Zuordnung von raumbezogenen Daten - am Beispiel der Datenmodelle ATKIS und GDF

Bei der Fakultät für
Bauingenieur- und Vermessungswesen
der Universität Stuttgart
zur Erlangung der Würde eines Doktor-Ingenieurs (Dr.-Ing.)
eingereichte Dissertation

vorgelegt von
Diplom-Informatiker Volker Walter
aus
Stuttgart

Hauptberichter: Prof. Dr.-Ing. habil. Dieter Fritsch
Mitberichter: Prof. Dr.-Ing Matthäus Schilcher

Institut für Photogrammetrie
Universität Stuttgart
Tag der mündlichen Prüfung: 11. 12. 1996

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einführung

Mit der wachsenden Leistungsfähigkeit von Geo-Informationssystemen steigt auch gleichzeitig der Bedarf an raumbezogenen digitalen Daten. Um diese Nachfrage zu befriedigen, werden von staatlichen und privaten Institutionen Daten in großem Umfang erfaßt. Die Erfassung und Fortführung der Daten ist zweckgerichtet und erfolgt abhängig von der Anwendung in unterschiedlichen Maßstäben und Datenmodellen. Das Ergebnis ist eine multiple Repräsentation derselben topographischen Objekte der Landschaft. *"Die Erfassung in unterschiedlichen Modellen dient dazu, komplizierte Wirklichkeit anschaulich zu machen, indem sie den Blick auf das Wesentliche lenken und unwesentliches außer acht lassen. Was jeweils wichtig und unwichtig an einem Gegenstand oder Sachverhalt ist, bestimmt das Ziel, das jemand mit seinem Modell verfolgt"* [Brüggemann 1990].

1.1 Integration von raumbezogenen Daten

Durch Mehrfacherfassung von raumbezogenen Daten ergeben sich Probleme sowie neue Anforderungen durch die Nutzer der Daten. Die Erfassung von raumbezogenen Daten ist kosten- und zeitintensiv. Daher ist es volkswirtschaftlich nicht sinnvoll, raumbezogene Daten mehrfach zu erfassen und fortzuführen, insbesondere da die Fortführungsproblematik noch nicht vollständig gelöst ist (siehe z.B. [Sellge 1992]). Je größer die Projekte sind, desto problematischer wird diese Situation. Die Datenmengen, die benötigt werden, um nationale und internationale Projekte durchzuführen, sind so umfangreich, daß sie nur mit einem hohen Investitionsaufwand erfaßt werden können. Es stellt sich daher die Forderung, daß Daten unterschiedlicher Herkunft und Genauigkeit integriert werden können, um dadurch das Anwendungspotential zu vergrößern, die Wiederverwendbarkeit der Daten zu steigern und den Erfassungs- und Fortführungsaufwand zu minimieren. Durch eine Integration können Fachdaten kombiniert, Objektklassen und Attribute ergänzt und die Genauigkeit der Geometrie verbessert werden.

Das Hauptproblem bei einer Integration besteht darin, daß Objekte der Landschaft in verschiedenen Datenmodellen unterschiedlich erfaßt werden. Selbst bei der mehrfachen Erfassung von Daten in ein und demselben Datenmodell entstehen durch unterschiedliche Diskretisierung der Koordinaten und unterschiedliche Interpretation der Objekte keine deckungsgleichen Datensätze. Bei raumbezogenen Daten aus unterschiedlichen Anwendungen bzw. Datenmodellen wird eine Integration zusätzlich durch die verschiedenen Weltsichten der Anwender sowie durch unterschiedliche Datenqualitätsmerkmale erschwert. Ein erster Schritt zur Integration von raumbezogenen Daten ist eine geometrische Integration, was bedeutet, daß geometrische Elemente, die einander ähnlich sind, "verschmolzen" werden. Ohne eine geometrische Integration können GIS-Analysen und Verschneidungsberechnungen gar nicht oder nur sehr umständlich und mit hohen Ungenauigkeiten durchgeführt werden. Erste Ansätze zur geometrischen Integration sind heute schon in Geo-Informationssystemen realisiert. Hierbei handelt es sich jedoch vor allem um Algorithmen, welche Inkonsistenzen in der Topologie, die beispielsweise nach Verschneidungsoperationen von Datensätzen aus unterschiedlicher Herkunft auftreten können, beseitigen. *"Inzwischen rücken jedoch zunehmend Ansätze in den Blickpunkt, die neben eventuellen Diskrepanzen in der Geometrie auch die semantischen Unterschiede aufgrund von abweichenden Modellverständnis in Betracht ziehen"* [Illert 1995]. Dies ist besonders dann von Bedeutung, wenn Daten für interdisziplinäre Anwendungen benötigt werden. Dieses Problem macht [Kophstahl 1994] deutlich: *"Interdisziplinäre Untersuchungen wie z.B. Umweltverträglichkeitsprüfungen, die auf die Datenlieferung vieler Dienststellen angewiesen sind, sind häufig uneffektiv und wenig aussagekräftig, weil die gelieferten Informationen nicht immer integriert und zu einer fachübergreifenden Analyse und Wertung zusammengeführt werden können."*

Um die Integration von raumbezogenen Daten zu fördern, werden Standards im Bereich der Geo-Daten etabliert und Basisinformationssysteme aufgebaut. Verschiedene Anwendungen, die auf den gleichen Basisdaten aufbauen, sind untereinander bereits geometrisch "kompatibel" [Kophstahl 1994] und unterscheiden sich nur in den zusätzlich erfaßten Fachinformationen. In Deutschland wurde daher 1986 von der Arbeitsgemeinschaft der Vermessungsverwaltungen der Bundesrepublik Deutschland (AdV) der Aufbau des Amtlichen Topographischen Kartographischen Informationssystem (ATKIS) beschlossen. Auch in anderen europäischen Staaten werden

Basisinformationssysteme für Geo-Daten aufgebaut. Neben diesen nationalen Standards werden international z.B. im Standard Geographic Data File (GDF) raumbezogene Daten nahezu flächendeckend in Westeuropa erfaßt. Das GDF-Datenmodell wurde speziell für Anwendungen der Fahrzeugnavigation entworfen und umfaßt daher vor allem Objekte aus dem Bereich des Straßenverkehrs. Wegen der hohen Bedeutung des Straßenverkehrs in Europa werden dieselben Objekte jedoch auch in den nationalen Basisinformationssystemen erfaßt. Dies bedeutet eine Mehrfacherfassung gleicher topographischer Objekte im großem Ausmaß. Bei einer näheren Betrachtung von ATKIS und GDF erkennt man, daß in beiden Datenmodellen eine Vielzahl von identischen Objekten erfaßt werden und sich die Modellierung sehr ähnelt. Es stellt sich daher die Frage, inwieweit diese beiden Modelle miteinander integriert werden können.

1.2 Integration als Zuordnungsproblem

In dieser Arbeit wird die Integration von raumbezogenen Daten auf ein Zuordnungsproblem abgebildet. Eine Integration bedeutet in einem ersten Schritt, daß einander entsprechende Primitive in den Datensätzen zugeordnet werden müssen. Der Begriff *Primitiv* kann hierbei für ein geometrisches Element stehen, aber auch ebenso für Objekte oder Objektstrukturen. Hierzu wird ein Primitiv in einem der Datensätze identifiziert, um anschließend das entsprechende Primitiv im anderen Datensatz zu suchen und zuzuordnen. Dies bedeutet, daß aus der Menge der Primitive des zweiten Datensatzes dasjenige zugeordnet werden soll, welches dem Ausgangselement am ähnlichsten ist, sofern solch ein Primitiv überhaupt existiert. Um dies durchführen zu können, muß die Ähnlichkeit der Primitive mit Hilfe von Ähnlichkeitsmaßen berechnet werden. Die Berechnung dieser Maße erfolgt als Funktion der Attribute der Elemente (merkmalsbasierte Zuordnung) und/oder der Relationen zwischen den Elementen untereinander (relationale Zuordnung). Attribute von raumbezogenen Daten können geometrischer Art sein, wie z.B. *Länge, Form,* oder *Lage* oder auch semantischer Art, wie z.B. *Objektart* oder *Objektname*. Relationen zwischen raumbezogenen Daten sind z.B. *verbunden mit* oder *größer als*. Nach einer Zuordnung zwischen den Primitiven zweier Datensätze kann die eigentliche Integration durchgeführt werden. Dies kann eine Kombination von Fachdaten, eine Ergänzung von Objektklassen und Attributen oder eine Verbesserung bzw. Homogenisierung der Geometrie sein.

1.3 Zielsetzung der Arbeit

Das Ziel der vorliegenden Arbeit besteht in der Entwicklung eines Zuordnungsverfahrens, welches die Integration von raumbezogenen Daten aus unterschiedlichen Datenmodellen ermöglicht. Der hier gewählte Ansatz basiert auf dem Prinzip der relationalen Zuordnung. Bei der relationalen Zuordnung werden die Daten durch eine relationale Beschreibung dargestellt und ihre Ähnlichkeit mit einem Abstandsmaß gemessen. Als Abstandsmaß wird ein Maß aus der Informationstheorie verwendet, welches in [Vosselman 1992] vorgestellt wurde. Es handelt sich um einen theoretisch fundierten Ansatz, welcher auf statistischen Auswertungen zwischen den zuzuordnenden Datensätzen basiert. Ein großer Vorteil dieses Ansatzes besteht darin, daß er unabhängig von den Datenmodellen definiert ist und keine Startwerte, Schwellwerte oder Gewichtsfaktoren (Tuning-Faktoren) benötigt. Mit Hilfe des Abstandsmaßes kann nach der besten Zuordnung zwischen zwei Datensätzen gesucht werden. Hierzu muß der Suchraum des Zuordnungsproblem dargestellt und ausgewertet werden.

Das von Vosselman vorgeschlagene Verfahren zur Auswertung des Suchraumes ist auf $1:1$ Zuordnungen beschränkt. Für die Zuordnung raumbezogener Daten gilt es nun, das Verfahren für beliebige $n:m$ Zuordnungen zu erweitern. Dies kann erreicht werden, indem verschiedene Elemente eines Datensatzes zu einem logischen Element zusammengefaßt werden. Hierzu wird ein Verfahren vorgestellt, welches mit Hilfe von Puffern $n:m$ Zuordnungen zwischen zwei Datensätzen aufstellt.

Als besonderes Problem bei der Durchführung der Zuordnung stellt sich die Größe des Suchraumes dar. Da dieser exponentiell zu der Anzahl der zuzuordnenden Elemente anwächst, werden a-priori Abschätzungen benötigt, die es ermöglichen, nur Teile des Suchraumes auszuwerten, und dennoch gleichzeitig eine optimale Lösung gewährleisten. Es wird gezeigt, wie diese Abschätzungen speziell für die Zuordnung von raumbezogenen Daten berechnet werden können. In der Arbeit von Vosselman wird ein relationales Zuordnungsverfahren zur Lokalisierung dreidimensionaler Objekte aus Rasterbildern mit Hilfe ihrer Modellbeschreibung präsentiert. Obwohl die Objekte nur aus ca. 50 geometrischen Elementen bestehen, entstehen Rechenzeiten, die z.T. bis zu mehrere Stunden betragen. Bei der Zuordnung von raumbezogenen Daten müssen jedoch typischerweise nicht nur einige wenige Elemente zugeordnet werden, sondern sehr große, zusammenhängende Vektordatensätze. Ein GDF-Datensatz der Stadt Stuttgart besteht beispielsweise u.a. aus über 20.000 linienförmigen Objekten. Die Situation heute,

gegenüber der Situation während der Arbeit von Vosselman, hat sich dadurch verbessert, daß inzwischen wesentlich leistungsstärkere Rechner zur Verfügung stehen. Diese Leistungssteigerung reicht jedoch nicht aus, um den exponentiellen Anstieg des Suchraumes zu bewältigen.

Bei der Untersuchung der Zuordnung von raumbezogenen Daten wird jedoch ersichtlich, daß eine starke Lokalität der Zuordnungen vorliegt. Es ist nicht notwendig einen Suchraum aufzuspannen, der alle möglichen Kombinationen von Zuordnungen umfaßt, sondern es muß lediglich eine Teilmenge betrachtet werden. Zum einen können schon bei dem Aufstellen von potentiellen Zuordnungspaaren geometrische und semantische Beschränkungen genutzt werden, die nicht erlaubte Zuordnungspaare eliminieren, und zum anderen kann auch die Auswertung des Suchraumes zum Teil lokal erfolgen und damit das exponentielle Wachstum eindämmen. Es wird eine heuristische Methode vorgestellt, die den Suchraum in voneinander unabhängig optimierbare Cluster aufteilt.

Aufgrund der unterschiedlichen semantischen Modelle von verschiedenen Datensätzen ist es nicht möglich, ein Abstandsmaß zu finden, welches die Elemente absolut fehlerfrei einander zuordnet. Dies bedeutet, daß das Ergebnis des Zuordnungsprozesses nicht gewollte Zuordnungen enthält, welche in einem nachfolgenden Schritt identifiziert und eventuell interaktiv bearbeitet werden müssen. Daher wird abschließend diskutiert, wie solche Fehlzuordnungen automatisch gefunden werden können.

Das vorgestellte Verfahren wird anhand von Datensätzen der Datenmodelle von ATKIS und GDF getestet und bewertet. Hierzu ist es notwendig, statistische Untersuchungen der Zuordnungen zwischen diesen Datensätzen durchzuführen. Um eine Bewertung der Ergebnisse durchführen zu können, werden Referenzzuordnungen benötigt. Hierzu wurde ein Werkzeug zur manuellen Zuordnung von ATKIS- und GDF-Daten erstellt.

1.4 Aufbau der Arbeit

In Kapitel 2 werden die verschiedenen Aspekte der Integration von raumbezogenen Daten diskutiert. Am Beispiel von Straßenverkehrsdaten wird die Mehrfacherfassung von raumbezogenen Daten in Deutschland dargestellt. Anschließend erfolgt eine Klassifikation verschiedener Stufen der Integration von raumbezogenen Daten. Diese Klassifikation ist zum einen von den betrachteten Datensätzen abhängig und zum anderen von der Vorgehensweise der Integration. Abschließend werden die Vorteile einer automatischen Zuordnung von GDF und ATKIS dargestellt.

Ein umfassender Überblick über den Standard GDF wird in Kapitel 3 gegeben. Das GDF-Datenmodell erlaubt eine einfache Modellierung von komplexen Zusammenhängen. Die hierfür entwickelten Konzepte werden an Beispielen aufgezeigt. Anschließend erfolgt eine kritische Diskussion der Qualitätsaussagen in der GDF-Gesamtdokumentation sowie eine Darstellung des GDF-Datenaustauschformates. In Deutschland werden GDF-Daten derzeit von den beiden Firmen Bosch und EGT erfaßt. Obwohl die Daten im gleichen Datenmodell vorliegen, entstehen hierbei unterschiedliche Datensätze. Diese Unterschiede werden an Beispielen diskutiert. Abschließend erfolgt eine Darstellung der Abbildung des GDF-Datenmodells auf das GIS-Produkt SICAD/open.

In Kapitel 4 wird die Konzeption von ATKIS diskutiert. Die Erfassung der Objekte der Landschaft erfolgt mit Hilfe eines digitalen Landschaftsmodells (DLM). Die Daten des DLM sollen den GDF-Daten zugeordnet werden. Daher wird das DLM eingehend untersucht. Neben dem DLM existieren in ATKIS noch das digitale kartographische Modell (DKM) und das digitale Geländemodell (DGM). Das DGM wird zukünftig integraler Bestandteil des DLM sein, wird aber derzeit noch eigenständig vorgehalten. Der Inhalt und die Modellierung dieser Modelle wird kurz dargestellt. Der Datenaustausch von ATKIS-Daten wird mit Hilfe der Einheitlichen Datenbankschnittstelle (EDBS) durchgeführt. Da im Zusammenhang mit ATKIS immer die EDBS betrachtet werden muß, werden die Vor- und Nachteile des Datenaustausches mit Hilfe der EDBS diskutiert.

Die Kapitel 3 und 4 bilden die Grundlage für einen Vergleich des GDF- und ATKIS-Datenmodells. Im Kapitel 5 werden die Unterschiede zwischen den beiden Datenmodellen erarbeitet. Für den Zuordnungsalgorithmus ist vor allem die unterschiedliche Erfassung und Modellierung der Objekte der Landschaft von großer Bedeutung. Anhand von Beispielen werden diese Unterschiede aufgezeigt. Weiter erfolgt eine Gegenüberstellung der Austauschformate und Erfassungsvorlagen.

Die Vorgehensweise der Zuordnung von raumbezogenen Daten wird in Kapitel 6 diskutiert. Hierbei handelt es sich um eine Darstellung der theoretischen Zusammenhänge, ohne dabei anwendungsspezifische Fragestellungen zu betrachten. Nach einer Vorstellung bestehender Arbeiten wird ein Überblick über die Vorgehensweise von Zuordnungsalgorithmen gegeben. Anschließend werden in einzelnen Unterkapiteln die verschiedenen Aspekte der Zuordnung von raumbezogenen Daten aufgezeigt. Da das Verfahren auf informationstheoretischen Maßen aufbaut, wird im Anhang ein Exkurs in die Informationstheorie gegeben. Abschließend wird aufgezeigt, daß die

in der Arbeit verwendeten Maße für die lokale und globale Darstellung der Qualität der Zuordnungen verwendet werden können.

In Kapitel 7 wird die eigentliche Zuordnung von GDF- und ATKIS-Daten beschrieben. Um das in Kapitel 6 beschriebene Verfahren durchführen zu können, müsssen zuerst statistische Auswertungen zwischen den Daten der beiden Datenmodelle durchgeführt werden. Die Auswahl der betrachteten Merkmale und die Ergebnisse der statistischen Auswertungen werden aufgezeigt. Danach erfolgt eine Darstellung der Vorgehensweise bei der automatischen Zuordnung. Anschließend werden die Ergebnisse diskutiert. Hierbei sollen die Vor- und Nachteile des Verfahrens aufgezeigt werden. Ob ein Verfahren in der Praxis einsetzbar ist, hängt unter anderem von seinem Zeitkomplexitätsverhalten ab. Daher wird das Zeitverhalten der einzelnen Teilschritte untersucht.

Eine Zusammenfassung der Ergebnisse wird in Kapitel 8 dargestellt. Abschließend wird diskutiert, welche neue Fragestellungen sich durch diese Arbeit ergeben haben und damit Schwerpunkt zukünftiger Forschung sein sollten.

Kapitel 2

Integration raumbezogenener Daten

In diesem Kapitel werden verschiedene Aspekte der Integration von raumbezogenen Daten diskutiert. Um in die Problematik einzuführen, wird am Beispiel von Straßenverkehrsdaten die Mehrfacherfassung von Objekten in unterschiedlichen Datenmodellen dargestellt. Aufgrund des dichten Straßennetzes und dem hohen Stellenwert des Individualverkehrs in unserer Gesellschaft spielen Daten des Straßenverkehrs eine wesentliche Rolle und werden von mehreren Institutionen in unterschiedlichen Datenmodellen erfaßt. Es wird daher eine Übersicht über Datenmodelle gegeben, in denen Straßenverkehrsdaten vorliegen. Dies ermöglicht eine Klassifikation in verschiedene Stufen der Integration von raumbezogenen Daten. Danach werden weitere Unterscheidungen zwischen geometrischer und semantischer Integration sowie zwischen Top-Down- und Bottom-Up-Vorgehensweisen durchgeführt. Abschließend werden mögliche Vorteile einer Integration der Datenmodelle GDF und ATKIS untersucht.

2.1 Straßenverkehrsdaten

Die in dieser Arbeit entwickelten Methoden wurden am Beispiel von Straßendaten aus den Datenmodellen GDF und ATKIS eingehend untersucht und verifiziert. Straßenverkehrsdaten spielen eine wichtige Rolle in unserer Gesellschaft. Neben Anwendungen, die Straßendaten vor allem wegen ihrer guten räumlichen Referenz nutzen, erhalten Anwendungen aus dem Bereich der Fahrzeugnavigation zunehmend eine wichtige Bedeutung. *„Die Entwicklung der Autonavigation mit all ihren Nebenaspekten wird für die Zukunft so bedeutend sein, daß gerade dieses Beispiel als ein hochinteressantes Anwendungsgebiet von Fachdaten in Verbindung mit Geometrie (Automated Mapping and Facilities Management - AM/FM) gesehen werden. Es ist ein wichtiger Baustein für die sich in der Entwicklung befindlichen Informationsgesellschaft“* [Barwinski 1988]. Die hohe Bedeutung von Straßenverkehrsdaten erklärt auch, warum in diesem Bereich Mehrfacherfassungen in großem Ausmaß durchgeführt werden.

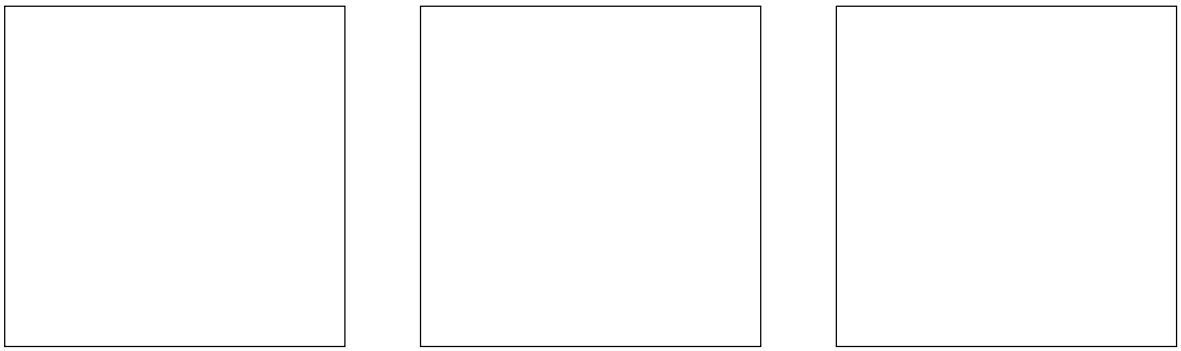
Im folgenden werden verschiedene Datenmodelle vorgestellt, in denen Straßenverkehrsdaten in Deutschland erfaßt werden. Bei einer detaillierten Betrachtung der Modelle erkennt man sogleich, daß die Objekte der Landschaft in verschiedenen Datenmodellen unterschiedlich erfaßt werden. So kann ein und dasselbe Landschaftsobjekt in einem Datenmodell durch ein Objekt dargestellt werden und in einem anderen Datenmodell durch mehrere Objekte. Die Gründe hierfür liegen vor allem in den unterschiedlichen Anwendungsgebieten, für die die Datenmodelle konzipiert wurden. Im folgenden werden die Unterschiede anhand von typischen Beispielen dargestellt. Eine ausführliche Diskussion der Datenmodelle GDF und ATKIS wird in den nächsten beiden Kapiteln präsentiert.

2.1.1 Geographic Data File

Das Geographic Data File (GDF) ist der Standard für den Austausch von digitalen Straßenverkehrsdaten in Europa [Heres, Berthet, Claussen und Hiestermann 1991]. GDF-Daten werden in nahezu ganz Westeuropa erfaßt und liegen bereits flächendeckend für Deutschland vor. Die Erfassung wird gleichzeitig von den zwei konkurrierenden Konsortien EDRA (European Digital Road Association; in Deutschland vertreten durch die Firma Bosch/Teleatlas) und EGT (European Geographic Technologies) unabhängig voneinander durchgeführt. Durch diese Doppelerfassung entstehen ähnliche, im Detail jedoch unterschiedliche Datensätze. Die Gründe hierzu sind verschiedene Erfassungsvorgaben und unterschiedliche Interpretation der GDF-Erfassungsvorschriften durch verschiedene Operateure. Abbildung 2.1 a) zeigt diese Situation an einem Beispiel. Insbesondere die verschiedene Interpretation des Kreuzungsbereiches führt zu Unterschieden in den Daten. Jedoch ist ebenso auszumachen, daß selbst komplexe Kreuzungsbereiche sehr ähnlich erfaßt werden.

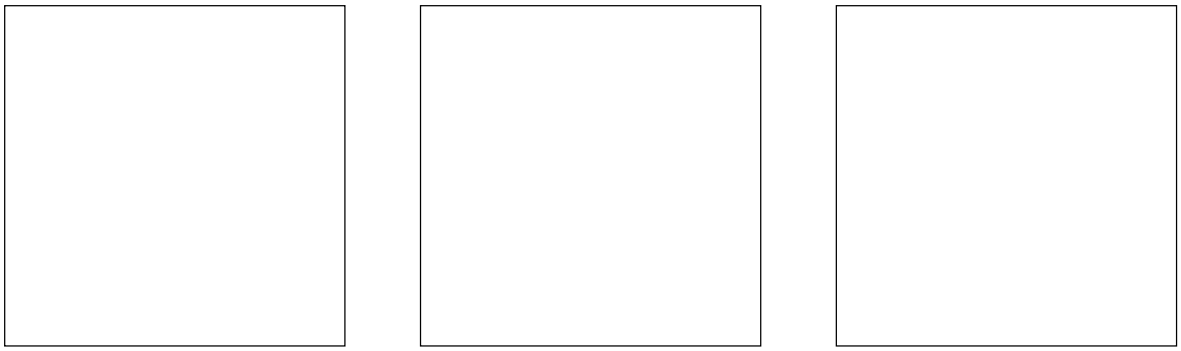
a)

GDF (Bosch) GDF (EGT) GDF (Bosch) + GDF (EGT)



b)

GDF (Bosch) ATKIS GDF (Bosch) + ATKIS



c)

ALK ATKIS ALK + ATKIS

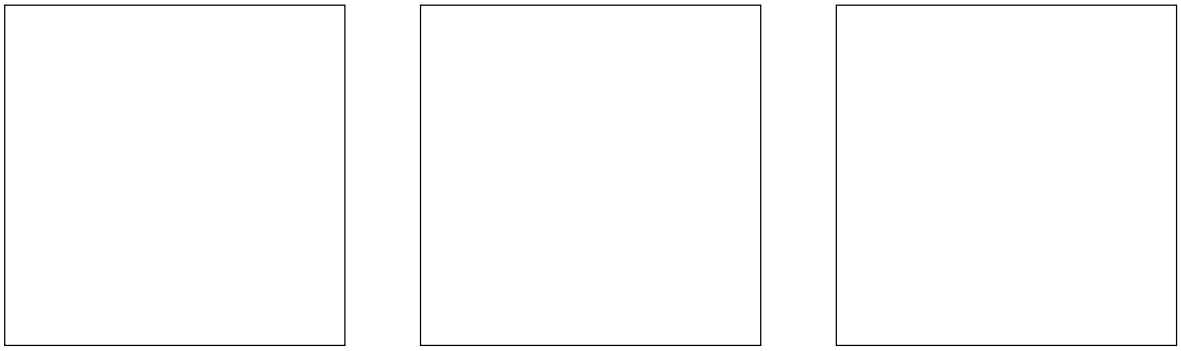


Abbildung 2.1: Unterschiedliche Erfassung des Straßennetzwerkes

2.1.2 Amtliches Topographisch-Kartographisches Informationssystem

ATKIS stellt das topographische Basisinformationssystem für raumbezogene Daten in Deutschland dar [AdV 1988]. Neben dem Maßstab 1:25.000 werden Daten in den Maßstäben 1:200.000 und 1:1.000.000 erfaßt. GDF und ATKIS korrespondieren am stärksten im Maßstab 1:25.000. Da ATKIS und GDF für unterschiedliche Anwendungszwecke entwickelt wurden, bilden die gemeinsam zu erfassenden Objektklassen und deren Attribute nur eine Schnittmenge. Während das Ziel von ATKIS darin besteht, Anwender mit einem Grundstock von raumbezogenen Daten zu versorgen, wurde GDF speziell für Anwendungen der Fahrzeugnavigation und -zielführung entwickelt. Jedoch bilden die Objekte des Straßen- und Schienenverkehrs sowie der Wasserwege eine Schnittmenge, da ATKIS und GDF in diesen Bereichen die gleichen Objekte der Landschaft erfassen. Einer der Hauptunterschiede zwischen der Erfassung von Straßen in GDF und ATKIS liegt darin, daß in GDF insbesondere Kreuzungsbereiche oftmals detailierter erfaßt werden als in ATKIS. Abbildung 2.1 b) zeigt die unterschiedliche Erfassung in GDF und ATKIS. Bei der Überlagerung von ATKIS und GDF finden sich typischerweise Gebiete in denen die Daten nahezu deckungsgleich sind, aber auch Gebiete in denen lokal stark unterschiedlich erfaßt wurde.

2.1.3 Automatisierte Liegenschaftskarte

Das Liegenschaftskataster bildet den gesetzlichen Nachweis der Liegenschaftskarten in Deutschland. Hierzu wird die Landschaft in den Maßstäben 1:500 bis 1:2.000 digital erfaßt und als „Automatisierte Liegenschaftskarte“ gespeichert [ALK 1986]. Die Fertigstellung der ALK wird zwar noch viele Jahre in Anspruch nehmen, jedoch sind bereits jetzt schon umfangreiche Datenbestände verfügbar. Aufgrund der großmaßstäblichen Erfassung der ALK-Daten besitzen diese eine sehr hohe Detailgenauigkeit. Abbildung 2.1 c) zeigt die unterschiedliche Erfassung von ALK- und ATKIS-Daten.

2.2 Stufen der Integration

Wie im vorhergehenden Abschnitt dargestellt wurde, erfolgt die Erfassung von Straßen in verschiedenen Datenmodellen in zum Teil sehr unterschiedlicher Art und Weise. Der Erfolg einer Integration zweier Datensätze hängt stark davon ab, wie “ähnlich“ die Datenmodelle einander sind. Im folgenden wird eine Klassifikation in verschiedene Stufen der Integration von raumbezogenen Daten durchgeführt [Walter und Fritsch 1996].

Stufe 1: Es sollen zwei Datensätze A' und A'' integriert werden, welche beide von ein und demselben ursprünglichen Datensatz A abstammen, aber beide unabhängig voneinander fortgeführt wurden. Die Daten werden beispielsweise von einem Datenerfasser geliefert und dann in bestimmten Zeitabständen fortgeführt. Während dieser Zeit beginnt der Benutzer seine Daten selbst fortzuführen oder digitalisiert zu den vorhandenen Daten weitere Fachdaten hinzu. Durch diese Änderungen an den Daten ist es nicht möglich, das nächste Update des Datenlieferanten zu übernehmen, da sonst die selbst digitalisierten Fortführungen (welche nicht unbedingt im neuen Update bereits enthalten sind) und die mit den Daten verknüpften Fachdaten verloren gingen. Abbildung 2.2 zeigt das Problem an einem Beispiel. Es kann nicht eindeutig festgestellt werden, wie der Datensatz A fortgeführt werden soll. Auf dieser Stufe ist jedoch die Integration einfach, da die Daten in demselben Datenmodell vorliegen und zum größten Teil sogar identisch sind. Daher reichen für diesen Zweck Programme aus, die identische Objektstrukturen in den beiden Datensätzen suchen. Dadurch können die Änderungen identifiziert und nachgeführt werden. Noch einfacher ist in diesem Fall, eine solche Situation ganz zu vermeiden und eindeutige ID's und Protokollmechanismen für die Fortführung zu nutzen.

Stufe 2: Zwei Datensätze sollen integriert werden, welche zwar im gleichen Datenmodell vorliegen, jedoch von unterschiedlichen Operateuren erfaßt wurden. Dies ist z.B. die Situation beim Vergleich von GDF-Daten der Firmen Bosch/Teleatlas und EGT. Die Unterschiede in der Erfassung ergeben sich hier zum einen durch eine unterschiedliche Diskretisierung der Koordinaten und zum anderen durch unterschiedliche Interpretation desselben Objekts durch verschiedene Operateure. Die maximal erreichbare Erfassungsgenauigkeit hängt von der *Zielgenauigkeit* der zu erfassenden Karte sowie des *persönlichen Einstellfehlers* und der *Reproduzierbarkeit des Digitizers* ab. Bei einer Erfassung von topographischen Karten gibt [Rappe 1995] als Formel zur Berechnung der maximalen Erfassungsgenauigkeit ds an:

$$ds = 2 * 10^{-4} * m = \frac{m}{5000} \tag{2.1}$$

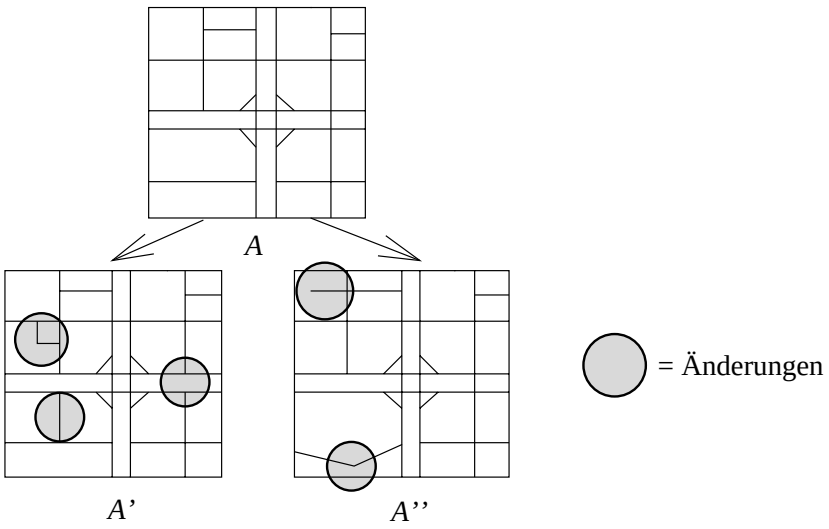


Abbildung 2.2: Zwei unterschiedliche Datensätze abstammend von demselben Datensatz

wobei m die Maßstabszahl des Kartengrunds ist. Dies bedeutet, daß bei der Erfassung der Daten von einer TK 10 mit einer maximalen Erfassungsgenauigkeit von 2 m zu rechnen ist. Die unterschiedliche Interpretation von Objekten ist zum einen von den verschiedenen Operateuren abhängig, kann zum anderen auch aufgrund verschiedener Digitalisiervorschriften erfolgen. So ist beispielsweise die Erfassungstiefe von EGT-Daten höher als die von Bosch-Daten, was bedeutet, daß kleinere Wege in den EGT-Daten noch erfaßt werden, in den Bosch-Daten jedoch nicht vorhanden sind. Auf der anderen Seite kann bei den Bosch-Daten festgestellt werden, daß Kreuzungsbereiche von komplexen Kreuzungen mit höherer Detailliertheit erfaßt sind als in den EGT-Daten. Ein weiterer, nicht vernachlässigbarer Grund für unterschiedliche Datensätze sind individuelle Digitalisierfehler. Dies können beispielsweise fehlende Objekte, fehlerhafte Attributwerte oder auch falsch erfaßte Topologie sein. Obwohl durch alle diese Fehlerquellen z.T. große Unterschiede in den verschiedenen Datensätzen auftreten können, ist eine Integration in diesem Fall immer noch einfach durchführbar. Da die Daten in demselben Datenmodell vorliegen, kann durch einen Vergleich der Objekte und Attribute bereits ein Großteil der Daten integriert werden.

Stufe 3: Die Datensätze liegen nicht im gleichen Datenmodell vor, jedoch sind die Datenmodelle einander ähnlich. Dies ist z.B. der Fall bei dem Vergleich von GDF- und ATKIS-Daten. Es ist damit zu rechnen, daß Integrationsprobleme der Stufe 3 in der Praxis häufiger vorkommen als Integrationsprobleme der Stufe 2, da das mehrfache Erfassen von Daten in ein und demselben Datenmodell typischerweise nur dann vorkommt, wenn zwei Datenerfasser als Konkurrenzunternehmen auftreten. In Stufe 3 treffen alle in Stufe 2 genannten Ursachen für Inkonsistenzen zwischen den Datensätzen ebenfalls zu. Durch die unterschiedlichen zugrunde liegenden Datenmodelle können jedoch weitere Differenzen auftreten. Dies können beispielsweise sehr unterschiedliche Objektstrukturen sein, die das Auffinden gleichartiger Objekte in den verschiedenen Datensätzen nicht erlauben. Beispiele hierzu sind andersartige Attributierung oder Objektstrukturierung gegenüber Layerstruktierung. Um in diesem Fall eine Integration durchführen zu können, müssen mehr Informationen in den Integrationsprozess einfließen. Da beide Datensätze die gleichen Objekte der Landschaft beschreiben, können, mit Hilfe des Raumbezugs der Objekte, identische Strukturen lokalisiert werden. Dies bedeutet, daß falls über die Objektstrukturen alleine keine Integration durchgeführt werden kann, die geometrische Repräsentation der Objekte genutzt werden muß.

Stufe 4: Die Datensätze stammen aus verschiedenen Datenmodellen, welche sich stark unterscheiden. Diese Situation liegt z.B. bei ALK- und GDF-Daten vor. Während in GDF Straßen nahezu ausschließlich durch linienhafte Objekte erfaßt werden, liegt in der ALK eine flächenhafte Erfassung vor. In diesem Fall ist eine direkte Zuordnung nicht möglich, sondern es muß zuerst eine Vorverarbeitung bei einem der Datensätze durchgeführt werden. Dies könnte z.B. eine Extraktion der Mittelachsen der Fahrbahnen in den ALK-Daten sein. Nach einer geeigneteten Vorverarbeitung erhält man im wesentlichen ein Integrationsproblem der Stufe 3. Da die ALK-Daten mit einer wesentlich höheren Genauigkeit erfaßt werden als die GDF-Daten, muß jedoch damit gerechnet werden, daß viele Objekte der ALK-Daten nicht in den GDF-Daten erfaßt wurden.

2.3 Geometrische Integration

Bei einer Integration von raumbezogenen Daten kann zwischen einer rein geometrischen Integration, bei der nur die Geometrie der Daten betrachtet wird, und einer semantischen Integration, welche die Objektsicht berücksichtigt, unterschieden werden. Die Möglichkeit Daten geometrisch zu integrieren, ist heute schon teilweise in Geo-Informationssystemen realisiert. Hierbei handelt es sich u.a. um Algorithmen, welche mit Fangkreisen versuchen, einander entsprechende Punkte miteinander zu verschmelzen. In der Abbildung 2.3 ist eine typische Anwendung für diesen Ansatz dargestellt. Es existieren zwei Layer *Straße* und *Landnutzung*, die unabhängig voneinander erfaßt wurden. Bei einer Überlagerung der Layer kann gesehen werden, daß die Linienzüge nicht deckungsgleich diskretisiert wurden. Dies kann bei Verschneidungsoperationen zu vielen kleinen nicht sinnvollen Polygonen (Sliver-Polygone) führen. Um dies zu vermeiden, werden mit Fangkreisen naheliegende Punkte miteinander verschmolzen (Homogenisierung). Bei diesem Verschmelzungsprozess können die Punkte je nach Layer gewichtet werden, um die Punkte, bei denen bekannt ist, daß sie mit einer höheren Genauigkeit erfaßt wurden, stärker einzubringen. Nach diesem Verschmelzungsprozess sind die alten Koordinaten jedoch nicht mehr zugänglich, es sei denn, sie wurden vorher explizit abgespeichert. Diese Vorgehensweise ist geeignet um Daten aus verschiedenen Quellen so aufzubereiten, daß raumbezogene Analysen zu sinnvollen Ergebnissen führen. Für eine echte Integration zweier Datensätze reicht eine Homogenisierung der Geometrie nicht aus, sondern es müssen einander entsprechende geometrische Elemente zugeordnet und diese Zuordnungen für eine Weiterverarbeitung verfügbar gemacht werden.

