the biggest handicaps. With the exception of data about residential buildings and particularly monuments, it is an unsolved problem to determine the total number of buildings. Thus the main issue of this article is the description of an appropriate estimation procedure. This procedure relies on 12430 communes and refers to data from the Cadaster of Real Estates and the Federal Office for Building and Regional Planning (BBR). The estimation is based on statistical data from well known and easily accessible institutions. The number of buildings is estimated for communes with missing data. Using methods from the, so called, Urban Data Mining approach, unsuspected relationships are found in the urban data. These relationships are valuable for the estimation. The quality of the estimation is analyzed by training and test data sets. Information optimization leads to the conclusion that 20% of the communes hold 80% of all buildings. For an improvement of the estimation it is essential to refine the amount and quality of data in the larger communes.

5. Patrick Erik Bradley. *Families of Dendrograms.* 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007. [16]

   Abstract. A conceptual framework for cluster analysis from the viewpoint of p-adic geometry is introduced by describing the space of all dendrograms for n datapoints and relating it to the moduli space of p-adic Riemannian spheres with punctures using a method recently applied by Murtagh (2004). This method embeds a dendrogram into the Bruhat-Tits tree associated to the p-adic numbers as a subtree used by Cornelissen et al. (2001) in p-adic geometry. After explaining the definitions, the concept of Bayesian classifiers is discussed in the context of moduli spaces, and upper bounds for the number of hidden vertices in dendrograms are given.

6. Patrick Erik Bradley. *Mumford dendrograms and discrete p-adic symmetries.* p-ADIC MATHPHYS.2007, Steklov Mathematical Institute, Moscow, Russia, October 1–6, 2007. [17]

   Abstract. In this talk, we present an effective encoding of dendrograms by embedding them into the Bruhat-Tits trees associated to p-adic number fields. As an application, we show how strings over a finite alphabet can be encoded in cyclotomic extensions of $\mathbb{Q}_p$ and discuss p-adic DNA encoding. The application leads to fast p-adic agglomerative hierarchic algorithms similar to the ones recently used e.g. by A. Khrennikov and others. From the viewpoint of p-adic geometry, to encode a dendrogram X in a p-adic field K means to fix a set S of K-rational punctures on the p-adic projective line $\mathbb{P}^1$. To $\mathbb{P}^1 \setminus S$ is associated in a natural way a subtree inside the Bruhat-Tits tree which recovers X, a method first used by F. Kato in 1999 in the classification of discrete subgroups of $\mathrm{PGL}_2(K)$.

   Next, we show how the p-adic moduli space $\mathfrak{M}_{0,n}$ of $\mathbb{P}^1$ with n punctures can be applied to the study of time series of dendrograms and those symmetries arising from hyperbolic actions on $\mathbb{P}^1$. In this way, we can associate to certain classes of dynamical systems a Mumford curve, i.e. a p-adic algebraic curve with totally degenerate reduction modulo p.

   In the end, we indicate some of our results in the study of general discrete actions on $\mathbb{P}^1$, and their relation to p-adic Hurwitz spaces.

4.3 Buchbeiträge

1. Martin Behnisch. *Urban Data Mining - Eine Methode zur Quantifizierung von Gebäudebeständen.* Erscheint in: Uta Hassler, Niklaus Kohler (Hrsg.) *Materialien zur Langfristst abilität.* [1]

4.4 Studien- und Diplomarbeiten, Dissertationen, Habilitationen, Berichte, sonstige Publikationen

1. Martin Behnisch. *Urban Data Mining.* Dissertation (2008). [4]

Literatur

[1] Behnisch, Martin: *Urban Data Mining — Eine Methode zur Quantifizierung von Gebäudebeständen.* In: Uta Hassler, Niklaus Kohler (Herausgeber): *Materialien zur Langfristst abilität.* Erscheint demnächst.

[2] Behnisch, Martin: *Urban Data Mining — spatiotemporal exploration of multidimensional data.* Präpublikation.

[3] Behnisch, Martin: *Spatial similarities in urbanisation and regional diversity.* In: Möller, A. und M. Schreiber (Herausgeber): *Proc. 22nd Int. Conf. Informatics for Environmental Protection*, Leuhphana Universität Lüneburg, 2008.

[4] Behnisch, Martin: *Urban Data Mining.* Doktorarbeit, Universität Karlsruhe (TH), 2008.

[5] Behnisch, Martin und Nguyen Xuan Thinh: *Spatial Data Mining.* In Arbeit.

[6] Behnisch, Martin und Alfred Ultsch: *Are there groups (cluster) of communities with the same dynamic behaviour?* Präpublikation.

