Universität Trier, Abteilung Kartographie

SPATIAL DATA MINING, GEOSTATISTIK UND GEOVISUALISIERUNG ALS WERKZEUGE ZUR DATENEXPLORATION BEI PLANUNGSAUFGABEN

Anja Matatko

Zusammenfassung: Der Lösung von Planungsaufgaben liegen Geodaten zugrunde, die Informationen über räumliche Strukturen in sich bergen. Die Umwandlung der Geodaten in georäumliches Wissen erfolgt in der Planungspraxis jedoch meist zufällig. Daher ist es Ziel der vorliegenden Arbeit, ein Ablaufschema zu erstellen, mit dem Planungsaufgaben standardisiert mit Hilfe von Analysemethoden aus den Bereichen Spatial Data Mining, Geostatistik und Geovisualisierung beantwortet werden können. Problematisch beim Einsatz solcher Methoden ist, dass die Planer nicht unbedingt mit deren Anwendung vertraut sind. Daher soll das aufgestellte Ablaufschema Methoden nutzen, deren Ergebnisse möglichst intuitiv interpretiert werden können. Somit können auch Statistik-Laien aus ihren Geodaten im Sinne der Datenexploration neue Erkenntnisse gewinnen. Interpretationsmöglichkeiten und konkrete Analyseergebnisse werden anhand eines Anwendungsbeispiels aus dem Bereich der Sozialplanung (basierend auf Matatko 2008) demonstriert.

Schlüsselwörter: Datenexploration, Geovisualisierung, Spatial Data Mining, Räumliche Entscheidungsunterstützung

// SPATIAL DATA MINING, GEOSTATISTICS AND GEOVISUALIZATION AS TOOLS FOR DATA EXPLORATION IN PLANNING TASKS

// Abstract: The solution of planning tasks is based on geodata, which hide information about spatial structures. However, the transformation from geodata to geospatial knowledge happens more or less at random. Therefore, the aim of the present research project is to establish a standardized workflow in order to solve planning tasks by using methods of analysis from spatial data mining, geostatistics and geovisualization. The problem is that planners are not necessarily accustomed to their application. That's the reason why the workflow should use methods which allow even users inexperienced in statistics to gain new information from their geodata in terms of data exploration. Interpretation facilities and detailed analysis results are presented by a practical example of social area analysis (taken from Matatko 2008).

Keywords: Data exploration, geovisualization, spatial data mining, spatial decision support

Anschrift der Autorin

Dipl.-Geogr. Anja Matatko
Universität Trier
Abteilung Kartographie
Behringstr.
D-54286 Trier
E: matatko@uni-trier.de

1. EINFÜHRUNG

Geodaten verbergen oft Informationen in sich, die bei der Entscheidungsfindung in Planungsfragen hilfreich sein können. Aufgrund der Multidimensionalität der Daten können diese Informationen jedoch nicht unmittelbar erschlossen werden. Daher existieren Methoden der räumlichen Statistik und des Spatial Data Minings, die unbekannte Zusammenhänge aufdecken und daraus resultierend Entscheidungen unterstützen. Die Ergebnisse können kartographisch präsentiert und verändert werden. Damit verbunden ist eine Funktionsänderung von Karten, die vermehrt als interaktives Analysewerkzeug anstatt als statischer Informationsspeicher gesehen werden. Es ist zu hinterfragen, welchen Nutzen die Methoden der räumlichen Statistik und des Spatial Data Minings verbunden mit Geovisualisierung für Entscheidungen in räumlichen Planungsprozessen bringen können. Dabei ist zunächst zu klären, an welchen Stellen in einem Planungsprozess Entscheidungen automatisiert werden können, und welche Analysemethoden jeweils bei bestimmten räumlichen Fragestellungen zum Einsatz kommen können. Anschließend wird anhand eines Praxisbeispiels aufgezeigt, welchen Mehrwert interaktive explorative Datenanalyse bei der Interpretation räumlicher Sachverhalte liefern kann.

2. DATENEXPLORATION AN DER SCHNITTSTELLE VON GEOVISUALISIERUNG, RÄUMLICHER STATISTIK UND SPATIAL DATA MINING

2.1 KARTEN: VON DER INFORMATION ZUR EXPLORATION

Karten wurden viele Jahre lang ausschließlich in ihrer Funktion als Informationsspeicher und Kommunikationsmittel betrachtet. Bereits MacEachren (1994) stellt mit Hilfe seines kartographischen Würfels (Abbildung 1) dar, dass kartographische Medien, in Abhängigkeit vom Interaktionsgrad und der Öffentlichkeit der Daten, unterschiedliche Funktionen wahrnehmen, die sich im Spannungsfeld zwischen Kommunikation und Visualisierung befinden, wobei ersteres dazu dient, bereits Bekanntes darzustellen, während bei letzterem Unbekanntes aufgedeckt werden soll.

Visualisierung mit hohem Interaktionsgrad und dem Ziel, Unbekanntes aufzudecken, wird von anderen Autoren als Datenexploration (Slocum et al. 2009) oder Geovisualisierung (Müller 2005, S. 239) bezeichnet. Die Funktion der Karten als Datenexplorationsmedium findet sich auch bei Dickmann (2007). Er betrachtet die Visualisierung von Geodaten mit Hilfe von Karten als fundamentale geographische Methode.

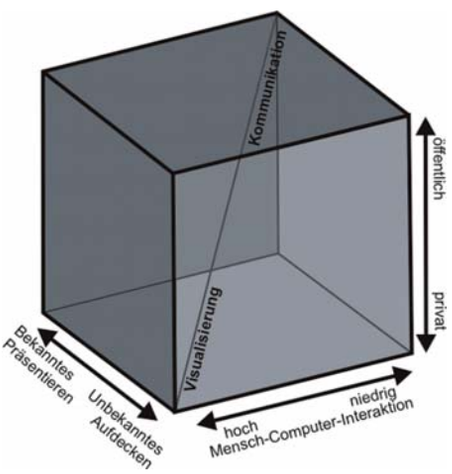

Abbildung 1: Der Kartographische Würfel (nach MacEachren 1994)

Interaktive Karten im Sinne der Weiterentwicklung von statischen Karten durch die Erweiterung um Rückkopplungsfähigkeiten zwischen Kartenersteller /-nutzer und Karte sieht Dickmann (2007) zum Einen als Suchhilfe nach Geoinformationen an, zum Anderen als Analyseinstrument. Als noch zu untersuchende Felder benennt er wahrnehmungspsychologische Aspekte und die Ebene der praktischen Anwendung. Lechthaler, Todor (2009) beschreiben beispielhaft den Einsatz kartographischer Informationssysteme zum Zwecke der Entscheidungsunterstützung bei Planungsfragen. Slocum et al. (2009, S. 408 f.) differenzieren vier Hauptziele der Datenexploration:

- Identifikation des räumlichen Musters eines bestimmten Attributes zu einer bestimmten Zeit
- Vergleich räumlicher Muster von zwei oder mehr Attributen zu einer bestimmten Zeit
- Identifikation von Veränderungen des räumlichen Musters eines bestimmten Attributes im Zeitverlauf
- Vergleich räumlicher Muster von zwei oder mehr Attributen im Zeitverlauf.