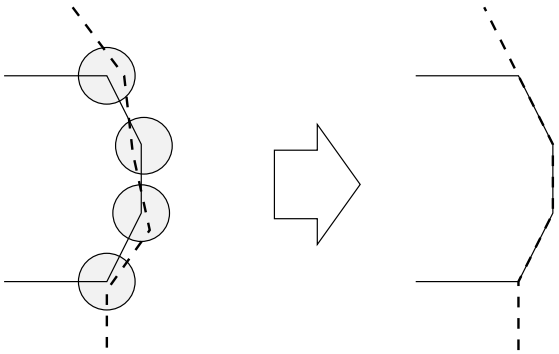


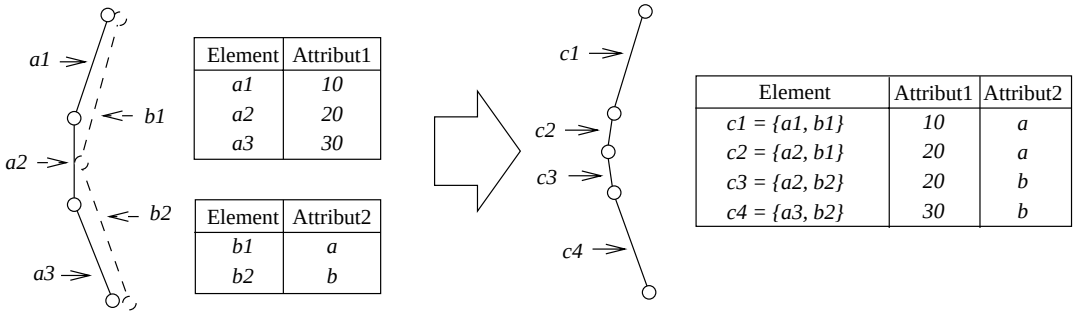
Abbildung 2.3: Homogenisierung der Geometrie

2.4 Semantische Integration

Eine semantische Integration geht einen Schritt weiter. Es werden neben der geometrischen Repräsentation auch die Sachdaten und Objektstrukturen integriert. Weiter werden die Sachdaten und Objektstrukturen selbst, als betrachtete Merkmale, zur Durchführung der Integration verwendet. Abbildung 2.4 zeigt dies an einem Beispiel. Es sind zwei Straßenobjekte dargestellt, welche in zwei unterschiedlichen Modellen erfaßt wurden. Neben einer unterschiedlichen Diskretisierung der Geometrie liegt auch eine unterschiedliche Attributierung vor. Es erfolgt eine Homogenisierung der Geometrie sowie eine Integration der Sachdaten, die im Beispiel aus zwei unterschiedlichen Attributtypen bestehen. Aufgrund der unterschiedlichen Attributierung erhöht sich die Anzahl der Liniensegmente auf Vier. Daraus wird ersichtlich, daß die geometrische Repräsentation der Objekte z.B. aufgrund unterschiedlicher Attributierung auch von der semantischen Bedeutung abhängt.

2.5 Top-Down- vs. Bottom-Up-Ansatz

Während in den vorhergehenden Abschnitten diskutiert wurde, ob eine Integration der Geometrieelemente erfolgt oder ob auch die Sachdaten integriert werden, wird in diesem Abschnitt die Richtung der Integration betrachtet. Es kann zwischen einem Bottom-Up- und einem Top-Down-Ansatz unterschieden werden. Beim Bottom-Up-Ansatz wird zuerst versucht, entsprechende geometrische Elemente zu finden (Geometrische Integration), um anschließend die zu diesen Geometrieelementen gehörenden Objektstrukturen zuzuordnen (Semantische Integration). Beim Top-Down-Ansatz werden zuerst identische Objektstrukturen gesucht, mit



Element	Attribut1
a1	10
a2	20
a3	30

Element	Attribut2
b1	a
b2	b

Element	Attribut1	Attribut2
c1 = {a1, b1}	10	a
c2 = {a2, b1}	20	a
c3 = {a2, b2}	20	b
c4 = {a3, b2}	30	b

Abbildung 2.4: Semantische Integration

einer anschließenden Möglichkeit der Homogenisierung der Geometrie. Abbildung 2.5 zeigt die zwei Vorgehensweisen schematisiert auf.

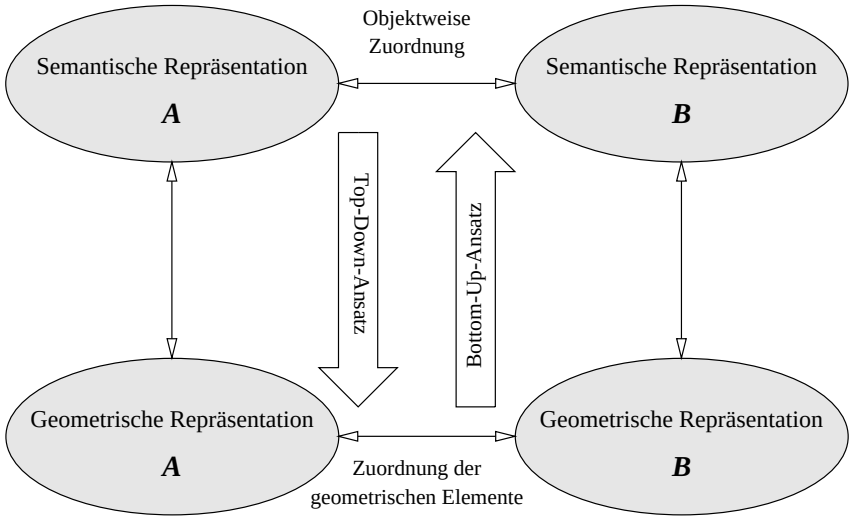


Abbildung 2.5: Top-Down- vs. Bottom-Up-Ansatz

Welche dieser beiden Techniken anzuwenden ist, hängt von den verwendeten Datenmodellen ab. Sollen z.B. die Objektklassen *Straße* aus zwei Datenmodellen zugeordnet werden, in denen das Attribut *Straßenname* gespeichert ist, ist ein Top-Down-Ansatz geeigneter, da die Straßennamen bereits eine eindeutige Identifizierung der Straßenobjekte darstellen und somit eine sehr leichte Zuordnung zwischen den Objekten ermöglichen. Es können auch Zwischenformen definiert werden, bei denen beispielsweise zuerst versucht wird, Linienelemente, welche Straßen repräsentieren, einander zuzuordnen, jedoch hierzu gleichzeitig Information nutzen, welche weiter höher in der Objektstruktur liegen, wie z.B. die Straßenbedeutung. Grundsätzlich sollte immer versucht werden, soviel Information wie möglich zu betrachten, um den Zuordnungsprozess zu erleichtern. Liegen jedoch keine einander entsprechenden Attribute von Objekten vor, muß ein Bottom-Up-Ansatz gewählt werden.

2.6 Integration von GDF und ATKIS

Die Gründe dafür, daß Straßendaten in Deutschland in mehreren Datenmodellen erfaßt werden, sind einerseits in unterschiedlichen fachlichen Interessen und andererseits in terminlichen Vorgaben zu suchen. *"Die Erfassung der EDRM-Informationen (European Digital Road Map) erfolgt bedauerlicherweise parallel zur ATKIS-DLM 25 Erfassung, da nur wenige Bundesländer die zeitgerechte Bereitstellung der Straßeninformationen aus ATKIS zusagen konnten. Da die Modellierung ATKIS-EDRM sich sehr ähneln, ist zu gegebener Zeit die Aktualisierung der EDRM auf ATKIS-Basis zu prüfen"* [Kophstahl 1994]. Im folgenden sollen die möglichen Vorteile einer Integration von GDF- und ATKIS-Daten diskutiert werden. Die Durchführbarkeit einer solchen Integration ist

neben technischen Fragen auch von politischen Fragen abhängig, die jedoch nicht Gegenstand dieser Arbeit sein können. Daher soll hier nur das prinzipielle Potential einer solchen Integration diskutiert werden.

Damit Straßenverkehrsdaten für Navigationszwecke verwendet werden können, müssen sie höchsten Ansprüchen an Qualität, Homogenität und Aktualität genügen [Wagner 1995]. Um das Straßennetz von Deutschland zu digitalisieren, ist ein enormer Aufwand an Zeit und Geld notwendig. So mußten für die GDF-Daten der Firma Bosch mehr als 2 Millionen Straßensegmente mit durchschnittlich sieben Attributen digitalisiert werden [Claussen 1995]. Nachdem die Daten nun flächendeckend vorliegen, sind sie fortzuführen. Es wird davon ausgegangen, daß sich Straßenverkehrsdaten im Mittel um 5 - 10 Prozent im Jahr ändern [Claussen 1995]. Eine noch höhere Veränderungsrate schätzt [Helmle 1995] mit 10 bis 15 Prozent. Daraus kann ersehen werden, welcher Aufwand mit der Fortführung dieser Daten verbunden ist. Der jährliche Pflegeaufwand der GDF-Datenbasis beträgt 20 bis 30 Prozent der Ersterfassung [Claussen 1995]. Die gleiche Situation tritt bei der Fortführung des ATKIS-Datenbestands auf. Aus wirtschaftlichen Gründen stellt sich die Frage, ob es möglich ist, die Daten nur in einem Datenmodell fortzuführen und durch eine Integration diese Änderungen in das andere Datenmodell zu übernehmen.

Es ist umstritten, ob dies nach der heutigen gesetzlichen Lage durchführbar ist: *"In der Bundesrepublik Deutschland sind die Grundlagenvermessung, die topographische Landesaufnahme und die Herstellung, Fortführung und Herausgabe der topographischen Landeskartenwerke öffentliche Aufgaben der Länder, die durch entsprechende Landesgesetze den Landesvermessungsbehörden zugewiesen sind"* [Herdeg 1994]. Inzwischen wird jedoch zunehmend die Notwendigkeit von Kooperationen deutlich. So wurde auf der Fachkonferenz "Aktualisierung von Geodaten" [DDGI 1996], bei der die Vertreter der wichtigsten Datenerfasser in Deutschland anwesend waren, beschlossen, daß Kooperationsmöglichkeiten zur Aktualisierung des Grunddatenbestandes zu prüfen und zu initiieren sind. Ebenso denkbar ist eine Fortführung von GDF-Daten mittels ATKIS-Daten, vor allem da ATKIS, neben der Erfüllung eigener Aufgaben, insbesondere auch den Bedarf Dritter an digitalen Informationen über die Topographie der Erdoberfläche befriedigen soll [Rossol 1988]. Hierbei würde sich der Fortführungsaufwand der GDF-Datenanbieter wesentlich verkleinern, sofern die Landesvermessungsbehörden in der Lage sind, die Qualitätsanforderungen an die Daten zu erfüllen.

Als weniger problematisch stellt sich ein Austausch von Attributen oder Objektklassen dar, die in einem Modell erfaßt werden, jedoch nicht in dem anderen. So sind z.B. in GDF die Straßennamen erfaßt, jedoch nicht in der derzeitigen Ausbaustufe von ATKIS. In der endgültigen Ausbaustufe sollen die Straßennamen ebenfalls erfaßt werden. Hier würde es sich anbieten, anstatt die Straßennamen erneut zu erfassen, sie aus den GDF-Daten zu extrahieren. In der anderen Richtung existiert auch bei den GDF-Datenanbietern Bedarf nach ATKIS-Daten. Dies könnte z.B. das Straßennetz von kleinen bis mittleren Ortschaften sein, die derzeit in den GDF-Datensätzen nicht digitalisiert sind, oder Objektklassen, die nicht aus dem Objektbereich Verkehr stammen, um damit das Anwendungsspektrum der GDF-Daten zu erweitern.

2.7 Schwerpunkt dieser Arbeit

Abbildung 2.6 zeigt den inhaltlichen Schwerpunkt dieser Arbeit am Beispiel der Integration von ATKIS und GDF. Die Ausgangssituation ist in Abbildung 2.6 a) dargestellt. Obwohl im Bereich des Straßenverkehrs dieselben Objekte erfaßt werden, liegen ATKIS- und GDF-Daten vollständig unabhängig voneinander vor. Um eine Integration zu realisieren erfolgt eine Zuordnung der Daten. Diese Zuordnungen sind die Voraussetzung für einen Import/Export von Attributen und Objekten (siehe Abbildung 2.6 c)) oder für eine vollständige Integration, bei der die beiden Datensätze homogenisiert und zu einen einzigen Datensatz integriert werden (siehe Abbildung 2.6 d)). Diese Arbeit widmet sich schwerpunktmäßig der Zuordnung von raumbezogenen Daten und zeigt die hierzu notwendigen Verfahren auf.

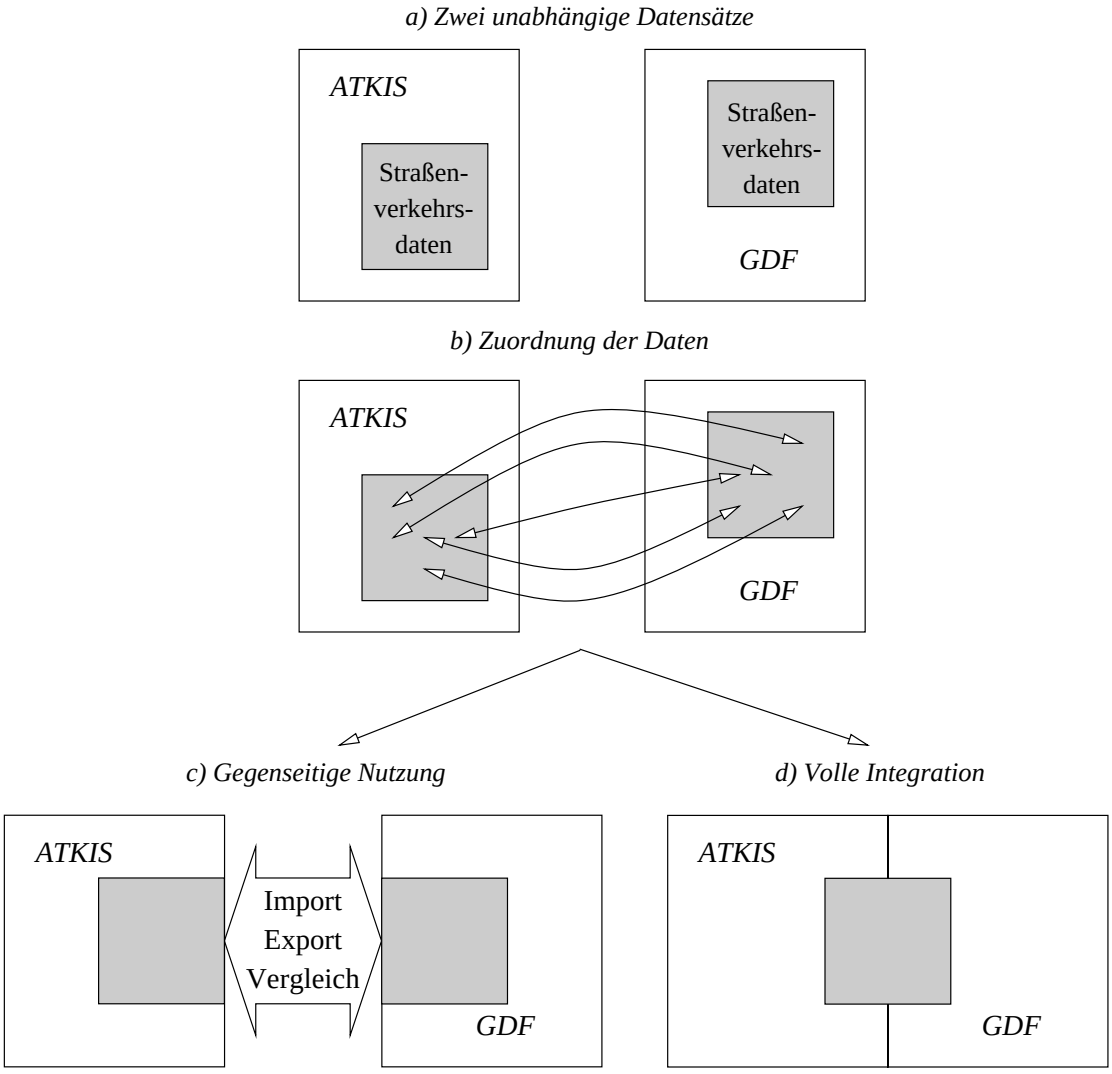


Abbildung 2.6: Integration von ATKIS und GDF

Kapitel 3

Geographic Data File

In diesem Kapitel wird ein umfassender Überblick über den Standard Geographic Data File (GDF) präsentiert. GDF wurde für Anwendungen aus dem Bereich der Fahrzeugnavigation entwickelt und hat viele Konzepte der Datenmodellierung speziell auf die daraus resultierenden Anforderungen an das Datenmodell zugeschnitten. Dies bedeutet jedoch nicht, daß diese Konzepte nicht auch für andere Anwendungen eingesetzt werden können. Insbesondere das Attributkonzept von GDF erlaubt eine leichte Modellierung komplizierter Sachverhalte. Da in GDF eine Vielzahl von Objektklassen definiert sind, die nicht nur Objekte aus dem Straßenverkehr umfassen, ist es möglich, GDF-Daten als Basisdaten für Anwendungen aus den unterschiedlichsten Bereichen zu verwenden. Damit haben GDF-Daten in Deutschland einen ähnlichen Stellenwert wie die Daten des Basisinformationssystem ATKIS. Umgekehrt gilt jedoch auch, daß in ATKIS ebenfalls Objekte aus dem Straßenverkehrsbereich erfaßt werden. Daher ist es notwendig, diese beiden Datenmodelle zunächst einem Vergleich zu unterziehen. Dieses Kapitel soll zusammen mit dem nächsten Kapitel, in dem das Datenmodell von ATKIS ausführlich diskutiert wird, eine Grundlage bilden, um die Unterschiede und Gemeinsamkeiten von GDF und ATKIS erarbeiten zu können.

Im folgenden wird zuerst ein Überblick über die Historie von GDF gegeben. Anschließend erfolgt eine Diskussion des konzeptionellen Datenmodells. Nach dieser übersichtsartigen Betrachtung wird das Attributkonzept und das Relationenkonzept im Detail diskutiert. Die Modellierung von komplexen Objekten wird an einem Beispiel beschrieben. Zur Untersuchung der GDF-Daten wurde eine Arbeitsumgebung auf der Basis von SICAD/open eingerichtet. Die Abbildung des GDF-Datenmodells auf dieses GIS-Produkt wird im letzten Teil dieses Kapitels beschrieben.

3.1 Historie

Ein erster Entwurf von GDF wurde im EUREKA-Projekt DEMETER (Digital Electronic Mapping of European Territory) von der Firma Bosch (Deutschland) und der Firma Philips (Niederlande) Mitte der achtziger Jahre entwickelt. EUREKA ist eine Initiative zur Unterstützung der Wettbewerbsfähigkeit von Unternehmen durch Förderung der Zusammenarbeit. Zu dieser Zeit war das einzig verfügbare nationale Austauschformat das britische National Transfer Format (NTF) [Salgé und Brüggemann 1992]. Das GDF-Austauschformat basiert auf dem NTF-Standard und wurde im Hinblick auf Straßenverkehrsdaten erweitert [Claussen 1989]. Obwohl der Standard der USA SDTS (Spatial Data Transfer Standard) zu diesem Zeitpunkt nur in einer Serie von Draft-Standards verfügbar war, flossen viele Ideen von SDTS in das GDF-Konzept ein [Heres et al. 1991]. Die erste Version GDF 1.0 wurde im Oktober 1988 veröffentlicht.

Im Jahr 1989 wurde von der Europäischen Kommision DRIVE I (Dedicated Road Infrastructure for Vehicle use and safety in Europe) ins Leben gerufen. Die Aufgaben waren die Verbesserung von Straßensicherheit, Transport-Effizienz und Umweltqualität. Im Rahmen des DRIVE I Programms wurden etwa 70 Projekte durchgeführt. Zwei dieser Projekte waren: Task Force EDRM (European Digital Road Map) und PANDORA (Prototyping A Navigation Database Of Road-network Attributes). Die Aufgaben dieser Projekte waren die Untersuchung von geeigneten Datenerfassungstechniken für RTI-Anwendungen (Roads Transports Informatics) sowie die Möglichkeiten der Nutzung existierender Datenquellen und die Brauchbarkeit des "draft Standards" GDF.

PANDORA war ein Parallelprojekt zu Task Force EDRM, jedoch begrenzt auf die Britischen Inseln [Claussen 1992]. In PANDORA und EDRM wurde der GDF 1.0 Standard intensiv getestet. Zu diesem Zeitpunkt wurde die Anzahl der Partner in der GDF-Standardisierungsgruppe erweitert. Die Wünsche der neuen Partner (speziell im Bereich der Touristeninformation und Transport-Logistik) resultierten in einer signifikanten Erweiterung des GDF-Feature- und Attribut-Kataloges [Heres et al. 1991]. Das erweiterte GDF wurde in der Version GDF 2.0 vorgestellt. Der französische Standard EDIGéO (Echange de Données Informatisées GéOgraphiques) war eine entscheidende Grundlage zur Weiterentwicklung des GDF-Standards von der Version 1.0 zur Version 2.0 [Brüggemann 1992].

Im Jahr 1992 wurde DRIVE II gestartet. Die Aufgabe von DRIVE II war die Vorbereitung einer Implementation eines ATT-Systems (Advanced Transport Telematics). Innerhalb von DRIVE II wurde das Nachfolgeprojekt EDRM II ins Leben gerufen. Die Aufgaben von EDRM II waren: Einführung von GDF als Daten-Austausch-Format, Entwicklung von Werkzeugen zur Generierung von GDF-Daten und Entwicklung von Methoden zur Integration von statischen Kartendaten mit dynamischer Verkehrsinformation. Die beteiligten Projekt-Partner setzten sich zusammen aus: Bosch, Philips, Daimler-Benz, Intergraph, Teleatlas und SAGEM. Die CERCO (Comité Européen des Responsables de la Cartographie Officielle) war als Partner zu Bosch eingebunden, mit der Möglichkeit den Standardisierungsprozeß zu beeinflussen. Das Ergebnis war der Standard GDF 2.1, der sich von GDF 2.0 vor allem im Konzept der segmentierten Attribute und der Möglichkeit der Verwendung von Zeitdomänen unterscheidet.

Die Europäische Kommision entschied, die eigentlichen Standardisierungsarbeiten den Technischen Kommisionen (TC) der CEN (Comité Européen de Normalisation) CEN/TC 278 und CEN/TC 287 zu übergeben [Salgé und Brüggemann 1992]. Die Zielrichtungen des TC 287 sind mehr allgemeiner Natur als die des TC 278 [Brüggemann 1995]. Während sich die Aktivitäten des CEN/TC 278 (Road Transport Telematic) vor allem auf das Anwendungsfeld des Straßenverkehrs richten, beschäftigt sich die CEN/TC 287 (Geographical Information) mit allgemeineren Standards auf dem Gebiet des Geoinformationswesen [Brüggemann 1992]. Momentan befindet sich GDF in den Normungsausschüssen des CEN und wird auf internationaler Ebene in der ISO (International Standardization Organization) diskutiert [Claussen 1995]. Die Version 3.0 liegt derzeit zur Abstimmung als europäischer Standard vor. In Bezug auf ATKIS ist zu erwarten, daß die terminologische Diskussion im CEN/TC 287 nicht nur den GDF-Standard definiert, sondern auch den ATKIS-Sprachgebrauch langfristig beeinflussen wird [Brüggemann 1994]. Als weiterer Standardisierungsschritt wurde eine vollständige Abbildung des GDF-Datenmodells auf das Datenaustauschformat ISO 8211 vorgenommen [Portele 1993]. ISO 8211 ist ein internationales Datenaustauschformat, welches zusätzlich zu den Daten auch das dazugehörige Datenmodell überträgt, und somit den Datenaustausch zwischen Rechnern ermöglicht, ohne daß die verwendeten Konverter das Datenmodell im voraus kennen müssen [ISO 1985].

3.2 Übersicht

Eine ausführliche Beschreibung des GDF-Standards befindet sich in der GDF-Gesamtdokumentation [Heres et al. 1991], welche insgesamt aus 8 Teilen besteht. Im ersten Teil wird eine Einführung und ein Überblick über das konzeptionelle Datenmodell (siehe Kapitel 3.3.1) sowie Definitionen der GDF-Fachausdrücke gegeben. Im zweiten Teil, dem *Feature Catalogue*, werden die zu erfassenden Objekte (Features)[1] beschrieben. Ein Feature ist ein Objekt der realen Welt und muß folgenden Kriterien genügen [Heres und Wood 1992]:

- Es muß sich umgangssprachlich durch ein Hauptwort ausdrücken lassen.
- Es muß räumlich exakt begrenzt sein.
- Zwei Features der gleichen Klasse müssen disjunkt sein.
- Features müssen in Form und Ort zeitbeständig sein, also z.B. keine Fahrzeuge.

Es werden vor allem solche Feature-Klassen definiert, welche Objekte aus dem Bereich der Verkehrsnavigation oder Verkehrsmanagementsysteme beschreiben, wie Straßen, Verkehrszeichen oder Parkflächen. Für jede Feature-Klasse wird eine kurze Definition angegeben. Einen Überblick über den Inhalt des *Feature Catalogue* wird in Anhang D aufgezeigt. Die Eigenschaften der Features werden im *Attribute Catalogue* definiert. Richtlinien für GDF-Attribute sind z.B. [Heres und Wood 1992]:

- Ein Attribut wird umgangssprachlich meist mit einem Adjektiv bezeichnet.
- Attribute dürfen keine räumliche Ausprägung haben.
- Attribute sollen nur Eigenschaften von Features ausdrücken und nicht Beziehungen zwischen ihnen.

GDF hat ein mächtiges Attributkonzept, welches die Bildung von komplexen, segmentierten und zeitabhängigen Attributen ermöglicht. Das Attributkonzept wird ausführlich in Kapitel 3.3.2 diskutiert.

[1] In der GDF-Namensgebung werden Objekte mit den Namen *Features* bezeichnet. Im folgenden wird immer der Name Feature verwendet, wenn ein Objekt im GDF-Datenmodell gemeint ist.

Im *Relationship Catalogue* werden die semantischen Relationen zwischen Features definiert. Es gibt viele Sachverhalte, welche sich mit Relationen wesentlich natürlicher abbilden lassen als mit Attributen (z.B. *Abbiegeverbot von Straße A nach Straße B*). Die semantischen Relationen in GDF lassen sich in unterschiedliche Klassen aufteilen [Heres und Wood 1992]. Eine Klasse beschreibt die *Teil-von* Beziehungen, wie z.B. *Straßenabschnitt in Stadt* oder *Stadt in Bundesland*. Weitere Klassen sind die Adjazenzrelationen (z.B. *Gebäude liegt an Straße*) und die Über- bzw. Unterführungsrelationen. Die letzte Klasse enthält die Abbiegerelationen wie z.B. *Abbiegen verboten* oder Vorfahrtsbeziehungen. Das Datenmodell für Relationen wird im Kapitel 3.3.3 diskutiert.

Die Regeln, wie Features mit kartographischen Primitiven (Punkt, Linie und Fläche) darzustellen sind, können im *Feature Representation Scheme* gefunden werden. Diese Regeln wurden bewußt aus dem Feature Catalogue herausgenommen, um alternative Schemata für andere Anwendungen definieren zu können. So kann eine Kreuzung in einer Anwendung durch einen Punkt und in einer anderen Anwendung durch eine Fläche dargestellt werden.

Der *Global Data Catalogue* enthält die Definitionen der globalen Daten, wie z.B. Erfassungsdatum, Genauigkeitsangaben oder geodätische Informationen. Die Notwendigkeit, nicht nur die Daten zu erfassen, sondern auch Metadaten über die eigentlichen Daten zu speichern, wird von immer mehr Anwendern gefordert. Insbesondere bei Anwendungen, welche mit Datensätzen von unterschiedlichen Datenquellen arbeiten, sind Metadaten von großer Wichtigkeit. Zu den Metadaten gehören auch Aussagen über die Qualität der Daten. Die Qualitätsanforderungen an GDF-Daten sind in der *Data Content Specification* definiert. Es werden Forderungen an die Vollständigkeit, Genauigkeit, Richtigkeit und Aktualität der Daten aufgestellt (siehe auch Kapitel 3.3.5).

Im letzten Teil der Dokumentation, den *Media Record Specifications*, wird beschrieben, wie die Daten physikalisch mit Hilfe von Datensätzen und -feldern im GDF-Austauschformat übertragen werden (siehe Kapitel 3.3.6). Neben diesem Datenübertragungsformat besteht auch die Möglichkeit, die Daten im internationalen Austauschformat ISO 8211 zu übermitteln.

3.3 Datenmodell

Im folgenden werden die Konzepte des GDF-Datenmodells diskutiert. Um die Datenmodelle beschreiben zu können, werden NIAM-Diagramme (Nijsen Information Analysis Method [Nijsen und Halpin 1989]) verwendet. Es handelt sich hierbei um eine graphische Notation zur Darstellung von Datenmodellen. NIAM-Diagramme zeichnen sich durch eine leichte Lesbarkeit und ein hohes Maß an Ausdrucksfähigkeit aus. Um zu vermeiden, daß durch die Übersetzung der GDF-spezifischen Ausdrücke in die deutsche Sprache falsche Interpretationen entstehen, werden im folgenden immer die in der englischen Sprache definierten Namen benutzt.

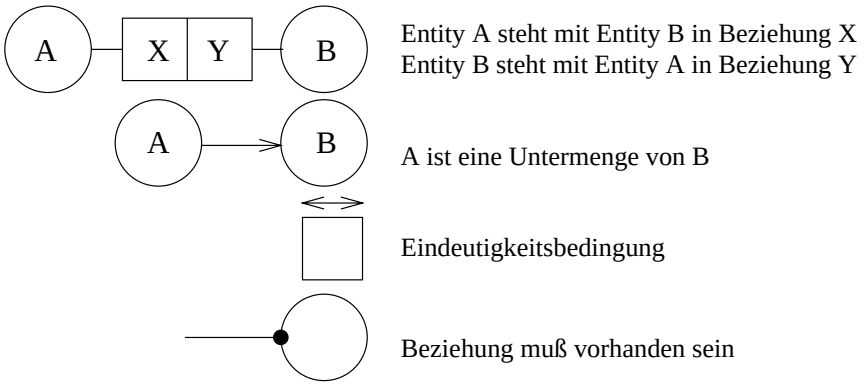


Abbildung 3.1: Erklärung der NIAM-Symbole

Abbildung 3.1 zeigt die wichtigsten Symbole der NIAM-Diagramme. Entities werden durch Kreise repräsentiert. Ein Entity ist ein Repräsentant eines Objektes oder Sachverhaltes, der datentechnisch repräsentiert werden muß [Rautenstrauch und Moazzami 1990]. Die Beziehungen zwischen den Entities können durch Pfeile und Rechtecke dargestellt werden. Pfeile stehen für Teilmengenbeziehungen und Rechtecke für Relationen. Die Mächtigkeit der Relationen können durch Punkte und Doppelpfeile dargestellt werden. Befindet sich an der Verbindung zwischen dem Entity und der Relation ein Punkt, so muß diese Beziehung mindestens einmal vorhanden sein (Mandatory). Wird über dem Rechteck ein Doppelpfeil gezeichnet, so darf die Beziehung höchstens einmal bestehen (Uniqueness). Die Kombination beider Symbole stellt eine Relation mit genau einem Partner dar.

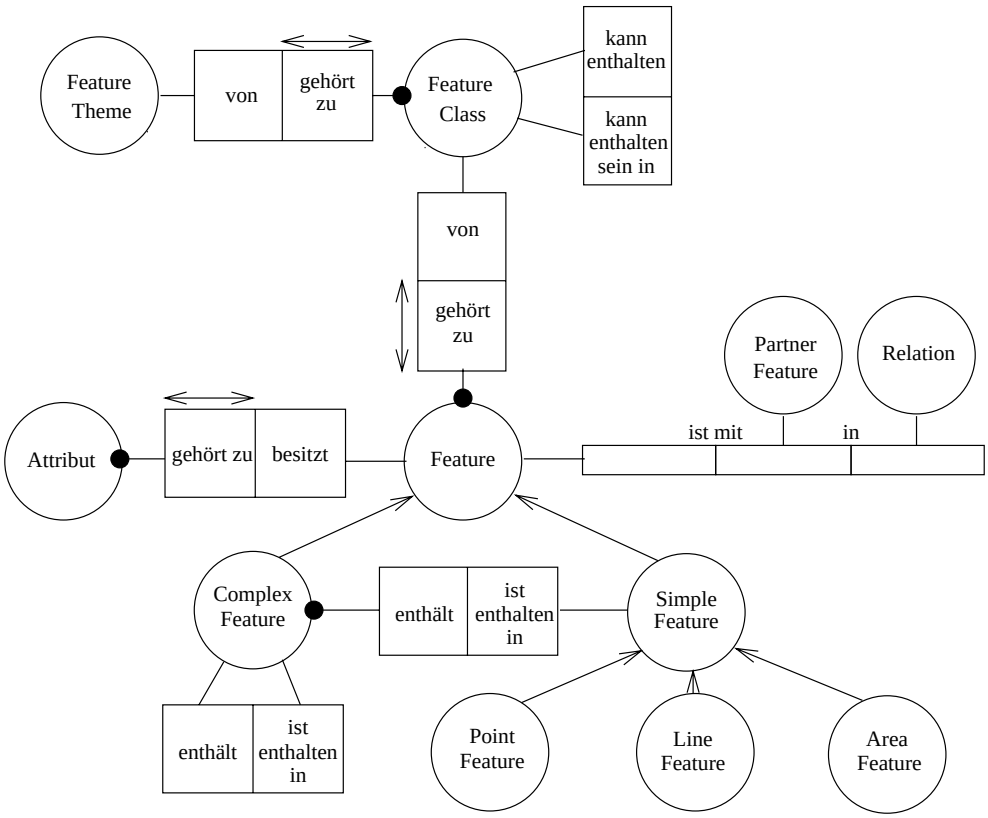


Abbildung 3.2: Konzeptionelles Datenmodell von GDF

3.3.1 Konzeptionelles Datenmodell

Abbildung 3.2 gibt einen Überblick über das konzeptionelle Datenmodell von GDF. Im Zentrum des Diagramms steht das Feature. Features können punktförmig, linienförmig, flächenförmig oder komplex sein. Komplexe Features bestehen selbst aus beliebigen Features. Die Teilmengenbeziehungen zwischen den verschiedenen Feature-Arten sind durch Pfeile dargestellt. Point Features, Line Features und Area Features sind eine Teilmenge der Simple Features, welche selbst wieder eine Teilmenge der allgemeinen Features ist.

Jedes Feature gehört zu *höchstens einer* Feature Class. Dies wird durch den Pfeil neben dem Rechteck dargestellt, welches die Relation zwischen Feature und Feature Class repräsentiert. Der Punkt an dem Kreis, welcher das Feature darstellt, gibt an, daß jedes Feature *mindestens* einer Feature Class angehört. Die Kombination von höchstens Eins und mindestens Eins resultiert in *genau Eins*. Daher handelt es sich um eine $1 : n$ Beziehung zwischen Feature Class und Features. Die Definition der Feature Class ist rekursiv und ermöglicht dadurch die Bildung von beliebigen Klassenhierarchien. Jede Feature Class gehört genau einem Feature Theme an. Folgende Feature Themes sind im GDF Katalog definiert: *Roads and Ferries, Administrative Areas, Settlements, Land Use Units, Brunnels*[2], *Railways, Waterways, Road Furniture* und *Services*.

Features können in semantischen Relationen mit beliebig vielen anderen Features stehen und eine beliebige Anzahl von Attributen besitzen. Auf diese Konzepte wird weiter unten noch genauer eingegangen.

3.3.2 Attributkonzept

Jedem Feature und jeder Relation in GDF können beliebig viele Attribute zugeordnet werden. GDF bietet ein mächtiges Attributkonzept an, welches erlaubt segmentierte, komplexe und zeitabhängige Attribute zu bilden. Abbildung 3.3 zeigt das Datenmodell für GDF-Attribute. Jedes Attribut gehört zu genau einem Typ und besitzt mindestens einen Wert. Durch die rekursive Definition der Attribute können komplexe Attribute gebildet werden. In der Abbildung wird ein neues NIAM-Symbol verwendet. Der gestrichelte Doppelpfeil mit einem **X** steht für

[2] Brunnel ist ein Kunstwort aus Bridge und Tunnel und bezeichnet allgemein Unter- und Überführungen.

Exklusion und bedeutet, daß ein Subattribut eines komplexen Attributes niemals direkt zu einem Feature gehören darf. Dies führt dazu, daß komplexe Attribute immer nur mit dem in der Hierarchie am höchsten stehenden Attribut mit einem Feature verbunden sind. Mit den beiden Werten *From Position* und *To Position* können segmentierte Attribute gebildet werden.

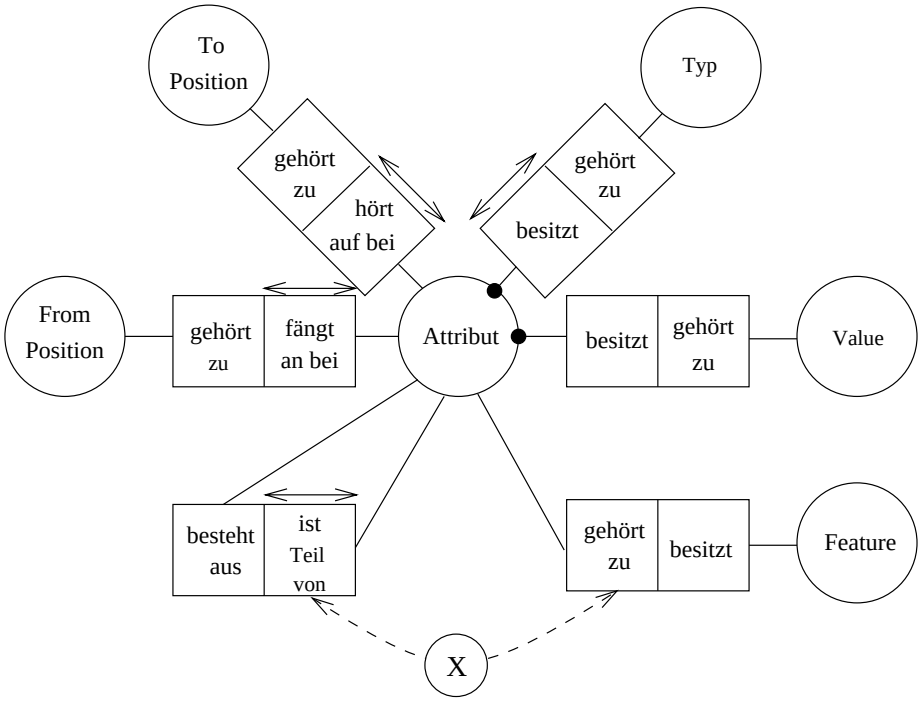


Abbildung 3.3: Datenmodell für Attribute in GDF

Eine Segmentierung von Attributen kann nur bei linienförmigen Features erfolgen. Eine Segmentierung bedeutet dabei die Angabe eines Bereiches, in dem das Attribut gültig ist. So kann z.B. eine Straßenverengung mit Hilfe von segmentierten Attributen derart dargestellt werden, daß das Attribut Straßenbreite die Straße in mehrere Segmente aufteilt, in denen die Straßenbreite durchgängig gültig ist.

Abbildung 3.4 zeigt diese Situation an einem Beispiel. In der schematischen Darstellung der Straße *L1204* kann gesehen werden, daß sich im mittleren Bereich der Straße eine Straßenverengung befindet. Dies bedeutet, daß dem Objekt Straße *L1204* nicht durchgängig ein Attribut Breite zugewiesen werden kann. Mit Hilfe von segmentierten Attributen ist es nicht nötig, die Straße in drei Teilobjekte aufzuteilen, was z.B. aus der Sicht eines Autofahrers unnatürlich wäre. Im Beispiel wurden daher diesem Objekt vier Attribute zugewiesen: drei Attribute um die Straßenbreite darzustellen und ein Attribut, welches die Straßenbezeichnung angibt und durchgängig für das gesamte Objekt Straße gültig ist. Diese Art der Modellierung eignet sich insbesondere für temporäre Änderungen im Straßennetz, wie sie z.B. durch Baustellen entstehen. Zur Modellierung einer Baustelle müssen keine neuen Objekte gebildet werden, sondern es ist ausreichend, entsprechend segmentierte Attribute an das bereits bestehende Objekt anzuhängen. Kurzzeitige Änderungen im Straßenverkehr lassen sich dadurch ohne geometrische und topologische Strukturänderungen in die Datenbank einfügen und wieder löschen.

Komplexe Attribute bestehen aus Attributen, die selbst wieder komplex sein können. Dadurch ist die Bildung beliebiger Attributhierarchien möglich. Ein komplexes Attribut in GDF ist z.B. *Traffic Sign Information*, welches aus den Teilattributen *Traffic Sign Class, Direction, Symbol on Traffic Sign, Textual Content of a Traffic Sign* und *Value on Traffic Sign* bestehen kann.

