

Universität Karlsruhe, Fakultät für Informatik

Institut für Programmstrukturen und Datenorganisation

Lehrstuhl für Systeme der Informationsverwaltung

Seminar Imperfektion und Datenbanken

Imperfekte Daten in GIS-Anwendungen

Olga Gromova

Betreuer: Jutta Mülle

WS2003/2004

# Kurzfassung:

Moderne geographische Informationssysteme können große Datenmengen speichern und verwalten. Sie haben hunderte von Funktionen und sind bedeutende Werkzeuge für die Hilfe bei Entscheidungsprozessen.

Der größere Teil von GIS basiert heutzutage auf der booleschen Logik. Das führt zu künstlicher Präzision in der Abbildung von geographischen Daten, die unpräzise sind, und löst fundamentale Probleme in der Repräsentation, Manipulation und Analyse von Information aus.

Diese Arbeit befasst sich mit den verschiedenen Ansätzen, die bei der Repräsentation, Speicherung und Bearbeitung von unpräzisen und vagen Daten in geographischen Informationssystemen verwendet werden.

# Inhalt:

1	Einleitung.....	3
2	Motivation.....	4
3	Ansätze.....	4
3.1	Quantitative Reasoning .....	5
3.2	Qualitative Reasoning .....	6
3.3	Fuzzy Logik.....	7
3.4	Rough Sets.....	10
4	Zusammenfassung.....	11
5	Literatur .....	12

# 1 Einleitung

„Geographische Informationssysteme (GIS) sind computergestützte Werkzeuge und Methoden, mit denen man flächenbezogene geographische Daten erheben, verwalten, ändern und auswerten kann“ (BUHMANN, 1996).

Geographische Informationssysteme unterscheiden sich von anderen Informationssystemen durch den Raumbezug der Daten. Während andere Informationssysteme meistens nur thematische Daten beinhalten, werden bei GIS auch raumbezogene Daten gespeichert. Auch die Beziehungen zwischen den thematischen und geometrischen Daten sind wesentliche Bestandteile des Geo-Informationssystems.

Zu den geographischen Daten gehören räumliche Informationen, wie Breite, Länge, Fläche und Nachbarschaftsbeziehungen sowie beschreibende Attribute, die Eigenschaften geometrischer Daten enthalten, wie zum Beispiel zusätzliche Informationen zur Landschaft in einem Landschaftsplanungssystem.

Die Speicherung von geometrischen Daten in der Datenbank ermöglicht eine leichtere Bearbeitung von Daten und kann eine Entscheidungshilfe zu vielen Anwendungsbereichen sein.

Geodaten beschreiben einzelne Objekte der Landschaft. Diese sind über den Raumbezug miteinander verknüpft.

Die Form und die relative Lage von Objekten werden meistens als Koordinaten in einem bestimmten Bezugssystem beschrieben. Die Beschreibung der Objekte kann in Vektor- oder Rasterform sein. Darüber hinaus wird die Topologie berücksichtigt. Die Topologie beschreibt, wie verschiedene Objekte miteinander in Beziehung stehen.

Bei der Rasterform sind Pixel in Zeilen und Spalten regelmäßig angeordnet. Zwischen den einzelnen Pixeln bestehen jedoch keine logischen Verbindungen. Bei der Rasterform wird zwischen Punkt, Linie und Fläche nicht unterschieden. Im Gegensatz zur Rasterform sind bei der Vektorform Punkte und Linien graphische Grundstrukturen. Dabei entsteht eine logische Datenstrukturierung.

Je nach Anwendung und je nach Daten werden Vektor- oder Rasterdaten verwendet. Vektordaten finden sich meist in Karten mit großem Maßstab, Rasterdaten in Karten mit mittlerem und kleinem Maßstab.

Eine Umwandlung von Vektordaten in Rasterdaten ist einfach möglich, umgekehrt aber mit größerem Aufwand verbunden. Aus diesem Grund werden häufig hybride Geoinformationssysteme verwendet, die die Daten in Raster- als auch in Vektorform speichern.