In der vorliegenden Arbeit werden insbesondere die ersten beiden Punkte behandelt. Als Methoden der Datenexploration betrachten Slocum et al. (2009) unter anderem die Manipulation der Rohdaten, Änderungen der Objekt-Zeichen-Zuordnung, Veränderung der Position des Betrachters, Hervorhebung von Teilen der Daten und Animationen. Im weiteren Verlauf der Arbeit wird gezeigt, welche Methoden aus dem Bereich der Manipulation der Rohdaten sich im Sinne einer weitgehenden Standardisierung des Datenexplorationsprozesses besonders gut zur Entscheidungsunterstützung eignen.

2.2 RÄUMLICHE STATISTIK, SPATIAL DATA MINING UND ENTSCHEIDUNGSUNTERSTÜTZUNG

Der Begriff Data Mining (DM) wird uneinheitlich definiert. Für die Diskussion der Definitionsversuche wird verwiesen auf Mitra, Acharya (2003) und Miller, Han (2001). DM kann verstanden werden als *„Prozess zur automatischen Auffindung interessanter versteckter Daten aus umfangreichen digital verfügbaren Datensammlungen"* (Umstätter 2005, o.S.) oder als *„teilweise automatisierte Suche nach versteckten Mustern in großen, multidimensionalen Datenbanken"* (Geoforschungszentrum Potsdam 2002, o.S.). DM stellt somit eine wesentliche Methode zur Generierung von neuen Informationen im Rahmen der Entscheidungsunterstützung dar. Spatial Data Mining (SDM) ist die Erweiterung von DM um die räumliche Dimension. Nach Slocum et. al. (2009, S. 492) sind die Methoden des SDM zwar komplexer als herkömmliche thematische Karten. Ihr Potential beim Aufdecken von Mustern, die ansonsten nicht sichtbar werden, sei jedoch enorm. Geostatistik (auch: räumliche Statistik) stellt nach Shekhar et. al. (2003) einen Teilbereich des SDM dar.

Der Gesamtprozess, in den SDM eingebunden ist, wird als Geographic Knowledge Discovery (GKD) bezeichnet. Der Ablauf des GKD wird in nachfolgender Abbildung 2 dargestellt. Die Autorin geht in ihrer Untersuchung auf die rot unterlegten Elemente Datenauswahl und -aufbereitung, SDM und Interpretation ein. Die zur Verfügung gestellten Daten werden zwar vor Analysebeginn aufbereitet und bereinigt, allerdings erfolgt dies nicht auf Basis eines Spatial Data Warehouses. Auch nach Kuonen (2006) ist ein solches keine zwingende Voraussetzung für DM. Die Ableitung von Wissen aus den Analyseergebnissen ist Aufgabe der

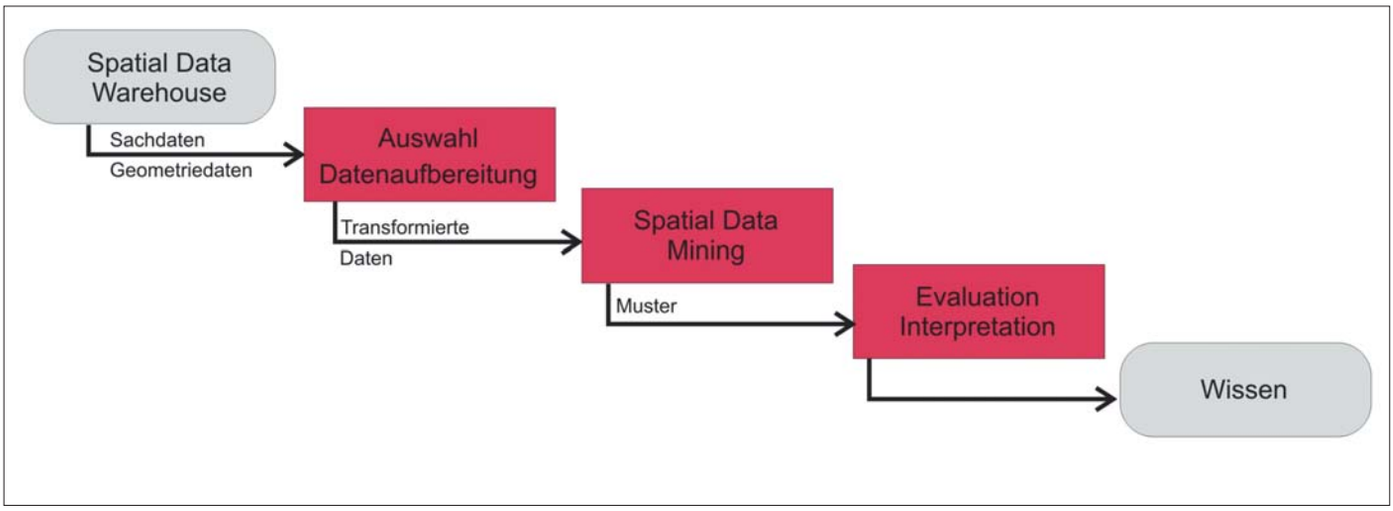

Abbildung 2: GKD-Prozess (Matatko 2008 verändert nach Mitra, Acharya 2003 und Miller, Han 2001)

Planungsexperten und somit ebenfalls nicht Bestandteil der Untersuchung.

Bei Entscheidungsunterstützungssystemen (engl.: Decision Support Systems, kurz DSS) handelt es sich nach Power (1998) um interaktive, computer-basierte Systeme, die Managern helfen sollen, Entscheidungen zu treffen. Er subsumiert unter dem Begriff zahlreiche Systeme, Hilfsmittel und Technologien, zudem diverse Software-Produkte, darunter Geographische Informationssysteme (GIS). Die Erweiterung eines DSS um die räumliche Dimension wird als Räumliches Entscheidungsunterstützungssystem (engl.: Spatial Decision Support System, kurz SDSS) bezeichnet. Clarke, Clarke (1995) kritisieren, dass ein konventionelles GIS den Bedürfnissen von Unternehmensfragen nicht gerecht werde und daher SDSS nötig seien, die sich in den Unternehmensprozess eingliedern. In Kapitel 3 dieses Artikels wird ein Ablaufschema vorgestellt, das als Grundlage zur Implementierung in einem SDSS genutzt werden kann.

2.3 INTERPRETATIONSPROBLEME BEI GEODATEN

Bei der Interpretation von räumlichen Daten treten zwei Probleme auf, die in diesem Kapitel näher erläutert werden: das Spatial Simpson's Paradox und das Modifiable Area Unit Problem (MAUP). Zu deren Erläuterung ist zunächst der Begriff der Granularität zu definieren.

Galton (2000) differenziert bei der Betrachtung von Granularität zwei verschiedene Ebenen: intrinsische und abbildende Granularität. Letztere bezieht sich auf die Korngröße einer photographischen Emulsion oder die Pixelgröße eines gedruckten Bildes und ist in dieser Arbeit nicht relevant. Erstere bezeichnet die verschiedenen Maßstäbe, in denen ein Sachverhalt eine signifikante Struktur aufzeigt. Sie ist bei räumlichen Fragestellungen in der Planung insofern von Interesse, dass Entscheider zunächst ein globales Bild von der Situation benötigen. Beim Auffinden unerwarteter Trends oder Variationen müssen sie Analysen auf niedrigeren räumlichen Aggregationsebenen durchführen, um Details und Ursachen aufzudecken. Auch eine inverse Vorgehensweise ist möglich: lokale Muster werden entdeckt, und Rückschlüsse darüber, wie sich das globale Bild darstellt, sollen gezogen werden. Die feinste Granularität bezieht sich auf das niedrigste Level der Datenaggregation in der Datenbank bzw. das detailreichste Informationslevel (Bédard et al. 2001). Die hierarchischen räumlichen Gliederungsstufen werden oft als *„Levels of Detail"* (LOD) bezeichnet.