Zeitabhängige Attribute sind die häufigste Form der komplexen Attribute. Jedem Attribut kann ein Attribut des Typs Zeitdomäne untergeordnet werden, in dem es gültig ist. Abbildung 3.5 zeigt ein Beispiel eines komplexen Attributes. In der zu modellierenden Situation soll einem Objekt Straße eine Höchstgeschwindigkeit in einem bestimmten Zeitintervall zugeordnet werden. Hierzu wird an das Objekt Straße ein komplexes Attribut gehängt, welches selbst aus zwei Subattributen besteht. Eines dieser Subattribute enthält die Zeitdomäne, welche das Intervall angibt, in dem die Höchstgeschwindigkeit gültig ist und in dem anderen wird der Wert der Höchstgeschwindigkeit gespeichert. Die Zeitdomänen werden mit Hilfe Boolescher Ausdrücke beschrieben und ermöglichen die Modellierung beliebiger Zeitangaben. Diese Form der komplexen Attribute erlaubt die

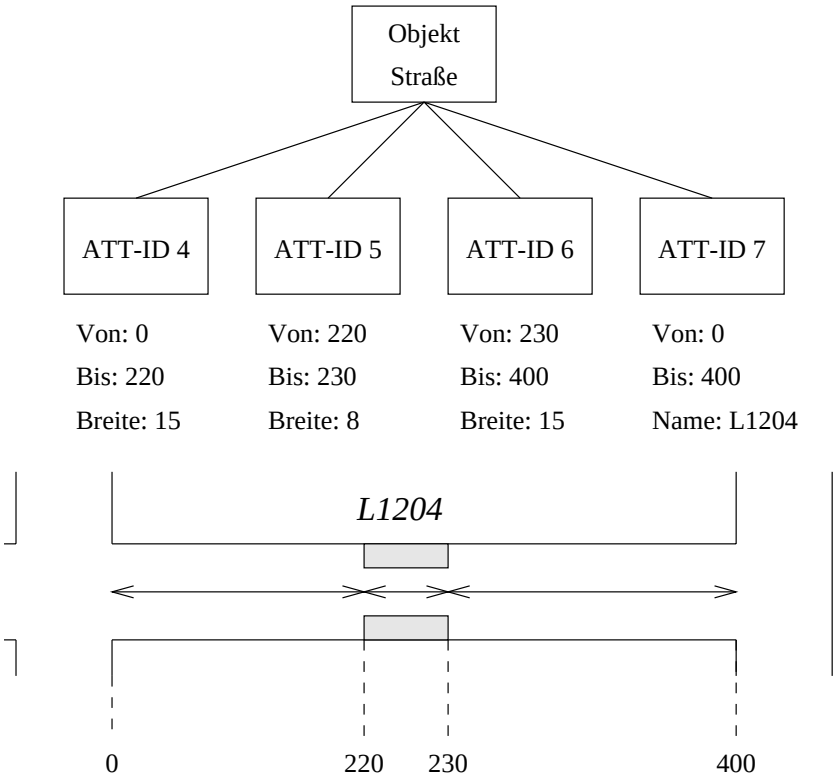


Abbildung 3.4: Segmentiertes Attribut in GDF

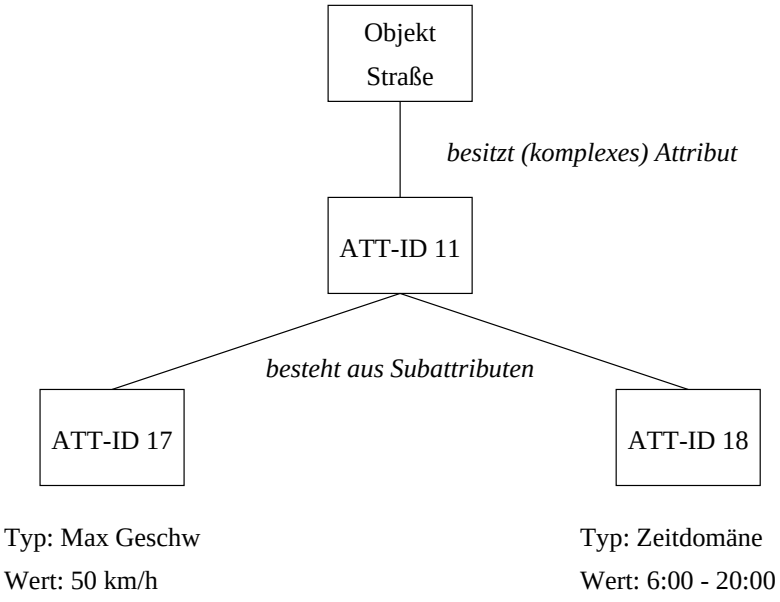


Abbildung 3.5: Komplexes Attribut in GDF

Verknüpfung von beliebigen Attributen mit einer Zeitdomäne. Attribute können in GDF auch an semantische Relationen gehängt werden. Dies ermöglicht eine einfache Darstellung komplexer Sachverhalte, wie z.B. *Abbiegeverbot von Straße A nach Straße B im Zeitintervall t.*

Eine sehr wichtige Information für die Fahrzeugnavigation ist die erlaubte Fahrtrichtung für Fahrzeuge. Auch diese Information wird in GDF mit Hilfe von Attributen modelliert. Die Topologie des Straßenverlaufs wird hierzu als gerichteter Graph abgespeichert. Zu jedem linienförmigen Straßenabschnitt existiert ein Attribut mit

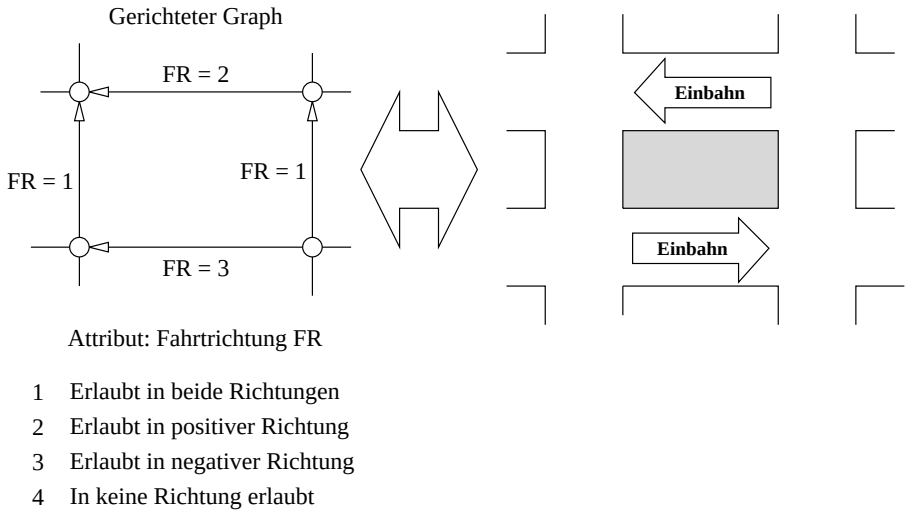


Abbildung 3.6: Modellierung des Verkehrsflusses

der Information, ob der Straßenabschnitt in beide, nur in positiver bzw. negativer Richtung oder gar nicht befahren werden darf. Abbildung 3.6 zeigt ein Beispiel für die Modellierung des Verkehrsflusses in GDF an einem kleinen Ausschnitt aus einem Straßennetzwerk.

3.3.3 Semantische Relationen

Eine semantische Relation ist eine bedeutungsvolle Verbindung zwischen zwei Features. Abbildung 3.7 zeigt das Datenmodell für Relationen. Relationen sind zur Darstellung bestimmter Informationen geeigneter als die Modellierung mit Attributen. So läßt sich die Information *Straße A liegt in Stadt B* viel natürlicher mit einer Relation als mit einem Attribut darstellen. Jede Relation in GDF besitzt einen eindeutigen Code sowie einen eindeutigen Namen, der eine umgangssprachliche Beschreibung der Relation angibt (z.B. Abbiegeverbot).

Die häufigsten Relationen sind binäre Relationen. Manche Relationen benötigen jedoch mehr Partner. So wird z.B. das Abbiegeverbot mit einer Relation mit drei beteiligten Features dargestellt (ein Straßenelement, eine Verbindung zwischen zwei Straßenelementen und ein weiteres Straßenelement). Mit Hilfe von Attributen, die an Relationen gehängt werden, ist es möglich, auch sehr komplexe Sachverhalte einfach zu modellieren.

3.3.4 Darstellung auf verschiedenen Ebenen

GDF bietet ein besonderes Konzept zur Darstellung von Features und komplexen Features an. Diese Darstellung soll an einem Beispiel in Abbildung 3.8 erläutert werden. In dem Rechteck links oben sind symbolisch die zu erfassenden Objekte der Landschaft angedeutet. Es ist ein Straßennetzwerk sowie eine administrative Grenze (z.B. durch zwei Landkreise) zu sehen.

Auf der Geometrie- bzw. Topologieebene (Ebene 0) besteht noch kein direkter Objektbezug. Alle erfaßten Knoten und Kanten sind als planarer Graph dargestellt. In der Abbildung sind Knoten durch Kreise und Zwischenpunkte durch Quadrate mit Kreuzen dargestellt. Die Kanten des Graphs werden gerichtet abgespeichert. Es ist zu erkennen, daß bei den Straßen, welche über die Grenze verlaufen, zusätzliche Knoten eingefügt wurden, um die Planaritätsbedingung zu erfüllen. Auf diesem Level sind exakte topologische Anfragen möglich (z.B.: *In welchem Landkreis befinde ich mich im Moment?*).

Jedes Feature Theme der Ebene 0 wird in einem eigenen Layer der Ebene 1 dargestellt. In der Abbildung ist das Feature Theme *Straßennetzwerk* dargestellt. Man kann sehen, daß die Knoten, welche durch den Schnitt der Straße mit der administrativen Grenze entstanden sind, in dieser Ebene nicht mehr dargestellt werden. In dieser Ebene wird die optimale Routenfindung durchgeführt.

Auf der Ebene 2 werden nur noch komplexe Features abgespeichert. Komplexe Features sind z.B. Kreuzungen, welche aus Straßenstücken und Verbindungsstücken bestehen. Die Aggregation von Straßenelementen und Straßenverbindungsstücken führt zu einer synthetischen Sicht auf das Straßennetzwerk, so wie es der Sicht

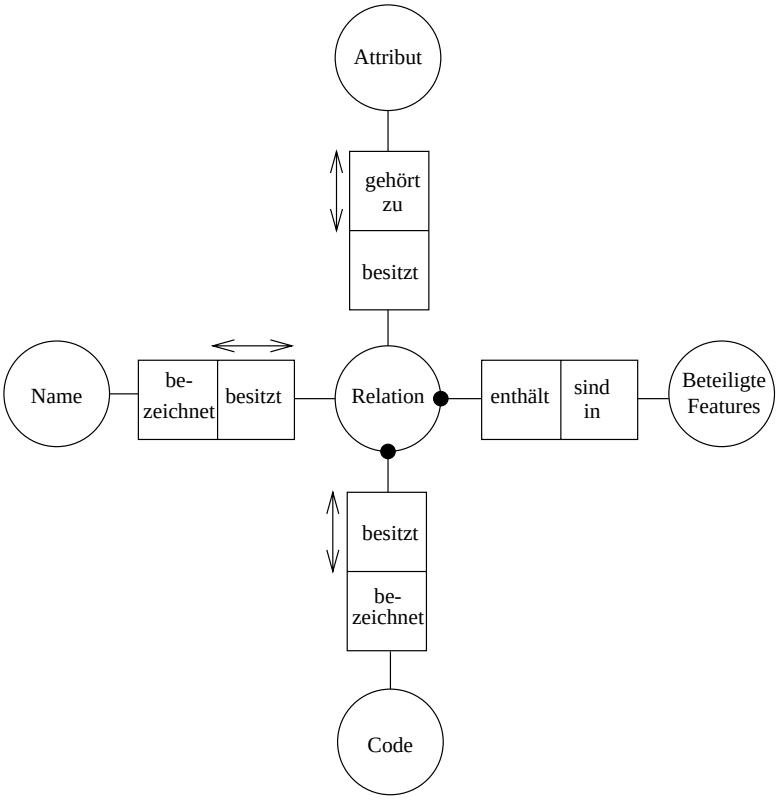


Abbildung 3.7: Datenmodell für Relationen in GDF

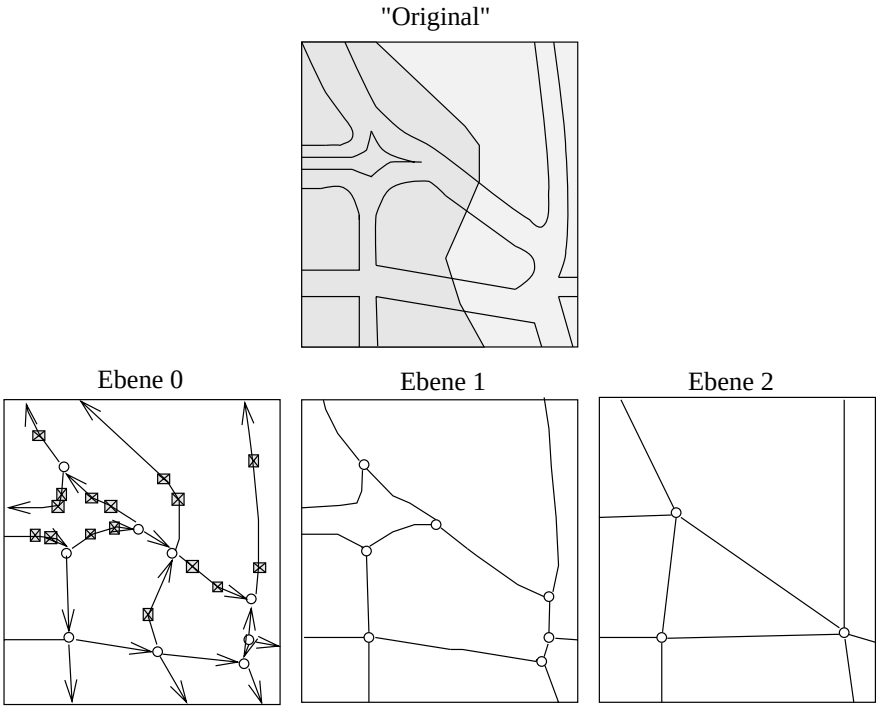


Abbildung 3.8: Darstellung in verschiedenen Ebenen

eines Autofahrers entspricht [Portier 1994]. In dieser Ebene kann die Verkehrsführung auf einem On-Board Display stattfinden. Speichert man zu den komplexen Features noch die Verbindungsrelationen zwischen den

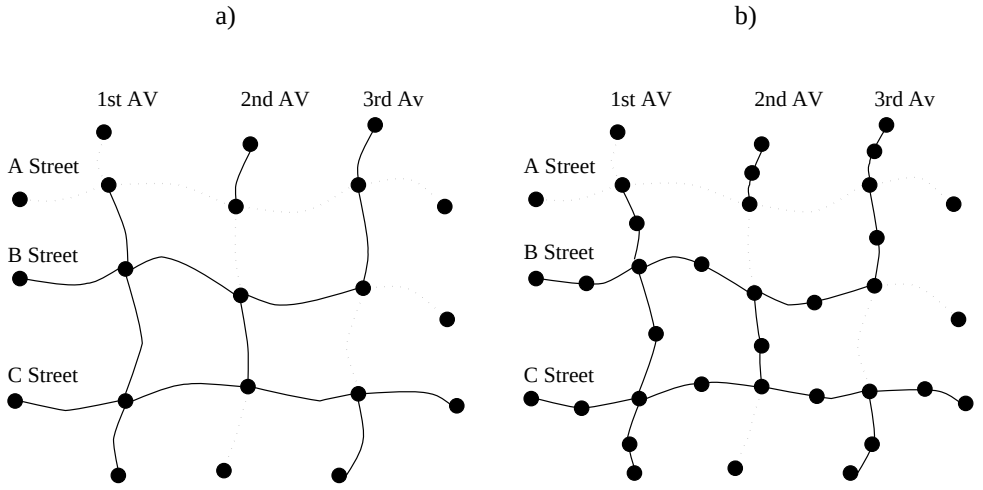
Kanten und einen Kostenfaktor für die Kanten ab, so ist eine optimale Routenfindung auf einem generalisierten Niveau möglich, ohne daß der Graph eine geometrische Repräsentation besitzen muß.

3.3.5 Qualitätsanforderungen

Um raumbezogene Daten verarbeiten zu können, müssen bestimmte Eigenschaften der Daten bekannt sein. Vor allem Aussagen über die Qualität der Daten sind von großer Bedeutung. Damit ein Benutzer raumbezogene Daten interpretieren kann, müssen die *sechs C's* [Killick 1992] einer digitalen Karte vorliegen:

- Was ist der Inhalt der Daten? *Content*
- Was ist der Raumbezug der Daten? *Coverage*
- Wie aktuell sind die Daten? *Currentness*
- Wie vollständig sind die Daten? *Completeness*
- Wie konsistent sind die Daten? *Consistency*
- Wie korrekt bzw. akkurat sind die Daten? *Correctness*

Während die Angaben für die ersten drei Fragen relativ einfach mit den Daten mitgeliefert werden können, entstehen bei den restlichen drei Fragen Probleme. In der *Data Content Specification* des GDF-Kataloges werden Angaben der folgenden Art gemacht: *99,99 % der Straßen der Klasse 3 bis 5 müssen erfaßt sein.* Auf den ersten Blick scheint es, daß sich die obigen Fragen auf Grund dieser Informationen beantworten lassen. Jedoch haben solche Qualitätsangaben keine Bedeutung, solange nicht exakt beschrieben ist, wie diese Angaben evaluiert werden können.



Anz. Straßensegmente	24
Anz. Straßensegmente mit Namen	15
Anz. eindeutige Straßennamen	6
Anz. eindeutige Straßennamen in der DB	5
% eindeutige Straßennamen in der DB	83%
% Straßensegmente mit Namen in der DB	62%

Anz. Straßensegmente	39
Anz. Straßensegmente mit Namen	30
Anz. eindeutige Straßennamen	6
Anz. eindeutige Straßennamen in der DB	5
% eindeutige Straßennamen in der DB	83%
% Straßensegmente mit Namen in der DB	77%

Abbildung 3.9: Unterschiedliche Qualitätsmaße für einen Datensatz (aus [Killick 1992])

Dieses Problem soll mit einem Beispiel (Abbildung 3.9) aus [Killick 1992] verdeutlicht werden. In dem Beispiel errechnen sich für zwei qualitativ gleichwertige Datensätze zwei unterschiedliche Qualitätsmaße. Diejenigen Straßen, für die kein Straßenname erfaßt wurde, sind gepunktet dargestellt. Es wird der Prozentsatz von Straßen mit Straßennamen in der Datenbank gesucht. Während bei der Berechnung der Vollständigkeit der Straßenelemente mit Namen in der Datenbank in Abbildung a) ein Prozentsatz von 62% erhalten wird, ist der Prozentsatz

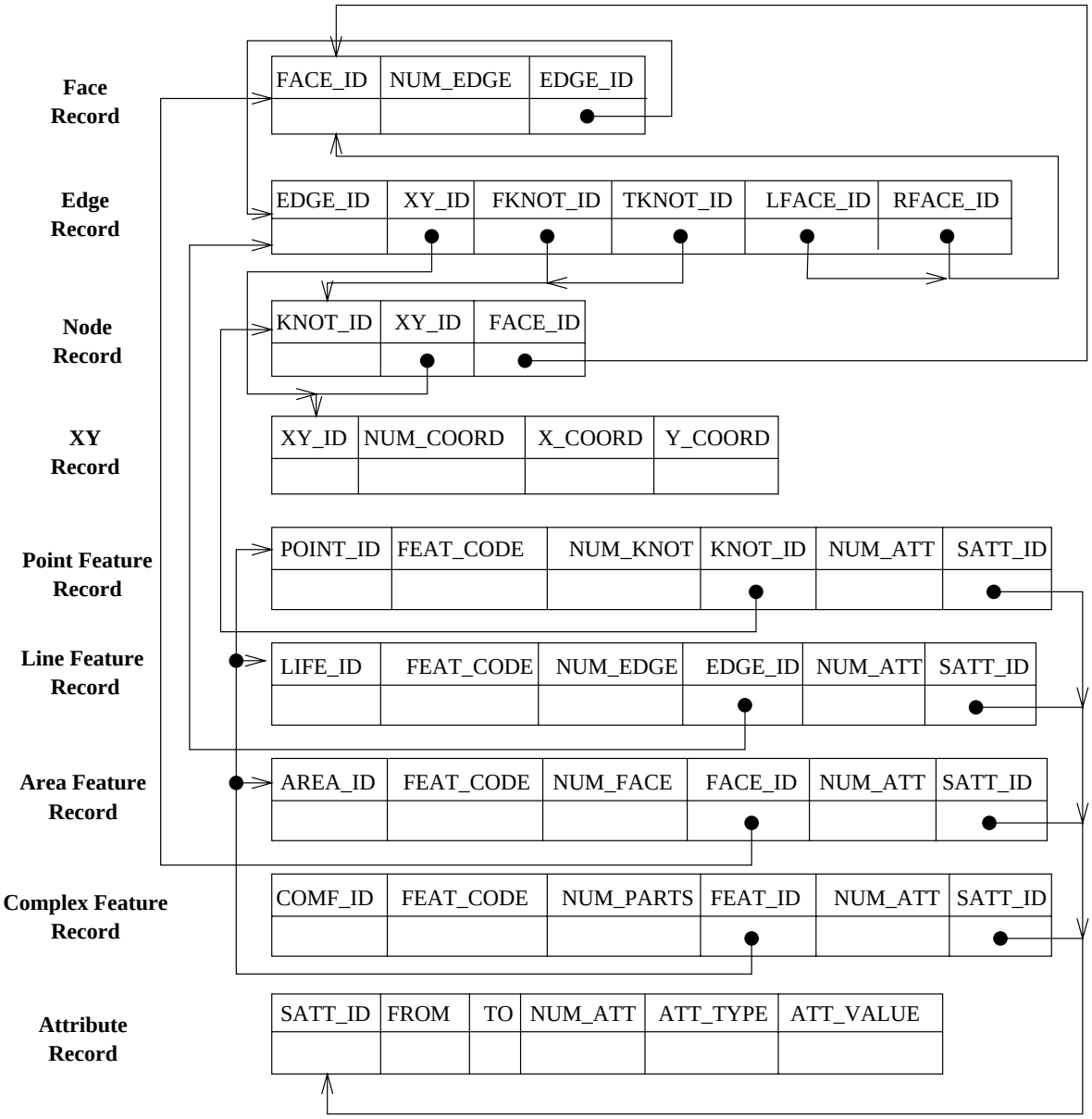


Abbildung 3.10: Struktur des Austauschformat für GDF-Daten

in Abbildung b) 77%, da hier einfach die Anzahl der Straßensegmente mit Namen durch Einführung zusätzlicher Knoten künstlich erhöht wurde.

Dieses Beispiel soll verdeutlichen, daß die alleinige Angabe von Qualitätsmaßen nicht ausreichend ist, sondern daß diese Zahlen erst dann Bedeutung bekommen, wenn die Art der Berechnung dieser Maße eindeutig nachvollzogen werden kann. Daher stellt sich als eine Forderung an die GDF-Daten, daß die Qualitätsaussagen genauer gefaßt werden müssen.

3.3.6 Austauschformat

Die GDF-Dokumentation enthält eine komplette Beschreibung zur physikalischen Übertragung der Daten mit Hilfe von Datensätzen und -feldern. Abbildung 3.10 zeigt die Struktur des Austauschformates. Aus Übersichtsgründen ist nur ein Teil der möglichen Datensätze und Datenfelder dargestellt. Eine GDF-Datei besteht aus einer Aufzählung von Datensätzen. Die einzelnen Datensätze sind durch Zeiger miteinander verknüpft. Die Verknüpfung der Datensätze erfolgt so, daß im Datenaustauschformat exakt die Struktur des Datenmodells abgebildet wird.

Als Beispiel soll ein linienförmiges Feature betrachtet werden. Ein linienförmiges Feature wird in einem eigenen Record abgespeichert. Dieser Record beinhaltet einen Pointer auf einen Kanten-Record, der selbst wieder auf zwei Knoten-Records und einen XY-Record zeigt. In dem XY-Record werden, falls vorhanden, die Koordinaten der Zwischenpunkte der Kante abgespeichert. Die zwei Knoten-Records zeigen ebenfalls auf je einen XY-Record, der die Koordinaten der Knoten enthält. Auf diese Art können die Daten redundanzfrei abgelegt werden.

3.4 Erfassungsstand

Die Erfassung von GDF-Daten wird in Europa von den zwei Konsortien EDRA (European Digital Road Association) und EGT (European Geographic Technologies) durchgeführt. Diese beiden Institutionen können als Konkurrenten angesehen werden, da sie beide unabhängig voneinander GDF-Daten flächendeckend in Europa erfassen. Das Konsortium EDRA ist ein Zusammenschluß der Firmen Bosch, Teleatlas und ETAK [Ireland 1994*a*]. Die Erfassung von EDRA-Daten in Deutschland wird von der Firma Bosch durchgeführt. EGT wurde von der Firma Philips sowie einer Vielzahl anderer Unternehmen gegründet, die alle im Bereich Verkehrsnavigation arbeiten [Ireland 1994*b*].

Bei einer Doppelerfassung von raumbezogenen Daten entstehen jedoch unterschiedliche Datensätze. Ein Grund dafür ist, daß die zu erfassenden Elemente bei mehrfachem Digitalisieren nie deckungsgleich, zumindest was die Erfassung der Koordinaten angeht, diskretisiert werden können. Weitere Unterschiede enstehen durch verschiedene Erfassungsvorgaben. EGT-Daten haben eine andere Erfassungstiefe als EDRA-Daten. Dies bedeutet, daß z.B. kleinere Seitenstraßen in einem Datensatz erfaßt werden, jedoch nicht in dem anderen. Als weiterer Unterschied ist die verschiedenartige Interpretation der zu erfassenden Objekte durch die verschiedenen Erfasser zu sehen. Insbesondere in komplexen Kreuzungsbereichen müssen sehr viele Informationen erfaßt werden. Diese Komplexität ist jedoch nur schwer durch eine Digitalisiervorschrift zu beschreiben und es bleibt ein Interpretationsspielraum für den Erfasser erhalten. Dies trifft auch bei weniger komplexen Objekten zu, wie z.B. bei der Wahl der Anzahl der Zwischenpunkte bei der Digitalisierung einer Kurve durch ein Polygon.

Abbildung 3.11 und 3.12 sollen die Unterschiede an Beispielen aufzeigen. Die Datensätze stammen aus dem Stadtgebiet Bonn. In Abbildung 3.11 a) sind Bosch-Daten (EDRA) und in Abbildung 3.11 b) EGT-Daten dargestellt. Bei einer genauen Betrachtung läßt sich feststellen, daß die Datensätze einander zwar sehr ähnlich sind, jedoch einige der Objekte des Bosch-Datensatzes nicht im EGT-Datensatz erfaßt sind und umgekehrt. In dem EGT-Datensatz waren insgesamt mehr Features (35.202) erfaßt als im Bosch-Datensatz (19.308). Bei einer direkten Überlagerung der beiden Datensätze (Abbildung 3.11 c)) wird ein systematischer Fehler deutlich. Hierbei handelt es sich um eine Translation der Größenordnung von ca. 50 Meter. Abbildung 3.11 d) zeigt die Überlagerung der beiden Datensätze nach der Rücktransformation dieser Translation.

Um die lokal unterschiedliche Erfassung darzustellen, zeigt Abbildung 3.12 eine Kreuzung mit EGT-Daten (Abbildung a) und Bosch-Daten (Abbildung b). In den Kreuzungsbereichen werden die Bosch-Daten mit einem höheren Detaillierungsgrad erfaßt als die EGT-Daten. Bei einer Überlagerung dieser Kreuzungsbereiche (Abbildung c) wird deutlich, daß selbst bei zwei Datensätzen, welche im gleichen Datenmodell erfaßt wurden, große Differenzen bestehen. Ein weiterer Unterschied zwischen den beiden Datensätzen ist in Abbildung d) dargestellt. Die Bildung von komplexen Objekten wurde nur im Bosch-Datensatz durchgeführt. Kreuzungsbereiche und Straßen mit zwei Fahrspuren sind im Bosch-Datensatz mit Hilfe von komplexen Objekten aggregiert. Die Bildung von komplexen Objekten entspricht der Ebene 2 Darstellung (siehe Kapitel 3.3.4). Abbildung 3.13 zeigt ein Beispiel für diese Darstellungsart der Objekte und Objektstrukturen mit Hilfe von Rauten und Linien. In der Abbildung ist eine Kreuzung zu sehen, von der in vier Richtungen Straßen mit zwei getrennten Fahrbahnen abgehen. Jedes Straßenstück wird mit einem linienförmigen Feature *Road Element* und jede Verbindung zwischen zwei Straßenstücken mit einem punktförmigen Feature *Junction* dargestellt. Zur Modellierung der Kreuzung wird ein komplexes Feature *Intersection* gebildet, mit dem die Features der Kreuzung aggregiert werden. Straßen, die mehr als eine Fahrspur besitzen, werden zu einem komplexen Feature *Road* aggregiert.

Abbildung 3.14 zeigt die Verfügbarkeit von GDF-Daten der Firma Bosch in Europa (entnommen aus [Bosch 1995]). Bei der Erfassung der Daten haben die Ballungsräume Priorität. In Deutschland stehen GDF-Daten seit 1995 flächendeckend zur Verfügung. Mit flächendeckend wird eine Tiefenerfassung (jede Straße über 3 Meter Straßenbreite) für die Ballungszentren und Städte mit mehr als 50.000 Einwohnern sowie die Erfassung des Überlandstraßennetzes bezeichnet [Wagner 1995]. Dies entspricht in Deutschland ca. 500.000 Straßenkilometern. Es ist damit zu rechnen, daß Ende 1997 Bosch-Daten für Westeuropa fast flächendeckend zur Verfügung stehen.

3.5 Abbildung auf SICAD/open

Im Rahmen dieser Arbeit wurde ein Konverter erstellt, welcher die Abbildung der Daten auf ein GIS-Produkt durchführt [Walter 1995*a*]. In GDF wird die zu erfassende Landschaft in Objekte strukturiert und erfaßt. Daher muß ein Werkzeug, mit dem die Daten bearbeitet werden sollen, die Möglichkeit der Objektbildung und -bearbeitung bereitstellen, um eine einfache Handhabung der Daten zu gewährleisten. Das GIS-Produkt SICAD/open der Firma Siemens Nixdorf AG [Siemens 1993*b*] besitzt die notwendige Funktionalität für die Abbildung des GDF-Datenmodells. Eine weitere Voraussetzung, welche an das GIS-Produkt gestellt wurde, war die Möglichkeit neben den GDF-Daten auch ATKIS-Daten bearbeiten zu können, da die Zuordnung von ATKIS- und GDF-Daten ein zentraler Bestandteil dieser Arbeit ist.

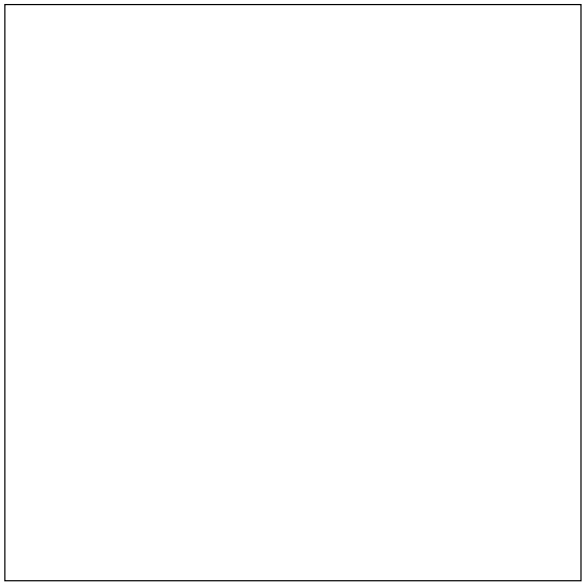
Hierzu wurden zwei verschiedene Wege der Abbildung des GDF-Datenmodells auf SICAD/open untersucht. Das Modul KRT1 [Siemens 1993*d*] bietet die Möglichkeit der Objektbildung und -bearbeitung. GDF-Features können direkt auf diese Objekte abgebildet werden. Die zu den Features gehörenden Sachdaten werden in relationalen Tabellen gespeichert, welche mit Hilfe von Pointern mit der Grafik verknüpft sind. Als alternative Möglichkeit wurde die Abbildung des GDF-Datenmodells auf das Modul ALK/ATKIS [Siemens 1993*a*] untersucht. Auch hier war eine vollständige Abbildung des Datenmodells möglich. Jedoch war diese Art der Abbildung etwas unnatürlich, da teilweise darauf geachtet werden mußte, daß Integritätsbedingungen, welche nur in ATKIS eine Rolle spielen, nicht durch GDF-Elemente verletzt wurden. Zwar bietet das Modul ALK/ATKIS eine höhere Funktionalität als das Modul KRT1, jedoch ist die Abbildung der GDF-Daten auf das Modul KRT1 direkter. Daher wurde als Werkzeug für die Bearbeitung von GDF-Daten das Modul KRT1 und zur Bearbeitung der ATKIS-Daten das Modul ALK/ATKIS verwendet.

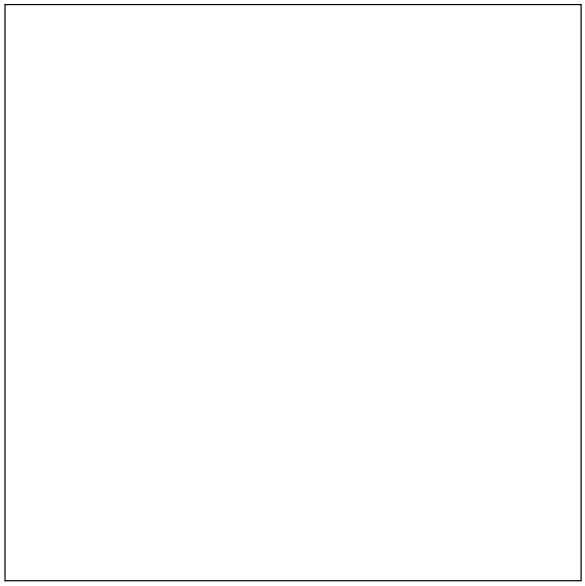
Um die Daten in SICAD einspielen zu können, werden sie zuerst in das Datenaustauschformat SQD umgewandelt. SQD ist das firmeneigene Datenaustauschformat von SICAD/open. Hierzu wurde ein Konvertierungsprogramm in der Rapid-Prototyping Sprache Python [Rossum 1994*a*, Rossum 1994*b*, Rossum 1994*c*] entwickelt. Python ist eine objektorientierte Interpretersprache und erlaubt es, sehr kompakte und leicht lesbare Programme in kurzer Entwicklungszeit zu erstellen. Python wurde am mathematischen Zentrum in Amsterdam entwickelt und ist als Public Domain Software für unterschiedlichste Plattformen erhältlich. Eine Beschreibung der Datenumsetzung, sowie der Definition der relationalen Tabellen findet sich in [Walter 1995*a*].