Als Beispiele für GIS kann man folgende Systeme nennen:

- Landschaftsplanungssysteme
- Biogeographie- und Botaniksysteme
- Kanalnetzsysteme
- Routenermittlungssysteme
- Transport- und Logistiksysteme

## 2 Motivation

Die vagen Begriffe und Daten sind ein wichtiger Bestandteil in der Forschung der Geographie. Solche vagen Informationen soll man auch in den Datenbanken abbilden können.

Der größere Teil von modernen GIS basiert auf der booleschen Logik, was zu einer künstlichen Präzision in der Abbildung von Geodaten führt.

Da die Geodaten oft unpräzise und unvollständig und die Abfragebedingungen oft vage sind, sollte man auch mit solchen Daten und Abfragen umgehen können.

Diese Seminararbeit befasst sich damit, wie man solche Daten speichern und bearbeiten kann.

## 3 Ansätze

In dieser Seminararbeit werden folgende Ansätze betrachtet:

- Quantitative Reasoning
- Qualitative Reasoning
- Fuzzy Sets
- Rough Sets

Außerdem werden Vor- und Nachteile von diesen Ansätzen anhand von Beispielen beschrieben.

### 3.1 Quantitative Reasoning

Quantitative Reasoning ist die traditionelle Form von Reasoning. Diese Form funktioniert sehr gut, wenn man mit präzisen Daten und Abfragebedingungen arbeitet.

Das quantitative Reasoning ist gut untersucht und weit verbreitet. Für diesen Ansatz stellen die modernen Datenbanksysteme viele verschiedene Werkzeugen zur Verfügung. Deswegen ist dieser Ansatz ein bedeutendes Mittel für die Datenverarbeitung. Der größte Teil von modernen GIS basiert auf diesem Ansatz.

Der Nachteil dieses Ansatzes ist, dass er die Abbildung von vagen Daten und die Abfragen mit unpräzisen Bedingungen nicht zulässt bzw. künstlich präzise macht.

#### **Beispiel:**

Als Beispiel betrachten wir folgende Abfragebedingungen:

Suche den Platz für das Bauen von einem Werk mit folgenden Bedingungen:

- Ca. 1000 m<sup>2</sup> groß
- Weit vom Trinkwasser
- Nah an der Strasse

Die Begriffe „weit“, „nah“, „ca.“ sind hier ungenau.

In geographischen Informationssystemen, die auf dem Quantitative Reasoning basieren, wird bei einer solchen Abfrage die boolesche Logik verwendet. Alle Bedingungen werden dabei in präzise Bedingungen umgewandelt.

In unserem Beispiel wäre eine mögliche Lösung, wenn man alle Plätze, die genau 1000 m<sup>2</sup> groß sind, selektiert, alle Plätze, die zum Beispiel 2000 m vom Trinkwasser entfernt sind und alle Plätze, die 500 m von der Fahrstrasse entfernt sind auswählt und dann eine große Menge berechnet. Dabei werden zum Beispiel die Plätze, die 501 m von der Fahrstrasse entfernt sind, nicht berücksichtigt. Es kann aber sein, dass dieser Platz unseren realen Anforderungen entspricht. Es kann auch sogar sein, dass dieser die einzige mögliche Lösung ist.

Diese künstliche Präzision ist der größte Nachteil von dem quantitativen Reasoning. Deswegen ist dieser Ansatz für die Bearbeitung von unpräzisen Abfragen nicht geeignet.

### 3.2 Qualitative Reasoning

Dieses Problem kann man lösen, indem statt quantitativen Daten die qualitativen verwendet werden. Das System kann dann solche vagen Begriffe wie „weit“, „gross“, „klein“ umgehen. Dies erreicht man durch die Einführung von linguistischen Variablen.

Die linguistische Variable ist eine Variable, die als Werte Worte oder Sätze in natürlicher Sprache annehmen kann. Die Werte, die die Variable haben kann, heißen linguistischen Werte.

#### **Beispiel:**

Als Beispiel für diesen Ansatz betrachten wir folgende Information:

„Das Objekt A gehört zum Stadtteil „Zentrum“.