Globale Muster müssen nicht mit denen auf lokaler Ebene übereinstimmen (sog. *„Spatial Simpson's Paradox"*). Daher empfiehlt sich die Anwendung der Technik des *„Spatial Slicing"*, eine Untersuchung räumlicher Muster auf verschiedenen Maßstabsebenen. Abbildung 3 zeigt zur Veranschaulichung des Sachverhalts eine Gegenüberstellung von lokalen und globalen Trends in räumlichen Daten. Die Erkenntnis der Maßstabsabhängigkeit räumlicher Muster bezeichnete Goodchild auf der UCGIS 2003 in Anlehnung an TOBLER'S *„Erstes Gesetz der Geographie"* als *„Zweites Gesetz der Geographie"* (Goodchild 2003 nach Shekhar, Zhang 2004, o.S.).

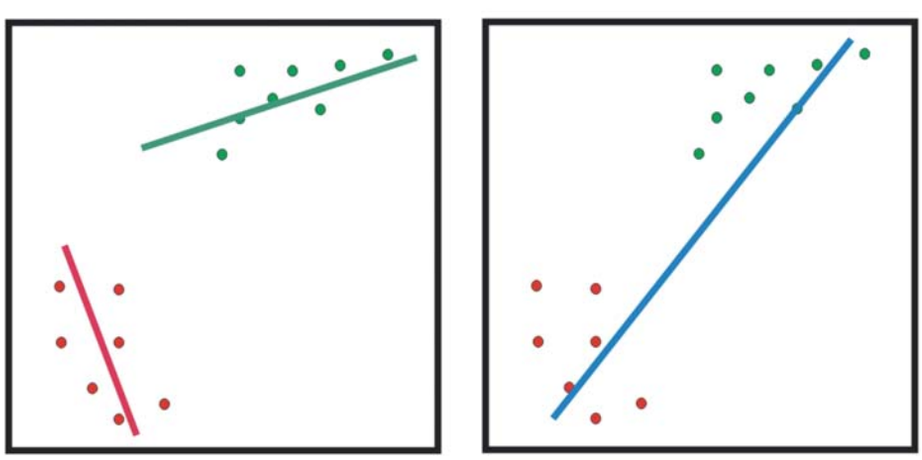

Abbildung 3: Regionale (links) und globale Trends (rechts) (Matatko 2008 verändert nach Shekhar et al. 2003)

Um eine Region in kleinere Raumeinheiten zu zerlegen, kann auf unterschiedliche Kriterien zurück gegriffen werden. Wird mit einer solchen Zerlegung weiter fortgefahren, entsteht im Ergebnis ein ineinander geschachteltes, hierarchisches, räumliches Gliederungssystem. Das Problem dieser Gliederungssysteme besteht darin, dass verschiedene Kriterien sich überschneiden können. Dadurch liegen zwei Zerlegungssysteme nicht exakt ineinander, sondern sie überlappen sich (Wong, Lee 2005 und Wrigley et al. 1996).

Wenn Daten auf unterschiedlichen hierarchischen Stufen eines räumlichen Gliederungssystems in einer räumlichen Analyse gemeinsam verwendet werden, können Ergebnisse entstehen, die nicht auf alle räumlichen Stufen übertragbar sind. Diese Inkonsistenz wird als *„scale effect"* bezeichnet. Wenn Daten aus verschiedenen hierarchischen Systemen eingesetzt werden, die zwar ungefähr die gleiche Anzahl an räumlichen Einheiten haben, aber deren Grenzen nach verschiedenen Kriterien bestimmt wurden, können die Ergebnisse ebenfalls Inkonsistenzen aufweisen, den sog. *„zoning effect"*. Der scale und der zoning effect ergeben gemeinsam das MAUP, das unter quantitativ orientierten Geographen schon lange diskutiert wird und in den 1990er Jahren ein wiederauflebendes Interesse erfuhr (Wong, Lee 2005).

Die Auswirkungen des MAUP auf Analyseergebnisse wurden bereits in zahlreichen Studien untersucht. Einen Überblick geben Fotheringham et al. (2000). So wächst beispielsweise der Korrelationskoeffizient mit der Größe der Raumeinheiten, weshalb in Frage gestellt werden kann, inwiefern er überhaupt auf räumlich aggregierte Daten angewendet werden kann.

Zur Lösung des MAUP existieren verschiedene Ansätze. Der einfachste beinhaltet die Erkenntnis, dass es das MAUP gibt, und dass daher Analysen auf mehreren Maßstabsniveaus durchgeführt werden sollten, um die Spannweite der Ergebnisse, die zu erhalten sind, aufzuzeigen. Andere Ansätze wollen räumliche Analysetechniken identifizieren oder einführen, die relativ maßstabsunempfindlich sind (Wrigley et al. 1996).

Neben dem MAUP identifiziert Openshaw (1996) das User Modifiable Area Unit Problem (UMAUP), da es den Datennutzern zunehmend möglich wird, die Zonierung der Daten selbst festzulegen. Brunner-Friedrich (2005) behandelt das UMAUP und die damit verbundene Gefahr der Fehlinterpretation von kartographischen Informationen bei interaktiven Anwendungen.

2.4 AUTOMATISIERUNG VON ENTSCHEIDUNGEN IN PLANUNGSFRAGEN

Planungsaufgaben sind gekennzeichnet durch unterschiedliche Schritte, in denen jeweils Entscheidungen getroffen werden müssen, die sich auf den weiteren Ablauf der Planung auswirken. Die einzelnen Schritte, die innerhalb des gesamten Planungsprozesses erfolgen müssen, ähneln sich bei räumlichen Fragestellungen stark. Daher ist es möglich, auf der Grundlage eines Modells des Planungsablaufs geeignete Methoden auszuwählen, die aus räumlichen Daten Mehrwert generieren können. Für den Einsatz im Rahmen eines SDSS sind insbesondere die Schritte interessant, die automatisiert abgebildet werden können. Abbildung 4 stellt den gesamten Planungsprozess schematisch dar.