Abbildung 3.15 zeigt ein Beispiel eines Demoprogrammes zur Darstellung von GDF-Daten in SICAD/open. In dem Fenster ist eine Straßenkreuzung sowie die Objekte und Objektbeziehungen mit Symbolen dargestellt. Da in GDF Features keine Objektkoordinate besitzen, werden zur Darstellung der Objektbeziehungen die Objektkoordinaten aus den Koordinaten der zugehörigen Elemente berechnet. Mit Hilfe von SICAD/open wurden statistische Auswertungen der Daten durchgeführt sowie die Präsentation der Ergebnisse erstellt.

a) Bosch-Daten

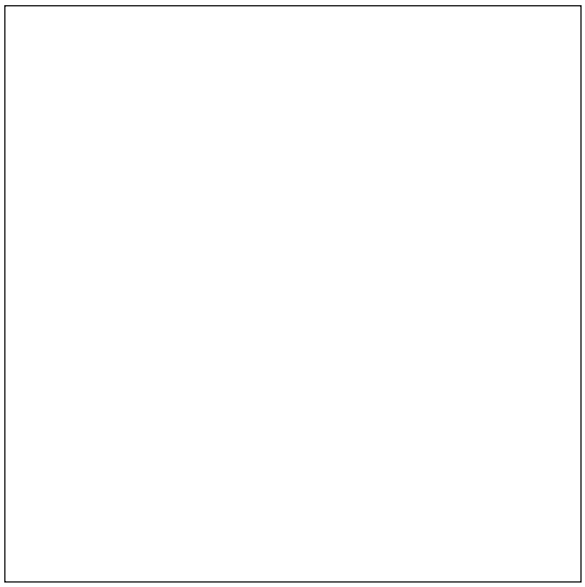
b) EGT-Daten





c) Überlagerung von a) und b)

d) Wie c), ohne globalen Fehler



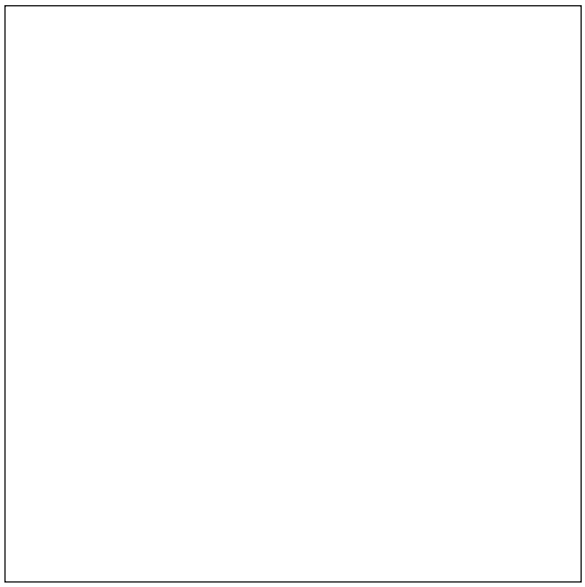
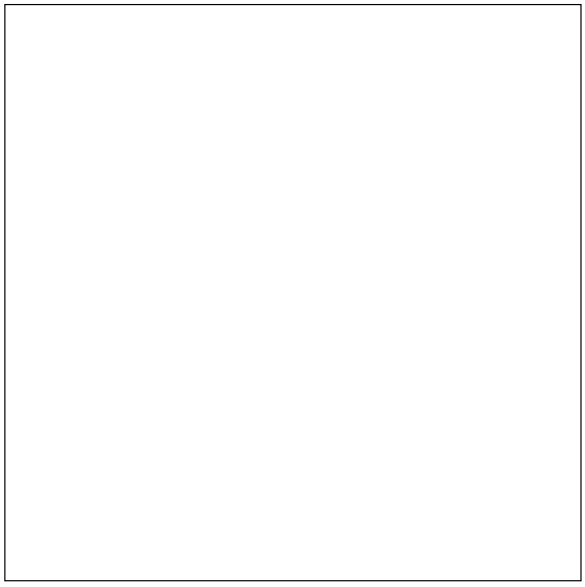
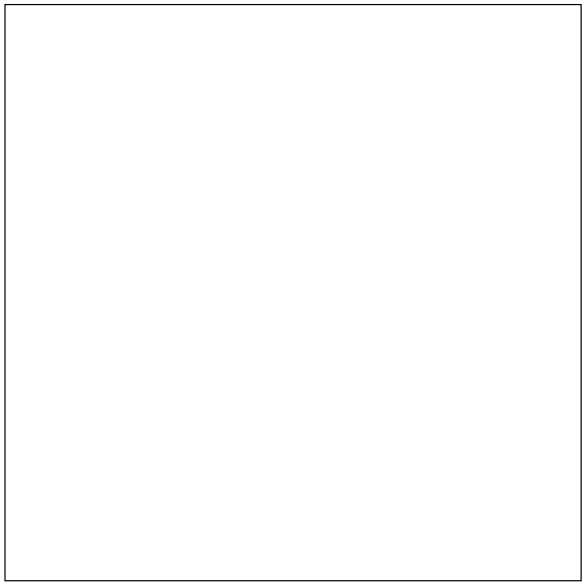


Abbildung 3.11: EGT und Bosch-Daten im Vergleich

a) EGT-Daten

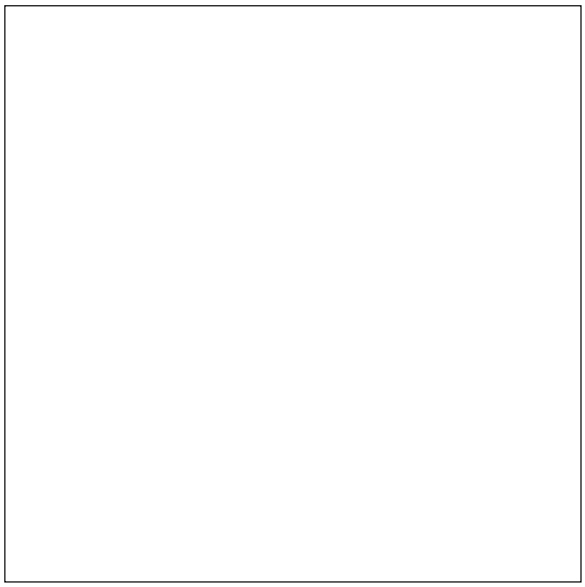
b) Bosch-Daten





c) Überlagerung von a) und b)

d) Darstellung der komplexen Objekte



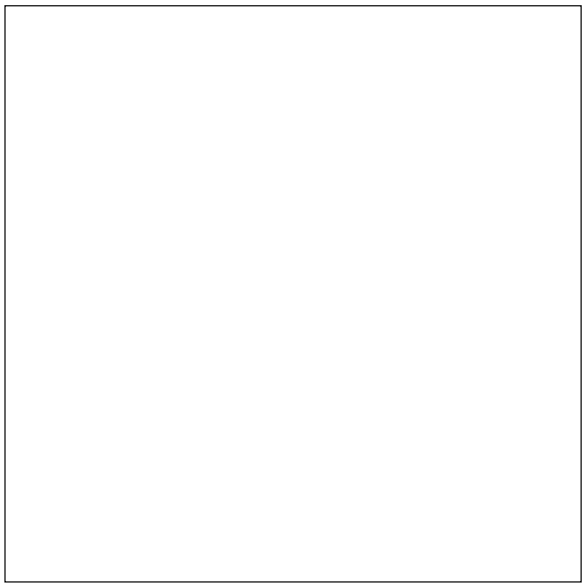


Abbildung 3.12: Kreuzungsbereich mit EGT- und Bosch-Daten

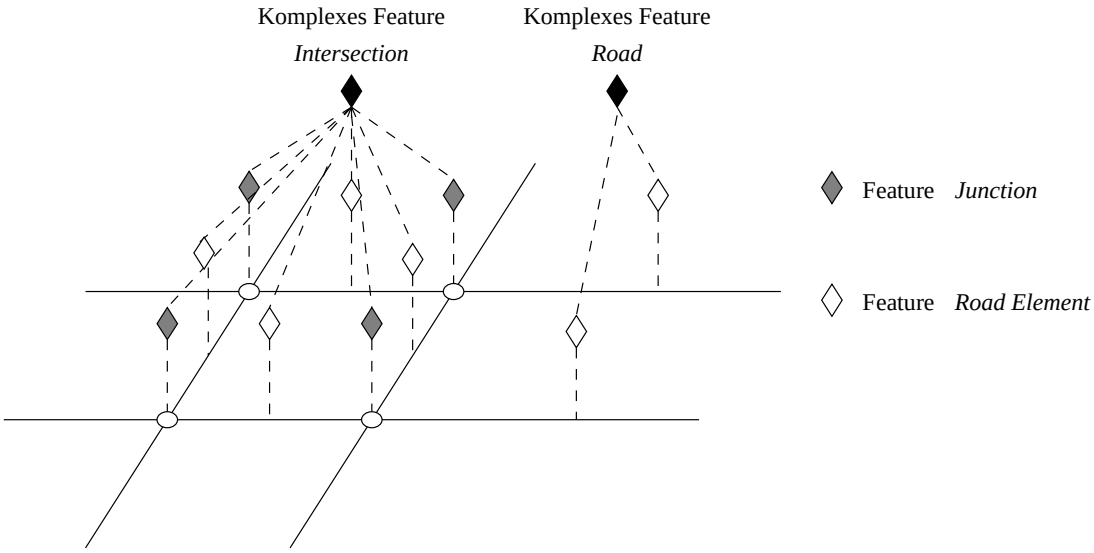


Abbildung 3.13: Bildung von komplexen Objekten

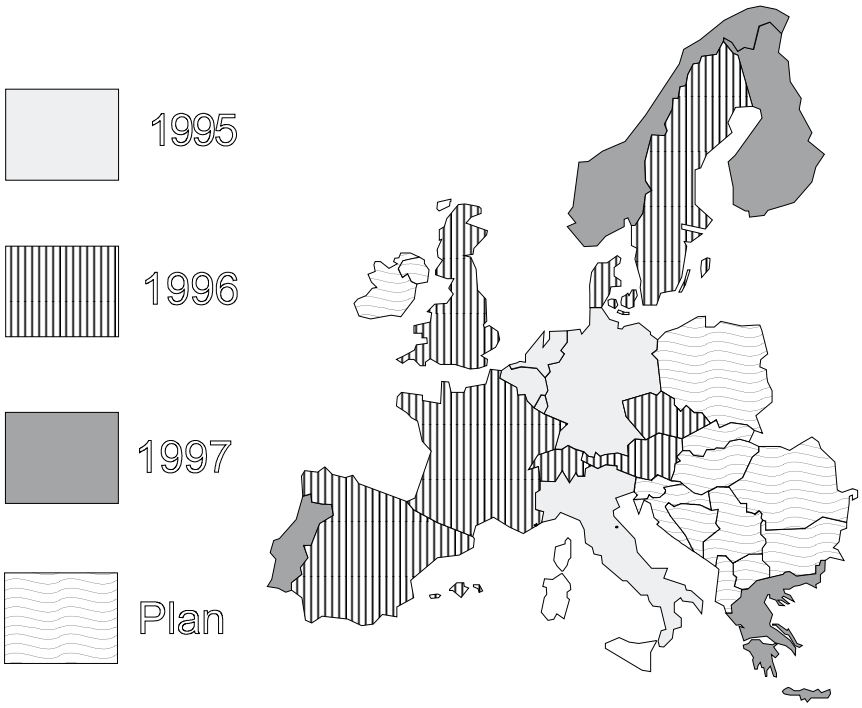


Abbildung 3.14: Erfassungsstand von GDF (aus [Bosch 95])

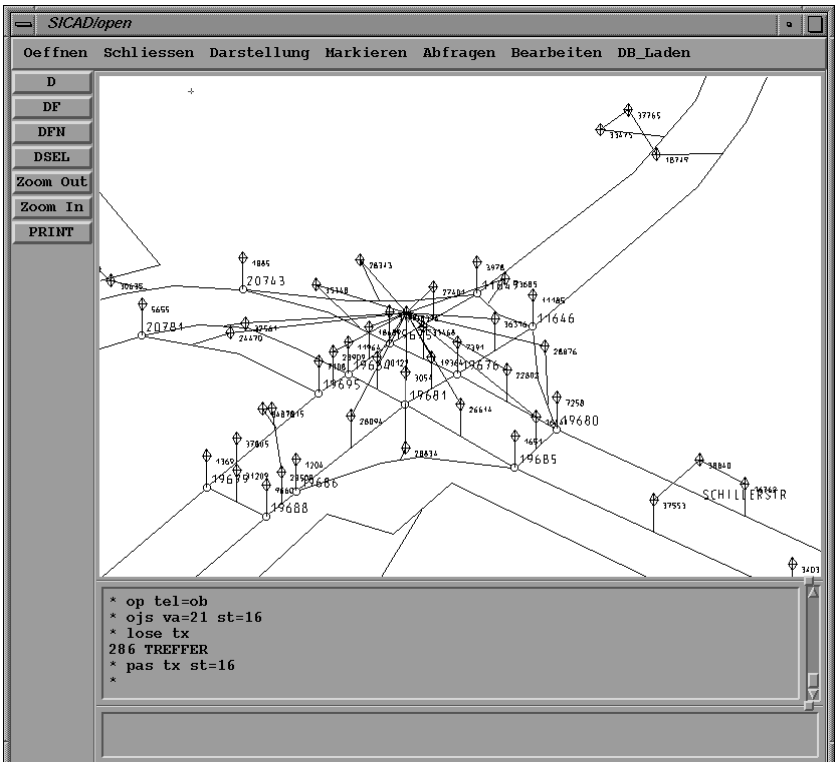


Abbildung 3.15: Demoprogramm für GDF-Daten

Kapitel 4

ATKIS

„Die Landesvermessungsbehörden der Länder der Bundesrepublik Deutschland haben die Aufgabe, die topographischen Erscheinungen der Landschaft und die Geländeformen, das Relief, aktuell zu erfassen und in den Topographischen Landeskartenwerken unterschiedlicher Maßstäbe und Informationsdichte nachzuweisen und darzustellen“ [Kophstahl 1988]. Durch fortschreitende Technologien werden diese Daten von Seiten der Benutzer immer mehr in digitaler Form gefordert. Solange diese Nachfrage nicht von den Landesvermessungsbehörden befriedigt werden kann, werden in vielen Unternehmen die Daten selbst digitalisiert. Hierdurch entsteht eine Mehrfacherfassung der Daten, die volkswirtschaftlich nicht sinnvoll ist. Daher ist es die Aufgabe der Landesvermessungsbehörden, einen bundeseinheitlichen digitalen Datenbestand zu erfassen, zu verwalten, und an Dritte weiterzugeben.

Aus diesen Gründen beschloß die Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV) im Oktober 1986 den Aufbau des Amtlichen Topographischen Kartographischen Informationssystem (ATKIS) [AdV 1988]. ATKIS ist als Basisinformationssystem zu verstehen, d.h. Anwender von raumbezogenen Daten sollen ATKIS-Daten als Basisdaten für ihre Anwendungen nutzen und falls erforderlich mit eigenen Daten anreichern. Inzwischen existieren bereits viele Anwendungen, die sowohl ausschließlich ATKIS-Daten als auch ATKIS als Basisdatensatz verwenden (siehe z.B. [Kophstahl 1995]).

ATKIS erfaßt im Modell DLM 25 Straßenverkehrsdaten mit einer Genauigkeit von ± 3 Meter. Diese Genauigkeit ist für die Fahrzeugnavigation ausreichend [Salgé und Brüggemann 1992]. Jedoch werden viele Informationen über das Straßennetz, wie z.B. Abbiegebeziehungen oder Einbahnstraßen, nicht in den Datenbestand aufgenommen. In diesem Zusammenhang stellt sich als wichtige Frage, inwieweit ATKIS-Daten als Basisdaten für Anwendungen aus der Fahrzeugnavigation überhaupt genutzt werden können.

In diesem Kapitel soll die Konzeption von ATKIS untersucht werden. Nach einer Darstellung der Zielsetzung von ATKIS wird eine Übersicht über die unterschiedlichen Erfassungsmodelle gegeben. In dieser Arbeit ist insbesondere das Digitale Landschaftsmodell (DLM) von Interesse und wird daher eingehend untersucht. Weiter existieren in ATKIS das Digitale Kartographische Modell (DKM) und das Digitale Geländemodell (DGM), auf die kurz eingegangen wird. Eine Übersicht über den Inhalt des Objektartenkataloges soll einen Eindruck über den Umfang des ATKIS-Projektes geben. Von der AdV wurde die Einheitliche Datenbankschnittstelle (EDBS) als Format für den Datenaustausch von ATKIS-Daten definiert [Ament 1993]. Dieses Austauschformat spielt bei der Bereitstellung von ATKIS-Daten eine zentrale Rolle und muß zusammen mit den ATKIS-Daten betrachtet werden. Abschließend wird der Erfassungsstand von ATKIS dargestellt.

4.1 Zielsetzung

„Interdisziplinäre Untersuchungen wie z.B. Umweltverträglichkeitsprüfungen, die auf der Datenlieferung vieler Dienststellen angewiesen sind, sind häufig uneffektiv und wenig aussagekräftig, weil die gelieferten Basis- und Fachinformationen nicht immer integriert und zu einer fachübergreifenden Analyse und Wertung zusammengeführt werden können“ [Kophstahl 1994]. Hier zeigt sich eines der Hauptziele von ATKIS als raumbezogenes Basisinformationssystem. Unterschiedlichste Fachinformationssysteme, welche auf den ATKIS-Daten basieren, sind geometrisch untereinander kompatibel, d.h., daß bei Verschneidungen keine Genauigkeitsverluste entstehen. Die Fachinformationssysteme weisen durch die Bereitstellung der Daten durch einen einzelnen Anbieter einen einheitlichen räumlichen Bezug auf, was den Datenfluß zwischen verschiedenen Anwendern erleichtert.

ATKIS hat daher eine integrierende Funktion für Anwendungen von raumbezogenen Daten. Dies ist insbesondere daher wichtig, da die Datenerfassung und -fortführung der kostenintensivste Aspekt eines Geo-Informationssystems ist [Bill und Fritsch 1991]. In der ATKIS-Gesamtdokumentation [AdV 1988] werden folgende Ziele definiert:

- Den Benutzern soll eine einfache und zuverlässige Beschaffung digitaler topographischer Daten ermöglicht werden, mit denen sie ihre Fachdaten verknüpfen können.

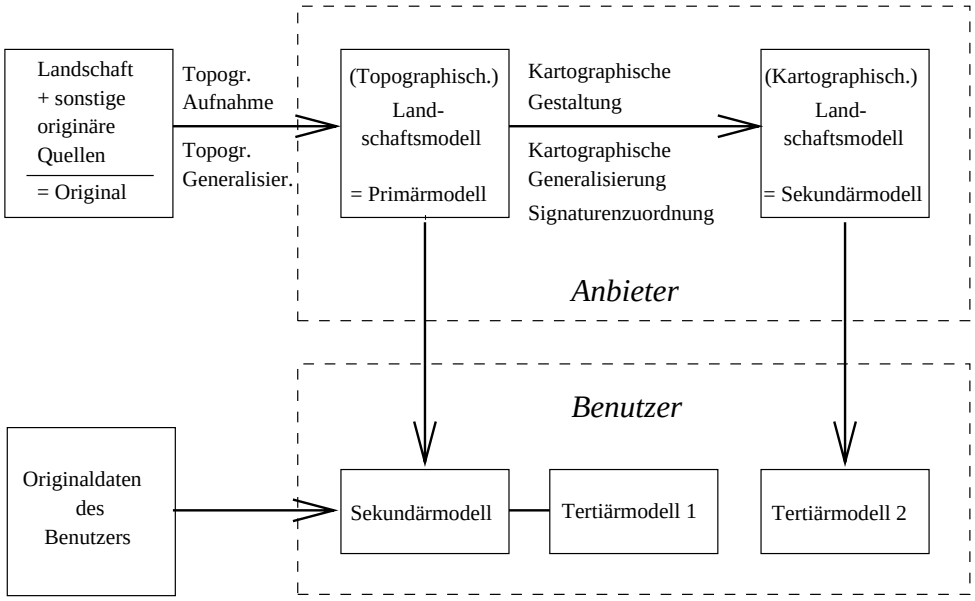


Abbildung 4.1: Kartographische Modelltheorie (vgl. [Hake 1982])

- Es soll sichergestellt werden, daß authentische und aktuelle topographische Informationen eingesetzt werden.
- Es soll die Wirtschaftlichkeit der Informationsgewinnung, -verarbeitung und -ausgabe gesteigert werden.

Durch die Erfassung von ATKIS-Daten durch die Landesvermessungsämter wird ein bundeseinheitlicher Datenbestand interessenneutral verwaltet. Die Erfassung und Abgabe von ATKIS-Daten wird als staatliche Dienstleistung verstanden [AdV 1988].

4.2 Übersicht

Um digitale raumbezogene Daten erfassen und speichern zu können, muß ein abstraktes Modell des Originals erstellt werden. Mit Hilfe dieses Modells werden im Original Objekte identifiziert, strukturiert und erfaßt. Das Original wird hierbei durch die Landschaft und sonstige originäre Quellen (z.B. Liegenschaftsbuch) gebildet. Die in der modernen Kartographie vertretene Modelltheorie [Hake 1982] beschreibt den Datenfluß bei der Erfassung und Weiterverarbeitung von raumbezogenen Daten (siehe Abbildung 4.1). Um ein Landschaftsmodell (Primärmodell) aufbauen zu können, wird die Landschaft (eventuell unter Hinzunahme weiterer originärer Quellen) in Objekte strukturiert. Die Entscheidung, welche Landschaftselemente zu welchen Objekten zusammengefaßt werden, hängt vom Anwendungszweck der Daten und vom verwendeten Maßstab ab. Bereits bei der topographischen Aufnahme erfolgt eine Generalisierung, da die Objektgrenzen im allgemeinen unscharf sind.

Im Primärmodell stehen alle digitalisierten Daten originär zur Verfügung und sind nicht durch kartographische Signaturen verschlüsselt. Sie können daher mit weiteren Anwenderdaten zu einem Sekundärmodell verknüpft und weiterverarbeitet werden. In einem nächsten Schritt werden den Objekten kartographische Signaturen zugeordnet. Hierbei müssen Generalisierungs- und Verdrängungsprozesse durchgeführt werden. Das hier entstandene Sekundärmodell kann in Form von digitalen und analogen Karten weitergegeben werden. Durch das Dekodieren der kartographischen Darstellungen stellt der Benutzer eine Analogie zum Original her und bildet dadurch das Tertiärmodell.

ATKIS setzt die kartographische Modelltheorie in die Praxis um. Abbildung 4.2 zeigt die Konzeption von ATKIS. Die Entscheidung, welche Landschaftselemente zu Objekten zusammengefaßt werden, wird im Objektartenkatalog festgelegt. Das durch die Objektbildung entstandene digitalisierte Primärmodell wird in ATKIS Digitales Landschaftsmodell (DLM) genannt. *„Wegen der noch nicht realisierten und der scheinbar auch nicht realisierbaren automatischen Objektgeneralisierung sind im ATKIS-Konzept mehrere DLM mit unterschiedlichen Maßstäben realisiert“* [Gran 1988].

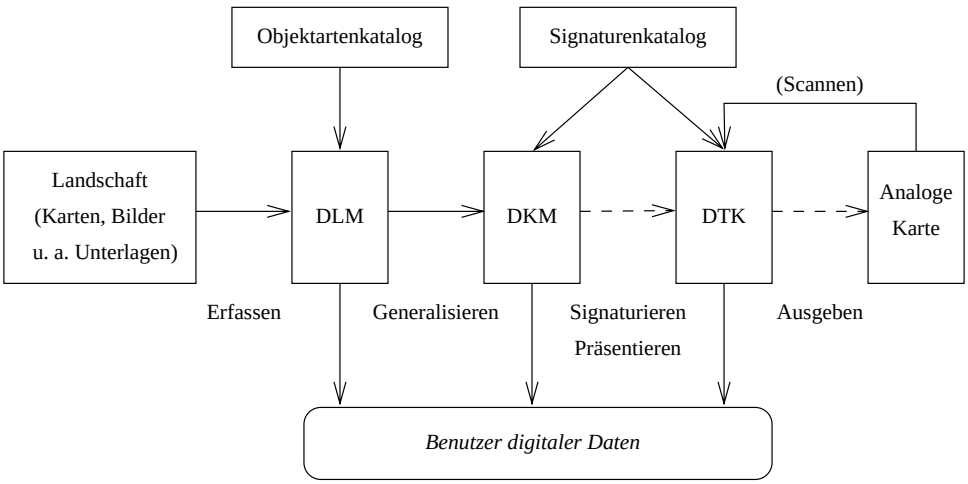


Abbildung 4.2: Konzeption von ATKIS

In der jetzigen Ausbaustufe werden drei unterschiedliche Landschaftsmodelle erfaßt. Das großmaßstäblichste und damit auch detaillierteste Landschaftsmodell ist das DLM 25. Es orientiert sich inhaltlich an der Topographischen Karte 1:25.000 (TK 25), mit einer Lagegenauigkeit von ± 3 m für die Objekte des Verkehrs- und Gewässernetzes, im übrigen ± 10 m. Das DLM 25 wird von den Landesvermessungsbehörden erfaßt. Der Einsatzbereich des DLM 25 ist grundsätzlich nicht an einen Kartenmaßstab gebunden. Das Anwendungsspektrum des DLM 25 als Basis für andere raumbezogene Fachinformationssysteme umfaßt näherungsweise die Verwendungsbereiche der Topographischen Landeskarten der Maßstäbe 1:5000 bis 1:100.000 [Kophstahl 1988]. Als weitere Erfassungsstufen werden das DLM 200 und DLM 1000 vom Institut für Angewandte Geodäsie (IfAG) in Frankfurt bearbeitet. Das DLM 200 entspricht inhaltlich der TÜK 200 mit einer Genauigkeit von ± 30 m, während das DLM 1000 mit der internationalen Weltkarte (IWK) mit einem Maßstab von 1:1.000.000 verglichen werden kann.

Ausgehend vom Landschaftsmodell wird mit Hilfe des Signaturenkataloges das Digitale Kartographische Modell (DKM) erstellt. In der ursprünglichen Konzeption [AdV 1988] war das DKM als digitales Endprodukt vorgesehen. Jedoch wurde bisher keine brauchbare Möglichkeit gefunden, das DKM automatisch so aus dem DLM abzuleiten, daß eine Karte in der Qualität der bisherigen analogen Landkartenwerke entsteht. Es stehen zwar von verschiedenen Software-Anbietern bereits Teillösungen zur Verfügung, um kartenähnliche Ausgaben zu erzeugen, jedoch sind diese Lösungen von der topographischen Karte noch weit entfernt und nicht standardisiert. Es handelt sich hierbei um Editierprogramme, welche durch Signaturentabellen gesteuert werden, und ungeneralisierte kartenähnliche Ausgaben erzeugen. Daher wurde 1995 beschlossen, als Endprodukt die Digitale Topographische Karte (DTK) zu standardisieren und den Produktionsweg, wie diese über das DKM zu erreichen ist, den Software-Entwicklern zu überlassen. Das Endprodukt DTK kann auf dem vorgesehenen Produktionsweg zur Zeit noch nicht realisiert werden. Da jedoch jetzt schon eine große Nachfrage nach einer DTK besteht, wird diese Nachfrage durch Scannen der bisherigen analogen Kartenwerke befriedigt.

4.3 Datenmodell

„Das ATKIS-Datenmodell ist die Grundlage für die Definition der topographisch-kartographischen Wirklichkeit aus datentechnischer Sicht, für die systemunabhängige Realisierung des ATKIS-Standards und für den Datenaustausch“ [Brüggemann 1990]. Im folgenden sollen die Aspekte der unterschiedlichen Ausprägungen des ATKIS-Datenmodells sowie der Inhalt des Objektartenkataloges und das Datenaustauschformat EDBS diskutiert werden. Der Schwerpunkt liegt dabei auf dem Digitalen Landschaftsmodell. Die Darstellung der Datenmodelle erfolgt mit Hilfe von NIAM-Diagrammen, welche in Kapitel 3.3 eingeführt wurden.

4.3.1 Digitales Landschaftsmodell

Abbildung 4.3 zeigt das konzeptionelle Datenmodell von ATKIS. Objekte der Landschaft werden durch ATKIS-Objekte und -Objektteile erfaßt. Objekte werden hierarchisch in Objektarten, Objektgruppen und Objektbe-

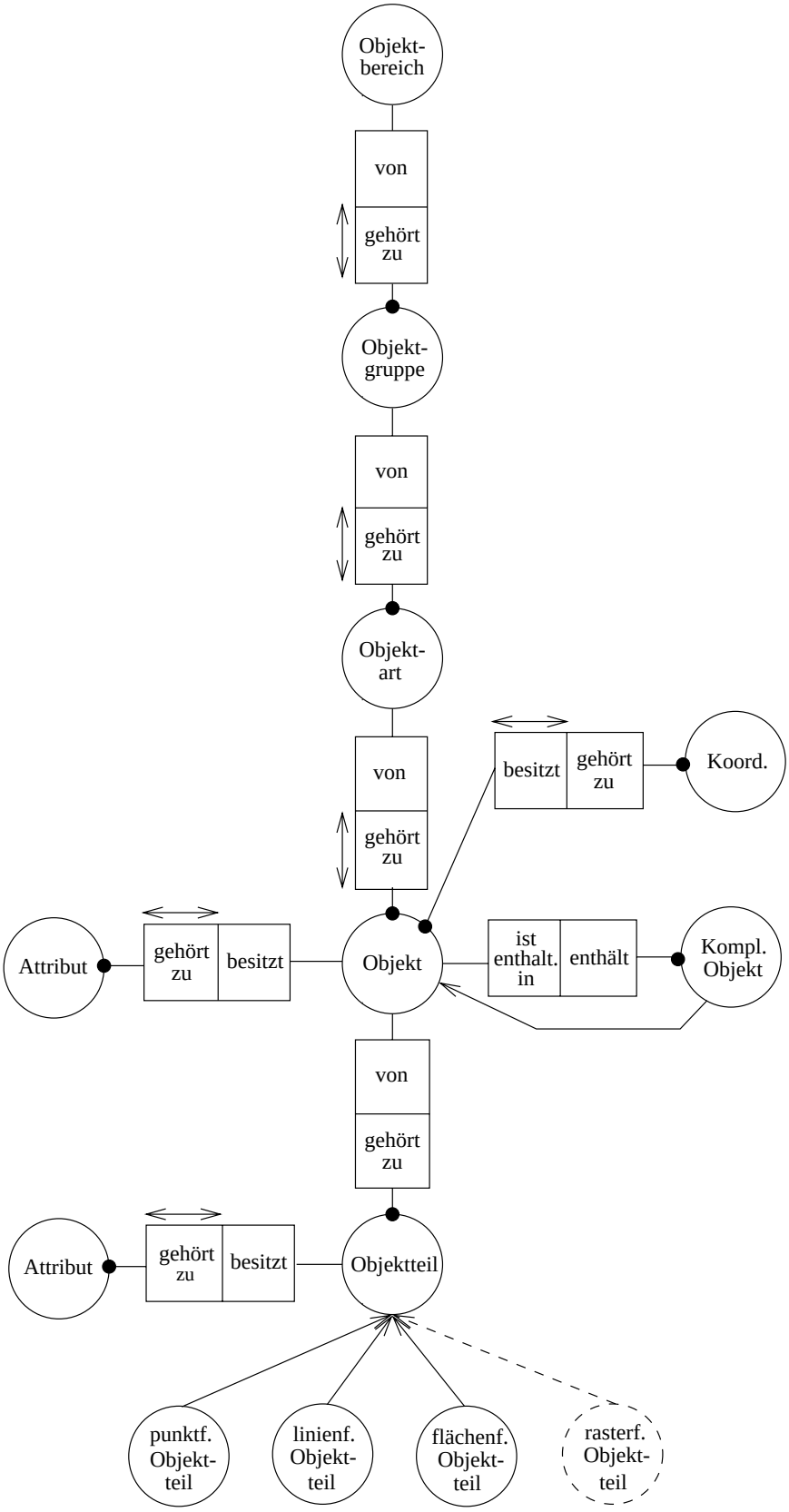


Abbildung 4.3: Konzeptionelles Datenmodell von ATKIS

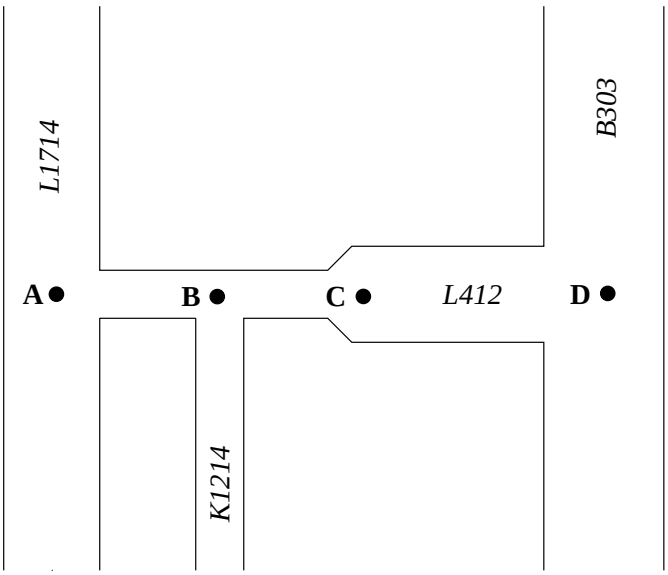


Abbildung 4.4: Bildung von Objekten und Objektteilen in ATKIS

reiche zusammengefaßt. Die Objektart ist die zusammenfassende Bezeichnung für gleichartige topographische Objekte zum Zwecke der Katalogisierung [AdV 1988]. Die hierarchische Gliederung entspricht einer Objektklassenbildung, jedoch findet zwischen den verschiedenen Objektklassen keine Vererbung statt. Objekte können zu komplexen Objekten aggregiert werden und beliebig viele Attribute besitzen. Komplexe Objekte dürfen als Attribute nur Namensattribute verwenden. Jedes Objekt besitzt genau eine Objektkoordinate. Die Objektkoordinate gibt an, an welcher Position das Symbol für das Objekt dargestellt werden soll und ermöglicht so die Selektion eines Objektes über ein Symbol am Bildschirm.

Der ATKIS-Objektartenkatalog ist attributorientiert, d.h. die Landschaft wird durch Objektarten grob und mit Hilfe von Attributen fein strukturiert. Objekte werden nach sachlogischen Gesichtspunkten gebildet. So werden am Beispiel der Abbildung 4.4 vier Objekte der Objektart *Straße* erfaßt, da vier Straßen mit unterschiedlicher Bezeichnung vorhanden sind (*L1714, K1214, L412* und *B303*). Die Objekte werden mit den Attributen Straßenname, Nutzung, usw. belegt. Das Attribut Fahrbahnbreite im Objekt *L412* kann nicht durchgängig mit dem gleichen Wert belegt werden. Daher ist das Objekt in mehrere Objektteile aufzuteilen. Die Objektteile sind in ATKIS die eigentlichen Träger der geometrischen Information. Daher besteht ein ATKIS-Objekt immer aus mindestens einem Objektteil, sofern es kein komplexes Objekt ist. Attribute, welche für das gesamte Objekt gelten, werden beim Objekt gespeichert, und Attribute, die für Objektteile unterschiedliche Werte besitzen, werden bei den Objektteilen gespeichert.

Ein Objektteil ist ein konkreter, geometrisch begrenzter und durch einheitliche Attribute und Relationen gekennzeichneter Gegenstand der Landschaft als Teil eines Objektes [AdV 1988]. Beim Wechsel eines Attributes oder einer topologischen Relation entsteht also immer ein neues Objektteil. Im obigen Beispiel werden 3 Objektteile für das Objekt *L412* benötigt (AB, BC, CD). Die Gründe hierfür sind die abgehende Straße im Punkt B (Änderung der topologischen Beziehung) sowie die Änderung der Straßenbreite im Punkt C. Objektteile können entweder punkt-, linien- oder flächenförmig sein[1]. Die Objektteile werden durch sogenannte Vektorelemente dargestellt. Ein Vektorelement beschreibt dabei entweder einen Punkt oder eine Linie. Für Linien können Interpolationsparameter und Zwischenpunkte angegeben werden. Zu jedem Punkt kann neben der Lage auch die Höhe abgespeichert werden, was bei der derzeitigen Erfassung von ATKIS-Daten jedoch nicht realisiert ist. Zwischen den Objektteilen werden topologische Relationen abgespeichert. Als einzige weitere Relation stehen Unter- bzw. Überführungsrelationen zur Verfügung. Semantische Relationen zwischen Objekten oder Objektteilen werden nicht erfaßt.

[1] Es ist auch möglich Objektteile mit Hilfe von Rastern zu bilden, was zur Zeit jedoch nicht genutzt wird.

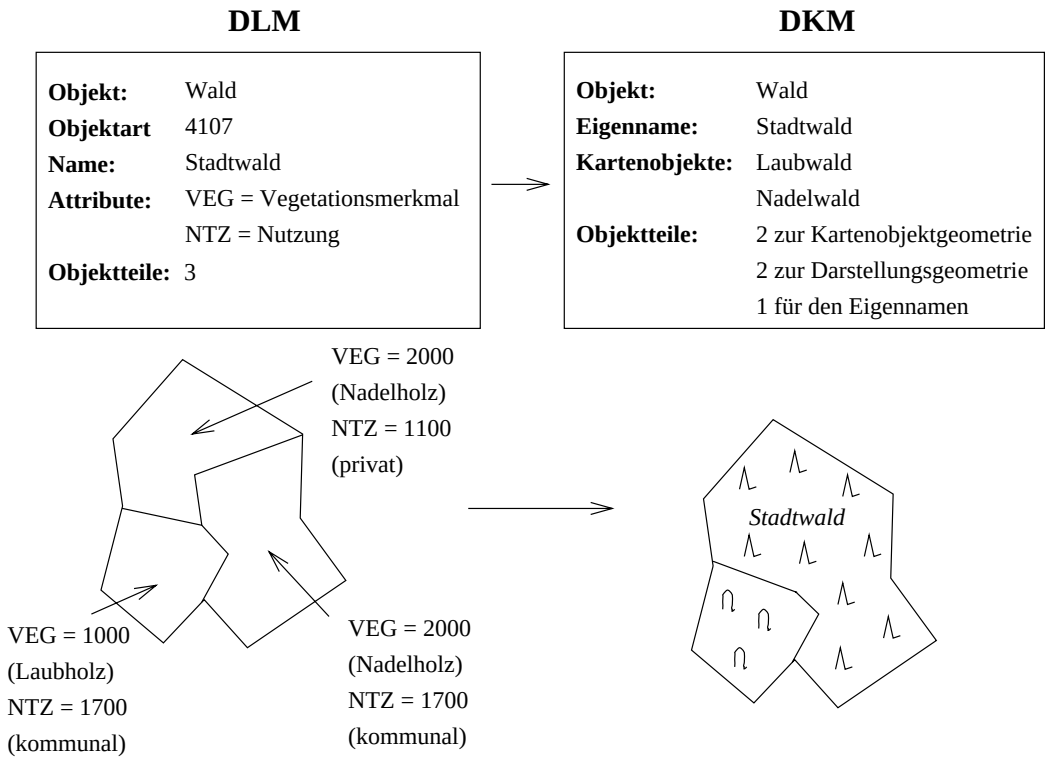


Abbildung 4.5: Ableitung des DKM aus dem DLM (vgl. [Vickus 1991])

4.3.2 Digitales Kartographisches Modell

Je größer der Maßstabsunterschied zwischen DLM und DKM ist, umso komplexer werden die Generalisierungsvorschriften. Daher hat man beschlossen, für jedes DLM ein eigenes DKM abzuleiten. Zwischen den Objekten in DLM und DKM besteht eine 1:1 Beziehung [Harbeck 1994]. Ein DKM-Objekt besteht aus DKM-Objektteilen, bei denen unterschieden werden kann zwischen Objektteilen zur Definition eines Objektes (Kartenobjektgeometrie), Objektteilen zur graphischen Ausgestaltung eines Objektes (Darstellungsgeometrie) und textförmige Objektteile für Eigennamen und Schriftzusätze. Das Datenmodell für DKM-Objekte entspricht dem Datenmodell des Digitalen Landschaftsmodells. Der Unterschied liegt jedoch in der Objektstrukturierung. Während die Objektstrukturierung im DLM durch die Objekteigenschaften, also den Attributen der Objekte, geprägt ist, ist im Gegensatz dazu die Objektstrukturierung des DKM grafikbezogen [Vickus 1991]. Abbildung 4.5 zeigt ein Beispiel zur Ableitung des DKM aus dem DLM. Das Objekt, welches dargestellt werden soll ist ein Wald, der aufgrund von drei unterschiedlichen Attributpaaren (Nadelholz/privat, Nadelholz/kommunal und Laubholz/kommunal) in drei Objektteile aufgeteilt ist.

Das Attribut Nutzung ist jedoch für die DKM-Bildung nicht relevant. Das DLM-Objekt *Wald* wird auf ein DKM-Objekt *Wald* abgebildet, welches aus zwei Kartenobjekten (*Laubwald* und *Nadelwald*) besteht. Die zwei DLM-Objektteile *Nadelholz* mit unterschiedlicher Nutzung werden zu einem Kartenobjekt zusammengefaßt. Weiter werden zwei Objektteile zur Darstellungsgeometrie, welche die Waldsignaturen repräsentieren, und ein Objektteil zur Darstellung des Namens gebildet. Hierbei handelt es sich jedoch nur um ein Roh-DKM. Weitere kartographische Bearbeitungen, wie Generalisierung oder endgültige Schriftplazierung müssen zur Zeit noch interaktiv erfolgen.

4.3.3 Digitales Geländemodell

Einer der Objektbereiche, welcher in der Endfassung von ATKIS geplant ist, ist das *Relief*. Hierbei handelt es sich um eine 2,5 dimensionale Beschreibung der Erdoberfläche. Dieser Objektbereich wird zur Zeit nicht erfaßt, sondern nur in Form von eigenständigen Digitalen Geländemodellen (DGM) geführt [Barwinski 1994]. Es stehen DGMs mit einer Rasterweite von 10 Meter bis 50 Meter bei den Landesvermessungsämtern zur Verfügung. In der nächsten Ausbaustufe wird ein DGM zur Beschreibung des Reliefs mit einer Rasterweite von 12,5 m und

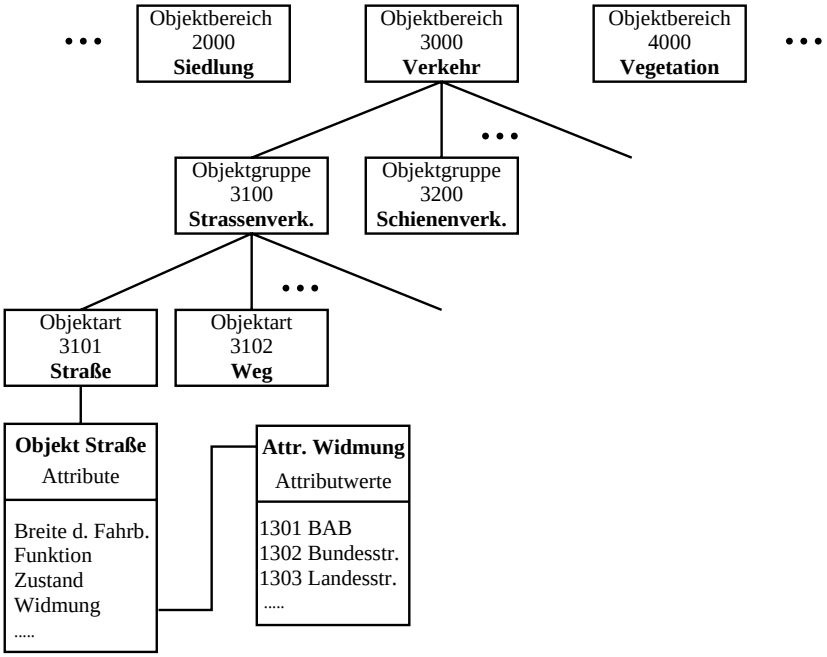


Abbildung 4.6: Ausschnitt aus dem ATKIS Objektartenkatalog

einer Höhengenauigkeit von ± 0,5 m erfaßt. Es wird somit als ein eigener Objektbereich integraler Bestandteil des ATKIS-DLM 25. Mit Hilfe dieser Geländemodelle besteht die Möglichkeit, die Koordinaten anderer Objektbereiche dreidimensional zu erfassen [Grünreich 1995]. Eine Vorstufe hierzu kann durch eine Interpolation der Höhen gewonnen werden. Für eine vollständige Integration ist eine Homogenisierung von identischen linienförmigen Strukturen im DGM und DLM (z.B. Bruchkanten des DGM müssen mit Böschungskanten des DLM übereinstimmen) nötig [Grünreich 1995].

4.3.4 Inhalt Objektartenkatalog

„Der ATKIS-Objektartenkatalog (ATKIS-OK) hat zum Ziel, die Landschaft zu strukturieren, die topographischen Erscheinungsformen und Sachverhalte der Landschaft zu klassifizieren und den Inhalt der digitalen Landschaftsmodelle festzulegen“ [Gran 1988]. Im ATKIS-OK wurden in der ersten Fassung 168 Objektarten sowie 111 Attributtypen mit 664 möglichen Attributwerten definiert [Grünreich 1990]. Inzwischen wurde eine Überarbeitung durchgeführt, bei der durch die praktische Erfassung auftauchende Fragestellungen gelöst werden sollen [Scholl 1995]. Hierbei handelt es sich jedoch nur um einige wenige Änderungen gegenüber der ursprünglichen Fassung.

Die zu erfassenden Objekte sind in sieben Objektbereiche und neunzehn Objektgruppen gegliedert. Die Objektbereiche im einzelnen sind: Festpunkte, Siedlung, Verkehr, Vegetation, Gewässer, Relief und Gebiete. Der ATKIS-OK ist nach folgenden Ordnungsmerkmalen aufgebaut: Die Reihenfolge der Objektarten innerhalb einer Objektgruppe entspricht der Sichtweise, wie sie beim Annähern eines Beobachters an die Erdoberfläche entsteht. Zunächst sind großflächige Erscheinungsformen aufgeführt (z.B. Ortslage), dann erfolgt eine weitergehende Strukturierung (z.B. Wohnbauflächen, Industrie und Gewerbeflächen usw.), und schließlich sind einzelne Bestandteile benannt (z.B. Gebäudeblöcke) [Gran 1988]. Dies erleichtert die Möglichkeit, Informationen über die Landschaft strukturiert nach der Erscheinungsform zu entnehmen.