[7] Behnisch, Martin und Alfred Ultsch: *Urban Data Mining Using Emergent SOM.* In: Preisach, Chr., H. Burkhardt, L. Schmidt-Thieme und R. Decker (Herausgeber): *Proc. 31st Ann. Conf. GfKl*, Studies in Classification, Data Analysis, and Knowledge Organization, Seiten 311–318, Freiburg i. Br., 2008. Springer.

[8] Behnisch, Martin und Alfred Ultsch: *Estimating the unumber of buildings in Germany.* In: *Proc. 32nd Ann. Conf. GfKl*, Studies in Classification, Data Analysis, and Knowledge Organization, Hamburg, Im Druck.

[9] Benois-Pineau, J. und A.Yu. Khrennikov: *Significance Delta Reasoning with p-Adic neural networks: application to shot change detection in video.* The Computer Journal. Im Druck.

[10] Benois-Pineau, J., A.Yu. Khrennikov und N.V. Kotovich: *Segmentation of Images in p-Adic and Euclidean metrics.* Dokl. Math., 64:450–455, 2001.

[11] Bergsdal, H., H. Brattebø, R.A. Bohne und D.B. Müller: *Dynamic flow analysis for Norway's dwelling stock.* Building Research & Information, 35:557–570, 2007.

[12] Bradley, Patrick Erik: *Mumford dendrograms.* The Computer Journal. Im Druck. DOI: 10.1093/comjnl/bxm088.

[13] Bradley, Patrick Erik: *On p-adic classification.* Präpublikation. arXiv:0903.2870v1 [cs.AI].

[14] Bradley, Patrick Erik: *An ultrametric interpretation of building related event data.* Präpublikation.

[15] Bradley, Patrick Erik: *Degenerating families of dendrograms.* Journal of Classification, 25(1):27–42, 2008. DOI: 10.1007/s00357-008-9009-5.

[16] Bradley, Patrick Erik: *Families of dendrograms.* In: Preisach, Chr., H. Burkhardt, L. Schmidt-Thieme und R. Decker (Herausgeber): *Proc. 31st Ann. Conf. GfKl*, Studies in Classification, Data Analysis, and Knowledge Organization, Seiten 95–102, Freiburg i. Br., 2008. Springer.

[17] Bradley, Patrick Erik: *Mumford dendrograms and discrete p-adic symmetries. p-*Adic Numbers, Ultrametric Analysis and Applications, 1(2):118–127, 2009. Im Druck. DOI: 10.1134/S2070046609020010.

[18] Bradley, Patrick Erik und Martin Behnisch: *Building stocks in perspective of states and transition probabilities.* In Arbeit.

[19] Bradley, Patrick Erik und Niklaus Kohler: *Methodology for the survival analysis of urban building stocks.* Building Research & Information, 35(5):529–542, 2007. DOI: 10.1080/09613210701266939.

[20] Bradley, Patrick Erik und Norbert Paul: *Using the relational model to capture topological information of spaces.* The Computer Journal. Im Druck. DOI: 10.1093/comjnl/bxn054.

[21] Bradley, Patrick Erik und Norbert Paul: *Topologie als Grundlage für Gebäudeinformationssysteme.* Technischer Bericht 2008-01, Institut für Industrielle Bauproduktion, Universität Karlsruhe, 2008.

[22] Bradley, Patrick Erik, Norbert Paul und Martin Behnisch: *A topological database for urban space-time data.* In Arbeit.

[23] Breunig, Martin, Patrick Erik Bradley, Norbert Paul und Andreas Thomsen: *Modellierung und Verwaltung der Topologie für Gebäudeinformationsmodelle unter besonderer Berücksichtugung von Planungsalternativen und Versionen.* Antragstext zu BR 2128/12-1 und BR 3513/3-1, 2008.

[24] Itard, Laure, Frits Meijer, Evert Vrins und Harry Hoiting: *Building Renovation and Modernisation in Europe: State of the art review.* Erabuild, Abschlussbericht, 2008.

[25] Khrennikov, A.Yu. und B. Tirozzi: *Algorithm of Learning of p-adic neural networks.* Präpublikation.

[26] Linde, Yoseph, Andrés Buzo und Robert M. Gray: *An algorithm for vector quantizer design.* IEEE Trans. Comm., 28:84–95, 1980.

[27] Meinel, Gotthard, Robert Hecht, Hendrik Herold und Georg Schiller: *Automatische Ableitung von stadtstrukturellen Grundlagendaten und Integration in einem Geographischen Informationssystem.* Bundesamt fr Bauwesen und Raumordnung, Bonn, 2008.

[28] Sartori, I., H. Bergsdahl, D.B. Müller und H. Brattebø: *Towards modelling of construction, renovation and demolition activities: Norwegian's dwelling stock.* Building Research & Information, 36:412–425, 2008.