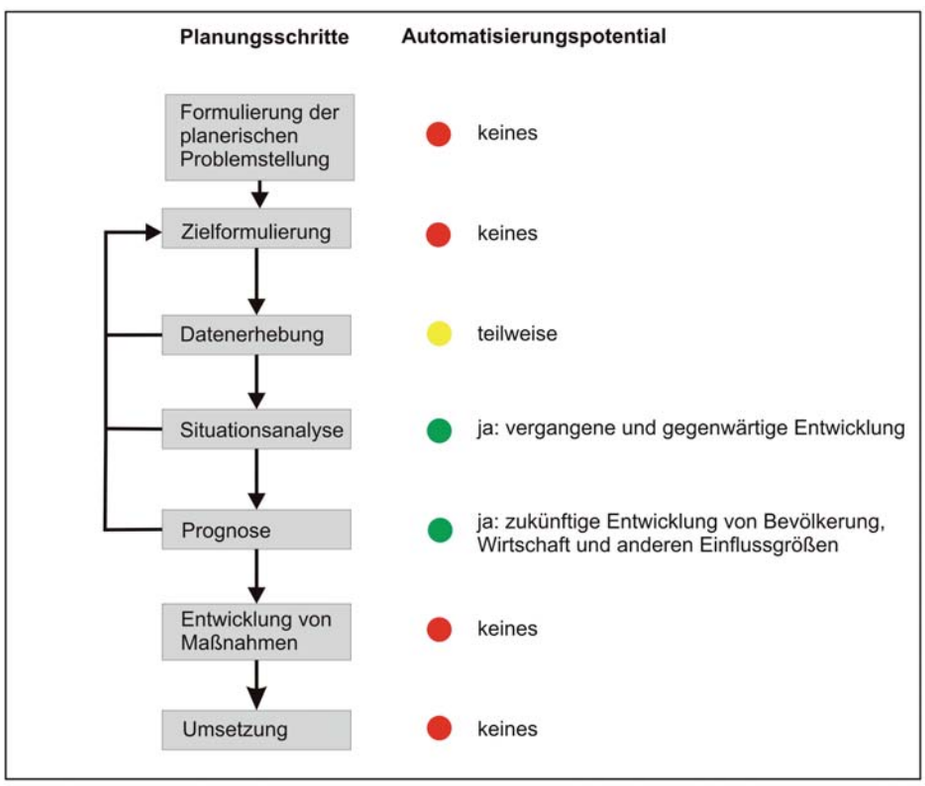

Abbildung 4: Ablaufschema eines Planungsprozesses und Automatisierungspotentiale (verändert nach Scholles *2000 und Konrad-Adenauer-Stiftung e.V. 2001)*

Kein Automatisierungspotential besteht bei den Planungsschritten, bei denen subjektive Einschätzungen und Werturteile handlungsleitend sind. Dies betrifft insbesondere die Frage, welche Probleme überhaupt als planungsbedürftig angesehen werden, sowie die mit der Planung verfolgten Ziele. Ebenso sind konkrete Handlungsmaßnahmen und deren Umsetzung nicht zu automatisieren. Der Bereich der Situationsanalyse und Prognose hingegen kann als automatisierbar im Sinne eines SDSS betrachtet werden. Denn bei der Analyse von gegenwärtiger und vergangener Verteilung von räumlichen Sachverhalten sowie Prognosen von deren zukünftiger Entwicklung kommen Analyseverfahren mit identischen Zielen zum Einsatz. Zu beachten ist, dass Zielformulierungen im Planungsprozess laufend überprüft und daraufhin Fragestellungen ggf. neu formuliert werden müssen. Neuformulierungen von Zielen können aus Analyseergebnissen resultieren. Da die Zieldefinition aber stark subjektiv geprägt ist, besteht auch bei deren Aktualisierung kein Automatisierungspotential.

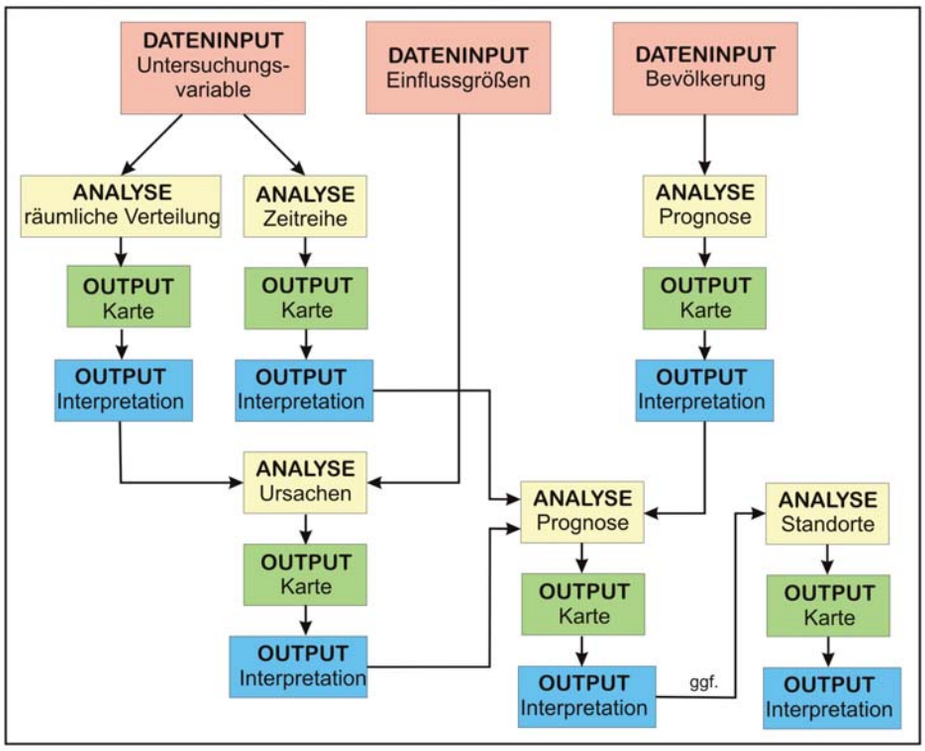

Abbildung 5: Analyseschema (verändert nach Matatko 2008)

2.5 ANALYSESCHRITTE

Das Analyseschema besteht aus drei Blöcken von Dateninputs. Hierbei handelt es sich um die eigentliche Untersuchungsvariable, verschiedene Einflussgrößen, die Auswirkungen auf die Untersuchungsvariable haben (könnten), sowie demographische Grundlagendaten zum jeweiligen Gebiet. Mit diesen Daten sollen entsprechend der zuvor festgelegten Automatisierungspotentiale in Planungsprozessen mehrere Analysen durchgeführt werden: die Darstellung der aktuellen sowie der vergangenen räumlichen Verteilung, die Korrelationen der Einflussfaktoren mit der Untersuchungsvariable und die Prognose von Bedarf und Zielgruppengrößen unter Einbeziehung demographischer Prognosedaten. Als Analyseoutput entsteht jeweils ein kartographisches Medium, das interpretiert werden muss. Abbildung 5 zeigt das komplette Ablaufschema aus Dateninput, Analyseschritten und Output. Das Anwendungsbeispiel in Kapitel 3 gibt nähere Erläuterungen zu den Bereichen Analyse der räumlichen Verteilung und Analyse der Ursachen. Für die Ausführungen zu den anderen Analyseschritten wird verwiesen auf Matatko (2008).