Abbildung 4.6 zeigt einen Ausschnitt aus dem ATKIS-Objektartenkatalog. Die oberste Ebene zeigt drei der sieben möglichen Objektbereiche. Für ein Objekt Straße wird aufgezeigt, wie es in der Hierarchie eingeordnet ist. Weiterhin sind einige der möglichen Attribute für ein Objekt Straße sowie ein Ausschnitt des Wertebereiches des Attributes Widmung angegeben.

Um ATKIS-Daten den Benutzern möglichst frühzeitig zur Verfügung stellen zu können, wurde die Erfassung in mehreren Realisierungsschritten geplant. Die erste Erfassungstufe DLM 25/1 umfaßt 59 Objektarten mit 18 Attributtypen sowie 42 möglichen Attributwerten [Grünreich 1990]. Diese Erfassungstufe wird demnächst

flächendeckend für Deutschland vorliegen (siehe Kapitel 4.3.5). Der Anteil der erfaßten Objekte in der zweiten Erfassungstufe DLM 25/2 wird 62% des gesamten ATKIS-Objektartenkataloges ausmachen [Scholl 1995].

4.3.5 Austauschformat

Der Anwender kann Datenauszüge erhalten, die nach seinen fachlichen Anforderungen selektierte Objekte enthalten. Die geometrische Begrenzung der Datenauszüge ist nicht an Kartenblattschnitte gebunden. Es ist auch eine landesweite oder geometrisch auf bestimmte Landesteile (z.B. Landkreis, Gemeinde) bezogene Auswahl möglich. Für den ATKIS-Datenaustausch hat die AdV die Einheitliche Datenbankschnittstelle (EDBS) auf der Basis der logischen Datenstruktur der ALK-Grundrißdatei festgelegt. Bei der EDBS handelt es sich um eine systemunabhängige und herstellerneutrale Schnittstelle. Sie ermöglicht den Datenaustausch von hierarchisch aufgebauten objektweise gespeicherten Daten. Hierzu stehen Befehle zur Verfügung, die es ermöglichen, eine Vielfalt von Übergabemöglichkeiten zu nutzen. Eine Gesamtdokumentation zum EDBS-Datenaustausch findet sich in [ALK 1986]. Für den ATKIS-Datenaustausch wird jedoch nur eine Teilmenge der möglichen Funktionalität benötigt. Daher wurde eine spezielle Dokumentation zum ALK/ATKIS-Datenaustausch [ALK 1993] von der AdV herausgegeben. In der Praxis wird nur die eingeschränkte Funktionalität des ALK/ATKIS-Datenaustausches verwendet.

Die EDBS kann nicht nur Daten für den Neueintrag liefern, sondern auch mit Hilfe von EDBS-Dateien Daten nach bestimmten Suchkriterien von der ALK/ATKIS-Datenbank anfordern und auch Daten fortzuführen [Ament 1993]. Die Fortführung von Daten, welche aus einem Basisinformationssystem stammen, ist für die Benutzer von großer Bedeutung und momentan in keinem anderen standardisierten Austauschformat definiert. Um die Daten mittels der EDBS übertragen zu können, werden sie auf die logische Datenstruktur der Grundrißdatei abgebildet. Die Grundrißdatei ist hierarchisch aufgebaut und ermöglicht die redundanzfreie Speicherung von Geometrie- und Objektdaten. Die hierarchische Struktur der Grundrißdatei läßt sich eindeutig linearisieren und ermöglicht auch eine eindeutige Rückabbildung der Daten in die Datenbank. Die linearisierten Daten werden in EDBS-Sätze eingebettet, die Informationen darüber enthalten, welche Daten übermittelt werden und wie die Daten verarbeitet werden sollen. Abbildung 4.7 zeigt die Struktur der Grundrißdatei.

Die EDBS als Medium zum Datenaustausch von ATKIS-Daten hat sich inzwischen in Deutschland etabliert [Harbeck 1995]. In der Praxis treten jedoch Probleme auf, die vor allem mit geringfügigen Unterschieden in der Definition der EDBS in den unterschiedlichen Bundesländern zusammenhängen[2]. Ein weiteres Problem entsteht, wenn Nutzer ihre Daten mit selbst erfaßten Daten anreichern. Da der Datenaustausch nur für das ATKIS-Modell realisiert ist, können erweiterte Datenmodelle nicht anderen Nutzern mit der EDBS verfügbar gemacht werden. Doch gerade die Erweiterung der Basisinformationen durch den Anwender ist ein wichtiges Konzept von ATKIS.

4.4 Erfassungsstand von ATKIS-Daten

Während der Aufbau des DLM 25/1 in den alten Bundesländern kurz vor der Fertigstellung steht, befinden sich die neuen Bundesländer noch bei der Ersterfassung. Die bereits erfaßten Daten sind bei den zuständigen Landesvermessungsämtern erhältlich. Eine vollständige Erfassung der Daten wird je nach Bundesland zwischen 1996 und 1997 erwartet. Anschließend soll mit der Erfassung der zweiten verdichteten Stufe DLM 25/2 begonnen werden, deren vollständige Erfassung mit dem Jahr 2002 geschätzt wird [Scholl 1995].

[2] Dieses Problem wird insbesondere dann deutlich, wenn der Nutzer Daten anfordert, welche über die gemeinsame Grenze von zwei Bundesländern gehen.

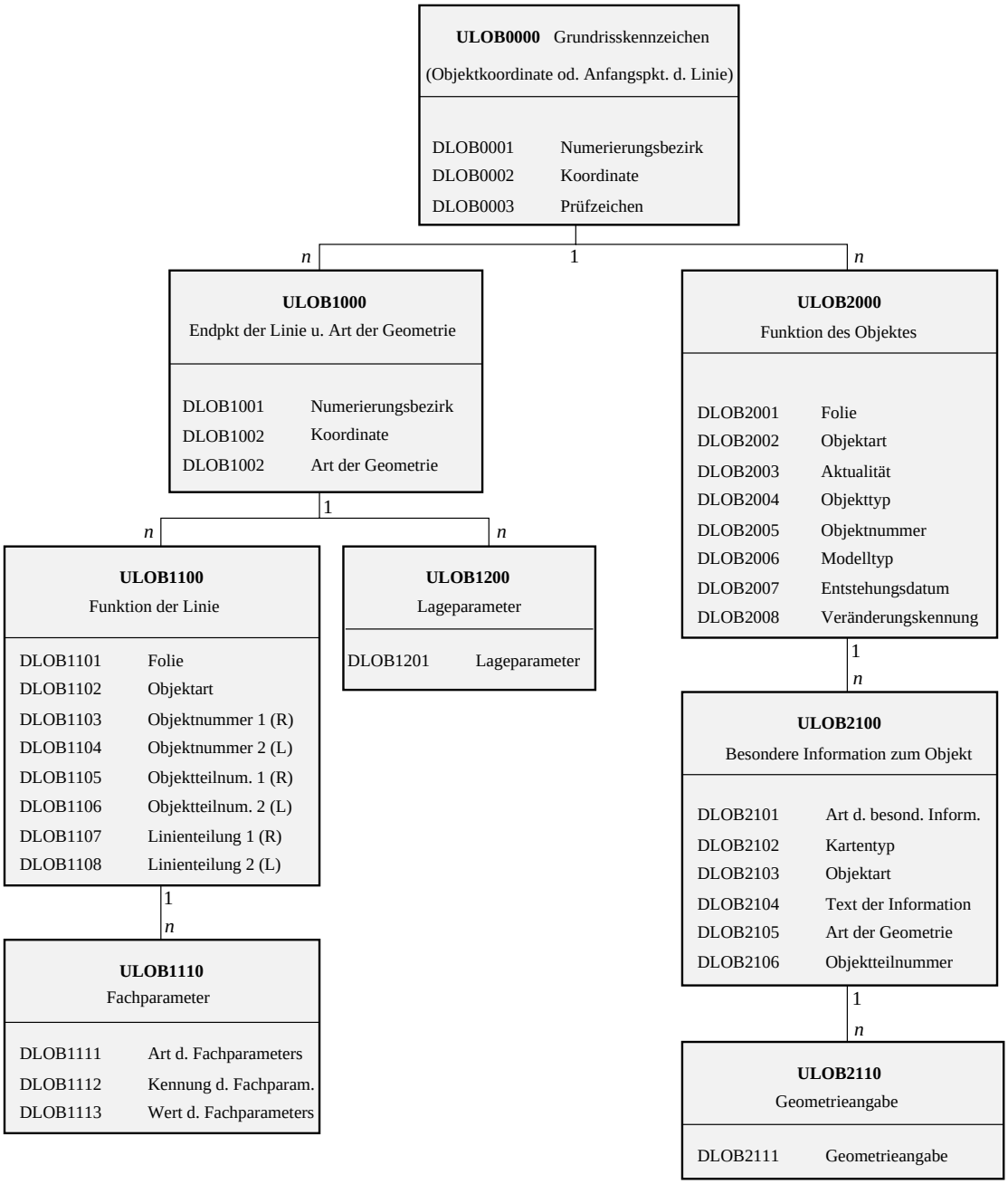


Abbildung 4.7: Logische Datenstruktur der Grundrißdatei

Kapitel 5

Gegenüberstellung GDF und ATKIS

In den letzten beiden Kapiteln wurden die Datenmodelle von GDF und ATKIS ausführlich dargestellt. In diesem Kapitel werden die Unterschiede zwischen den beiden Datenmodellen diskutiert. Hierzu reicht es nicht aus, nur die eigentlichen Datenmodelle zu untersuchen, sondern es müssen ebenso die Datenkataloge betrachtet werden. In den Datenkatalogen werden die Landschaftsobjekte definiert, die in der Datenwelt mit Hilfe des Datenmodells modelliert werden. Es werden semantische und geometrische Integritätsbedingungen definiert und die Objektbildung natürlichsprachlich dargestellt. Bei der geometrischen Qualität der Daten spielen die Erfassungsvorlagen eine große Rolle. Die Aktualität der Daten wird durch den Fortführungszyklus bestimmt.

Im folgenden werden diese Punkte in einzelnen Unterkapiteln untersucht. Wie bereits in Kapitel 3 dargestellt, werden GDF-Daten von zwei verschiedenen Institutionen unabhängig voneinander erfaßt. Die hier dargestellten Untersuchungen basieren auf GDF-Daten, welche von der Firma Bosch/Teleatlas erfaßt wurden. Dies hat jedoch nur Auswirkungen auf quantitative Aussagen über die Daten, wie z.B. die Anzahl der erfaßten Objektklassen. Bei den Datenmodellen gibt es keinen Unterschied zwischen den beiden Datenanbietern. Abschließend werden in diesem Kapitel anhand der in dieser Arbeit verwendeten Testdatensätze typische Unterschiede in der Modellierung von ATKIS- und GDF-Daten aufgezeigt.

5.1 Datenmodell

Das GDF-Datenmodell ist dem ATKIS-Datenmodell sehr ähnlich. Die meisten der Unterschiede entstehen aus der Tatsache, daß sich der Schwerpunkt bei ATKIS auf die Darstellung der Topographie bezieht, wogegen GDF für verkehrsrelevante Informationen steht. Der folgende Vergleich der Datenmodelle bezieht sich auf einen Vergleich des ATKIS-DLM 25 mit dem GDF-Datenmodell. In beiden Datenmodellen werden Objekte des Straßenverkehrs per Definition mit einer Genauigkeit von ± 3 Meter erfaßt. Eine Korrespondenz des Digitalen Kartographischen Modells (DKM) gibt es in GDF nicht.

5.1.1 Konzeptionelles Datenmodell

Die konzeptionellen Datenmodelle von GDF und ATKIS sind sich in ihrem Aufbau sehr ähnlich. Der Hauptunterschied der Modelle besteht darin, daß in GDF die Modellierung der Welt mit Objekten bzw. komplexen Objekten durchgeführt wird, wogegen in ATKIS eine weitere Unterscheidung zwischen Objekten und Objektteilen stattfindet. Bei einer Betrachtung der Daten kann gesehen werden, daß es keine 1 : 1 Entsprechung zwischen ATKIS-Objekten und GDF-Objekten bzw. ATKIS-Objektteilen und GDF-Objekten gibt (Beispiele siehe Kapitel 5.3). Die Bildung von komplexen Objekten ist in beiden Datenmodellen möglich. ATKIS-Objekte werden in einer Klassenhierarchie, bestehend aus Objektart, Objektgruppe und Objektbereich organisiert, wogegen im GDF-Datenmodell beliebig viele hierarchisch angeordnete Objektklassen und ein übergeordneter Themenbereich gebildet werden können. Dabei handelt es sich jedoch in beiden Modellen lediglich um eine statische Hierarchie, bei der keinerlei Vererbung zwischen den verschiedenen Stufen stattfindet [Fritsch und Anders 1996]. Durch die rekursive Definition der Objektklassen des GDF-Datenmodells ist es prinzipiell möglich, beliebige, nicht hierarchische, Klassennetzwerke zu bilden. Dies wird jedoch im Datenkatalog von GDF nicht genutzt. Als weiterer Unterschied kann gesehen werden, daß ATKIS-Objekte, im Gegensatz zu GDF-Objekten, eine Objektkoordinate besitzen. Diese Koordinate wird verwendet, um ein Symbol am Bildschirm darzustellen, welches das Objekt repräsentiert.

5.1.2 Attributkonzept

Das GDF-Attributkonzept ist dem ATKIS-Attributkonzept überlegen. Während es in GDF möglich ist, zeitabhängige, komplexe und segmentierte Attribute zu bilden, können in ATKIS nur einfache Attribute

verwendet werden. Durch die Verwendung der EDBS-Schnittstelle (siehe Kapitel 5.7) beim Datenaustausch kommt es sogar zu einer weiteren Einschränkung der Attributwerte auf eine maximale Länge von sieben Byte.

5.1.3 Relationenkonzept

Bei Relationen kann zwischen topologischen und semantischen Relationen unterschieden werden. Das ATKIS-Datenmodell existiert in zwei Varianten [AdV 1988]. Die Variante A verzichtet auf die Abspeicherung von topologischen Relationen, wogegen in der Variante B die topologischen Relationen explizit abgespeichert sind. Im GDF-Datenmodell werden topologische Relationen ebenfalls explizit gespeichert.

Als einzige semantische Relation sind im ATKIS-Datenmodell Überführungs- und hierarchische Relationen möglich. Hierarchische Relationen dienen zur Bildung von komplexen Objekten. Alle anderen semantischen Relationen zwischen topographischen Objekten sind nur implizit in der DLM-Geometrie enthalten. Im GDF-Datenmodell können semantische Relationen mit beliebig vielen Partnern gebildet werden. Weiter ist es möglich an Relationen Attribute zu hängen, um somit z.B. Relationen zu modellieren, welche nur in bestimmten Zeitintervallen gültig sind.

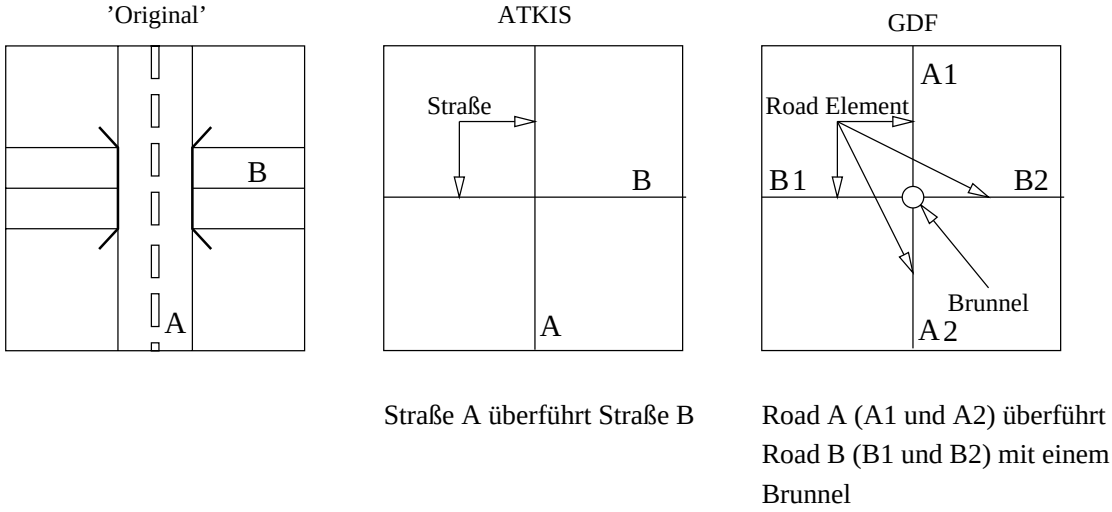


Abbildung 5.1: Modellierung von Überführungen in ATKIS und GDF

Über- bzw. Unterführungen werden im GDF-Datenmodell mit einer eigenen Featureklasse *Brunnel* abgebildet. Dies führt dazu, daß, im Gegensatz zu ATKIS, die Geometrie immer als planarer Graph erfaßt wird. Abbildung 5.1 zeigt die unterschiedliche Modellierung an einem Beispiel. In ATKIS wird die Überführung mit einer zweistelligen Relation dargestellt. Die GDF-Modellierung erfolgt mit einem punktförmigen Feature und einer dreistelligen Relation.

5.2 Datenkatalog

Neben den Features, die in der GDF-Gesamtdokumentation [Heres et al. 1991] definiert sind, existieren noch Erweiterungen für andere Anwendungen, wie z.B. für das Projekt STORM (Stuttgart Transport Operation by Regional Management) [Hiestermann 1992] oder eine Erweiterung für Flughäfen [Daimler 1994]. Die aktuelle Erfassung der Daten unterscheidet sich in einigen Teilen von den ursprünglichen Vorgaben des GDF-Kataloges. Im Anhang D ist eine Zusammenstellung der Feature Themes und Feature Classes, wie sie derzeit von Bosch/Teleatlas erfaßt werden [van Essen 1994], dargestellt. Das Straßennetzwerk wird in dem Feature Theme *Roads and Ferries* erfaßt. Die Zusammenfassung von Straßen und Fähren innerhalb des gleichen Feature Theme erfolgt aus Gründen der in GDF existierenden Weltsicht aus dem Blickwinkel des Transportes von Gütern und Personen.

Nummer	*Objektart*	*Bedeutung nach Objektartenkatalog*
3101	Straße	Befestigter, in erster Linie dem Kraftfahrzeugverkehr dienender Verkehrsweg
3102	Weg	Befestigter oder unbefestigter Geländestreifen, der zum Befahren und/oder Begehen vorgesehen ist
3103	Platz	Verbreiteter Straßenraum in Ortslagen, der als Platz bezeichnet wird oder ebene, befestigte oder unbefestige, unbebaute Fläche, die bestimmten Zwecken dient
3104	Straße (komplex)	Vollzählige Erfassung der für den öffentlichen Verkehr zugelassenen Straßen, soweit die Modellierung komplex erfolgen muß
3105	Straßenkörper	Funktional geprägte Fläche, die in erster Linie einen befestigten, dem Kraftfahrzeugverkehr dienenden Verkehrsweg aufzunehmen bestimmt ist
3106	Fahrbahn	Befestigter und in Fahrstreifen unterteilter Bereich innerhalb eines Straßenkörpers, der in erster Linie dem fließenden Verkehr dient

Tabelle 5.1: Definition der Objektarten des Objektbereiches Straßenverkehr in ATKIS

5.3 Objektbildung

Die Modellierung des Straßennetzwerkes in GDF erfolgt mit den Feature-Klassen *Road Elements, Junctions, Roads* und *Intersections*. *Road Elements* sind die kleinsten unabhängig voneinander zu erfassenden Einheiten. Sie müssen eine eindeutige Menge von Attributen besitzen und sind an beiden Enden mit einem Feature *Junction* verbunden. *Junctions* sind die Verbindung von zwei oder mehreren *Road Elements*. Mit Hilfe der (komplexen) Features *Roads* und *Intersections* können *Road Elements* und *Junctions* zu Einheiten aggregiert werden (Beispiel siehe Abbildung 3.13).

Die Unterscheidung, zu welcher Klasse eine Straße gehört, erfolgt in GDF mit Hilfe des Attributes *Functional Class (FC)*. In ATKIS dagegen erfolgt diese Unterscheidung durch die Bildung verschiedenener Objektklassen. Tabelle 5.1 zeigt die verschiedenen Objektklassen der Objektgruppe Straßenverkehr in ATKIS und Tabelle 5.2 gibt eine Aufschlüsselung der Bedeutung des Atributes *FC* in der GDF-Datenerfassung. Eine direkte Entsprechung zwischen den verschiedenen Objektarten in ATKIS und der Aufteilung nach dem Attribut *Functional Class* in GDF läßt sich nicht finden. Eine Erfassung des Straßenkörpers (Objektart 3105) findet in GDF jedoch nicht statt. Straßen mit getrennten Fahrspuren[1] werden in der Regel in ATKIS mit Hilfe zweier Fahrbahnen (Objektart 3106) und einem komplexen Objekt Straße (Objektart 3104) modelliert. Straßen mit nicht getrennten Fahrspuren werden normalerweise mit der Objektart Straße (3101) erfaßt. Eine Unterscheidung dieser Art findet in GDF nicht statt. Je nachdem, ob eine Straße getrennte Fahrspuren besitzt oder nicht, wird sie mit einem oder zwei *Road Elements* erfaßt.

Im folgenden wird an einigen Beispielen die unterschiedliche Modellierung von Straßen in GDF und ATKIS diskutiert. Abbildung 5.2 a) zeigt eine Straße, deren Straßenbreite sich verändert. Sowohl in ATKIS als auch in GDF wird an der Position, an der sich die Straßenbreite verändert, ein topologischer Knoten eingeführt (In GDF könnte diese Situation auch mit Hilfe eines segmentierten Attributes modelliert werden; siehe hierzu Kapitel 3.3.2). Unter der Voraussetzung, daß sich an dieser Position nicht auch der Straßenname ändert, werden in ATKIS die beiden Teile der Straße mit zwei Objektteilen erfaßt und als Bestandteil eines Objektes abgespeichert. Die Modellierung in GDF erfolgt in der gleichen Weise, jedoch mit dem Unterschied, daß Knoten im Straßennetzwerk als eigene Features des Typs *Junction* erfaßt werden.

In Abbildung 5.2 b) ist eine Straße mit zwei physikalisch getrennten Fahrspuren abgebildet. In ATKIS werden hierzu zwei Objekte für die Fahrbahnen (Objektart 3106) sowie ein Objekt, welches die Mittelachse des Straßenkörpers beschreibt (Objektart 3105), gebildet. Da zwischen einem Objekt und der Geometrie des Objektes nach dem ATKIS-Datenmodell immer mindestens ein Objektteil stehen muß, wird zu den drei Objekten je ein Objektteil gebildet. Die Fahrbahnen und der Straßenkörper bilden eine Einheit und werden daher zu einem

[1] Mit getrennten Fahrspuren sind hier in erster Linie physikalisch getrennte Fahrspuren oder Straßen mit durchgezogenen Mittellinien gemeint

Functional Class	*Bedeutung*
0	Alle Straßen, die als "Autobahn", "Mehrbahnige Autostraße" oder "Fernstraße" im 1 : 4.500.000 V.A.G Atlas klassifiziert sind
1	Alle Straßen, die als "Autobahn", "Vier- oder mehrspurige Straße" oder "Bundesstraße" im 1 : 200.000 V.A.G Atlas klassifiziert sind
2	Alle Straßen, die als "Hauptstraße" im 1 : 200.000 V.A.G Atlas klassifiziert sind
3	Alle Straßen, die als "Nebenstraße" in den Kartenblättern 1 : 50.000 der Landesvermessungsämter klassifiziert sind
4	Alle Straßen, die als "Befestigter Fahrweg" in den Kartenblättern 1 : 50.000 der Landesvermessungsämter klassifiziert sind und mit einer 1/2 mm breiten Linie dargestellt werden
5	Alle Straßen, die als "Befestigter Fahrweg" in den Kartenblättern 1 : 50.000 der Landesvermessungsämter klassifiziert sind und mit einer 1/4 mm breiten Linie dargestellt werden

Tabelle 5.2: Definition des Attributes *Functional Class* in GDF

komplexen Objekt Straße (Objektart 3104) aggregiert. Die Erfassung in GDF ist in diesem Fall einfacher. Die beiden Fahrbahnen entsprechen zwei Features *Road Element*, welche zu einem komplexen Feature *Road* gehören.

Das Beispiel in Abbildung 5.2 c) veranschaulicht, daß schon bei einfachen Kreuzungen sehr stark unterschiedliche Objektstrukturen in den beiden Datenmodellen entstehen.

5.4 Objektinterpretation

Die Datenkataloge enthalten die Definitionen über die zu erfassenden Objekte. Während es möglich ist, die Syntax eines Austauschformates oder eines Datenmodells mit einer formalen Sprache eindeutig zu definieren, werden zur Beschreibung der Erfassung der Objekte natürlichsprachliche Definitionen verwendet. Dies kann auch bei geschulten Operateuren zu unterschiedlichen Interpretation führen. Beispiele für natürlichsprachliche Beschreibungen aus dem ATKIS-Objektartenkatalog sind z.B. *... in erster Linie dem Kraftfahrzeugverkehr dienender Verkehrsweg ...* oder *... kleinere Böschungen ...* . Durch diese teilweise unscharfen Ausdrucksweisen können bei Grenzfällen unterschiedliche Interpretationen entstehen. Dies kann zu Inkonsistenzen zwischen ATKIS- und GDF-Daten führen, die sich nicht durch die Erfassungsregeln oder Datenmodelle erklären lassen. Von besonderer Bedeutung sind in diesem Zusammenhang ebenfalls die Erfassungsvorlagen, deren Qualität den Entscheidungsprozess, zu welcher Objektart ein Objekt gehört bzw. ob es überhaupt erfaßt werden muß, wesentlich beeinflußt.

5.5 Erfassungsvorlagen

Sowohl ATKIS als auch GDF erfüllen nationale (GDF sogar internationale) Aufgaben. Daher ist es von großer Wichtigkeit, daß die Daten homogen mit gleichem Inhalt und gleicher Qualität erfaßt werden [Claussen 1995]. Während GDF-Daten von einem Anbieter flächendeckend für Deutschland erfaßt werden, erfolgt die Erfassung der ATKIS-Daten aufgrund der föderalen Struktur durch die jeweiligen Landesvermessungsämter der Bundesländer. Die zentrale Erfassung an einer Stelle hat den Vorteil, daß dadurch die Homogenität der Daten gesichert werden kann. Für die geometrische Qualität der Daten sind u.a. die Erfassungsquellen ausschlaggebend. Die Erfassung der GDF-Daten erfolgt typischerweise aus großmaßstäbigen Karten 1:5.000 oder 1:10.000 [Claussen 1995]. In Überlandgebieten erfolgt die Digitalisierung aus Karten des Maßstabes 1:25.000 [Claussen 1995]. Die Erfassung von ATKIS-Daten erfolgt je nach Bundesland aus den unterschiedlichsten Quellen und auch mit unterschiedlicher Erfassungssoftware (siehe z.B. [Barwinski 1994]). Da, zumindest in Stadtbereichen, GDF- und ATKIS-Daten aus großmaßstäblichen (oftmals auch identischen) Quellen erfaßt werden, spielen Generalisierungen bei einem Vergleich der Datensätze kaum keine Rolle. In Tabelle 5.3 sind die verschiedenen Erfassungsvorlagen der einzelnen Bundesländer dargestellt (entnommen aus [Harbeck 1995]). In den verschiedenen Bundesländern ist aufgrund der unterschiedlichen Erfassungsquellen und -software mit Inhomogenitäten in den ATKIS-Datenbeständen zu rechnen.

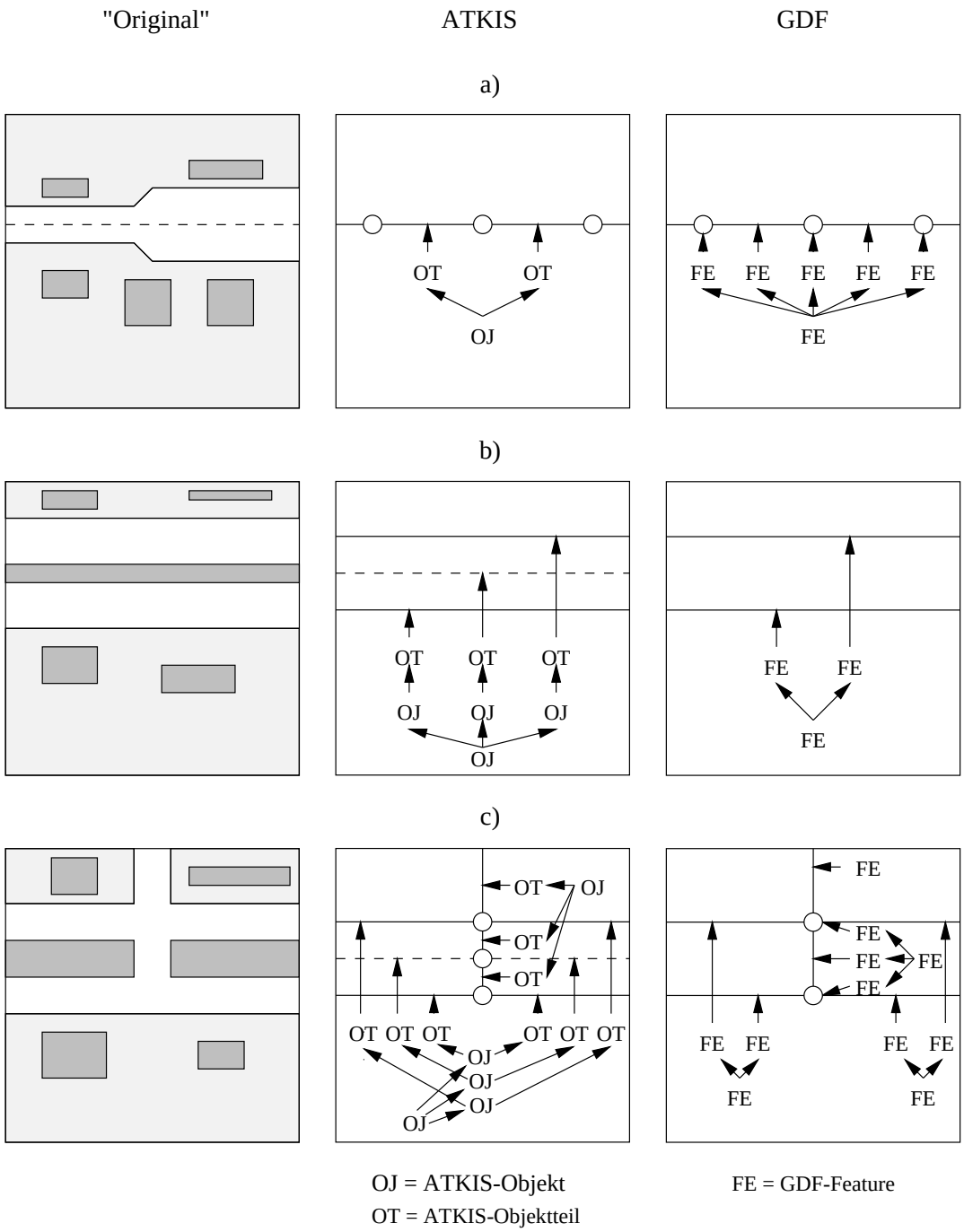


Abbildung 5.2: Modellierung von Straßen in ATKIS und GDF

Bundesland	*Datenquellen*
Baden-Württemberg	TK 25, Ortophotos
Bayern	TK 25
Berlin	Stadtkarte 5
Brandenburg	TK 10
Bremen	DGK 5
Hamburg	DGK 5, Luftbilder
Hessen	TK 5, Orthophotos
Mecklenburg-Vorpommern	TK 10
Niedersachsen	DGK 5
Nordrhein-Westfalen	DGK 5, Luftbilder
Rheinland-Pfalz	DGK 5, Orthophotos
Saarland	DGK 5, ALK
Sachsen	TK 10
Sachsen-Anhalt	TK 10
Schleswig-Holstein	DGK 5
Thüringen	TK 10

Tabelle 5.3: Erfassungsquellen von ATKIS (entnommen aus [Harbeck 1995])

	FC = 0	FC = 1	FC = 2	FC = 3	FC = 4	FC = 5
Prohibited Turn	1	1	1	1	1	2
Special Traffic Restriction	2	2	2	2	2	2
Direction of Traffic Flow	1	1	1	2	2	2
Road Class	1	1	1	2	2	2
Form of Way	1	1	1	2	2	2

Tabelle 5.4: Fortführungszyklus von Attributen des Feature Klasse *Roads* (in Jahre)

5.6 Fortführung/Aktualität

Neben der Definition des Datenmodells und dem Inhalt des Objektartenkataloges spielt auch die Frage der Aktualität beim Vergleich von raumbezogenen Daten eine große Rolle. Bei ATKIS-Daten muß sichergestellt sein, daß alle 5 Jahre eine generelle Durchsicht erfolgt sowie bei wesentlichen Objektarten, wie z.B. Verkehrsobjekten, eine Fortführung im halbjährlichen bis jährlichen Zyklus durchgeführt wird [Scholl 1995]. Noch weiter geht die Forderung von [Barwinski 1994]: "ATKIS lebt von seiner Aktualität und verlangt daher, nicht periodisch fortgeführt zu werden, sondern kontinuierlich". Die Realisierung einer Fortführung wird in den einzelnen Bundesländern derzeit angedacht bzw. schon getestet [Scholl 1995]. Da ATKIS-Daten seit 1989 erfaßt werden und erst jetzt mit der Fortführung begonnen wird, sind die derzeitigen Daten bis zu 5 Jahre alt.

In der *Data Content Specification* der GDF-Dokumentation [Heres et al. 1991] wird für die verschiedenen Feature-Arten und deren Attribute und Relationen genau angegeben, in welchem Turnus sie fortzuführen sind. Tabelle 5.4 gibt hierzu ein Beispiel für die geforderte Aktualität von Attributen der Feature Klasse *Roads*. Die Straßen werden in GDF in *Functional Classes* (FC) eingeteilt. Diese Klassifizierung ist eine Angabe über die Bedeutung der Straße (siehe Tabelle 5.2).

Es muß damit gerechnet werden, daß sich bei Vorliegen eines GDF- und ATKIS-Datensatzes desselben Gebietes, diese Daten in ihrer Aktualität unterscheiden. Da die Veränderungen im Straßenverkehr jährlich mit bis zu 10 Prozent geschätzt werden [Claussen 1995], können auch bei einem Aktualitätsunterschied von wenigen Wochen bereits signifikante Änderungen in den Daten auftreten. Als sehr positiv ist anzusehen, daß in ATKIS bei jedem Objekt das Erfassungsdatum abgespeichert wird und sich damit die Aktualität genau verifizieren läßt. In GDF besteht diese Möglichkeit der objektweisen Speicherung der Aktualität nicht, sondern hier wird lediglich ein globaler Parameter für eine Menge von Objekten mitgeführt.

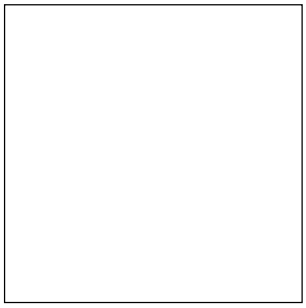


Abbildung 5.3: Testdaten ATKIS und GDF; Beispiel 1

Da GDF-Daten vor allem für die Verkehrsnavigation genutzt werden, müssen sie Informationen darüber enthalten, wie Kreuzungsbereiche befahren werden können. Dies bedeutet, daß zusätzlich zu den Fahrspuren auch sämtliche Abbiegespuren zu erfassen sind. Da bei Kreuzungsbereichen, in denen sich mehrere Fahrspuren treffen, eine Vielzahl von Abbiegemöglichkeiten bestehen, führt dies zu einer entsprechend komplexen Erfassung. Abbildung 5.4 zeigt die unterschiedliche Erfassung einer Kreuzung mit mehreren Fahrspuren in ATKIS und GDF. An diesem Beispiel kann gesehen werden, daß die GDF-Erfassung detaillierter ist und eine Obermenge der ATKIS-Elemente bildet.

Während im Beispiel in Abbildung 5.4 gesehen werden kann, daß eine Teilmenge der Elemente der beiden Datensätze ähnlich erfaßt wurde, zeigt das Beispiel in Abbildung 5.5 einen Kreuzungsbereich, der sehr stark unterschiedlich digitalisiert worden ist. Hier sind die Gemeinsamkeiten der beiden Datensätze so klein, daß kaum Entsprechungen zwischen den Datensätzen gefunden werden können.

Abbildung 5.6: Testdaten ATKIS und GDF; Beispiel 4

Abbildung 5.7: Testdaten ATKIS und GDF; Beispiel 5

Kapitel 6

Zuordnung von raumbezogenen Daten

Eine Integration von raumbezogenen Daten aus verschiedenen Datenmodellen erfordert in einem ersten Schritt, daß Elemente in den Datensätzen, welche die gleichen topographischen Objekte beschreiben, identifiziert werden können. Aufgrund der Anforderungen der verschiedenen Anwendungen werden gleiche topographische Objekte jedoch durch unterschiedliche Elemente und in unterschiedlichen Abstrahierungsgraden beschrieben. So werden die Objekte der Objektklasse *Straße* in einem Datenmodell, welches z.B. Daten für Verkehrsnavigationsysteme beschreibt, wesentlich detaillierter erfaßt, als in einem Datenmodell, welches für metereologische Voraussagen verwendet wird. Um zu einem gegebenen Objekt, das durch bestimmte Attribute, abhängig vom zugrunde liegenden Datenmodell, beschrieben wird, ein korrespondierendes Objekt in einem anderen Datenmodell zu finden, können Zuordnungsalgorithmen angewandt werden. Aufgabe eines solchen Zuordnungsalgorithmus ist es, mit Hilfe der Beschreibung eines Objektes im Ausgangsdatensatz das Objekt im anderen Datensatz zu finden, dessen Beschreibung am ähnlichsten ist. Im folgenden wird speziell die Vorgehensweise von Zuordnungstechniken beschrieben, die den Raumbezug der Daten ausnutzen. Diese Art von Zuordnungstechniken wird dann benötigt, wenn die Objektstrukturen in den zuzuordnenden Datensätzen für einen direkten Vergleich der Objekte zu unterschiedlich sind. Nach dem Klassifikationsschema des Kapitels 2.2 handelt es sich um ein Zuordnungsproblem der Stufe 3.

In diesem Kapitel wird die Vorgehensweise für die Zuordnung von raumbezogenen Daten diskutiert. Der Schwerpunkt liegt auf den theoretischen Grundlagen, ohne dabei anwendungsspezifische Fragestellungen oder bestimmte Datenmodelle zu betrachten. Nach einer Vorstellung von bestehenden Arbeiten wird ein Überblick über die Lösung von Zuordnungsproblemen gegeben. Anschließend wird ein Ansatz zur automatischen Zuordnung von raumbezogenen Daten vorgestellt. Die hierzu notwendigen Teilschritte werden in den einzelnen Unterkapiteln diskutiert.

6.1 Bestehende Arbeiten

Zuordnungsprobleme treten häufig im Bereich der digitalen Bildverarbeitung auf. Dies können Problemstellungen aus dem Bereich der Low-Level Verarbeitung sein, wie z.B. das Auffinden von homologen Bildpunkten in einem Stereobildpaar [Förstner 1986], aber auch Problemstellungen aus der High-Level Verarbeitung, wie das Erkennen von Werkstücken in einem digitalen Bild [Grimson 1990]. Diese Zuordnungsprobleme basieren auf der gleichen Grundidee. Aus einer Menge von Elementen, die durch eine bestimmte Art beschrieben werden, soll das Element gefunden werden, welches einem gegebenen am meisten ähnelt. Die Beschreibung der Elemente wird in der Regel durch Attribute wie Form, Farbe, Größe oder Koordinaten angegeben. Je höher der Grad der Beschreibung ist, desto leichter läßt sich die richtige Zuordnung finden.

Wird der Grad der Ähnlichkeit ausschließlich über die Werte der Attribute der Elemente berechnet, so spricht man von merkmalsbasierter Zuordnung (feature based matching). In vielen Anwendungen sind aber nicht nur die Attribute der Elemente bekannt, sondern es existieren auch Relationen zwischen den Elementen eines Datensatzes untereinander. Dies können z.B. topologische Relationen wie Adjazenz oder Inzidenz sein oder quantitative Relationen wie z.B. Größenbeziehungen. Werden diese Relationen der Elemente untereinander im Zuordnungsprozeß berücksichtigt, spricht man von relationaler Zuordnung (relational matching). Die relationale Zuordnung kann als eine Erweiterung der merkmalsbasierten Zuordnung angesehen werden. Eine umfassende Einführung in relationale Zuordnungstechniken findet sich in [Ballard und Brown 1982].