Man kann für die qualitative Beschreibung von Koordinaten in der Stadt eine Variable  $x$  definieren, die die Werte {„Zentrum“, „Ost“...} annehmen kann. Die Domain von dieser Variable wird dann

$$L(x) = \text{„Zentrum“} + \text{„Ost“} + \dots$$

Wenn das Objekt A im Zentrum ist, kann die entsprechende Restriktion so formuliert werden:

$$R(x) = \text{„Zentrum“}$$

Räumliche Relationen zwischen Objekten können auch durch die Komposition von linguistischen Variablen definiert werden.

Zum Beispiel kann die räumliche Relation zwischen zwei Objekten A und B durch die Variable  $(x_A, x_B)$  ausgedrückt werden. Die Domain wird dann

$$L(x_A, x_B) = L(x_A) \times L(x_B)$$

Die entsprechende Restriktion lautet :

$$R(x_A, x_B) \subseteq L(x_A) \times L(x_B)$$

Man unterscheidet zwischen interaktiven und noninteraktiven linguistischen Variablen. Zwei Variablen sind noninteraktiv, wenn die Relation auf  $(x_A, x_B)$  gleich des kartesischen Produkts von Restriktionen auf  $x_A$  und auf  $x_B$  ist.

Normalerweise sind die Relationen in GIS durch Restriktionen von Werten der interaktiven Variablen definiert.

**Beispiel:**

Die Distanz zwischen zwei Objekten A und B kann durch Relation  $(x_A, x_B)$  definiert werden, wobei  $x_A, x_B$  Positionen von A und B sind.  $x_A, x_B$  sind hier interaktiven Variablen.

Linguistische Variablen ermöglichen also qualitative Relationen auszudrücken. Für die Integration von quantitativen und qualitativen Daten reicht es aber nicht. Dies ermöglicht nur die Verwendung von Fuzzy Logik.

### 3.3 Fuzzy Logik

Fuzzy Logik wurde im Jahr 1965 von L.A. Zadeh vorgestellt. Allgemein ist die Fuzzy Logik eine Erweiterung von Boolesche Logik. Mit Hilfe von der Fuzzy Logik kann man mit unpräzisen Daten in meisten GIS umgehen.

	General Arguments for vagueness	Fuzzy semantic relation model	Fuzzy semantik import model
Climate	Moraczewski	McBratney and Moore	Leung
Vegetation		Dale, Roberts	
Soil/Land evaluation		McBratney and De Gruijter	Burrough, Lagacherie, Wand, Hall
Remote sensing		Fischer and Pathirana Foody Robinson and Strahler Wang, Wilkinson	
Landscape	Fischer and Wood	Usery, Wood	
Natural language		Altman, Fischer, Orf Robinson, Wang	

Tab. 1 Forschungsbeispiele im Bereich Fuzzy Ansatz in GIS [aus 5]

Im Gegensatz zur booleschen Logik ermöglicht Fuzzy Logik auch die Abbildung unpräziser und vager Daten. Fuzzy Logik verwendet man zum Beispiel bei der Bearbeitung von Klassifikationsfehlern oder bei Ungenauigkeiten der Objektrahmen. Im Jahr 1985 wurde auch eine Repräsentationssprache entwickelt, die Abfragen in natürlicher Sprache verarbeitet und auf der Fuzzy Logik basiert. Einige Beispiele für Forschungen im Bereich des Fuzzy Logik Ansatzes in GIS sind in Tabelle 1 angezeigt.

Die Fuzzy Untermenge für Domain  $D$  wird als eine Menge von Paaren  $\langle d, \mu(d) \rangle$  definiert, wobei  $d \in D$  und  $\mu$  eine Zugehörigkeitsfunktion  $\mu: D \rightarrow [0,1]$  ist. Es gibt verschiedene Ansätze, wie man  $\mu(d)$  finden kann. In GIS werden folgende Ansätze verwendet:

1. Fuzzy Semantik Relationsmodel: basiert auf der Clusteranalyse; dabei werden die Mengen von Daten nach Mustern untersucht.
2. Semantisches Import Model: dabei wird die Zugehörigkeitsfunktion durch ein Expert Model berechnet.