3. ANWENDUNGSBEISPIEL STUDIERENDE IN TRIER

3.1 BEGRÜNDUNG DER AUSWAHL DES BEISPIELS

Als Anwendungsbeispiel aus dem Bereich der Sozialplanung wird in dieser Arbeit die Verteilung der Studierendenwohnsitze in der Stadt Trier verwendet. Problemstellungen ergeben sich dabei insbesondere aus der Wohnungssituation. Daher ist von Interesse, ob sich die Studierenden in bestimmten Stadtteilen clustern, und ob es dabei zwischen den Studierenden verschiedener Studienrichtungen Differenzierungen gibt (die Universität besteht aus 2 Standorten). Des Weiteren werden im Rahmen der Sozialplanung die Statistischen Bezirke analysiert, um herauszufinden, ob besondere Gebiete existieren, in denen höherer sozialplanerischer Bedarf besteht, d.h. Gebiete mit überdurchschnittlichen Kinder-, Migranten-, Arbeitslosenzahlen oder ähnlichem. Konfliktpotential im Stadtraum besteht, falls es Cluster sozial schwacher Bevölkerungsschichten gibt, und diese mit Clustern von Studierenden übereinstimmen. Planerischer Handlungsbedarf entsteht, weil dadurch Konkurrenz im Wohnungsmarkt um ähnliche Wohnungstypen zu erwarten ist. Im Bereich der Sozialplanung wäre ein Einsatz von GIS-gestützten SDSS auch sinnvoll, um Bedarfsplanungen (z.B. Kindertagesstätten, Schulen) durchzuführen. Der Stadtverwaltung Trier liegen jedoch keine Anschriften der Kindergartenkinder bzw. Schüler vor. Daher werden die Studierenden betrachtet, bei denen die Bedarfsplanung sich jedoch nicht auf die Einrichtung Universität bezieht, sondern auf die von Studierenden benötigte sonstige Infrastruktur. Insgesamt gesehen ist der sozialplanerische Bedarf bei der Zielgruppe Studierende deutlich niedriger als bei anderen Bevölkerungsschichten. Da für Studierende jedoch eine georeferenzierbare Datenbasis verfügbar ist, werden sie exemplarisch genutzt, um die Vorgehensweise einer dem zuvor beschriebenen Workflow entsprechenden Analyse zu demonstrieren.

3.2 DATENGRUNDLAGEN

Die Anschriften der Studierenden der Universität Trier liegen tabellarisch vor. Die Tabelle beinhaltet die von Studierenden angegebene Kontaktadresse, die nicht immer mit dem Wohnsitz in Trier identisch sein muss. Von den 13.789 vorhandenen Anschriften liegen 9.828 im PLZ-Bereich 53, 54 oder 55 (entspricht dem Postleitzahlbereich Trier und Umland). Von diesen können 9.630 per Geokodierung räumlich verordnet werden. Dies entspricht einer Trefferquote von 98 %. 6.564 dieser geokodierten Anschriften befinden sich direkt im Trierer Stadtgebiet und werden daher in die Analyse einbezogen. Es ist davon auszugehen, dass die reale Anzahl der Studierenden, die einen Wohnsitz in Trier haben, ca. 30 % über den in dieser Analyse verwendeten Zahlen liegt. Dies führt dazu, dass Quotientenbildungen (z.B. *„Studierende pro Einwohner"*) nur Vergleiche der Stadtteile oder Statistischen Bezirke untereinander erlauben, nicht jedoch in ihrer Absolutheit bewertet werden dürfen. Ersteres erfolgt unter der Annahme, dass die nicht vorhandenen 30% der Fälle sich räumlich analog zu den erfolgreich geokodierten Trierer Anschriften verhalten. Auch dies ist jedoch nicht mit Gewissheit zu sagen und müsste, falls eine solche Analyse Grundlage von Planungen sein soll, genauer geprüft werden.

Die sozioökonomischen Beispieldaten, die für Korrelations- und Regressionsanalysen genutzt werden, stammen von der

Stadtverwaltung Trier. Es handelt sich einerseits um die Grenzen der Stadtteile und Statistischen Bezirke, die als Shapefile vorliegen. Daneben werden tabellarisch Arbeitslosen-, Ausländer-, Kinderzahlen, Altersverteilungen sowie Haushaltsgrößen zur Verfügung gestellt, die sich auf die Statistischen Bezirke bzw. Stadtbezirke beziehen.

3.3 SOFTWAREEINSATZ

Für die in Abbildung 5 dargestellten fünf Analyseschritte sind geeignete Berechnungsverfahren auszuwählen sowie Software, in der entsprechende Methoden implementiert sind. In Tabelle 1 werden jedem Analyseschritt typische Verfahren der räumlichen Statistik oder des Spatial Data Mining zugeordnet. In der Untersuchung wird nach dem Prinzip des Prototyping jeweils eine Methode / Software auf die Beispieldaten angewendet, und je nach Ergebnis sachlogisch entschieden, ob weitere Methoden zu testen sind oder das Ergebnis zufriedenstellend ist und mit dem nächsten Analyseschritt fortgefahren werden kann.

3.4 ANALYSEERGEBNISSE

In diesem Kapitel werden einige Analyseergebnisse präsentiert, die mit Hilfe der zuvor genannten Methoden entsprechend des Ablaufschemas erzielt werden konnten. Es handelt sich dabei um eine Auswahl der in Matatko (2008) dokumentierten Untersuchungsergebnisse.

Die Verteilung der Studierenden im Stadtgebiet wird zunächst mit Hilfe von Dichtekarten analysiert. Dabei interessiert neben dem Anteil der Studierenden an den Einwohnern auch die Studierendenzahl absolut, in Abbildung 6 dargestellt mit Hilfe der Kreisdiagramme. Deutlich wird, dass die Studierenden vor allem in der Innenstadt, den angrenzenden Stadtbezirken sowie in Nachbarschaft der Universität (diese befindet sich an der Grenze zwischen Neu-Kürenz und Tarforst) leben. In Filsch leben zwar relativ wenige Studierende, diese stellen dort jedoch einen hohen Anteil an den Einwohnern. Zu beachten ist, dass bei der Geokodierung nicht alle Studierenden erfasst wurden, und somit die Anteile an der Bevölkerung in der Realität noch höher liegen. Der Vergleich zwischen Statistischen Bezirken und Stadtbezirken macht kleinräumige Disparitäten sichtbar. So leben in einem Neubaugebiet in direkter Nachbarschaft der Universität unterdurchschnittlich wenige Studierende. Zudem wird deutlich, dass in einem der Statistischen Bezirke des Stadtbezirkes Matthias (am südlichen Rand der Innenstadt) überdurchschnittlich viele Studierende leben. Diese Ergebnisse zeigen, wie bereits in Kapitel 2.3 beschrieben, dass eine Betrachtung der räumlichen Sachverhalte auf mehreren räumlichen Ebenen erforderlich ist.

Analyseschritt	Verfahren aus Statistik, Spatial Data Mining	Software	Implementierte Methoden
Räumliche Verteilung	Explorative Datenanalyse: Maße für Streuung und Zentraltendenz, Histogramme	ArcGIS	Histogramme, Statistiken, Choroplethen (Graduated Colors)
		R	Package GeoXp: interaktive explorative Datenanalyse: Scatterplot, Histogramm
		Excel	Diagramme: Häufigkeiten, Anteile
	Cluster / Autokorrelation	ArcGIS	Nearest Neighbor, Moran's I, General G, Gi*, Anselin Local Moran's, Ripley's K
		Cluster-Seer	Morans I, Ripley's K, Getis Ord Local G, Besag & Newell, Anselin Local Moran's, ….
		CrimeStat	Nearest Neighbor, Moran's I, Geary's C, Moran's correlogram, Ripley's K, Anselin Local Moran's, …
		R	Nearest Neighbor, Moran's I, Hierarchische Clusterung, K-means, …
		STIS	Gi, Gi*, Local Moran, Turnbull
Ursache	Regression	R	Lineare Regression, GWR
Zeitreihe	Zeitreihenanalyse: Streudiagramm, Polygonzug, Trendextrapolation, gleitender Durchschnitt	ArcGIS	Diagramme (charts)
		Cluster-Seer	Zeitliche oder raum-zeitliche Cluster: Grimson's, Knox's, Kulldorf's, …
		CrimeStat	Knox, Mantel
		STIS	Visualisierung
Standorte	Standortanalyse: Potentialanalyse, Wettbewerbsanalyse	ArcGIS	Datenverschneidung
Prognose	Zeitreihenanalyse: Trendextrapolation, gleitender Durchschnitt	ArcGIS	Datenverschneidung
		CrimeStat	Correlated walk analysis