Erste Ansätze zur Zuordnung von raumbezogenen Daten sind heute schon teilweise in Geo-Informationssystemen implementiert. Hierbei handelt es sich jedoch vor allem um Algorithmen, welche versuchen, ähnliche Geometrieelemente miteinander zu verschmelzen (Homogenisierung). Weitere Ansätze versuchen in den Datensätzen identische Objektstrukturen zu finden und für einen Attributaustausch oder Geometrieangleich zu nutzen. Diese Ansätze arbeiten jedoch nur mit Daten, die jeweils in demselben Datenmodell erfaßt wurden, da sonst keine identischen Objektstrukturen gefunden werden können. “Inzwischen rücken jedoch zunehmend Ansätze in den

Blickpunkt, die neben eventuellen Diskrepanzen in der Geometrie auch die semantischen Unterschiede aufgrund von abweichendem Modellverständnis in Betracht ziehen“ [Illert 1995].

Als einer der ersten Ansätze in der Literatur findet sich die Arbeit des Bureau of the Census in Washington DC [Rosen und Saalfeld 1985, Saalfeld 1988]. Motivation war hier, die Vermessungskarten des United States Geological Survey (USGS) mit den Karten des Census Bureaus zu verknüpfen. Das Census Bureau hatte von Hand Karten von über 5 % des Landes digitalisiert, worauf 60 % der Bevölkerung wohnten. Ziel war es nicht eine neue Karte zu erzeugen, sondern beide ursprüngliche Karten zu verbessern und auch Fehler darin aufzudecken. So sollten bei den USGS-Karten Attribute wie Straßenarten, Straßennamen und Hausnummern hinzugefügt werden. Bei den Karten des Census Bureau sollten neu gebaute Häuser eingefügt und linienhafte Objekte der USGS-Karte übernommen werden. Hierzu wurde ein iteratives Verfahren entworfen. In einem ersten Schritt werden Zuordnungen zwischen punktförmigen Elementen aufgestellt. Die Zuordnungen können interaktiv am Bildschirm eingeben oder automatisch mit Hilfe von topologischen und geometrischen Merkmalen bestimmt werden. Diese Punkte werden als identisch betrachtet. Anschließend werden mit Hilfe einer Rubber-Sheet-Transformation [Gillmann 1985] die verbleibenden Punkte transformiert und das Verfahren beginnt von vorne. Man kann sich diese Transformation wie ein Gummituch vorstellen. Das Gummituch wird über eine der Karten gelegt. Danach werden die übereinstimmenden Punkte zur Deckung gebracht und befestigt. Die nicht zugeordneten Punkte werden dadurch lokal transformiert. Dieses Verfahren wird in der Mathematik auch als stückweise linearer Homomorphismus bezeichnet und führt insbesondere bei Datensätzen, die sich nicht durch eine globale Transformation aufeinander abbilden lassen, zu guten Ergebnissen. Linienförmige Elemente werden dann zugeordnet, wenn ihr Anfangspunkt und ihr Endpunkt zugeordnet werden konnten. Das Verfahren geht davon aus, daß sich die Datensätze nicht stark unterscheiden und betrachtet nur $1:1$ Zuordnungen zwischen den Elementen. Dieser Ansatz ist daher nur für ähnliche Datensätze geeignet, da ansonsten auch $n:m$ Beziehungen zwischen den Datensätzen betrachtet werden müssen (siehe Kapitel 6.4).

In einer aktuellen Arbeit [Gabay und Doytsher 1994] wird untersucht, wie sich Karten, die sich nur gering geometrisch und topologisch unterscheiden, zuordnen lassen. Unter der Annahme, daß einer der beiden Datensätze mit einer höheren Genauigkeit erfaßt wurde, wird anschließend die Geometrie des anderen Datensatzes mit Hilfe der Zuordnungen verbessert [Gabay und Doytsher 1995]. Hierbei handelt es sich um einen iterativen Ansatz, bei dem ausgehend von Zuordnungen, welche durch eine direkte Überlagerung der zwei Datensätze gefunden werden, weitere Zuordnungen durch Verfolgung von linienhaften Objekten zu suchen sind. Dieser Ansatz kann $n:m$ Zuordnungen aufstellen; jedoch hängt die Anzahl der gefundenen Zuordnungen davon ab, welche Zuordnungen durch die direkte Überlagerung als Startzuordnungen gefunden werden. Die Zuordnungen werden durch einen Vergleich der Attribute und topologischen Relationen der Elemente berechnet.

Ein weiterer iterativer Ansatz wird in [Kraft 1995] untersucht. Die Zuordnungen werden hier jedoch nicht durch Linienverfolgung, sondern durch ein heuristisches Verfahren berechnet. In einem ersten Schritt werden alle möglichen $1:1$ Zuordnungen aufgestellt und in einer Liste mit potentiellen Zuordnungen gespeichert. Es erfolgt eine Bewertung der Zuordnungen mit Hilfe von geometrischen und topologischen Merkmalen. Die Zuordnung, welche die beste Bewertung aufweist, wird in die endgültige Zuordnungsliste eingetragen. Danach werden alle die Zuordnungen aus der Liste der möglichen Zuordnungen entfernt, welche ein Element der gefundenen Zuordnung enthalten. Anschließend wird wieder die Zuordnung mit der besten Bewertung gesucht. Das Verfahren läuft solange, bis alle potentiellen Zuordnungen aus der Liste entfernt wurden. Dieser Ansatz versucht zwar die optimale Zuordnung zweier Datensätze zu finden, kann jedoch durch seinen lokalen Optimierungsansatz (siehe Kapitel 6.8) insbesondere bei parallelen Linienzügen zu Fehlzuordnungen führen. Ein Hauptmerkmal des Verfahrens ist, daß es die Zuordnung der Daten in zwei Schritten durchführt. In einem ersten Schritt wird mit Hilfe einer Ausgleichungsrechnung der globale Fehler zwischen den beiden Datensätzen minimiert, um anschließend die eigentliche Zuordnung der Elemente durchzuführen.

Welche Anforderungen Algorithmen zur Zuordnung von raumbezogenen Daten erfüllen müssen, beschreibt [Brown, Rao und Baran 1995]. In diesem Artikel wird vor allem auf die Bedeutung von $n:m$ Zuordnungen eingegangen und Beispiele für automatische Attributergänzungen gegeben.

In der vorliegenden Arbeit wird ein Ansatz vorgestellt, welcher auf einem relationalen Zuordnungsverfahren basiert. In [Vosselman 1992] wird die Theorie des relationalen Zuordnens beschrieben sowie darauf aufbauend ein Verfahren vorgestellt, welches Lage und Form von dreidimensionalen Objekten aus einem Grauwertbild berechnet. Aus dem Grauwertbild werden mit Hilfe von Bildverarbeitungsalgorithmen Kanten extrahiert und mit Zuordnungsalgorithmen auf eine Modellbeschreibung abgebildet. Der Vorteil dieses Ansatzes liegt darin, daß die Ergebnisse unabhängig von Startparametern sind und dadurch eine eindeutige Lösung gefunden wird. Das Verfahren basiert auf einem theoretisch fundierten Ansatz aus der Informationstheorie. Weitere Anwendungen relationaler Zuordnungstechniken finden sich z.B. bei [Vosselman und Haala 1992, Haala und Vosselman 1992],

die ein Verfahren zur Erkennung von topographischen Paßpunkten vorstellen oder in [Zilberstein 1992], der die Zuordnung überlappender Luftbilder beschreibt.

6.2 Problemeinführung

Um in zwei Datensätzen korrespondierende Elemente zu finden, werden Zuordnungstechniken (Matching-Techniken) benötigt. Sei $A = \{a_1, a_2, \ldots, a_n\}$ die Menge der Elemente des Datensatzes A und $B = \{b_1, b_2, \ldots, b_m\}$ die Menge der Elemente des Datensatzes B, so wird eine Zuordnungsfunktion $h(A \rightarrow B)$ gesucht, die die Elemente der Menge A möglichst gut auf die Menge B abbildet. Um die Zuordnungen zu bewerten, werden üblicherweise Kosten- bzw. Leistungsfunktionen eingeführt, welche angeben, wieviel eine Einzelzuordnung an Kosten bzw. Leistung an der Gesamtzuordnung beiträgt. Im folgenden wird immer der Begriff Leistungsfunktion verwendet, da in dieser Arbeit zur Bewertung der Zuordnungen eine Leistungsfunktion definiert wurde. Dies hat Vorteile, wie später noch gezeigt werden wird. Grundsätzlich gelten die aufgeführten Überlegungen analog auch für Kostenfunktionen. Der Wert der Leistungsfunktion wird je nach verwendeten Verfahren aus der Ähnlichkeit der Attribute (merkmalsbasierte Zuordnung) sowie zusätzlich aus der Ähnlichkeit der Relationen zwischen den Elementen (relationale Zuordnung) berechnet.

Die Gesamtleistung einer Zuordnung zweier Datensätze ergibt sich dann durch die Aufsummierung aller Einzelleistungen. Sei $l = f(a_n, b_m)$ die Leistungsfunktion einer Zuordnung, dann ist:

$$L = \sum_{z=1}^{Z} l(a_{za}, b_{zb}) \tag{6.1}$$

die Höhe der Leistung der Gesamtzuordnung. Die beste Zuordnung zweier Datensätze ergibt sich aus der Kombination von Einzelzuordnungen, bei der die Aufsummierung der Leistungen der Einzelzuordnungen das Maximum annimmt. Hier wird ersichtlich, daß es sich beim Zuordnen von zwei Datensätzen um ein kombinatorisches Problem handelt. Die Lösung kombinatorischer Probleme benötigt jedoch eine exponentiell zur Datenmenge ansteigende Berechnungszeit.

Sei n die Anzahl der Elemente der Menge A und m die Anzahl der Elemente der Menge B sowie ohne Beschränkung der Allgemeinheit $n \geq m$. Gesucht ist die Anzahl der Kombinationen von Zuordnungen, wobei gelten soll, daß jedem Element aus A genau ein Element aus B zugeordnet werden soll (nur $1:1$ Zuordnungen). Wählt man das erste Element aus A, so gibt es m mögliche Partner, die diesem Element zugeordnet werden können. Für das zweite Element existieren $m-1$ mögliche Partner, usw. Insgesamt gibt es dann $m!$ verschiedene Kombinationen von $1:1$ Zuordnungen. Die Anzahl der möglichen Kombinationen wird noch größer, wenn neben $1:1$ Zuordnungen allgemein auch $n:m$ Zuordnungen $(n, m \geq 0)$ zugelassen werden.

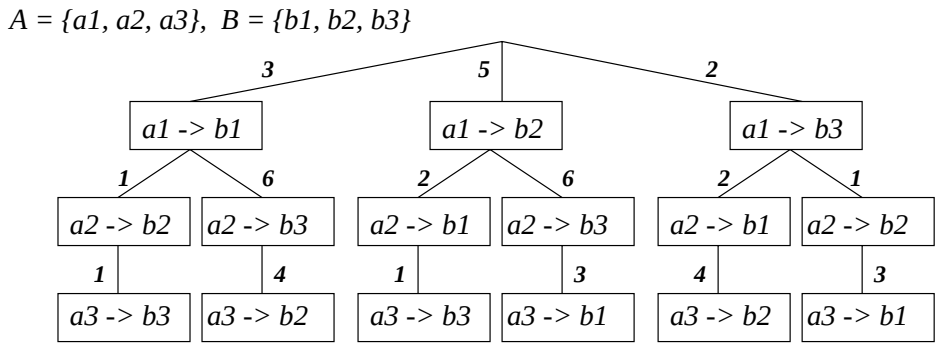


Abbildung 6.1: Zuordnungsbaum für $n = m = 3$

Abbildung 6.1 zeigt den Suchraum (d.h. sämtliche Kombinationen von Zuordnungen) für $n = m = 3$ mit Hilfe eines Zuordnungsbaumes dargestellt. Jede mögliche Kombination wird durch einen Pfad im Baum, der bei der Wurzel beginnt und einem Blatt endet, dargestellt. Die Anzahl der verschiedenen Kombinationen entspricht der Anzahl der Blätter des Baumes. An den Kanten im Baum sind die Werte der Leistungsfunktion der Zuordnungen in der Abbildung als fiktive Werte aufgetragen. Die beste Zuordnung wird durch einen Pfad von der Wurzel zu einem Blatt des Baumes dargestellt, bei dem die Aufsummierung der Werte den maximalen Wert ergibt. Um

diesen Pfad zu finden, werden Suchverfahren, die in der Informatik speziell im Bereich der künstlichen Intelligenz häufig angewandt werden, eingesetzt [Winston 1987]. Wird der gesamte Baum durchsucht, handelt es sich um eine Vollraumsuche. Da der Rechenaufwand von Vollraumsuchverfahren sehr groß sein kann, werden heuristische Verfahren genutzt. Diese Verfahren versuchen mit Hilfe von Heuristiken den Suchraum einzuschränken und damit das Zeitverhalten der Baumsuche zu verbessern.

Bei den Baum-Suchverfahren handelt es sich, auch bei der Verwendung von heuristischen Algorithmen, um Aufgaben mit einer sehr hohen Zeitkomplexität. Daher müssen Möglichkeiten gefunden werden, den Suchraum zu minimieren. Der einfachste Ansatz die Daten einander zuzuordnen wäre, nacheinander jedes Element des einen Datensatzes herauszugreifen und mit allen Elementen des anderen Datensatzes zu vergleichen. Dies führt zwar zur Lösung des Problems, ist wegen des exponentiellen Zeitbedarfs praktisch nicht durchführbar. Daher muß beim Aufstellen des Suchbaumes darauf geachtet werden, daß nur die Blätter des Baumes weiter expandiert werden, bei denen eine Lösung gefunden werden kann. Durch den Raumbezug der Daten ergibt sich eine natürliche Einschränkung des Suchbereiches. Abbildung 6.2 zeigt dies an einem Beispiel. Angenommen es ist bekannt, daß die Zuordnungen $a_1 \rightarrow b_1$ und $a_2 \rightarrow b_1$ keinen Sinn ergeben, da die Elemente a_1 und a_2 an völlig anderen Koordinaten liegen als das Element b_1, dann müssen die Blätter, die die Zuordnungen $a_1 \rightarrow b_1$ und $a_2 \rightarrow b_1$ repräsentieren, auch nicht expandiert werden, da diese Kombination nicht sinnvoll ist und daher nicht weiterverfolgt werden braucht. Dies führt zu einer wesentlichen Reduzierung des Suchraumes.

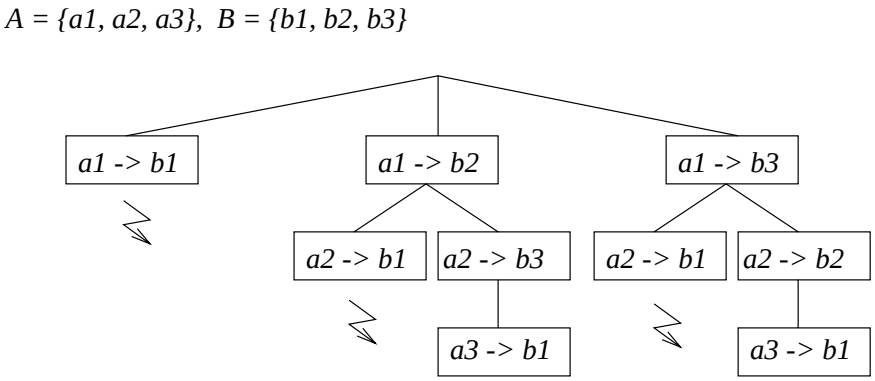


Abbildung 6.2: Ausnützen natürlicher Beschränkungen

Die dargestellten Probleme wurden bisher nur für 1 : 1 Zuordnungen betrachtet. Wie weiter unten gezeigt wird, müssen jedoch bei der Zuordnung von raumbezogenen Daten allgemein $n : m$ Zuordnungen in Betracht gezogen werden. Dies bedeutet zwar auf der einen Seite eine Vergrößerung des Suchraumes, auf der anderen Seite ergeben sich dadurch weitere Möglichkeiten, um den Suchraum wieder einzuschränken. In den folgenden Unterkapiteln werden die einzelnen Problemstellungen der Zuordnungsverfahren speziell an dem Beispiel der Zuordnung zweier raumbezogener Datensätze diskutiert.

6.3 Vorverarbeitung

Werden raumbezogene Daten mehrfach erfaßt, so unterscheiden sich diese zum einen durch systematische Fehler und zum anderen durch lokal unterschiedliche Erfassung. Systematische Fehler treten beispielsweise durch nicht korrekte Transformationen zwischen Digitalisiertischkoordinaten und Modellkoordinaten auf. Lokale Unterschiede ergeben sich durch die Erfassung in unterschiedlichen Datenmodellen sowie durch fehlerhafte Digitalisierung, wie z.B. nicht korrekte Koordinaten, Fehler in der Topologie oder fehlende Elemente.

Wie weiter oben bereits dargestellt wurde, ist es von großer Wichtigkeit, den Suchraum, der ausgewertet werden muß, zu minimieren. Durch den Raumbezug der Daten kann davon ausgegangen werden, daß das gesuchte Partner-Element im anderen Datensatz in etwa an den gleichen geometrischen Koordinaten gefunden werden kann. Es ist also ausreichend, nur die Daten im anderen Datensatz zu betrachten, welche in einem Puffer um das Element liegen. Die Größe des Puffers hängt zum einen vom globalen Fehler ab und zum anderen von lokalen Unterschieden. Daher wird die Zuordnung in zwei Schritten durchgeführt. Um die Breite des Puffers zu minimieren, werden in einem Vorverarbeitungsschritt zuerst die globalen Fehler mit Hilfe einer Transformation eliminiert. Die Berechnung der Parameter kann entweder durch interaktives Bestimmen von Paßpunkten oder automatisch durchgeführt werden. Bei einer automatischen Bestimmung der Parameter werden ähnliche Flächen

in den beiden Datensätzen gesucht [Kraft 1995]. Das Maß der Ähnlichkeit ergibt sich aus den geometrischen Attributen und topologischen Beziehungen der Flächen. Ähnliche Flächen werden über ihre Flächenschwerpunkte einander zugeordnet. Anschließend werden aus diesen Zuordnungen mit Hilfe der Ausgleichungsrechnung die gesuchten Parameter einer Affintransformation geschätzt.

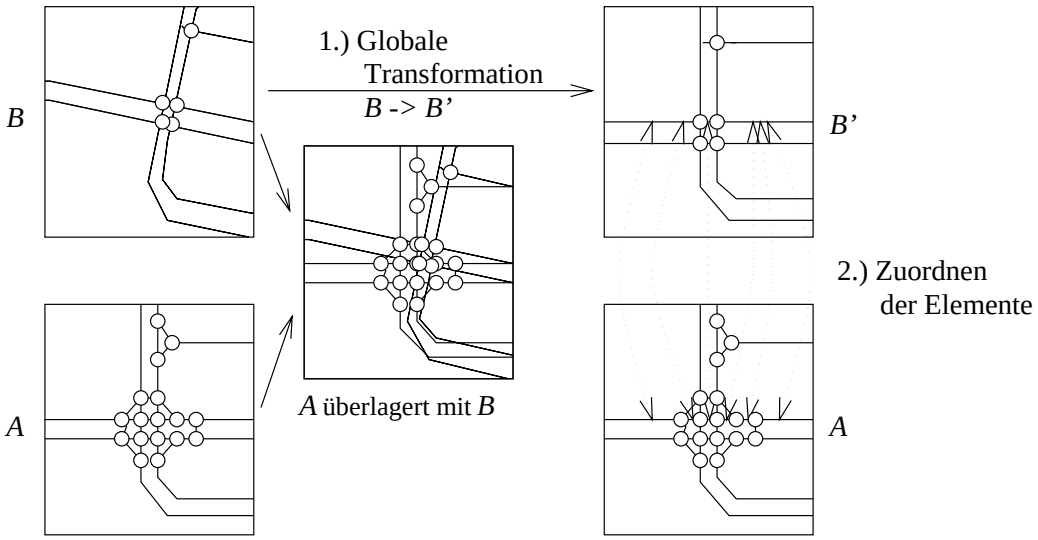


Abbildung 6.3: Vorverarbeitung der Daten

Abbildung 6.3 zeigt die Vorverarbeitung der Daten an einem Beispiel. A und B sind die beiden Datensätze, die einander zugeordnet werden sollen. Bei einer Überlagerung der beiden Datensätze kann gesehen werden, daß sie nicht deckungsgleich sind. Der systematische Fehler kann jedoch durch eine Transformation $B \rightarrow B'$ minimiert werden. Danach erfolgt die eigentliche Zuordnung der Elemente.

6.4 Buffer Growing

Nach dem Eliminieren des systematischen Fehlers wird um die Elemente des einen Datensatzes ein Puffer gelegt, in dem die Zuordnungspartner aus dem anderen Datensatz gesucht werden. Abbildung 6.4 zeigt die unterschiedlichen Fälle, die beim Zuordnen der Elemente auftreten können. Ein Element a_i kann genau einem Element b_j zugeordnet werden. Weiter ist die Zuordnung von einem Element a_i zu mehreren Elementen $\{b_{j1}, b_{j2}, \ldots b_{jn}\}$ erlaubt. Mehrere Elemente $\{a_{i1}, a_{i2}, \ldots a_{in}\}$ können mehreren Elementen $\{b_{j1}, b_{j2}, \ldots b_{jm}\}$ zugeordnet werden und es ist möglich, daß ein Element überhaupt nicht zugeordnet werden kann, was mit einer "Wildcard" $(*)$ dargestellt wird. Die Zuordnung in dem rechten Kasten stellt einen Sonderfall der $n : m$ Zuordnungen dar. Hier werden ein oder mehrere Elemente des einen Datensatzes zu mehreren nicht topologisch zusammenhängenden Elementen des anderen Datensatzes zugeordnet. Solche Zuordnungen kommen beispielsweise dann vor, wenn in einem der Datensätze Straßen durch ihre Mittelachsen und im anderen durch die Fahrspuren erfaßt werden. Auf diesen Sonderfall wird später eingegangen.

Mit Hilfe von Puffern ist es möglich $1 : 1$ und $1 : n$ Zuordnungen aufzustellen. Hierzu wird um ein gegebenes Element ein Puffer gelegt und es wird getestet, welches Element bzw. welche topologisch zusammenhängenden Elemente des anderen Datensatzes in diesem Puffer liegen. Um die $n : 1$ und $n : m$ Zuordnungen zu finden, wird ein Verfahren, welches im folgenden *Buffer Growing* genannt wird, angewendet. Abbildung 6.5 zeigt das Verfahren an einem Beispiel. In der obersten Situation ist ein Puffer um das Element a_2 gebildet worden. Es kann kein Element aus B gefunden werden, welches vollständig in diesem Puffer liegt. Daher kann a_2 lediglich der Wildcard $(*)$ zugeordnet werden. Danach wird der Puffer um ein Element verlängert. Die beiden Elemente a_2 und a_3 werden als ein logisches Element a_2a_3 zusammengefaßt und können dem Element b_1 zugeordnet werden, da dieses vollständig in dem Puffer liegt. Ebenso können b_1 und b_2 zu einem logischen Element b_1b_2 zusammengefaßt und dem logischen Element a_2a_3 zugeordnet werden. Nur solche Zuordnungen werden im folgenden weiter betrachtet, welche neue Informationen enthalten. So ist es z.B. nicht nötig, die Zuordnung $a_2a_3a_4 \rightarrow b_1b_2b_3$ zu betrachten, da dies durch die Kombination der Zuordnungen $a_2a_3 \rightarrow b_1b_2$ und $a_4 \rightarrow b_3$ abgedeckt ist.

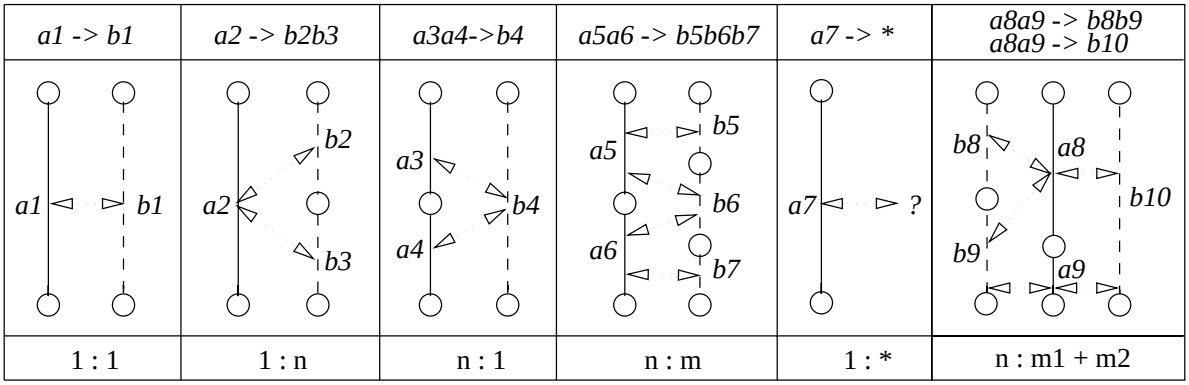


Abbildung 6.4: Kardinalität der Zuordnungen

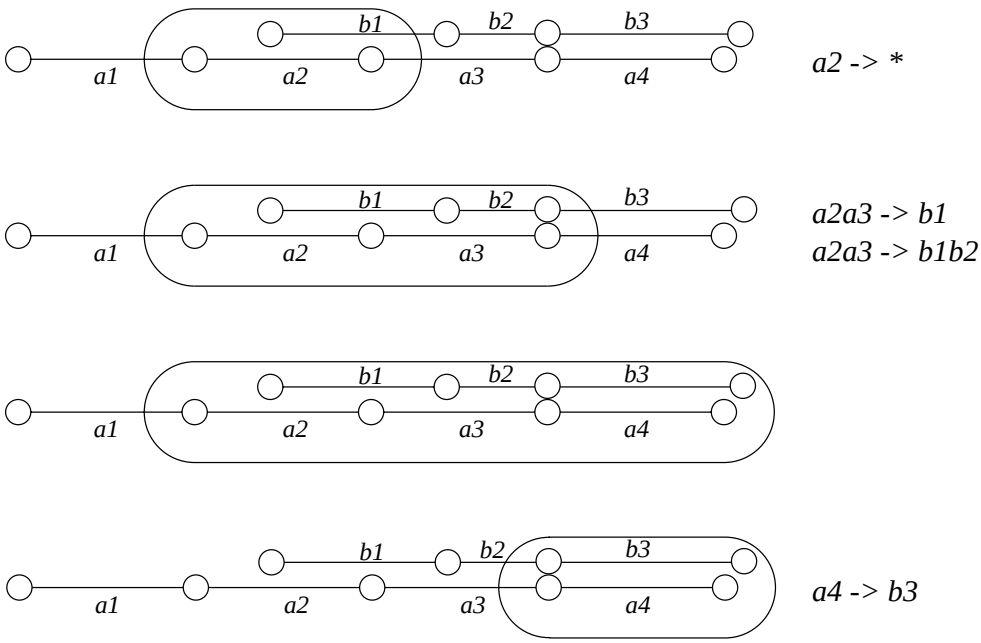


Abbildung 6.5: Buffer Growing

Bei einer Implementation des Buffer Growing müssen einige Besonderheiten beachtet werden. Damit alle möglichen $n : m$ Zuordnungen gefunden werden, ist es notwendig, daß die Puffer um jedes Element gebildet werden und sowohl auf die linke als auch auf die rechte Seite wachsen. Dies führt jedoch dazu, daß Zuordnungen mehrfach gefunden werden. Daher müssen alle gefundenen Zuordnungen in einer Liste abgespeichert werden. Beim Eintrag in diese Liste wird zuerst überprüft, ob die Zuordnung bereits enthalten ist oder ob sich diese Zuordnung durch eine Kombination "kürzerer" Zuordnungen bilden läßt. Ist dies der Fall, wird sie zurückgewiesen und der Puffer braucht in diese Richtung nicht weiter verlängert werden.

6.5 Ausnützen von Beschränkungen

Mit Hilfe des Buffer Growing werden alle potientiellen Zuordnungspaare zwischen den beiden Datensätzen aufgestellt. Mit Zuordnungspaaren werden auch die $n : m$ Zuordnungen bezeichnet, da in diesem Fall mehrere Elemente zu einem logischen Element zusammengefaßt werden (siehe vorheriger Abschnitt). Die Anzahl der potentiellen Zuordnungspaare entscheidet über den Aufwand, der benötigt wird, um aus dieser Menge diejenigen endgültigen Zuordnungspaare herauszufinden, welche eine optimale eindeutige Zuordnung zwischen den beiden

Datensätzen beschreiben. Daher ist es aus Performancegründen sehr entscheidend, die Anzahl der potentiellen Zuordnungspaare so gering wie möglich zu halten.

6.5.1 Geometrische Beschränkungen

Die Aufgabe besteht darin, diejenigen Zuordnungen zu finden, bei denen im voraus gesagt werden kann, daß sie sehr unwahrscheinlich sind. Beispielsweise ist es unwahrscheinlich, daß ein linienförmiges Element des einen Datensatzes einem linienförmigen Element des anderen Datensatzes zugeordnet wird, wenn sich diese beiden Elemente im Winkel von 90 Grad schneiden. Neben dem Winkel kann auch die Länge, die Form oder die Entfernung zwischen linienförmigen Elementen betrachet werden (siehe Abbildung 6.6 a). Dabei ist die Bedingung, daß die Entfernung zwischen zwei Elementen nicht ein bestimmtes Maß überschreiten darf, bereits implizit im Buffer Growing enthalten.

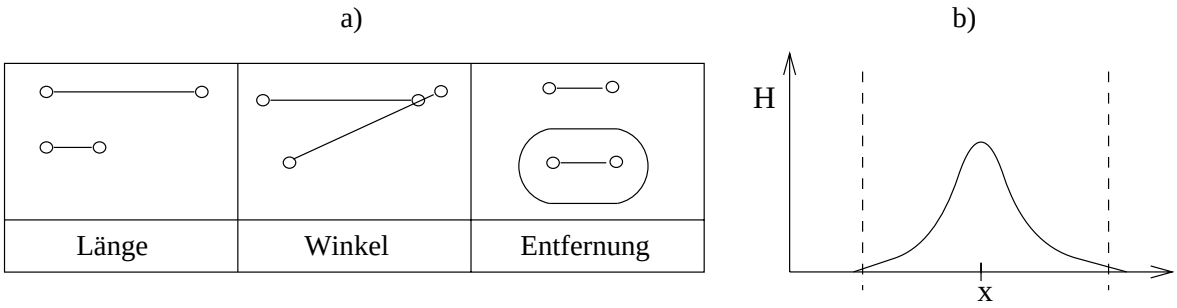


Abbildung 6.6: Geometrische Beschränkungen

Die Festlegung von Ober- und Untergrenzen, ab denen ein Zuordnungspaar nicht weiter betrachtet werden soll, stellt sich jedoch als ein Problem dar. Diese Grenzen sind von den beteiligten Datenmodellen abhängig und lassen sich nicht allgemein für unterschiedliche Datenmodelle formulieren. Jedoch können statistische Auswertungen der Datensätze zwischen zwei Datenmodellen Aussagen über die Wahrscheinlichkeit machen, ob ein potentielles Zuordnungspaar in der Endzuordnung enthalten ist. Hierzu kann eine Häufigkeitsverteilung anhand von Stichproben aufgestellt werden, die z.B. beschreibt, mit welcher Häufigkeit ein Element der Länge x einem Element der Länge y zugeordnet wird. Es wird nun nach Schranken gesucht, welche z.B. eine 99,9 prozentige Sicherheit bieten, daß alle die potentiellen Zuordnungspaare weiter betrachtet werden, die in der Endzuordnung vorkommen können. Kann eine Häufigkeitsverteilung gefunden werden, die eine Form wie die Häufigkeitsverteilung in Abbildung 6.6 b) besitzt, ist es möglich solche Schranken zu definieren. Im Falle von Straßenverkehrsdaten ist das Verfahren noch weiter zu verfeinern, wenn z.B. Stadtbereiche und ländliche Bereiche getrennt voneinander statistisch ausgewertet werden, da hier unterschiedliche Häufigkeitsverteilungen zu erwarten sind.

6.5.2 Thematische Beschränkungen

Neben den geometrischen Beschränkungen lassen sich auch Beschränkungen aus den thematischen Daten ableiten. Dies können die Attribute von Straßen, wie z.B. Breite, Funktion oder Name sein. Hierbei muß zwischen numerischen und alphanumerischen Attributen unterschieden werden. Während sich bei den numerischen Attributen dasselbe Verfahren wie bei den geometrischen Beschränkungen anwenden läßt, können alphanumerische Attribute nicht direkt miteinander verglichen werden. Abbildung 6.7 zeigt diese Situation an einem Beispiel. Eine Straße mit dem Straßennamen *Heilbronner Straße* soll einer anderen Straße zugeordnet werden. Die Wahrscheinlichkeit, daß die Straße *Hailbroner Straße* der Partner ist, erscheint höher als ein Partner mit dem Namen *Pragstraße*, jedoch stellt sich hier das Problem, eine Ähnlichkeitsfunktion zu definieren.

6.6 Merkmalsbasierte vs. relationale Zuordnung

Nach dem Aufstellen einer Liste aller Zuordnungspaare muß aus dieser mehrdeutigen Liste die Kombination von endgültigen Zuordnungen gesucht werden, die der besten Gesamtzuordnung der beiden Datensätze entspricht. Um diese berechnen zu können, müssen die einzelnen Zuordnungen bewertet werden. Hierzu kann die Ähnlichkeit

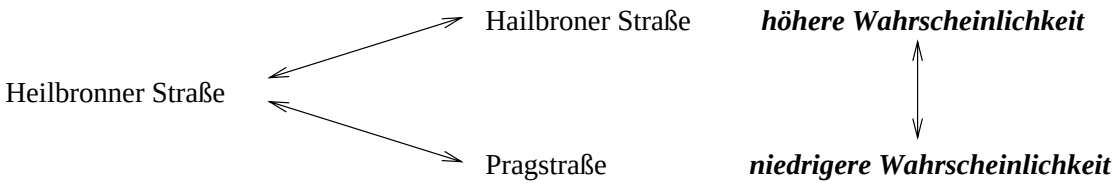
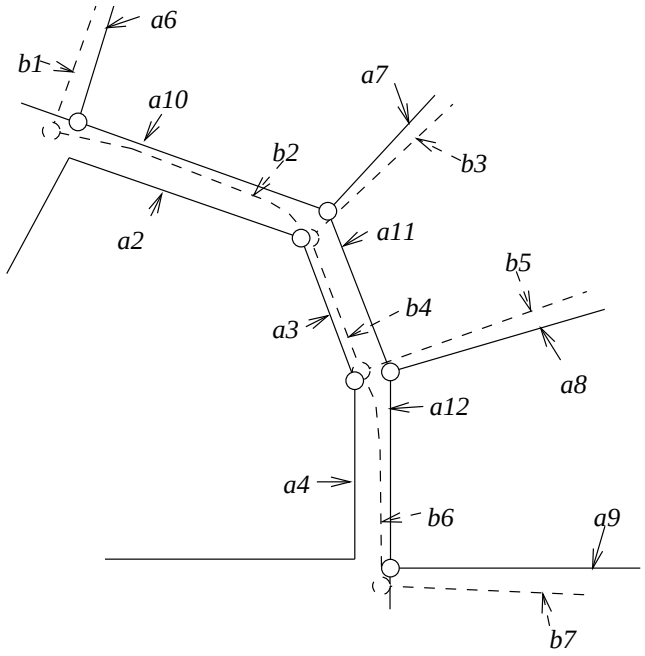


Abbildung 6.7: Thematische Beschränkungen

der Attribute (merkmalsbasierte Zuordnung) sowie zusätzlich die Ähnlichkeit der Beziehungen der Elemente untereinander (relationale Zuordnung) ausgewertet werden. Bei raumbezogenen Daten können zur Bewertung der Zuordnungen neben den Attributen insbesondere die topologischen Relationen herangezogen werden.



Feature Based Matching	Relational Matching
a6 -> b1	a6 -> b1
a10 -> b2	a10 -> b2
a3 -> b4	a11 -> b4
a12 -> b6	a12 -> b6
a7 -> b3	a7 -> b3
a8 -> b5	a8 -> b5
a9 -> b7	a9 -> b7

Abbildung 6.8: Merkmalsbasierte und relationale Zuordnung

Daß mit Hilfe der relationalen Zuordnung bessere Ergebnisse erzielt werden können als mit einer merkmalsbasierten Zuordnung, soll Abbildung 6.8 verdeutlichen. Die Elemente des Datensatzes A sind mit durchgezogenen Linien, die Elemente des Datensatzes B mit gestrichelten Linien dargestellt. Bei einem Feature Based Matching, welches beispielsweise die Attribute *Position* und *Länge* der einzelnen Elemente bewertet, würde dem Element a_3 das Element b_4 zugeordnet werden, da dies das Element des anderen Datensatzes ist, welches die ähnlichsten Attribute besitzt. Diese offensichtliche Fehlzuordnung kann mit einem relationalen Zuordnungsansatz vermieden werden, welcher zusätzlich zu den Attributen auch die topologischen Beziehungen der Elemente untereinander bewertet. Da alle vier topologischen Nachbarn von b_4 richtig zugeordnet wurden, kann bei einer Betrachtung der Topologie festgestellt werden, daß die Zuordnung $a_{11} \rightarrow b_4$ wahrscheinlicher ist als die Zuordnung $a_3 \rightarrow b_4$.

6.7 Berechnung der Leistungsfunktion

Die Zuordnung der raumbezogenen Daten erfolgt mit einem relationalen Zuordnungsansatz. Hierzu werden die Zuordnungen mit Hilfe der Attribute und Relationen der Elemente bewertet. Um die Ähnlichkeit zweier Elemente a_i und b_j darzustellen, wird häufig eine Kostenfunktion verwendet:

$$\text{Kosten}(a_i, b_j) = \sum_{n=1}^{N} \text{Kosten}(att_n(a_i), att_n(b_j)) + \sum_{m=1}^{M} \text{Kosten}(rel_m(a_i), rel_m(b_j)) \tag{6.2}$$

$P(a_i\|b_j)$	gerade	leicht gekrümmt	stark gekrümmt
gerade	0,8	0,2	0,0
leicht gekrümmt	0,2	0,6	0,2
stark gekrümmt	0,0	0,2	0,8

Tabelle 6.1: Eine mögliche Übergangsmatrix für die Linienform

wobei N die Anzahl der Attribute att_n und M die Anzahl der Relation rel_m ist. Das bedeutet, daß die Kostenfunktion sich aus der Summe der Kosten der Zuordnung der einzelnen Attribute und Relationen berechnet. Die Kosten, die bei der Zuordnung der Elemente zweier Datensätze A und B entstehen, ergeben sich aus der Summe der Kosten aller Einzelzuordnungen:

$$\text{Kosten}(A,B) = \sum_{k=1}^{M} \text{Kosten}(a_{i_k}, b_{j_k}) \tag{6.3}$$

Die beste Zuordnung ist diejenige, bei der die niedrigsten Gesamtkosten entstehen. Bei dieser Technik entstehen jedoch Probleme, wenn zugelassen wird, daß Elemente einer Wildcard zugeordnet werden dürfen (siehe Kapitel 6.4). Dieser Fall tritt bei der Zuordnung von raumbezogenen Daten dann auf, wenn ein Objekt der Landschaft in einem Datensatz erfaßt wurde, im anderen jedoch nicht. Da die Zuordnung zu einer Wildcard keine "echte" Zuordnung darstellt, entstehen auch keine Kosten. Dies bedeutet aber, daß die beste Zuordnung zweier Datensätze diejenige Zuordnung ist, bei der alle Elemente der Wildcard zugeordnet werden, da hier die Gesamtkosten ebenfalls Null wären. Um dieses Problem zu umgehen, werden üblicherweise "Strafkosten" für die Zuordnung eines Elementes zu einer Wildcard eingeführt. Diese Strafkosten sind jedoch datenabhängig und schwer zu bestimmen.

Dieses Problem kann mit Hilfe einer Leistungsfunktion vermieden werden. Die Leistungsfunktion ist ein Maß dafür, wieviel Unterstützung eine Einzelzuordnung zu einer Gesamtzuordnung beiträgt. Die beste Gesamtzuordnung ist die Zuordnung, bei der die Summe der Leistungen der Einzelzuordnungen ein Maximum ergibt. Da eine Zuordnung eines Elementes zu einer Wildcard die Gesamtzuordnung nicht verbessert, ist die Leistung dieser Zuordnung gleich Null.