**Beispiel (aus [3]):**

Wir definieren eine Fuzzy Menge für das Begriff „Zentrum“ als

$$R = \langle M 1, 1 \rangle + \langle M 2, 0.3 \rangle + \langle M 3, 0 \rangle + \dots$$

Wobei für jedes Objekt ein Zugehörigkeitswert definiert wird.

D.h. jedes Objekt ist mehr oder weniger „im Zentrum“. Die Objekte, deren Zugehörigkeitswert gleich 1 ist, sind definitiv zentral, die Objekte, die den Zugehörigkeitswert gleich 0 haben, sind entsprechend nicht in Zentrum.

Allgemein, wird in GIS jede Objektlokationsinformation wie Länge und Breite mit vagen Informationen, die durch eine Fuzzy Menge repräsentiert ist, assoziiert. Die Stelligkeit der Fuzzy Menge kann 1 oder auch größer als 1 sein. Im letzten Fall wird dadurch eine Relation, wie zum Beispiel Nachbarschaft zwischen Objekten, beschrieben.

Durch das Definieren von Fuzzy Menge und Verwendung von Inference Regeln kann man vage Abfragen bearbeiten.

Dieser Ansatz hat aber auch Nachteile. Einer davon ist die Relativität von Zugehörigkeitswerten. Wir betrachten folgende Beispiele solcher Relativität:

- Distanzrelativität
- Relativität, die durch verschiedene Maßstäbe verursacht wird

- Größe des Objektes
- andere Faktoren wie z.B. Reisekosten

**Distanzrelativität.** Die subjektive Distanzabschätzung zwischen zwei Objekte kann abhängig von einem dritten Objekt sein. Zum Beispiel kann die Entfernung zwischen London und Milan (s. Abb. 1) subjektiv als „groß“ eingeschätzt werden, weil Paris näher an London als Milan ist. Wenn man aber diese Karte ohne Paris betrachtet, dann kann es sein, dass die Distanz zwischen London und Milan als „nicht groß“ eingeschätzt wird.

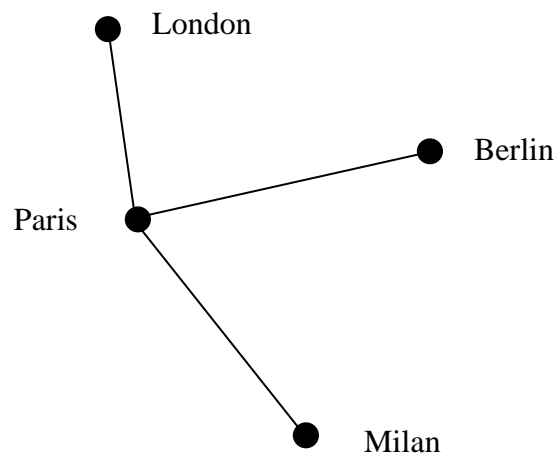


Abb. 1 Distanzrelativität Beispiel (aus [3])

**Relativität, die durch verschiedene Maßstäben verursacht wird.** Zwei Objekte, die auf einer Karte „weit“ entfernt sind, können auf einer anderen Karte als „nicht weit“ betrachtet werden. Zum Beispiel ist Paris „weit“ von Berlin auf der europäischen Karte, aber „nicht weit“ auf der Weltkarte.

**Größe des Objektes.** Sie kann die Abschätzung der Distanz beeinflussen; je kleiner die Objekte sind, desto „weiter“ scheinen sie entfernt zu sein. Zum Beispiel Distanz 1 km zwischen zwei Häusern kann man als „groß“ bewerten, während die gleiche Entfernung zwischen zwei Städten wahrscheinlich als „klein“ eingeschätzt wird.

**Andere Faktoren.** Auch andere Faktoren, wie zum Beispiel Reisekosten, können die subjektive Einschätzung der Distanz beeinflussen.