Tabelle 1: Softwareeinsatz für die einzelnen Analyseschritte (verändert nach Matatko 2008)

Eine vertiefende Betrachtung der räumlichen Verteilung der Studierenden kann mit einer Hot-Spot-Analyse in ArcGIS durchgeführt werden. Das Ergebnis zeigt Abbildung 7, wiederum auf Ebene der Stadtbezirke und der Statistischen Bezirke. In rot sind jeweils die Gebiete dargestellt, die bei einem Signifikanzniveau von 5% überdurchschnittlich hohe Werte aufweisen. Im Vergleich zu

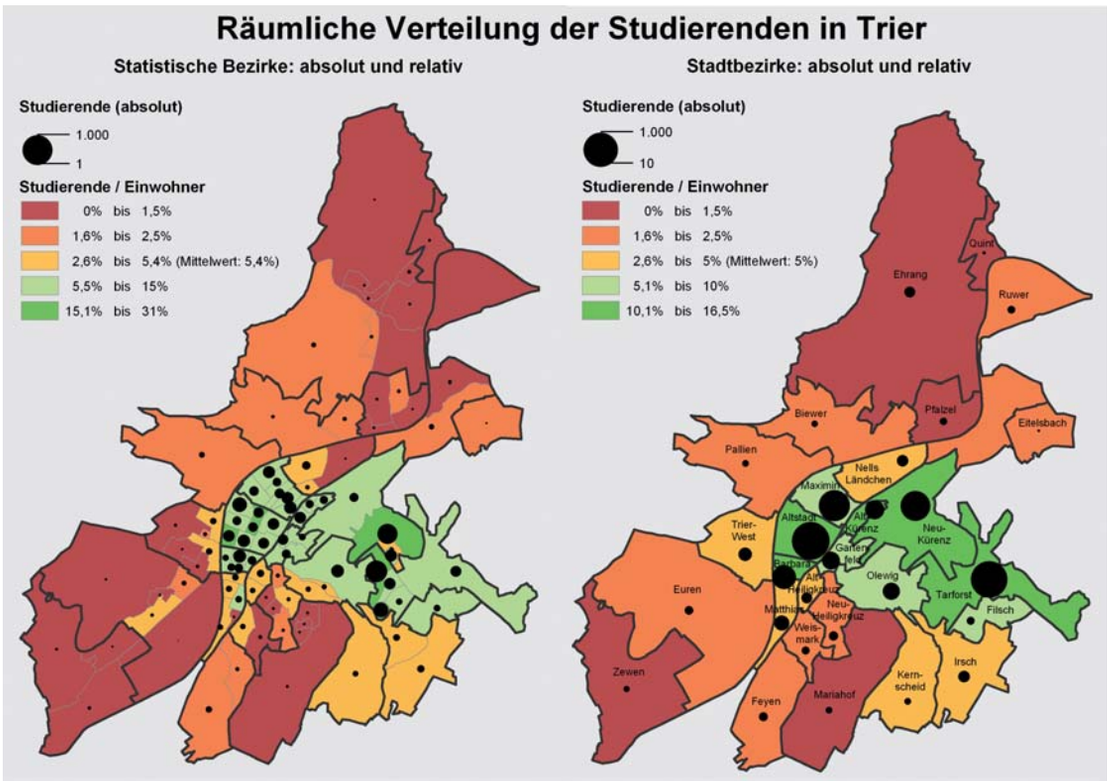

Abbildung 6: Verteilung der Studierenden in Trier (Datenquelle: Stadtverwaltung Trier 2006, Universität Trier 2004).

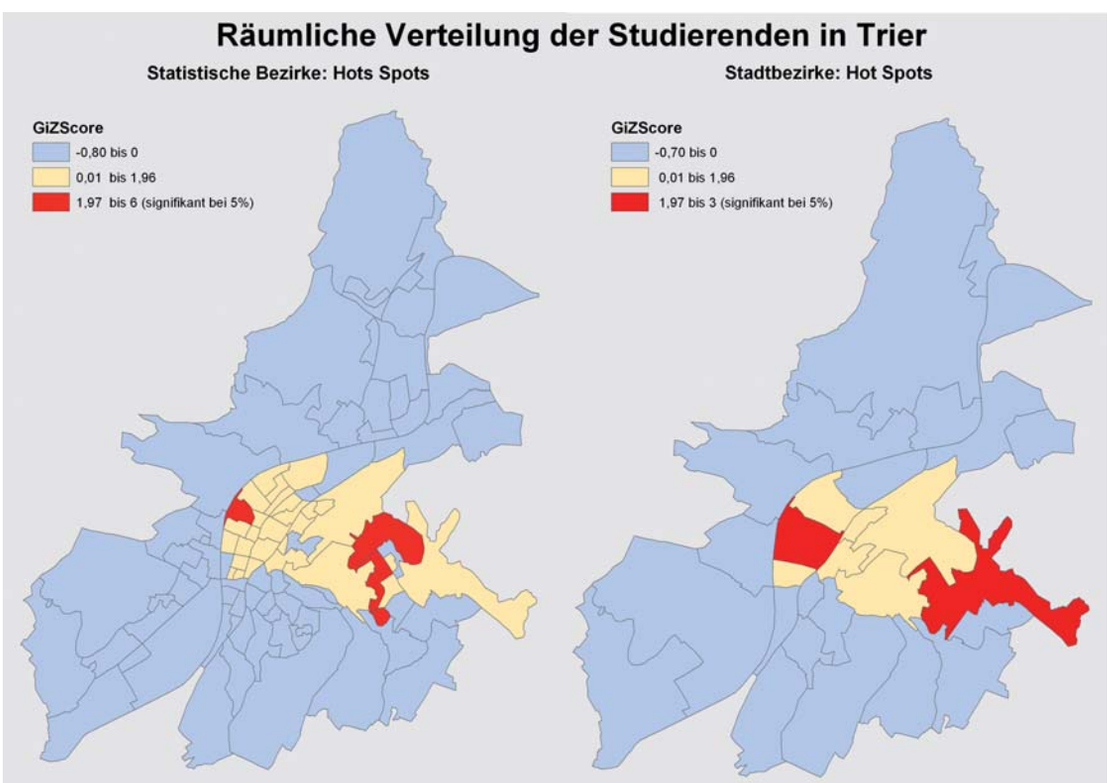

Abbildung 7: Hot Spots der Studierenden in Trier (Datenquelle: Stadtverwaltung Trier 2006, Universität Trier 2004).

Abbildung 6 ist festzuhalten, dass es sich dabei nur um eine kleine Teilmenge der Gebiete handelt, in denen die Relation Studierende pro Einwohner größer ist als der Mittelwert.