Im folgenden wird die Berechnung der Leistungsfunktion diskutiert. In [Vosselman 1992] wird ein Verfahren vorgestellt, welches aus der Informationstheorie abgeleitet ist, und das Berechnen einer Leistungsfunktionen für Zuordnungsprobleme beschreibt. Es ist theoretisch abgesichert und kommt ohne datenabhängige Werte (wie z.B. Schwellwerte oder Gewichtsfaktoren) aus. Dieses Verfahren soll für die Zuordnung von raumbezogenen Daten verwendet werden. Da zum Verständnis Begriffe aus der Informationstheorie benötigt werden, wird im Anhang A ein Exkurs in die Informationstheorie gegeben.

6.7.1 Zuordnung als Kommunikationssystem

Ein Kommunikationssystem überträgt Nachrichten von einem Sender über einen Kanal zu einem Empfänger. Je nach Eigenschaften des Kanals werden Unterschiede zwischen der versandten und der empfangenen Nachricht auftreten. Im Falle eines idealen Kanals entspricht die empfangene Nachricht exakt der gesendeten. Ist der Kanal jedoch gestört, d.h. es tritt Rauschen auf, unterscheidet sich die empfangene Nachricht von der gesendeten. Der beste Kanal ist der, bei dem die kleinste Differenz zwischen gesendeter und empfangener Nachricht auftritt.

Das Problem der Suche nach dem besten Kanal kann als äquivalent zum Zuordnungsproblem von zwei Datensätzen gesehen werden [Vosselman 1992, Boyer und Kak 1986, Boyer und Kak 1988]. Um die beste Abbildung zwischen Sender und Empfänger zu finden, wird diejenige Zuordnung gesucht, welche die gesendete Nachricht in die empfangene Nachricht überführt, welche ihr am ähnlichsten ist. Genauso kann das Zuordnungsproblem als ein Kommunikationssystem gesehen werden. Ein Sender versendet als Nachricht den Datensatz D_1, welcher beim Empfänger als D_2 empfangen wird. Der Unterschied zwischen D_1 und D_2 entsteht durch Rauschen im Kanal. Um dieses Rauschen im Kanal zu modellieren, müssen die Übergangswahrscheinlichkeiten ausgewertet werden.

Angenommen, es sollen zwei Datensätze mit linienförmigen Elementen einander zugeordnet werden. Eines der Attribute, welches zu den Elementen ausgewertet werden kann, ist z.B. die Krümmung der Linie. Es werden drei Klassen der Krümmung betrachtet: "gerade", "leicht gekrümmt" und "stark gekrümmt". Nun kann durch

$I(a_i\|b_j)$	gerade	leicht gekrümmt	stark gekrümmt
gerade	0,3 bit	2,3 bit	∞
leicht gekrümmt	2,3 bit	0,7 bit	2,3 bit
stark gekrümmt	∞	2,3 bit	0,3 bit

Tabelle 6.2: Matrix der bedingten Information

experimentelles Messen oder durch analytische Auswertung eine Matrix der Übergangswahrscheinlichkeiten für das Attribut Krümmung aufgestellt werden. Tabelle 6.1 zeigt eine mögliche Matrix für diesen Fall. Es kann gesehen werden, daß die Linien des Datensatzes A der Klasse "gerade" und "stark gekrümmt" mit einer Wahrscheinlichkeit von 0.8 einer Linie der gleichen Klasse im Datensatz B zugeordnet werden. Ein Wechsel der Klasse in die nächste nebenliegende Klasse erfolgt mit einer Wahrscheinlichkeit von 0.2. Ein Wechsel in die übernächste Klasse ist nicht möglich.

Die bedingten Wahrscheinlichkeiten der Übergangsmatrix geben also an, wie wahrscheinlich es ist, daß ein Element mit einem gegebenen Attributwert einem anderen Element und dessen Attributwert zugeordnet wird. Berechnet man nun aus diesen bedingten Wahrscheinlichkeiten die bedingte Information durch den negativen Logarithmus, so erhält man ein Maß für die Überraschung (siehe Anhang A), daß ein bestimmtes Attribut einem anderen Attribut zugeordnet wird. Tabelle 6.2 zeigt die bedingte Information aus der Tabelle 6.1 berechnet. Es kann gesehen werden, daß wenn ein Element mit Attribut "gerade" einem Element mit Attribut "stark gekrümmt" zugeordnet wird, die Überraschung unendlich groß ist. Dies ist jedoch gewünscht, da bei den Übergangswahrscheinlichkeiten definiert wurde, daß dieser Fall nicht vorkommen darf.

In einem Ansatz von [Boyer und Kak 1986] wurde die bedingte Information als Kostenfunktion zur Lösung des Zuordnungsproblem definiert. Die beste Zuordnung wird als die Zuordnung definiert, bei der die bedingte Information minimal ist. In der Arbeit von [Vosselman 1992] wird dieser Ansatz aufgegriffen und bewiesen, daß die Maximierung der gegenseitigen Information zum gleichen Ergebnis führt wie die Minimierung der bedingten Information. Jedoch werden einige Probleme, die beim Ansatz von Boyer und Kak vorhanden waren, vermieden. Insbesondere ist durch die Verwendung einer Leistungsfunktion die Möglichkeit der Zuordnungen von Wildcards ohne datenabhängige Strafkosten möglich.

Die Elemente werden nicht nur durch ihre Attribute, sondern auch durch ihre Relationen beschrieben. Daher müssen die Übergangsmatrizen auch für die Relationen aufgestellt werden. Jedoch haben nicht alle Attribute und Relationen den gleichen Stellenwert beim Zuordnungsproblem. So hat z.B. der Name einer Straße eine wesentlich stärkere Ausagekraft als die Länge der Straße. Daher sind verschiedene Attribute und Relationen auch getrennt zu betrachten. Dies bedeutet, daß die Abbildung des Zuordnungsproblem auf ein Kommunikationssystem als Mehr-Kanal-System erfolgen muß [Vosselman 1992].

6.7.2 Berechnung der Leistungsfunktion

Im folgenden wird ein Datensatz formal durch eine relationale Beschreibung D definiert (die Notation erfolgt nach [Shapiro und Haralick 1981]):

$$D = (P, R) \tag{6.4}$$

Dabei stehen $P = \{p_1, p_2, \ldots p_n\}$ für die Elemente des Datensatzes und $R = \{r_1, r_2, \ldots r_m\}$ für die Relationen zwischen diesen Elementen. Sei $D_1 = (P_A, R_A)$ und $D_2 = (P_B, R_B)$ die relationalen Beschreibungen zweier Datensätze, dann wird eine Abbildung $h : D_1 \rightarrow D_2$ gesucht, welche die beste Zuordnung zwischen den beiden Datensätzen beschreibt. Es wird eine Leistungsfunktion I_h mit Hilfe der gegenseitigen Information definiert. Diese Leistungsfunktion ist eine Funktion der Abbildung h und berechnet sich aus der Summe der gegenseitigen Information der Elemente $I_h(P_A; P_B)$ und der gegenseitigen Information der Relationen $I_h(R_A; R_B)$:

$$I_h(D_1; D_2) = I_h(P_A; P_B) + I_h(R_A; R_B) \tag{6.5}$$

Die gegenseitige Information wird aus der Differenz der Selbstinformation und der bedingten Information berechnet:

$$I_h(D_1; D_2) = I(D_1) - I_h(D_1|D_2) \tag{6.6}$$

Berechnung der Selbstinformation einer relationalen Beschreibung

Die Selbstinformation eines Attributes $attr_i$ kann durch die Wahrscheinlichkeit des Auftretens dieses Attributes errechnet werden:

$$I(attr_i) = -\log P(attr_i = w_i) \tag{6.7}$$

wobei $P(attr_i = w_i)$ die Wahrscheinlichkeit ist, daß das Attribut a_i den Wert w_i besitzt. Die Information aller Attribute eines Elementes p_k, unter der Voraussetzung, daß sie unabhängig voneinander sind, ergibt sich durch die Aufsummierung der Information aller Attribute:

$$I(p_k) = \sum_{l=1}^{L} I(attr_l) \tag{6.8}$$

wobei L die Anzahl der Attribute ist. Die gesamte Information der Elemente $P = \{p_1, p_2, \ldots p_K\}$ ergibt sich über die Aufsummierung über alle Elemente:

$$I(P) = \sum_{k=1}^{K} I(p_k) \tag{6.9}$$

$$= \sum_{k=1}^{K} \sum_{l=1}^{L_k} -\log P(attr_l = w_l) \tag{6.10}$$

wobei K die Anzahl der Elemente darstellt. Die Information einer relationalen Beschreibung ergibt sich nicht nur aus der Information der Elemente, sondern auch durch die Information der Relationen zwischen diesen Elementen. Um dies mit in die Berechnung einzubringen, werden alle Relationen als Tupel gespeichert [Vosselman 1992]. In diesen Tupeln stehen die an der Relation beteiligten Elemente sowie ein Attribut, welches angibt, ob die Relation besteht oder nicht (z.B. (p_4, p_5, wahr)). Falls die Relationen weitere Attribute haben, so können sie in das Tupel mit eingefügt werden. Zwei linienförmige Elemente, welche unter dem Winkel 30 Grad miteinander verbunden sind, würden z.B. als (p_9, p_3, wahr, 30.0) abgespeichert werden. Die Information eines einzelnen dieser Attribute einer Relation $relatt_l$ errechnet sich wieder aus der Wahrscheinlichkeit, daß dieses Attribut eintritt:

$$I(relatt_l) = -\log P(relatt_l = v_l) \tag{6.11}$$

wobei $P(relatt_l = v_l)$ die Wahrscheinlichkeit ist, daß das Attribut $relatt_l$ den Wert v_l annimmt. Die Information eines Tupels $tupel_n$ einer Relation errechnet sich durch die Aufsummierung der Einzelinformationen aller Attribute des Tupels (wieder unter der Voraussetzung, daß die Attribute voneinander unabhängig sind):

$$I(tupel_n) = \sum_{p=1}^{P} I(relatt_p) \tag{6.12}$$

Wobei P die Anzahl der Attribute des Tupels ist. Die Information aller Tupel $tupel_q$ einer Relation r_v ist dann:

$$I(r_v) = \sum_{q=1}^{Q} I(tupel_q) \tag{6.13}$$

wobei Q die Anzahl der Tupel ist. Die gesamte Information aller Relationen R ergibt sich dann aus der Aufsummierung über alle Einzelrelationen:

$$I(R) = \sum_{v=1}^{V} I(r_v) \tag{6.14}$$

$$= \sum_{v=1}^{V} \sum_{q=1}^{Q_v} I(tupel_q) \tag{6.15}$$

$$= \sum_{v=1}^{V} \sum_{q=1}^{Q_v} \sum_{p=1}^{P_q} I(relatt_p) \tag{6.16}$$

$$= \sum_{v=1}^{V} \sum_{q=1}^{Q_v} \sum_{p=1}^{P_q} -\log P(relatt_p = v_p) \tag{6.17}$$

Die Information einer relationalen Beschreibung D ergibt sich durch die Addition der Information der Attribute und der Information der Relationen:

$$I(D) = I(P) + I(R) \tag{6.18}$$

Berechnung der bedingten Information einer Beschreibung

Die bedingte Information einer Beschreibung errechnet sich analog zur Information einer relationalen Beschreibung. Jedoch werden anstatt der Wahrscheinlichkeiten, daß ein bestimmtes Attribut einen bestimmten Wert annimmt, die Übergangswahrscheinlichkeiten betrachtet. Die Aufsummierung der bedingten Informationen der Elemente muß nur über die Elemente erfolgen, die tatsächlich in der Abbildung h vorkommen:

$$I_h(D_2|D_1) = \sum_{(i,j)\in h} \sum_{l=1}^{L_{(i,j)}} -\log P(attr_{(l,i)}|attr_{(l,j)}) + \sum_{(i,j)\in h} \sum_{q=1}^{Q_{(i,j)}} \sum_{p=1}^{P_q} -\log P(relatt_{(p,i)}|relatt_{(p,j)}) \tag{6.19}$$

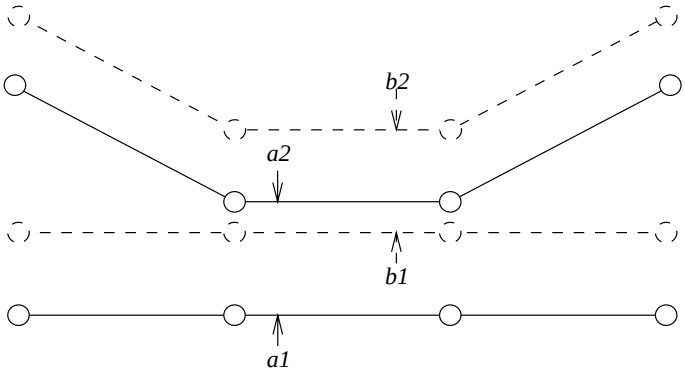
6.8 Globale vs. lokale Optimierung

Nachdem die Bewertungsfunktion der Zuordnungen definiert wurde, kann jetzt aus der mehrdeutigen Liste von potentiellen Zuordnungspaaren die eindeutige Kombination von Zuordnungen gesucht werden, die der besten Gesamtzuordnung entspricht. Wie bereits dargestellt, handelt es sich hierbei um ein Problem, dessen Suchraumgröße exponentiell zunehmen kann. Um die Zeitkomplexität zu verringern, versuchen existierende Verfahren oftmals in einem ersten Schritt Startzuordnungen zu finden (siehe Kapitel 6.1). Um die Startzuordnungen zu bestimmen, wird nach einzelnen Zuordnungen gesucht, deren Leistungsfunktion einen hohen Wert besitzt. Diese Zuordnungen dienen dann als Vorinformation, um die restlichen Zuordnungen zu bestimmen. Hierbei handelt es sich jedoch nur um einen lokalen Optimierungsansatz, da nur die Leistung einzelner Zuordnungen betrachtet wird, anstatt die des gesamten Datensatzes.

Dies kann zu Fehlzuordnungen führen. In Abbildung 6.9 soll dies an einem Beispiel verdeutlicht werden. Der Datensatz A ist mit durchgezogenen Linien und der Datensatz B mit gestrichelten Linien gezeichnet. An der rechten Seite sind fiktive Werte der Leistungsfunktion dargestellt, wie sie in diesem Fall auftreten können. Bei einer lokalen Optimierung würde die Zuordnung $a_2 \rightarrow b_1$ einen hohen Wert aufweisen, da die Länge, Form und Positionen der beiden Elemente fast identisch ist. Möglicherweise könnten auch die jeweils topologischen Nachbarn einander zugeordnet werden, und damit die Leistung der Zuordnung noch erhöhen. Diese Zuordnung ist jedoch offensichtlich falsch und würde verhindern, daß die Elemente a_1 und b_2 richtig zugeordnet werden können. Bei einem globalen Optimierungsansatz werden jedoch alle möglichen Kombinationen getestet. Dadurch ist das Verfahren unabhängig von Startparametern und führt zu einer eindeutigen Lösung.

6.9 Suchbäume

In [Vosselman 1992] wird ein Verfahren zur Aufstellung der Suchbäume vorgeschlagen. Dieses Verfahren betrachtet jedoch nur 1 : 1 Zuordnungen zwischen den Elementen der Datensätze und muß für die Zuord-



Lokal optimiert:

a2 -> b1	Leistung = 10
Gesamt:	10

Global optimiert:

a2 -> b2	Leistung = 7
a1 -> b1	Leistung = 7
Gesamt:	14

Abbildung 6.9: Globale und lokale Optimierung

nung von raumbezogenen Daten erweitert werden, da hier $n : m$ Zuordnungen zu betrachten sind. Abbildung 6.10 zeigt ein Beispiel für die obersten drei Ebenen eines Suchbaumes. Als Wurzel des Suchbaumes wird immer die "leere" Zuordnung $* \rightarrow *$ verwendet. Dies ist der Startzustand des Suchproblems und repräsentiert diejenige Gesamtzuordnung, bei der überhaupt keine Elemente einander zugeordnet wurden. Da diese Zuordnung keine Leistung beiträgt, ist die Gesamtleistung Null.

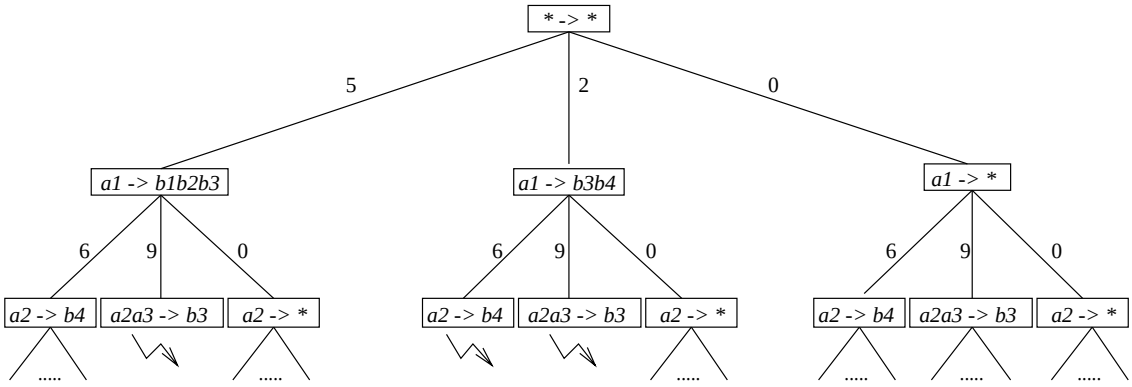


Abbildung 6.10: Aufstellen des Suchbaums

In der nächsten Ebene wird ein Element aus A herausgegriffen und alle $n : m$ Zuordnungen dargestellt, welche dieses Element beinhalten. In dem Beispiel wurde das Element a_1 gewählt. Die den Zuordnungen entsprechenden Leistungen werden an den Kanten aufgetragen. Die Zuordnung $a_1 \rightarrow *$ ergibt wiederum das Leistungsmaß Null, da dies keine echte Zuordnung ist, sondern bedeutet, daß das Element a_1 keinem Element aus B zugeordnet werden konnte. Auf der nächsten Ebene werden dann alle Zuordnungen dargestellt, die ein weiteres Element aus A enthalten, welches noch nicht verwendet wurde. Hierbei ist jedoch zu beachten, daß keine Widersprüche entstehen dürfen. Da das Ergebnis eine eindeutige Zuordnung ergeben soll, dürfen nur solche Elemente auf der linken und rechten Seite der Zuordnung auftauchen, welche nicht bereits höher im Zuordnungsbaum verwendet wurden. So braucht die Zuordnung $a_2 a_3 \rightarrow b_3$ im linken und mittleren Teilbaum nicht weiter betrachtet werden, da das Element b_3 bereits weiter oben zugeordnet wurde. Jeder Knoten im Baum, der zu einem Widerspruch führt, darf nicht weiter expandiert werden.

Die Forderung, daß die Endzuordnung eine eindeutige Liste aus Einzelzuordnungen darstellt, muß nicht in jedem Fall eingehalten werden. Liegen beispielsweise zwei Datensätze vor, bei denen im einen Datensatz Straßen durch ihre Mittelachse repräsentiert werden und im anderen Datensatz durch ihre Randbegrenzung, ist es sinnvoll, zuzulassen, daß den Elementen, welche die Mittelachsen darstellen, zwei Elemente gleichzeitig zugeordnet werden dürfen. Dies kann sehr einfach im Algorithmus implementiert werden. Ein Widerspruch ensteht in diesem Fall erst dann, wenn ein Element einer Zuordnung mehr als einmal bereits weiter oben im Baum vorkommt.

6.10 Aufteilung des Suchraumes

Nachdem der Suchbaum erstellt ist, kann mit Hilfe eines Suchverfahrens die Kombination von Zuordnungen gesucht werden, die der besten Gesamtzuordnung entspricht. Wegen des exponentiellen Wachstums des Suchraums kann dies zu langen Rechenzeiten führen. Bei einer näheren Betrachtung der Zuordnungen zeigt sich jedoch, daß die Menge der potentiellen Zuordnungspaare in voneinander unabhängig optimierbare Teilmengen aufgeteilt werden kann. Dies kann zum einen durch eine Aufteilung des zuzuordnenden Gebietes in Teilgebiete erreicht werden und zum anderen durch eine Technik, welches im folgenden mit dem Namen *Clusterbildung* bezeichnet wird.

6.10.1 Bildung von Teilgebieten

Wegen des Raumbezugs der Daten ist es nicht notwendig den gesamten Suchraum in einem Schritt zu optimieren. Hierzu wird das Gebiet, in dem die Daten liegen, in einzelne Teilgebiete (Patches) aufgeteilt, welche unabhängig voneinander zu optimieren sind. Bei der Bildung der Teilgebiete muß darauf geachtet werden, daß die Ausdehnung groß genug ist, damit die Zuordnung der Daten in einem Teilgebiet unabhängig von den anderen Teilgebieten ist. Diese Forderung ist jedoch nicht für die Ränder der Teilgebiete erfüllbar. Bei der Zuordnung von Daten in den Randbereichen der Teilgebiete kann es zu Fehlzuordnungen kommen, weil die Informationen aus den Nachbargebieten nicht verfügbar sind. Dieses Problem kann mit Hilfe von überlappenden Teilgebieten gelöst werden. In den Überlappungsgebieten werden die Daten mehrfach zugeordnet und anschließend gegeneinander abgeglichen. Abbildung 6.11 zeigt eine mögliche Aufteilung eines Gebietes in einzelne Teilgebiete.

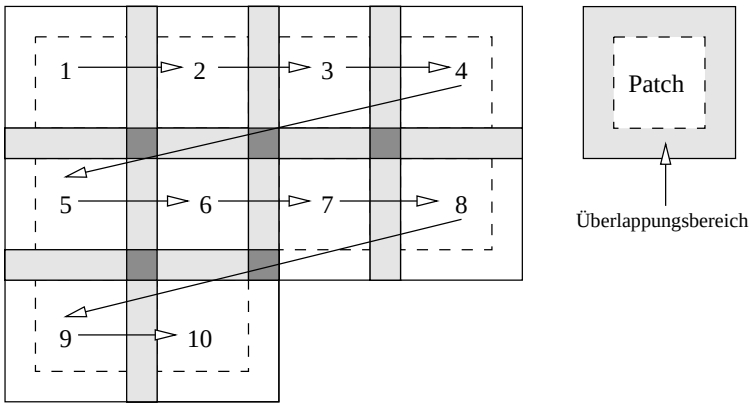


Abbildung 6.11: Aufteilung des Suchraumes in Teilgebiete

6.10.2 Clusterbildung

Ob ein Element a_i des Datensatzes A einem anderen Element b_j des Datensatzes B zuzuordnen ist, hängt aufgrund der relationalen Zuordnungstechnik nicht nur davon ab, welche Attribute es besitzt, sondern auch mit welchen anderen Elementen es in Relation steht. Als Beispiel soll eine Zuordnung zwischen linienförmigen Elementen betrachtet werden, deren relationales Ähnlichkeitsmaß aus der topologischen Relation "verbunden" berechnet wird. Dies bedeutet, daß das Ähnlichkeitsmaß I der Zuordnung $a_i \rightarrow b_j$ davon abhängt, ob die topologischen Partner der Elemente zugeordnet wurden:

$$I(a_i, b_j) = f(a_i, b_j) + g(Partner(a_i), Partner(b_j)) \tag{6.20}$$

Da die Zuordnung der topologischen Partner jedoch wieder von den Partnern der Partner abhängt usw., setzt sich die Abhängigkeit des Ähnlichkeitsmaßes immer weiter fort:

$$\begin{aligned} I(a_i, b_j) \quad = \quad & f(a_i, b_j) \\ & + g(Partner(a_i), Partner(b_j) \\ & + h(Partner(Partner(a_i)), Partner(Partner(b_j))) \\ & + \ldots \end{aligned} \tag{6.21}$$

Durch diese globale Abhängigkeit entsteht der exponentielle Anstieg des Zeitverhalten des Zuordnungsalgorithmus. Bei der Betrachtung von realen Daten, kann jedoch gesehen werden, daß diese globale Abhängigkeit zwar theoretisch besteht, daß jedoch eine z.T. lokale Optimierung unter bestimmten Voraussetzungen zum gleichen Ergebnis führen kann. Abbildung 6.12 zeigt einen Datensatz, in dem die verschiedenen Stufen der Abhängigkeit graphisch eingezeichnet sind. Der dunkelste Bereich zeigt ein Element a_i, zu dem ein Partnerelement zugeordnet werden soll. Stufe 1 und Stufe 2 geben jeweils die topologischen Nachbarn bzw. die Nachbarn der Nachbarn an. Je nachdem, wie die Struktur des Datensatzes ist, entstehen unterschiedliche Formen. Dies führt ebenfalls zu einer Teilgebietsbildung in den Datensätzen, jedoch ist hier die Form der Teilgebiete nicht fest vorgegeben, sondern eine Funktion der Daten selbst. Die Gebiete, die dabei entstehen, nennen wir *Cluster*, das Verfahren *Clusterbildung*.

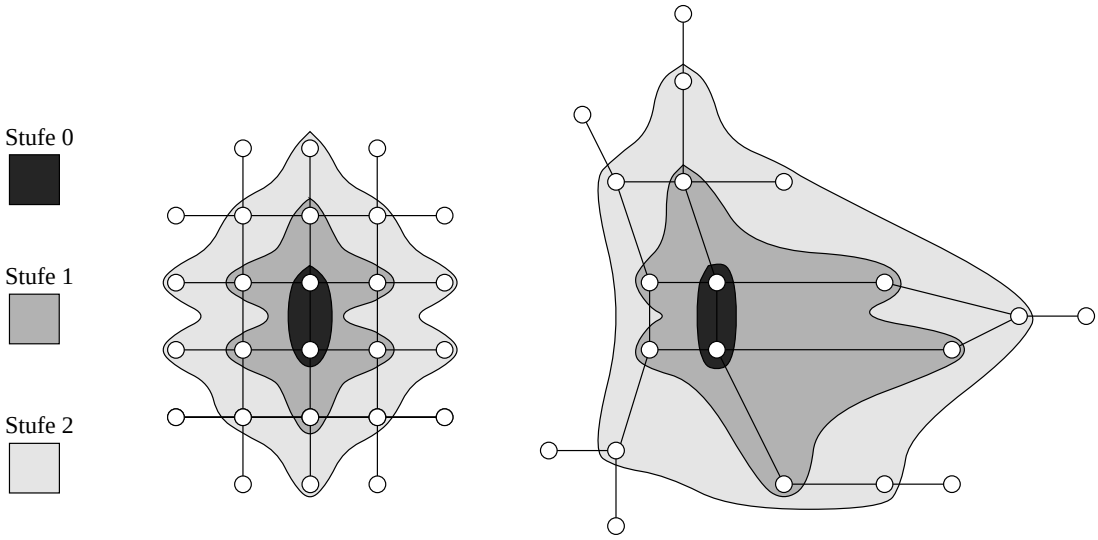


Abbildung 6.12: Clusterbildung

Bei der Zuordnung zwischen zwei Datensätzen ist es möglich, die optimale Zuordnung zu finden, wenn die Cluster beliebig groß werden können. Durch das exponentielle Zeitverhalten der Suchalgorithmen ist dies jedoch nicht praktisch durchführbar. Wir schlagen daher zur Lösung des Zuordnungsproblem vor, den Suchraum in einzelne Cluster der Stufe 2 aufzuteilen und sukzessive nacheinander zu optimieren. Hierbei handelt es sich um eine Heuristik, die nicht unbedingt in jedem Fall zur optimalen Lösung führen muß, jedoch den Suchraum sehr stark einschränkt und wie später aufgezeigt wird, zu guten Ergebnissen führt.

Es wird zu jedem Element des Datensatzes A ein Cluster der Stufe 2 gebildet. Mit Elementen eines Datensatzes werden dabei auch immer die logischen Elemente bezeichnet, die durch das Buffer Growing entstanden sind (siehe Kapitel 6.4). Dies bedeutet, daß ein Element a_i in mehreren Clustern vorkommen kann, sofern es in mehreren logischen Elementen enthalten ist.

$$Cluster_{\text{Stufe}_2}(a_i) = a_i \cup Partner(a_i) \cup Partner(Partner(a_i)) \tag{6.22}$$

Zu allen Elementen des $Cluster_{\text{Stufe}_2}(a_i)$ werden nun alle potentiellen Zuordnungspartner aus dem Datensatz B gesucht. Das Ergebnis ist eine Liste von potentiellen Zuordnungspaaren $a_{ik} \rightarrow b_{jk}$. Aus dieser Liste wird die optimale Zuordnung mit Hilfe eines Suchverfahrens (siehe unten) bestimmt. Bei dieser Optimierung gilt als Randbedingung, daß die ursprüngliche Zuordnung $a_i \rightarrow bj$ (also die Stufe 0 des Clusters) in der Lösung enthalten sein muß. Diese Vorgehensweise wird für alle Cluster des Datensatzes A durchgeführt. Abbildung 6.13 zeigt die Situation nach der Berechnung von vier Clustern.

Das Ergebnis der Optimierung eines Clusters wird in einer globalen Ergebnisliste abgespeichert. Sind alle Cluster berechnet, enthält diese Liste das Ergebnis der Gesamtzuordnung. Entsteht nun der Fall, wie in Abbildung 6.13, daß sich ein Cluster mit einem bereits berechneteten Cluster überschneidet, so werden alle Zuordnungen dieser Schnittmenge aus der globalen Ergebnisliste wieder entfernt, und dem Zuordnungsprozess erneut zugewiesen. Daher kann es vorkommen, daß ein bereits optimiertes Cluster komplett wieder verworfen wird, um zu testen ob ein anderes Cluster zu einem besseren Gesamtergebnis führt.

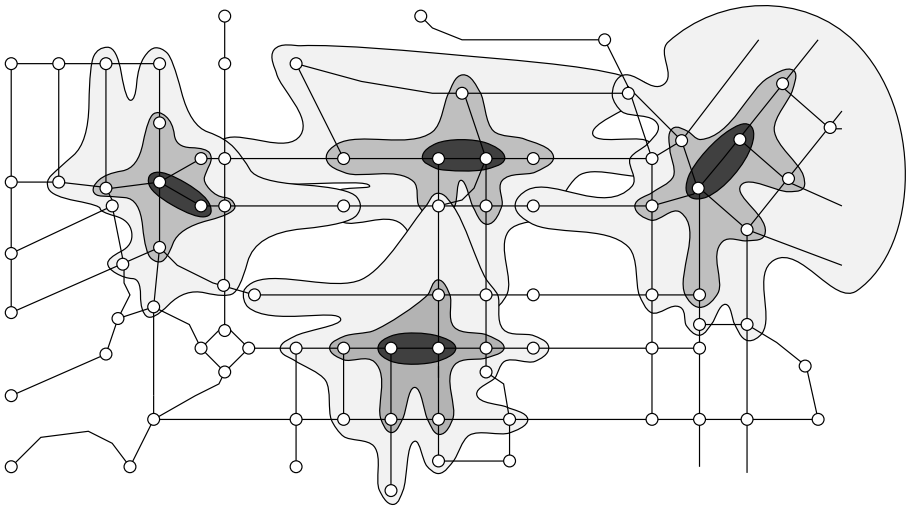


Abbildung 6.13: Sukzessive Berechnung der Cluster

Diese Vorgehensweise optimiert also die Zuordnung zwischen zwei Datensätzen in mehreren Teilschritten. Bei der Optimierung eines Teilschrittes wird versucht die Gesamtleistung der kompletten Zuordnung zu maximieren. Dadurch, daß ein bereits zugeordnetes Element wieder aus der Gesamtzuordnung entfernt und dem Zuordnungsprozess erneut unterworfen werden kann, funktioniert dieses Verfahren nach dem *Trial and Error* - Prinzip. Ob das Verfahren die optimale Lösung findet oder nur eine suboptimale Lösung, hängt unter anderem davon ab, in welcher Reihenfolge die Cluster optimiert werden. Als Reihenfolge bietet sich an, die Cluster nach der Größe zu sortieren und die kleinsten Cluster zuerst zu optimieren. Je kleiner das Cluster ist, umso wahrscheinlicher ist es, daß die gefundene Teiloptimierung der endgültigen Lösung entspricht. Da die Optimierung eines Clusters davon abhängt, wie die Nachbarn des Clusters zugeordnet sind, ist es vorteilhaft zuerst diejenigen Cluster zu optimieren, bei denen eine richtige Lösung wahrscheinlicher ist, um somit die Zuordnung der folgenden größeren Cluster zu erleichtern.

6.11 Suchverfahren

Der letzte Teil der Aufgabe ist es, aus den Clustern die Kombination von Zuordnungen zu berechnen, bei der die Leistungsfunktion den maximalen Wert annimmt. Es handelt sich hierbei um ein kombinatorisches Problem, welches mit einem Baumsuchverfahren gelöst werden kann. Hierzu wird ausgehend von einem Startknoten der Suchbaum in einer bestimmten Reihenfolge sukzessive aufgebaut. Bei der Wahl der Reihenfolge wird grundsätzlich zwischen Tiefensuche (depth-first search) und Breitensuche (breadth-first search) unterschieden [Charniak und McDermott 1985]. Bei der Tiefensuche wird der Knoten expandiert, welcher am weitesten links und am weitesten unten im Baum steht. Bei der Breitensuche werden dagegen immer zuerst alle Knoten eines Levels von links nach rechts expandiert. Abbildung 6.14 zeigt die beiden Verfahren einander gegenübergestellt. Die Zahl neben dem Knoten gibt die Reihenfolge der Expandierung an. Welches der beiden Suchverfahren geeigneter ist, hängt von der Problemstellung und der Form des Baumes ab. Ein Vorteil der Tiefensuche ist es jedoch, daß sie im allgemeinen weniger Speicherplatz benötigt als die Breitensuche, da hier weniger nicht expandierte Knoten abgespeichert werden müssen.

Oftmals ist es nicht nötig, einen Suchbaum vollständig aufzubauen, da bei bestimmten Knoten im Baum bereits im voraus berechnet werden kann, daß der unter diesem Knoten liegende Teilbaum keine optimale Lösung enthalten kann. Verfahren die solche Methoden bereitstellen, werden auch heuristische Verfahren genannt. Häufig angewandte Methoden sind z.B. Bergsteigen (hill-climbing search), Bestensuche (best-first search) oder Strahlensuche (beam search) [Winston 1987]. Diese Verfahren sind in der Lage eine Kostenfunktion zu minimieren, ohne dabei in jedem Fall den gesamten Baum aufzubauen. Die Verfahren expandieren die Knoten eines Baumes in einer bestimmten Reihenfolge, die abhängig ist von der Bewertung der Kanten. Wird dabei ein Blatt erreicht, werden die Gesamtkosten dieses Pfades abgespeichert. Bevor nun ein Knoten expandiert wird, wird zuerst getestet, ob die bis dahin angelaufenen Kosten höher sind als die Kosten eines bereits gefundenen vollständigen Pfades. Ist dies der Fall, braucht dieser Knoten nicht expandiert werden, da bereits der bis dahin erreichte

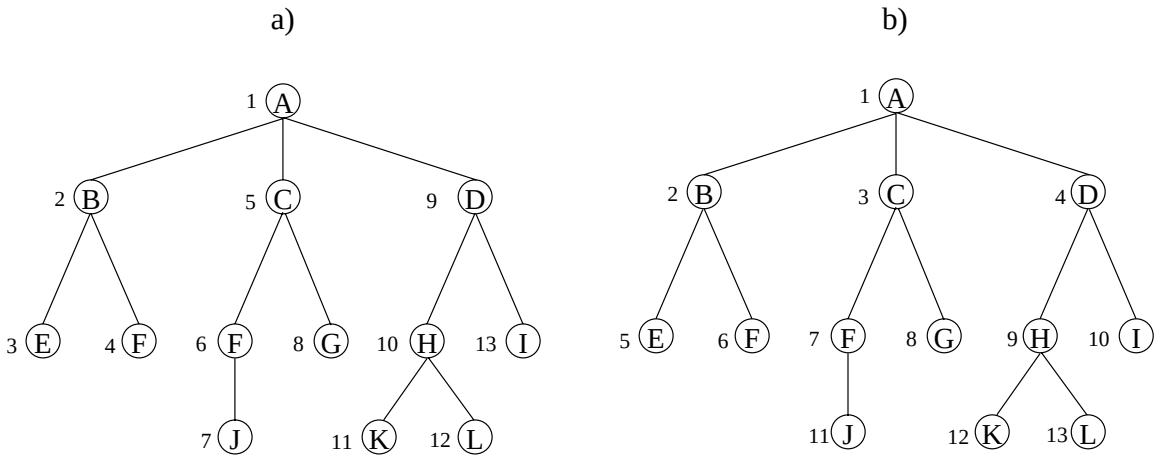


Abbildung 6.14: a) Tiefensuche und b) Breitensuche

Teilpfad höhere Kosten verursacht hat als ein bereits gefundener kompletter Pfad. Bei einer Leistungsfunktion kann diese Art von Verfahren jedoch nicht angewandt werden, da die Aussage, daß bei einem Teilknoten weniger Leistung angefallen ist als bei einem bereits gefundenen Pfad, keine Angabe darüber macht, wieviel Leistung im verbleibenden Restpfad noch erreicht wird.

Die oben genannten Verfahren können auch für Leistungsfunktionen verwendet werden, wenn eine Abschätzung der Leistung eines Teilpfades erfolgen kann. Im folgenden wird die Arbeitsweise eines Bergsteige-Algorithmus [Winston 1987] beschrieben, welcher so angepaßt wird, daß er für Leistungsfunktionen verwendet werden kann. Der Algorithmus verwendet die Leistungsfunktion $f(n)$, die die Leistung eines Pfades im Baum von der Wurzel bis zu einem Blatt angibt. Diese Leistungsfunktion kann für jeden Knoten in die zwei Funktionen $g(n)$ und $h(n)$ aufgeteilt werden. An einem bestimmten Knoten n gibt $g(n)$ die bereits erzielte Leistung von der Wurzel zum Knoten n an, und $h(n)$ das Maximum der Leistung des Restpfades. Während des Erstellens eines Suchbaumes kann jederzeit $g(n)$ berechnet werden, wogegen $h(n)$ nur mit $h^*(n)$ geschätzt werden kann. Ein Beispiel für die Berechnung der Abschätzung $h^*(n)$ wird in Kapitel 7 aufgezeigt. Mit Hilfe dieser Abschätzung kann zu jedem Knoten eine Abschätzung der Gesamtleistung erfolgen:

$$f^*(n) = g(n) + h^*(n) \tag{6.23}$$

Der Bergsteige-Algorithmus ist eine Sonderform der Tiefensuche. Er expandiert immer den Knoten im Baum, bei dem der Anstieg der Leistung $f^*(n)$ am größten ist. Wenn ein Pfad von der Wurzel zu einem Blatt berechnet wurde, kann die Gesamtleistung des Pfades exakt bestimmt werden. Bevor ein weiterer Knoten expandiert wird, wird zuerst getestet, ob die Abschätzung der Leistung f^* höher ist als die höchste Gesamtleistung eines Pfades, die bis dahin erzielt wurde. Nur wenn das der Fall ist, muß der Knoten expandiert werden. Voraussetzung hierzu ist allerdings, daß die Abschätzung der Leistung des Restweges nie unterschätzt werden darf, was bedeutet:

$$h^*(n) \geq h(n) \tag{6.24}$$

Nur falls diese Bedingung erfüllt ist, wird die optimale Lösung auf jeden Fall gefunden. Auf der anderen Seite soll jedoch $h^*(n)$ so wenig wie möglich überschätzt werden, da sonst zuviele Knoten expandiert werden, die zu keiner optimalen Lösung führen. Im folgenden wird eine formale Beschreibung dieses Algorithmus dargestellt. In Kapitel 6.9 wurde bereits der logische Aufbau eines Suchbaumes für das Zuordnungsproblem beschrieben. Es ist jedoch noch offen gelassen, in welcher Reihenfolge der Aufbau erfolgt und wie Widersprüche im Baum erkannt werden. Es handelt sich um einen rekursiven Algorithmus, der mit zwei Parametern aufgerufen wird: $Z = \{z_1, z_2, \ldots z_n\}$ ist die Menge der Zuordnungen eines Clusters (siehe Kapitel 6.10) und in L werden die Zuordnungen des gerade aktuellen Weges im Baum während der Berechnung gespeichert.