Daraus folgt, dass dieser Ansatz zu weiteren Imperfektionen in der Datenbank führen kann, wie zum Beispiel Unvereinbarkeit bei der Verwendung einer zweiten Datenbank oder bei der Weiterentwicklung einer existierenden Datenbank.

### 3.4 Rough Sets

Die Rough Sets wurden in der 80er Jahren von Zdzislaw Pawlak eingeführt.

In der Rough Sets Theorie werden statt der booleschen {True, False} drei Werte {T, M, F} verwendet. Dabei kann der Wert M unterschiedlich interpretiert werden, zum Beispiel als Grenzenfall in einer „vagen“ Interpretation, als unbekannter Wert, der True oder False sein kann, in eine „Fehler-basierte“ Interpretation oder so, dass manche Teile des Objektes als T klassifiziert werden können und manche als F („Unpresize“ Interpretation).

#### Beispiel (aus [2]):

Als Beispiel für die Verwendung von Rough Sets Analysen betrachten wir die Auswertung von Daten aus Tabelle 2.

Object	Consumer income	Luxury car showroom Site suitability
Store1	High	Good
Store2	Medium	Good
Store3	Medium	Poor
Store4	Low	Poor

Tab. 2 Beispiel Daten [aus 2]

Hier:

- Menge von Objekten

$$Z = \{\text{Store1}, \text{Store2}, \text{Store3}, \text{Store4}\}$$

- Wertmengen:

$$R_{\text{income}} = \{Y1, Y2, Y3\}, Y1 = \{\text{Store1}\}, Y2 = \{\text{Store2}, \text{Store3}\}, Y3 = \{\text{Store4}\}$$

$$R_{\text{site\_suitability}} = \{X1, X2\}, X1 = \{\text{Store1}, \text{Store2}\}, X2 = \{\text{Store3}, \text{Store4}\}$$

Als Ergebnis bekommen wir für die obere Approximation

$$S_o(Z) = \{\text{Store1}, \text{Store2}, \text{Store3}\}$$

Und für die untere Approximation

$$S_u(Z) = \{\text{Store1}\}$$

Die erhaltenen Daten kann man mit Hilfe von Rough Sets Analysen dann weiter bearbeiten.

Es ist auch möglich, die Rough Sets Analysen rekursiv zu verwenden. Ein Beispiel von solchen rekursiven Analysen wird in Abb. 2 deutlich.

Rough Sets Analysen ermöglichen die Arbeit mit Kategorien, die Arbeit mit vagen, intuitiven Begriffen und rekursive Analysen. Sie haben aber auch Nachteile, wie zum Beispiel die Probleme bei der Vereinigung von einigen Datenbanken. Im Vergleich mit dem Fuzzy Sets Ansatz ist die Rough Sets Analyse ziemlich begrenzt und ermöglicht lediglich die Abbildung von relativ einfachen Modellen von Imperfektion.

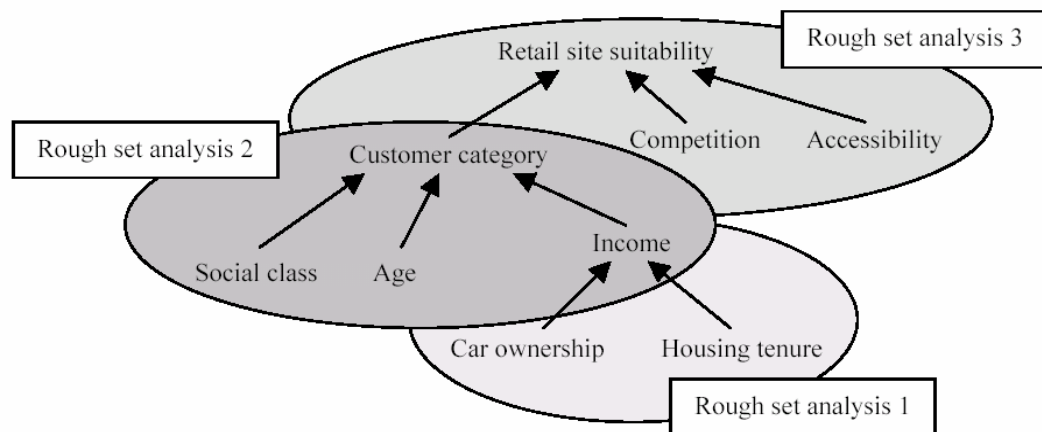


Abb. 2 Rekursive Rough Sets Analyse [aus 2]