Eine Regressionsanalyse zur Suche nach Ursachen der räumlichen Verteilung wird in R durchgeführt. Die Regressionsgleichung wurde zunächst mit allen vorhandenen Variablen erstellt. Anschließend wurden nur die signifikanten Regressionskoeffizienten beibehalten (das Signifikanzniveau liegt jeweils unter 5%): die Arbeitslosenzahl sowie die Ausländerzahl pro Einwohner. Zudem wurde der Stadtteil Eitelsbach aus dem Datensatz entfernt, da dieser in den ersten Durchläufen zu starken Verzerrungen der Ergebnisse führte. Wegen der geringen Einwohnerzahl sind dort aus Datenschutzgründen keine Angaben zu Arbeitslosenzahlen verfügbar. Der komplette Analyseausschluss konnte das Regressionsmodell erheblich verbessern und erbrachte auf der Ebene der Stadtbezirke als Ergebnis R^2 in der Höhe von 48%. Die Ausländerquote beeinflusst die Studierendenzahl positiv, die Arbeitslosenquote negativ.

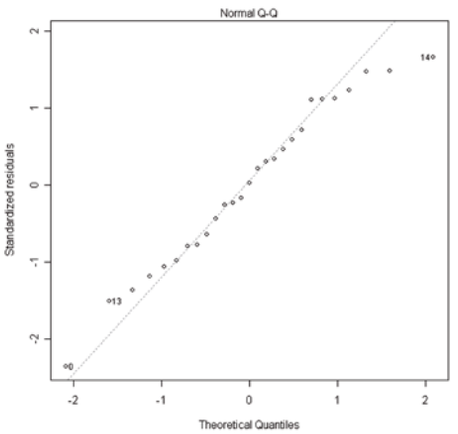

Abbildung 8: Q-Q-Plot der Residuen der Regressionsgleichung „Studierende" (Matatko 2008)

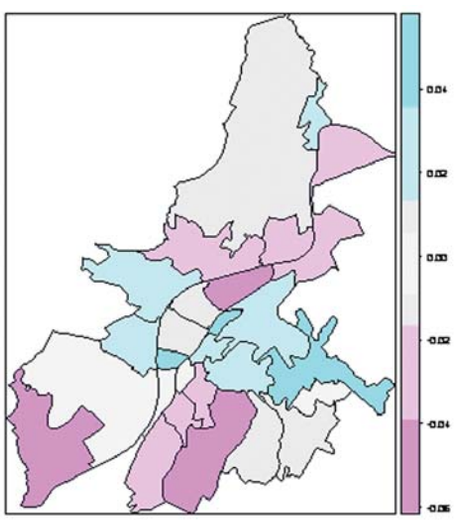

Abbildung 9: Residuen-Mapping der Regressionsgleichung „Studierende" (Matatko 2008)

Diverse Graphen werden in R mit dem Befehl plot() gezeichnet. An dieser Stelle ist der Q-Q-Plot der Residuen der Regressionsgleichung von Interesse (Abbildung 8). Es zeigt sich deutlich, dass sie fast alle auf einer Gerade liegen, jedoch am oberen und unteren Ende einige Ausreißer zu finden sind.

Die kartographische Darstellung der Residuen (Abbildung 8) gestaltet sich sehr heterogen. Auffallend ist der sehr hohe negative Wert im nördlichen Teil des Stadtteils Trier-Nord. Dieser ist darauf zurückzuführen, dass in diesem Gebiet hohe Ausländeranteile (Unterbringung in Heimen) zu finden sind, jedoch nur eine geringe Anzahl an Studierenden dort leben. Besonders hohe positive Werte befinden sich im direkten Umfeld der Universität. Dort befinden sich zahlreiche Studierendenwohnheime, es handelt sich jedoch um Wohngebiete mit - gemessen an der Einwohnerzahl - relativ niedrigen Ausländeranteilen und zum Teil relativ hohen Arbeitslosenquoten.

Um auch eine räumliche Variable in die Regressionsanalyse einzubeziehen, wurde die Entfernung des Mittelpunktes jedes Stadtbezirkes zur Universität berechnet. Die Regressionsgleichung, ohne Einbezug von Eitelsbach, führt zu R^2 von 56%, somit ist das Regressionsmodell besser zu bewerten als das vorhergehende. Allerdings sind die Regressionskoeffizienten kleiner. Die Entfernung beeinflusst die Anzahl der Studierenden pro Einwohner negativ, d.h., je größer die Entfernung des Mittelpunktes des Stadtbezirkes zur Universität, desto weniger Studierende leben dort. Die bereits zu Abbildung 6 aufgestellte Vermutung, dass die Nähe zur Universität wesentlich ist für die Wohnstandortwahl der Studierenden, kann somit auch statistisch belegt werden.

4 FAZIT

Die Untersuchung konnte insgesamt zeigen, dass ein vordefinierter Pool an Analysemethoden entsprechend eines standardisierten Ablaufs im Planungsprozess auf geläufige Planungsaufgaben angewendet werden kann und Geodaten im Sinne der kartographischen Datenexploration neue Erkenntnisse über räumliche Sachverhalte liefern können. Die Analyse der Ursachen muss statistisch gesehen weitere Kriterien berücksichtigen, wie z.B. die Multikollinearität der Regressionsvariablen. Der Workflow ist somit hinsichtlich der Voraussetzungen der Untersuchungsvariable und der Einflussgrößen weiter zu detaillieren. In Bezug auf die Methoden und die Software konnten beim Test des Workflows anhand der Beispiel-Anwendung Favoriten heraus gearbeitet werden, die als besonders praxistauglich empfunden werden. Die Ausführungen beziehen sich dabei auf die komplette Untersuchung aus Matatko (2008).

Nach Ansicht der Verfasserin erweist sich die Herstellung diverser Dichtekarten als positiv, da diese insbesondere durch das Übereinanderlegen von über- oder unterdurchschnittlichen Werten verschiedener Karten sehr zuverlässige Ergebnisse liefern konnten. Zu diesen fehlen zwar Angaben zu Signifikanzniveaus. Hierbei kann aber die vertiefende Betrachtung mit einer Hot-Spot-Analyse (Gi*) helfen. Letztere gibt, gerade wenn auf den ersten Blick die räumlichen Zusammenhänge nicht klar erkennbar sind, weitere Einsichten in die Daten. Wie ein Vergleich der Ergebnisse des Gi* mit denen der Dichtekarten gezeigt hat, sind die Muster zwar nicht exakt identisch, sie stimmen jedoch in den relevanten Minima und Maxima überein.

Die Regressionsanalyse liefert insbesondere durch das Residuen-Mapping für die räumliche Planung interessante Ergebnisse. Als weniger hilfreich erwiesen sich bei der Untersuchung räumlicher Verteilungen und Zusammenhänge die Analysewerkzeuge der Spatial Statistics Tools in ArcGIS (außer Gi*). Die Ergebnisse zeigen zwar statistische Signifikanzen, die inhaltlichen Aussagen sind jedoch für die Anwendung in der Planungspraxis wenig hilfreich, da Handlungsbedarf nicht nur von statistischen Signifikanzwerten abhängig gemacht wird, sondern stark subjektiv beeinflusst wird. Somit erscheinen kartographische Darstellungen hilfreicher als die Angabe einzelner Signifikanzmaße.

Zu kleinräumigen räumlichen Prognosen will die Verfasserin dringend raten. Durch räumliche Trendextrapolation können neue Einsichten über zukünftige Entwicklungen gewonnen werden, und negative Entwicklungen durch gezielte Maßnahmen aufgehalten werden. Dies konnte die Verfasserin in Matatko (2008) an einem Planungsbeispiel aus dem Geomarketing aufzeigen.

Da in ArcGIS erst wenige statistische Funktionen implementiert sind, aber die kartographischen Möglichkeiten als sehr hilfreich bei der Entscheidungsfindung bewertet werden, wird als Ergänzung zu ArcGIS-Analysen der Einsatz der OpenSource Statistik-Software R empfohlen. Derzeit sind die Standard-Kartenausgaben in R noch sehr einfach gehalten, und für kleinere Veränderungen am Karten-Layout ist ein vertieftes Studium der R-Kommandos nötig, weshalb die kartographische Visualisierung zur Präsentation der Analyseergebnisse im GIS einfacher umzusetzen ist. Weitere Software aus Tabelle 1 wurde getestet. Da das Ziel der Arbeit jedoch darin bestand, für die Planungspraxis sinnvolle Analyseschritte in eine standardisierte Abfolge zu bringen, wird zu der Kombination aus ArcGIS (oder einem sonstigen Geoinformationssystem) und R geraten, da dadurch ein Umstieg auf weitere, oft kostenintensive, und den Planern nicht vertraute Systeme vermieden wird. ◀

Literatur

Bédard, Y.; Merrett, T.; Han, J. (2001): Fundamentals of spatial data warehousing for geographic knowledge discovery. In: Miller, H. J.; Han, J. (Hrsg.): Geographic Data Mining and Knowledge Discovery. Routledge Chapman & Hall, S. 53-73.

Brunner-Friedrich, B. (2005): Interaktivität - Nachteil oder Potenzial bei der Erschließung unscharf abgrenzbarer Sachverhalte? In: Strobl., J.; Blaschke, T.; Griesebner, G. (Hrsg.): Angewandte Geoinformatik 2005. Beiträge zum 17. AGIT-Symposium Salzburg. Wichmann, S. 66-75.

Clarke, G.; Clarke, M. (1995): The development and benefits of customized spatial decision support systems. In: Longley, P.; Clarke, G. (Hrsg.): GIS for Business and Service Planning. GeoInformation International, S. 227-245.

Dickmann, F. (2007): Die Rolle moderner Geovisualisierungsmethoden in der Humangeographie. In: Dickmann, F. (Hrsg.): Geovisualisierung in der Humangeographie. Nutzung kartengestützter Informationssysteme in Wissenschaft und Praxis. Kirschbaum, S. 9-19.

Fotheringham, A. S.; Charlton, M.; Brunsdon, C. (2000): Quantitative Geography. Perspectives on Spatial Data Analysis. Sage Pubn Inc.

Galton, A. (2000): Qualitative Spatial Change. Oxford University Press.

Geoforschungszentrum Potsdam (2002): Spatial Mining for Data of Public Interest. (online: http://www.gfz-potsdam.de/pb2/pb21/spin_projekt/spin_projekt.html, Zugriff 05/2006).

Grimshaw, D. J. (1994): Bringing Geographical Information Systems Into Business. John Wiley & Sons.

Konrad-Adenauer-Stiftung e.V. (2001): E-learning Kommunalpolitik. Prozeß: Planungsablauf und Beteiligte. (online: http://www.kas.de/kommunal/e-learning/detail.php?graf=41, Zugriff 01/2010).

Kuonen, D. (2006): Data Mining Myths Versus Realities. (online: http://www.statoo.com/en/datamining/DMmyths.pdf, Zugriff 05/06).

Lechthaler, M.; Todor, R. (2009): Allgemeingültiges Konzept zur Unterstützung räumlicher Entscheidungen in unterschiedlichen Planungsbereichen. In: Kartographische Nachrichten, Nr. 6, 2009, Kirschbaum, S. 309-315.

MacEachren, A. M. (1994): Visualization in modern cartography: Setting the agenda. In: MacEachren, A. M.; Taylor, D. R. F. (Hrsg.): Visualization in Modern Cartography. Elsevier, S.1-12.

Matatko, A. (2008): Raumbezogene Statistik und Visualisierung zur Entscheidungsunterstützung in der Planung. GIS- und Spatial Data Mining-Methoden im Vergleich. Selbstverlag der Geographischen Gesellschaft Trier.

Miller, H. J.; Han, J. (2001): Geographic data mining and knowledge discovery: an overview. In: Miller, H. J.; Han, J. (Hrsg.): Geographic Data Mining and Knowledge Discovery. Routledge Chapman & Hall, S. 3-32.

Mitra, S.; Acharya, T. (2003): Data Mining. Multimedia, Soft Computing, and Bioinformatics. John Wiley & Sons.

Müller, A. (2005): Datenexploration und Wissenskommunikation in der Geovisualisierung. In: Kartographische Nachrichten, Nr. 5, 2005, Kirschbaum, S. 236-243.

Openshaw, S. (1996): Developing GIS-relevant zone-based spatial analysis methods. In: Longley, P.; Batty, M. (Hrsg.): Spatial Analysis: Modelling in a GIS Environment. John Wiley & Sons, S. 55-73.

Power, D. J. (1998): What is a Decision Support System? (online: http://dssresources.com/papers/whatisadss/, Zugriff 01/2010).

Scholles, F. (2000): Gesellschaftswissenschaftliche Grundlagen: Planungsmethoden. (online: http://www.laum.uni-hannover.de/ilr/lehre/Ptm/Ptm_WissArb.htm, Zugriff 01/2010).

Shekhar, S; Zhang, P. (2004): Spatial Data Mining: Accomplishments and Research Needs. (online: http://www-users.cs.umn.edu/~shekhar/talk/giscience04keynote.pdf, Zugriff 01/2010).

Shekhar, S; Zhang, P.; Huang, Y.; Vatasavai, R. R. (2003): What's Special about Spatial Data Mining? (online: http://www.cs.umn.edu/research/shashi-group/paper_ps/sdm_slide_0903.pdf, Zugriff 01/2010).

Slocum, T. A.; McMaster, R. B.; Kessler, F. C.; Howard, H. H. (2009): Thematic Cartography and Geovisualization. Pearson.

Stadtverwaltung Trier (2006): Geodaten Stadtbezirke und Statistische Bezirke, Attributdaten des Amtes für Stadtentwicklung und Statistik zu Stadtbezirken und Statistischen Bezirken.

Umstätter, W. (2005): Semiotischer Thesaurus. (online: http://www.ib.hu-berlin.de/~wumsta/wistru/definitions/hierdef01b.pdf, Zugriff 01/2010).

Universität Trier (2004): Studierendenwohnsitze. Anonymisierte Anschriften aus der Studierendendatenbank.

Wong, D. W.; Lee, J. (2005): Statistical Analysis of Geographic Information with ArcView GIS® and ArcGIS®. John Wiley & Sons.

Wrigley, N.; Holt, T.; Steel, D.; Tranmer, M. (1996): Analysing, modelling and resolving the ecological fallacy. In: Longley, P.; Batty, M. (Hrsg.): Spatial Analysis: Modelling in a GIS Environment. John Wiley & Sons, S. 25-40.