## 4 Zusammenfassung

Die vage Information ist ein wichtiger Teil von fachlichen Kenntnissen in der Geographie.

Da die geographischen Daten oft unpräzise und unvollständig sind und die Abfragebedingungen vage sind, muss man dies auch in der Datenbank abbilden können. Der größere Teil von GIS basiert heutzutage auf der booleschen Logik, was zur künstlichen

Präzision in der Abbildung von Daten führt und Probleme in Repräsentation, Manipulation und Analyse von Information auslöst.

Es gibt verschiedene Ansätze, die bei GIS für die Abbildung von vagen Daten verwendet werden. Die traditionelle Form von Reasoning ist das quantitative Reasoning. Dieser Ansatz ist weit verbreitet und gut untersucht, moderne Datenbanksysteme stellen dafür viele Werkzeuge zur Verfügung. Dieser Ansatz eignet sich sehr gut für die Abbildung von präzisen Daten und die Bearbeitung von genauen Abfragen. Die Bearbeitung von Abfragen mit unpräzise formulierten Bedingungen lässt dieser Ansatz jedoch nicht zu.

Dieses Problem wird durch Verwendung von qualitativen statt quantitativen Daten gelöst. Dies wird durch die Einführung von linguistischen Variablen erreicht.

Linguistische Variablen ermöglichen das Ausdrücken qualitativer Begriffe. Das reicht aber nicht für die Integration von quantitativen und qualitativen Daten. Dies ermöglicht hingegen die Verwendung von Fuzzy Logik.

Mit Hilfe der Fuzzy Logik kann man mit unpräzisen Daten in den meisten GIS umgehen. Der Vorteil von diesem Ansatz ist das Ermöglichen der Abbildung von Vagheit. Nachteile sind, dass es nicht immer leicht ist, die Zugehörigkeitsfunktion zu finden. Außerdem kann dieser Ansatz zu anderen Formen von Unperfektion führen, wie zum Beispiel Unvereinbarkeit verschiedener Datenbanken oder Relativität von Berechnung der Zugehörigkeitswerte.

Der Rought Sets Ansatz ermöglicht die Arbeit mit Kategorien, rekursiven Analysen und die Arbeit mit intuitiven Begriffen. Dieser Ansatz hat aber auch ebenso Nachteile, wie die Relativität und kann auch weitere Imperfektionen in der Datenbank verursachen. Im Vergleich mit dem Ansatz der Fuzzy Logik ist die Rough Sets Analyse weniger universell und ermöglicht die Abbildung nur von einfachen Modellen von Vagheit.

## 5 Literatur

[1] Matt Duckham, Uncertainty and geographic information: computational and critical convergence, Department of Computer Science, University of Keele, 2002

[2] Matt Duckham, Keith Mason, John Stell, Mike Worboys , A Formal approach to imperfection in geographic information, Department of Computer Science, University of Keele, 2002, Computer, Environment and Urban Systems v25 pp. 89-103

[3] Hans W. Guesgen, Jochen Albrecht, Imprecise reasoning in geographic information systems, , Department of Computer Science, University of Auckland, 1998, , "Fuzzy Sets and Systems", 113 (2000) 121-131

[4] Valeie Cross, Aykut Firat, Fuzzy objects for geographical information systems, University of North Carolina at Charlotte, 1998, , "Fuzzy Sets and Systems", 113 (2000) 19-36

[5] Peter Fischer, Sorites paradox and vague geographies, Department of Geography, University of Leicester, 1998, , "Fuzzy Sets and Systems", 113 (2000) 7-18