(0) **Prozedur Baumsuche**(Z, L)**:**
(1) **IF** $Z == \emptyset$:
(2) **IF** $f(L) > Max\_Leistung$:
(3) $Max\_Leistung = f(L)$
(4) $Beste\_Gesamtzuordnung = L$
(5) Beende Prozedur
(7) Wähle aus Z die Zuordnung $z_i = a_{i_1} a_{i_2} \ldots a_{i_n} \rightarrow b_{i_1} b_{i_2} \ldots b_{i_m}$
bei der die Abschätzung der Gesamtleistung $f^*(z_i)$ am stärksten ansteigt
(8) Entferne z_i aus Z
(9) $Z\_Kopie = Z$; $L\_Kopie = L$
(10) $L = L \cup z_i$
(11) $A_i = \{a_{i_1}, a_{i_2} \ldots a_{i_n}\}$
(12) $B_i = \{b_{i_1}, b_{i_2} \ldots b_{i_m}\}$
(13) **FOR** $z_j = a_{j_1} a_{j_2} \ldots a_{j_n} \rightarrow b_{j_1} b_{j_2} \ldots b_{j_m} \in Z$:
(14) $A_j = \{a_{j_1}, a_{j_2} \ldots a_{j_n}\}$
(15) $B_j = \{b_{j_1}, b_{j_2} \ldots b_{j_m}\}$
(16) **IF** $(A_i \cap A_j \neq \emptyset)$ **OR** $(B_i \cap B_j \neq \emptyset)$:
(17) Entferne z_j aus Z
(18) **IF** $f^*(Z) > Max\_Leistung$: **Baumsuche**(Z, L)
(19) $Z = Z\_Kopie$; $L = L\_Kopie$
(20) **IF** $f^*(Z) > Max\_Leistung$: **Baumsuche**(Z, L)

Der rekursive Algorithmus beginnt mit der Überprüfung der Abbruchbedingung (1). Falls die Menge der Zuordnungen abgearbeitet ist, hat die Baumsuche ein Blatt im Baum erreicht. Die durch den Pfad von der Wurzel bis zum Blatt erzielte Leistung wird mit der Maximal-Leistung, die während der Berechnung des Baumes bereits erreicht wurde, verglichen (2). Ist sie höher, wird diese Leistung als neue Maximal-Leistung gesetzt (3) und der Pfad von der Wurzel zum Blatt wird als die beste Gesamtzuordnung gespeichert (4).

Nun beginnt der eigentliche Aufbau des Suchbaumes. Aus der Menge der Zuordnungen Z wird diejenige Zuordnung z_i herausgegriffen, bei der die Abschätzung des Anstieges der Gesamtleistung $f^*(z_i)$ am höchsten ist (7). Dies entspricht einem Bergsteige-Algorithmus. Die Idee beim Bergsteigen ist es, daß das Maximum der Gesamtleistung in der Richtung des höchsten Zuwachses der Leistungsfunktion des nächsten Knotens liegt. Dies ist eine Heuristik und führt oftmals nur zu einem lokalen Maximum, jedoch ist die Wahrscheinlichkeit hoch, daß das globale Maximum bereits in einem frühen Stadium des Aufbaus des Suchbaumes gefunden wird. Da jede Zuordnung nur einmal verwendet werden darf, wird z_i aus der Menge der Zuordnungen Z entfernt (8). Eine Kopie der Menge aller Zuordnungen Z und der Menge der Zuordnungen, welche auf dem Pfad im Baum abgespeichert wurden, werden in den Variablen *Z_Kopie* und *L_Kopie* für einen späteren rekursiven Aufruf des Algorithmus gespeichert (9). Die gewählte Zuordnung z_i, wird in die Liste der Zuordnungen des aktuellen Weges L eingetragen (10). Der Suchraum wird dadurch eingeschränkt, daß jedes Teilelement einer Zuordnung nur einmal verwendet werden darf. So ist es z.B. nicht erlaubt, die Zuordnungen $a_1 \rightarrow b_1 b_2$ und $a_2 \rightarrow b_2 b_3$ gleichzeitig zu verwenden, da sonst das Element b_2 zweimal zugeordnet wird. Daher werden nun alle Zuordnungen z_j aus Z entfernt, die ein Teilelement besitzen, welches in der Zuordnung z_i enthalten ist (13-17).

Im letzten Teil ruft sich der Algorithmus zweimal rekursiv auf. Diese beiden Aufrufe entsprechen den beiden Fällen, daß die Zuordnung z_i in der Gesamtzuordnung vorkommt oder nicht. In (18) wird die Baumsuche mit den neu berechneten Werten Z und L aufgerufen (die Zuordnung z_i kommt in der Gesamtzuordnung vor) und in (20) mit den vorher gespeicherten Werten *Z_Kopie* und *L_Kopie* (die Zuordnung z_i kommt in der Gesamtzuordnung nicht vor). Der rekursive Aufruf erfolgt jedoch nur, wenn die geschätzte Leistung des dadurch erzeugten Knotens höher ist als die maximale Leistung, die bisher erreicht wurde. Da die Schätzung so definiert wurde, daß die Leistung zwar überschätzt, jedoch niemals unterschätzt werden darf, ist gewährleistet, daß nur solche Knoten nicht expandiert werden, die auf keinen Fall zu einem optimalen Pfad im Baum gehören.

Im folgenden wird die Arbeitsweise des Algorithmus an einem Beispiel diskutiert. Abbildung 6.15 zeigt den Suchbaum, der aus der Menge der Zuordnungen $Z = \{a_1 \rightarrow b_1 b_2 b_3,\ a_5 \rightarrow b_7,\ a_1 \rightarrow b_3 b_4,\ a_5 \rightarrow b_4\}$ erzeugt wurde. Die Knoten im Baum sind durch Rechtecke, die Blätter durch grau hinterlegte Rechtecke dargestellt. Die Zuordnungen sind mit einer Leistungsfunktion bewertet. Die Tabelle in der Abbildung gibt die Werte der Leistungsfunktion der einzelnen Zuordnungen an. Um das Beispiel zu vereinfachen, wurde ein merkmalsbasierter Zuordnungsansatz gewählt. Im Gegensatz zur relationalen Zuordnung kann bei der merkmalsbasierten Zuordnung die Leistungsfunktion schon vor dem Aufbau des Baumes für jede Zuordnung bestimmt werden, da die

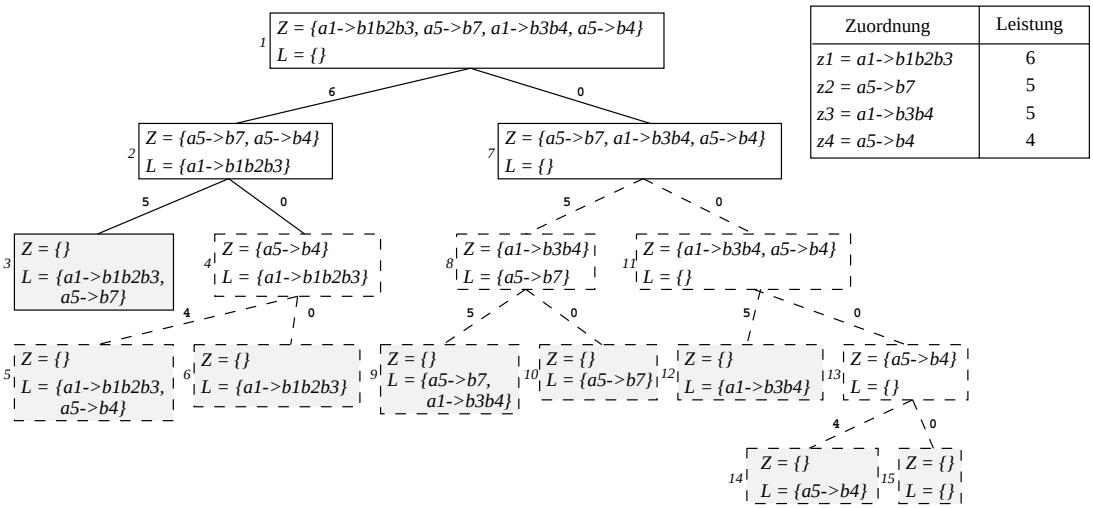


Abbildung 6.15: Beispiel für den Algorithmus zur Baumsuche

Belegung der Attribute unabhängig von der Kombination der Zuordnungen ist. In Kapitel 7 wird die Berechnung der Leistungsfunktion des relationalen Anteils der Daten an einem Beispiel beschrieben. Der Aufbau des Baumes ist unabhängig von der Art der Berechnung der Leistungsfunktion, jedoch läßt sich die verwendete heuristische Strategie leichter mit einem merkmalsbasierten Zuordnungsansatz erklären, da hier die Leistungen der Zuordnungen vor dem Erzeugen des Baumes berechnet werden können.

Der Aufbau des Baumes erfolgt von der Wurzel mit den Startparametern Z und L, wobei Z alle Zuordnungen eines Clusters enthält und L die leere Menge ist. In L werden die bereits gefundenen Zuordnungen des aktuellen Pfades gespeichert. Die kursive Zahl neben den Rechtecken gibt die Reihenfolge der Erzeugung der Knoten an. Aus Z wird die Zuordnung mit der höchsten Leistung gewählt, welche bei der Wurzel die Zuordnung $z_1 = a_1 \rightarrow b_1b_2b_3$ ist. Nun können zwei Fälle unterschieden werden. Der erste Fall ist der, daß diese Zuordnung in der Gesamtzuordnung vorkommt und entspricht dem linken Sohn des gerade aktuellen Knoten, wogegen beim zweiten Fall diese Zuordnung verworfen wird und nicht in der Gesamtzuordnung vorkommt (rechter Sohn). Wird die gewählte Zuordnung verwendet, so wird sie aus Z entfernt. Alle verbleibenden Zuordnungen aus Z, welche ein Element dieser Zuordnung enthalten, werden ebenfalls aus Z entfernt. So wird bei der Erzeugung des Knoten 2 die Zuordnung z_3 aus Z entfernt, da die Elemente a_1 und b_3 in der Zuordnung z_1 vorkommen. Die Leistung wird an der Kante aufgetragen und z_1 wird in L abgespeichert. Wird die Zuordnung nicht verwendet (rechter Sohn), so wird sie lediglich aus Z entfernt und nicht in L eingetragen. Die Leistung ist in diesem Fall gleich Null, da auch keine Zuordnung stattgefunden hat.

Durch diese Vorgehensweise wird immer die Zuordnung verwendet, bei der die Gesamtleistungsfunktion am stärksten ansteigt (Bergsteige-Baumsuche). Mehrfachzuordnungen von Elementen in der Gesamtzuordnung werden dadurch verhindert, daß jedesmal, wenn eine Zuordnung in L eingetragen wird, alle Zuordnungen aus Z entfernt werden, die gleiche Elemente enthalten. Der Algorithmus versucht immer den am weitesten links und am weitesten unten stehenden Knoten zu expandieren (Tiefensuche). Da Zuordnungen auch verworfen werden können, ist die Zuordnung von Elementen zu Wildcards implizit realisiert, ohne daß diese explizit in der Zuordnungsliste auftauchen. Je weiter rechts unten ein Knoten im Baum steht, desto mehr Wildcards wurden bereits zugeordnet. Daher ist es auch wahrscheinlich, daß Gesamtzuordnungen mit einer hohen Gesamtleistung links im Baum zu finden sind, und durch die Tiefensuche schnell gefunden werden können.

In der Abbildung ist der gesamte Suchbaum dargestellt. Mit Hilfe einer Abschätzung der Gesamtleistung ist es aber nicht notwendig, den kompletten Suchbaum aufzubauen, um die optimale Lösung zu finden. Da in dem Beispiel ein merkmalsbasierter Zuordnungsansatz verwendet wurde, kann die maximale Gesamtleistung, die von einem Knoten aus erreicht werden kann, leicht abgeschätzt werden. Die maximale Gesamtleistung eines Knotens errechnet sich aus der Summe der Leistungen von der Wurzel bis zu diesem Knoten zuzüglich des Maximums der unterhalb dieses Knoten stehenden Teilpfade. Dieses Maximum kann auf keinen Fall höher sein als die Summe der Leistungen der Zuordnungen in Z. Nachdem Knoten Nummer 3 erzeugt wurde, ist eine Lösung des Zuordnungsproblems sowie dessen Gesamtleistung bekannt. Im weiteren müssen nur noch die Knoten expandiert werden, deren Abschätzung der Gesamtleistung höher ist als die bereits erreichte Maximal-Leistung. Im Beispiel bedeutet dies, daß die Knoten 4, 8 und 11 nicht erzeugt werden müssen, da im voraus berechnet werden kann,

daß die maximale Gesamtleistung dieser Knoten geringer ist als die Leistung, die durch den Pfad von der Wurzel zu Knoten Nummer 3 erreicht wurde. Alle Knoten, die nicht erzeugt werden, sind in der Abbildung mit gestrichelten Linien dargestellt.Dies macht deutlich, wie stark der Suchraum eingeschränkt werden kann. Aus vier Zuordnungen z_1 bis z_4 sollte die Kombination gefunden werden, bei der die Gesamtleistung am höchsten ist. Hätte man alle Kombinationen (zuzüglich der Wildcard-Zuordnungen) erzeugt, hätte man insgesamt $2^5 - 1 = 31$ Knoten erhalten. Durch das Aufdecken von Widersprüchen durch Mehrfachzuordnungen müssen nur noch 15 Knoten erzeugt werden. Mit Hilfe einer Abschätzung der Gesamtleistung konnte die Evaluierung des Suchraums auf 4 Knoten begrenzt werden.

6.12 Qualitätsmaße der Zuordnungen

Die Zuordnung zweier raumbezogener Datensätze wurde auf ein Kommunikationssystem übertragen. Eine Zuordnung wird mit Hilfe der gegenseitigen Information bewertet:

$$I(a_i; b_j) = I(a_i) - I(a_i|b_j) \tag{6.25}$$

Eine Zuordnung wird dann als gut bewertet, wenn die gegenseitige Information hoch ist. Die Eigeninformation eines Elementes $I(a_i)$ ist unabhängig von der Zuordnung und kann direkt aus den Stichproben bestimmt werden. Daher ist die Eigeninformation eines Elementes konstant und eine niedrige bedingte Information steht für eine gute Zuordnung. Dies bedeutet, daß das Maß der Überraschung so klein wie möglich sein muß bzw. es soll so wenig Information wie möglich übertragen werden. Die bedingte Information kann daher als ein Qualitätsmaß für die Zuordnungen genutzt werden. Je niedriger die bedingte Information ist, umso höher ist die Wahrscheinlichkeit, daß die Zuordnung richtig ist.

Da es nicht auszuschließen ist, daß bei der automatischen Zuordnung von raumbezogenen Daten auch Fehlzuordnungen entstehen, muß eine Möglichkeit der interaktiven Nachkontrolle bereitgestellt werden. Je unterschiedlicher die zugrundeliegenden Datenmodelle sind, desto häufiger ist mit Fehlzuordnungen zu rechnen. Um diese Nachkontrolle zu erleichtern, kann die bedingte Information der Zuordnungen z.B. farblich kodiert am Bildschirm dargestellt werden. Dies ermöglicht einem Operateur, die Ergebnisse schnell zu kontrollieren, da bei den Zuordnungen mit Fehlern zu rechnen ist, bei denen ein hohes Maß an Überraschung entsteht.

Dieses Qualitätsmaß läßt sich auch für die gesamte Zuordnung berechnen. Hierzu wird dir durchschnittliche bedingte Information pro Zuordnung berechnet. Sie gibt den durchschnittlichen Verlust an Information pro Zuordnung an. Ein hoher Wert dieses Maßes ist ein Hinweis darauf, daß sich die Datensätze stark unterscheiden und daher keine gute Gesamtzuordnung gefunden werden konnte. Ist die durchschnittliche bedingte Information pro Zuordnung gleich Null, so liegt ein idealer Kanal vor und die beiden Datensätze enthalten exakt die gleiche Information. Durch die Betrachtung dieses Informationsmaßes ist es möglich, die Qualität der gesamten Zuordnung global zu bewerten.

Bis jetzt wurde immer nur die Qualität der Zuordnungen betrachtet. Oftmals stellt sich die Aufgabe, Datensätze nach ihrer Qualität zu beurteilen. Auch hier liefert die durchschnittliche bedingte Information pro Zuordnung ein geeignetes Maß. Wir nehmen an, daß wir eine Zuordnung zwischen zwei Datensätzen besitzen, die optimal ist. Diese erhalten wir dadurch, daß ein Operateur einen Datensatz, der bewertet werden soll, einem Referenzdatensatz zuordnet. Da die Zuordnung optimal ist, hängt der Verlust der Information im Kanal nur noch von dem zu bewertenden Datensatz ab. Dies ermöglicht, zwei Datensätze gegenüber einen Referenzdatensatz zu vergleichen und Aussagen darüber zu machen, in welchem Datensatz mehr Information über den Referenzdatensatz enthalten ist. Diese Aussagen beziehen sich immer nur auf die Menge der betrachteten Attribute und Relationen. Daher kann der Informationsverlust für einzelne Attribute und Relationen oder auch für Kombinationen von mehreren Attributen und Relationen betrachtet werden.

Typische Qualitätsmaße die zur Beurteilung von raumbezogenen Daten gefordert werden, sind z.B. Vollständigkeit, Richtigkeit oder Korrektheit (siehe Kapitel 3.3.5). Die Berechnung dieser Maße erweist sich jedoch als schwierig. Während ein Maß für die Vollständigkeit in den meisten Fällen eine Prozentangabe sein wird, ist z.B. für die Korrektheit nicht einmal klar, welche Maßeinheit verwendet werden soll. Das Gebiet der Qualitätsuntersuchung von raumbezogenen Daten ist ein derzeit sehr aktueller Forschungszweig. Die hier vorgestellten Maße reflekieren die Betrachtung der Attribute und Relationen aus einer informationstheoretischen Sicht und haben als Einheit das Maß der Information (bit). Dadurch, daß diese Maße durch die Zuordnung von zwei Datensätzen berechnet werden, sind sie für den Vergleich von raumbezogenen Daten gut geeignet, da sie die relativen Unterschiede zwischen verschiedenen Datensätzen betrachten. Ein weiterer Vorteil dieser Maße ist, daß es durch das Berechnen der Maße aufgrund ihrer Wahrscheinlichkeiten möglich ist, numerische und symbolische Attribute gemeinsam zu betrachten.

Kapitel 7

Zuordnung von ATKIS- und GDF-Daten

Im letzten Kapitel wurden die allgemeinen Grundlagen der Zuordnung von raumbezogenen Daten aufgezeigt, ohne dabei anwendungsspezifische Fragestellungen zu betrachten. In diesem Kapitel wird die Zuordnung von Daten aus den Datenmodellen ATKIS und GDF (der Firma Bosch/Teleatlas) diskutiert. Um die statistischen Auswertungen für die Berechnung der Leistungsfunktion durchführen zu können, wurden Testgebiete ausgewählt und manuell zugeordnet. Diese Zuordnungen dienen auch als Referenzzuordnungen zur Überprüfung des automatischen Verfahrens. Im folgenden wird zuerst die Vorgehensweise der manuellen Zuordnung beschrieben. Danach erfolgt eine Darstellung der einzelnen Teilschritte der automatischen Zuordnung in der gleichen Reihenfolge wie im letzten Kapitel. Die Ergebnisse des automatischen Verfahrens werden mit den manuell vorgenommenen Zuordnungen verglichen und diskutiert. Abschließend erfolgt eine Untersuchung des Zeitverhaltens des Verfahrens.

7.1 Manuelle Zuordnung von raumbezogenen Daten

Im Rahmen dieser Arbeit wurde mit dem GIS-Produkt SICAD/open [Siemens 1993*c*] ein interaktives Werkzeug zur manuellen Zuordnung von ATKIS- und GDF-Daten entwickelt [Muratoglu 1996]. Abbildung 7.1 zeigt eine Darstellung der Oberfläche mit ATKIS- (durchgezogene Linien) und GDF-Daten (gestrichelte Linien). Den Daten wurden eindeutige ID's zugewiesen, um die interaktive Exploration der Ergebnisse zu erleichtern. Es stehen Funktionen zur Zuordnung der Daten, zur graphischen und alphanumerischen Darstellung sowie zur Auswertung der Zuordnungen zur Verfügung. Neben diesen Methoden steht weiter die gesamte Funktionalität des GIS-Produkts zur Verfügung.

7.1.1 Vorüberlegungen

Entscheidend für die Aussagekräftigkeit der statistischen Auswertungen ist die Wahl der Testgebiete und der betrachteten Objektklassen. Je nachdem, ob Testgebiete aus dicht oder aus dünn besiedelten Landschaften vorliegen, ist mit unterschiedlichen Ergebnissen der statistischen Auswertungen zu rechnen. Da in weniger dicht besiedelten Gebieten die zu erfassenden Objekte oft einfacher strukturiert sind (z.B. in Kreuzungsbereichen), gibt es bei der Erfassung weniger Interpretationsfreiraum und es ist mit einer höheren Ähnlichkeit zwischen den Datensätzen zu rechnen.

Bei einer Zuordnung von ATKIS-Daten ist auch die Herkunft der Daten von Bedeutung. Da die Landesvermessungsämter mit unterschiedlicher Software und Erfassungsvorlagen (siehe Kapitel 5.5) arbeiten, muß mit Inhomogenitäten bei Daten aus verschiedenen Bundesländern gerechnet werden. Durch die zentralisierte Erfassung der GDF-Daten kann von einem homogenen Datenbestand in Deutschland ausgegangen werden[1].

Ebenso entscheidend für die statistischen Auswertungen ist die Auswahl der betrachteten Objektklassen. Es können nur Elemente zugeordnet werden, welche dieselben Objekte der Landschaft beschreiben. Daher werden durch eine Vorselektion die Objektklassen in den beiden Datensätzen ausgewählt, welche eine gemeinsame Schnittmenge bilden. Diese Schnittmenge kann zur Verfeinerung der statistischen Auswertungen weiter aufgeteilt werden. So ist bei der Zuordnung von Daten der Objektklasse "Autobahn" mit anderen Ergebnissen zu rechnen als z.B. bei der Objektklasse "Nebenstraße".

Eine Untersuchung aller dieser Fälle würde den Rahmen dieser Arbeit sprengen. Die folgenden Untersuchungen basieren auf vier Testgebieten, welche aus dem Stadtbereich Stuttgart gewählt wurden. Eine übersichtsartige Darstellung der Testgebiete sowie eine Auswertung der manuellen Zuordnungen befinden sich in Anhang C. Es handelt sich hierbei um dicht bebaute Gebiete aus dem Innenstadtbereich und dem Stadtrandbereich. Alle vier Gebiete haben eine Größe von $2 * 2 \ km^2$. Die Straßendichte ist in den vier Gebieten unterschiedlich.

[1] Dies trifft jedoch nicht im internationalen Maßstab zu. In verschiedenen Ländern werden GDF-Daten von unterschiedlichen Institutionen erfaßt.

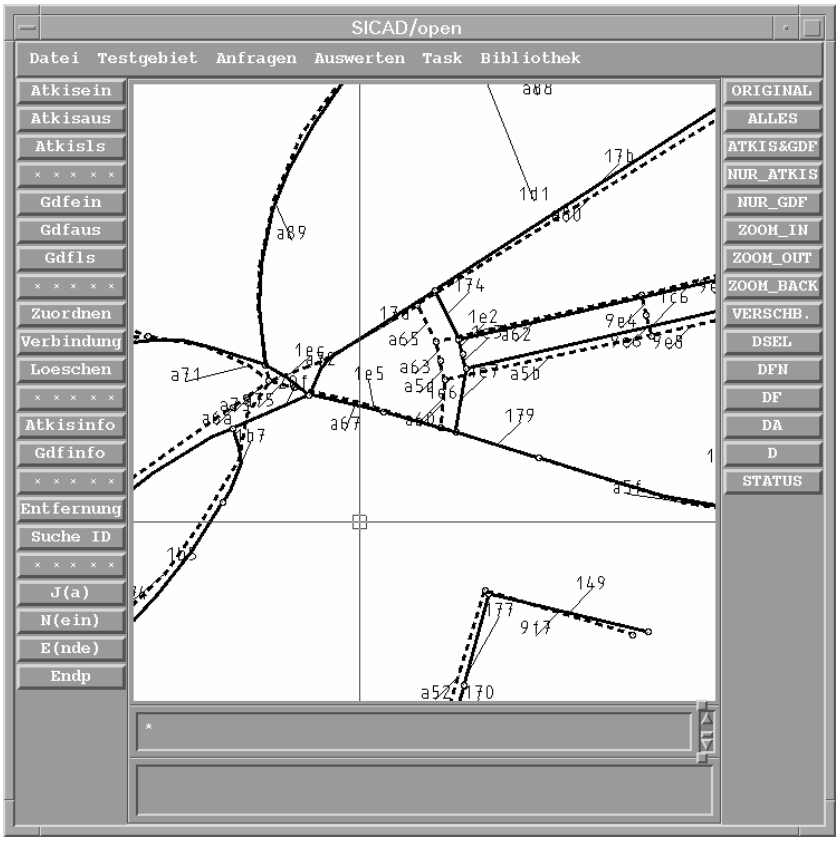


Abbildung 7.1: Interaktive Oberfläche zur manuellen Zuordnung der Daten

Die Anzahl der linienförmigen Elemente der ATKIS-Datensätze liegt im Bereich von 363 bis 640 Elemente, die der GDF-Datensätze von 435 bis 963 Elemente. Die betrachteten Objektarten der ATKIS-Daten sind *Straße* (Objektart 3101), *Platz* (Objektart 3103) und *Fahrbahn* (Objektart 3106). Die Objektart *Weg* (3102) wurde nicht mit in die Zuordnung aufgenommen, da diese Objekte nur zu einem kleinen Teil in den GDF-Daten erfaßt sind. Aus den GDF-Daten wurden alle Elemente der Feature Klassen *Road Element* und *Brunnel* ausgewählt. Das Attribut *Functional Class* (siehe Tabelle 5.2) ist bei den Daten der Testgebiete mit den Werten 2 bis 5 belegt. Dies bedeutet, daß keine Straßen in den Testgebieten vorhanden sind, die als "Autobahnen" oder "Bundesstraßen" klassifiziert werden.

7.1.2 Manuelle Zuordnung der ATKIS- und GDF-Daten

Bei der manuellen Zuordnung der Daten hat ein Operateur die Aufgabe, die beste Zuordnung zwischen zwei Datensätzen von Hand zu erstellen. Hierzu muß definiert werden, wann eine Zuordnung akzeptiert werden kann bzw. wann eine Zuordnung zurückzuweisen ist, weil die Attribute und Relationen der zuzuordnenden Daten zu unterschiedlich sind. Bestehen überhaupt keine Einschränkungen für die Zuordnungen, wäre es theoretisch möglich, beliebige Datensätze, die nichts miteinander zu tun haben, zumindest teilweise einander zuzuordnen.

Objekte aus dem Straßenverkehr werden in ATKIS und GDF mit einer Genauigkeit von ± 3 m erfaßt. Wird diese Genauigkeit eingehalten, darf die Entfernung zwischen zwei geometrischen Elementen, welche dasselbe Objekt beschreiben, maximal 6 m sein. Tatsächlich finden sich in lokalen Bereichen weitaus größere Unterschiede bis zu 15 m und mehr. Abbildung 7.2 zeigt die Vorgaben für die manuelle Zuordnung der Daten in dieser Arbeit. Da Straßen durch ihre Mittelachsen bzw. durch die Mittelachsen der Fahrbahnen repräsentiert werden, erfolgt eine Zuordnung der linienförmigen Elemente der Datensätze. Der Anfangs- bzw. Endpunkt eines linienförmigen Elementes muß in einem Fangkreis mit einem Radius von 15 m zum Anfangs- bzw. Endpunkt des zuzuordnenden Elementes liegen. Das linienförmige Element selbst, darf einen 15 m breiten Puffer um die zuzuordnende Linie nicht verlassen. Zuordnungen, deren Elemente mehr als 15 m voneinander entfernt sind, werden nicht zugelassen, da hier ein Attributaustausch oder eine Geometriehomogenisierung nicht mehr sinnvoll ist. Für Anwendungen,

bei denen die geometrische Genauigkeit eine wichtige Rolle spielt, ist es empfehlenswert, Puffer geringerer Breite zu definieren.

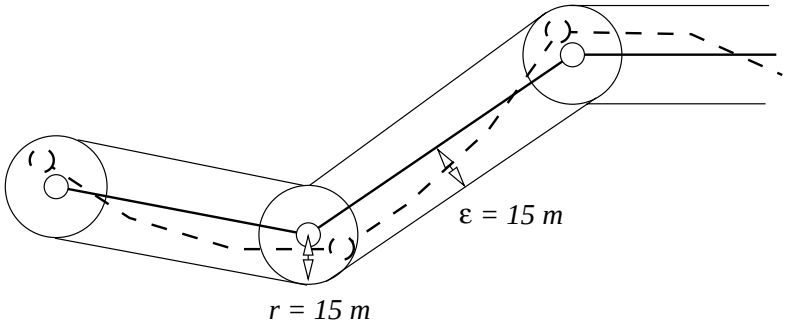


Abbildung 7.2: Vorgaben für die manuelle Zuordnung der Daten

Zwischen den Datensätzen dürfen beliebige $n : m$ Zuordnungen gemäß Abbildung 6.4 gebildet werden. Der Sonderfall $n : m_1 + m_2$ ist sowohl für ATKIS- als auch GDF-Elemente erlaubt. Dies bedeutet, daß ein ATKIS-Element zwei nicht topologisch miteinander verbundenen GDF-Elementen zugeordnet werden darf und umgekehrt. Dies entspricht der Situation, daß eine Straße in einem der Datensätze nur durch die Mittelachse und im anderen Datensatz durch zwei getrennte Fahrspuren erfaßt wurde.

7.1.3 Probleme

Während sich die manuelle Zuordnung der Daten in den meisten Gebieten problemlos durchführen läßt, kann es in Teilbereichen zu Schwierigkeiten kommen. Abbildung 7.3 zeigt ein typisches Beispiel einer Kreuzung, bei der keine eindeutige Zuordnung der Daten gefunden werden kann. Im ATKIS-Datensatz treffen sich vier Straßen einer Kreuzung in einem Punkt, wogegen im GDF-Datensatz ein zusätzliches Zwischenstück digitalisiert wurde. Die ATKIS-Elemente a_2 und a_4 lassen sich eindeutig zu b_2 und b_5 zuordnen sowie a_1 zu b_1 und a_3 zu b_4. Problematisch ist die Zuordnung des GDF-Elementes b_3, welches sich weder eindeutig zu a_1 noch zu a_3 zuordnen läßt. Wird das Element b_3 überhaupt nicht zugeordnet, ist die Toplogie der Elemente nur noch teilweise in den Zuordnungen abgebildet.

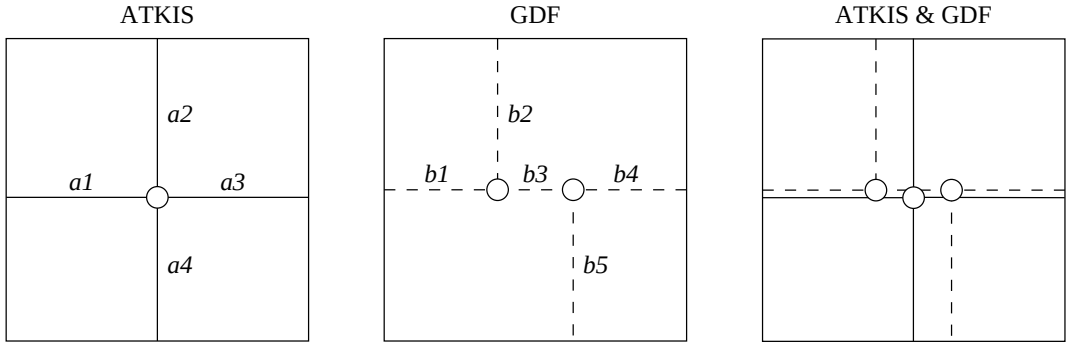


Abbildung 7.3: Probleme bei der manuellen Zuordnung; Beispiel 1

Ein ähnliches Problem ist in Abbildung 7.4 dargestellt. Hier wird eine Kreuzung in ATKIS wie im letzten Beispiel durch einen Punkt repräsentiert, wogegen im GDF-Datensatz dieselbe Kreuzung mit einem Kreisverkehr erfaßt wurde. Auch hier läßt sich keine eindeutige Zuordnung zwischen den linienförmigen Elementen finden.

Die hier dargestellten Probleme entstehen dadurch, daß in beiden Fällen Kreuzungsbereiche in einem Datensatz punktförmig und im anderen Datensatz durch eine oder mehrere Linien erfaßt wurden. Dadurch ist keine eindeutige Zuordnung zwischen den linienförmigen Elementen möglich. Da diese Fälle selten auftreten, werden die statistischen Auswertungen durch diese Kreuzungsbereiche kaum beeinflußt. Beim Vergleich zwischen den Ergebnissen der automatischen und der manuellen Zuordnung können hier jedoch Unterschiede entstehen.

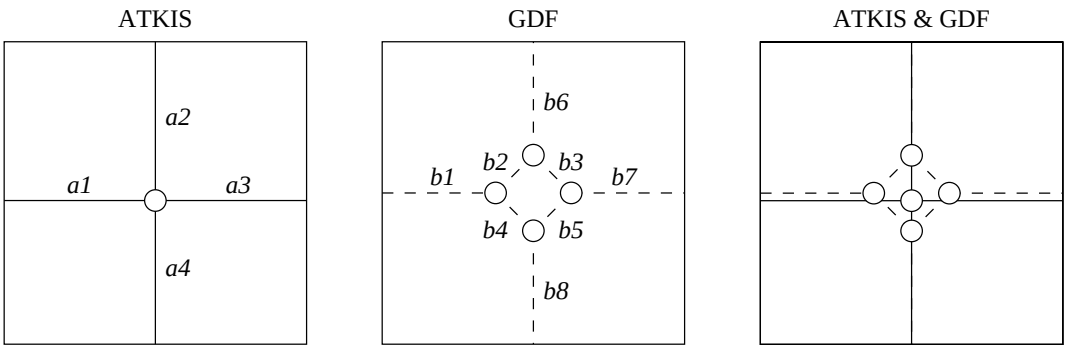


Abbildung 7.4: Probleme bei der manuellen Zuordnung; Beispiel 2

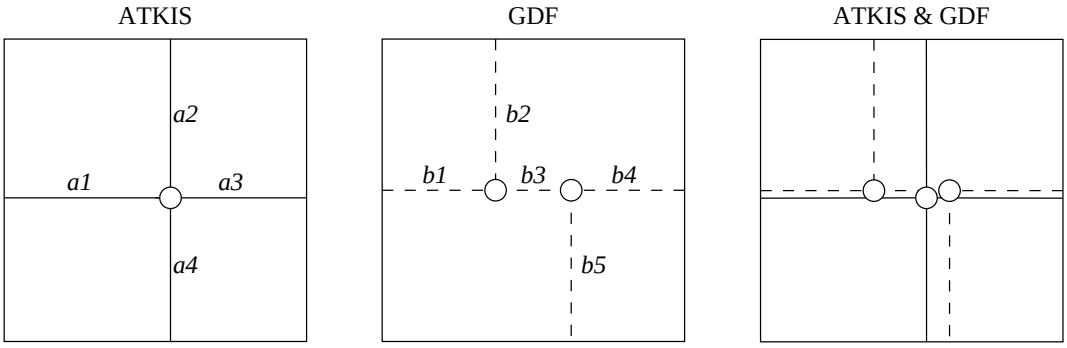


Abbildung 7.5: Probleme bei der manuellen Zuordnung; Beispiel 3

Abbildung 7.5 zeigt nochmals die gleiche Situation, wie in Abbildung 7.3, mit dem Unterschied, daß hier die ATKIS-Kreuzung nicht exakt zwischen den beiden Einmündungen der GDF-Straßen liegt, sondern nach rechts versetzt ist. In diesem Fall kann das GDF-Element b_3 zusammen mit b_1 dem ATKIS-Element a_1 zugeordnet werden. Dadurch bleibt die Information über die Topologie zumindest teilweise erhalten. In einem Nachbearbeitungsschritt können dann die Zuordnungen identifiziert werden, bei denen die topologischen Relationen stark abweichen. Je schwieriger die Entscheidung ist, wie die Elemente einander zuzuordnen sind, desto wahrscheinlicher ist es, daß Unterschiede in den Ergebnissen der manuellen und der automatischen Zuordnungen entstehen.

7.1.4 Auswertung der manuellen Zuordnungen

Eine tabellarische Aufstellung der Ergebnisse der manuellen Zuordnungen befindet sich in Anhang C. Insgesamt enthalten die vier Testgebiete 2.063 ATKIS-Elemente und 2.919 GDF-Elemente. Ca. 86 Prozent der ATKIS-Daten (1780 Elemente) und 74 Prozent der GDF-Daten (2186 Elemente) haben korrespondierende Elemente im anderen Datensatz.

Kardinalität *ATKIS : GDF*	Anzahl Zuordnungen	in Prozent	Anzahl beteiligte Elemente	in Prozent
$1 : *$	283	11.6	283	5.6
$* : 1$	733	30.2	733	14.7
$1 : 1$	781	32.7	1562	31.4
$1 : n$	346	14.2	1237	24.8
$n : 1$	144	5.9	473	9.5
$n : m$	134	5.5	694	13.9

Tabelle 7.1: Verteilung der Zuordnungen

Abbildung 7.6: ATKIS- und GDF-Daten mit a) und ohne b) globalem Fehler

Tabelle 7.1 zeigt die prozentuale Verteilung der Kardinalität der gefundenen Zuordnungen. Die 1 : 1 Zuordnungen machen mit ca. 32.7 Prozent den Großteil aller Zuordnungen aus und enthalten ca. 31.4 Prozent aller linienförmigen Elemente. Die restlichen Elemente sind Bestandteil von Zuordnungen zu keinem oder zu mehr als einem Element. Die Tabelle zeigt weiterhin, daß ein einzelnes ATKIS-Element mehreren GDF-Elementen zugeordnet wird, wesentlich häufiger eintritt, als der Fall, daß ein einzelnes GDF-Element mehreren ATKIS-Elementen zuzuordnen ist. Bei den Zuordnungen von Elementen zu einer "Wildcard" kann zwischen zwei Fällen unterschieden werden. Im ersten Fall wird zu einem Element überhaupt kein passendes Partnerelement im Puffer gefunden. Diese Elemente können sofort aus dem Zuordnungsprozeß herausgenommen werden. Im anderen Fall sind im Puffer potentielle Partner enthalten. Diese werden jedoch zu anderen Elementen oder gar nicht zugeordnet. Dieser Fall tritt vor allem in komplexen Kreuzungsbereichen auf.

7.2 Vorverarbeitung für automatische Zuordnung

Wie bereits in Kapitel 6.3 diskutiert, werden in einem ersten Schritt, globale geometrische Unterschiede zwischen den Datensätzen eliminiert. Bei einer Überlagerung der ATKIS- und GDF-Daten der vier Testgebiete kann in allen Fällen eine Translation in der Größenordnung von ca. 10 m festgestellt werden. Abbildung 7.6 zeigt die Überlagerung der Daten vor und nach der Vorverarbeitung. Die Parameter der Transformation wurden mit Hilfe von interaktiv am Bildschirm gemessenen Paßpunkten bestimmt. Als weiterer Vorverarbeitungsschritt empfiehlt sich eine Überprüfung der Topologie. Topologische Erfassungsfehler in den Datensätzen können zu Fehlzuordnungen führen. Im Bereich von Unterführungen, die in ATKIS lediglich mit einer Überführungsreferenz erfaßt sind (siehe Kapitel 5.1.3), ist eine Schnittbildung durchzuführen, damit beide Datensätze als planarer Graph vorliegen.

Als weiterer Bearbeitungsschritt wird eine Eliminierung von Knoten, welche für den Zuordnungsprozeß redundant sind, durchgeführt. Hierbei handelt es sich um sogenannte unechte Knoten, in denen exakt zwei Kanten enden. Unechte Knoten werden z.B. dann erfaßt, wenn ein linienförmiges Objekt ein Attribut besitzt, welches nicht durchgängig für das gesamte Objekt gültig ist. Abbildung 7.7 zeigt ein Beispiel für diese Situation. In der Abbildung ist eine Straße dargestellt, bei der sich, die beiden Attribute *Straßenbreite* und *Anzahl der Fahrbahnen* ändern, ohne daß eine topologische Änderung vorliegt. In diesen Fällen muß ein Knoten digitalisiert werden, da die Attribute nicht durchgängig für das gesamte Objekt gültig sind[2].

Abbildung 7.8 zeigt ein weiteres Beispiel für die Erfassung von unechten Knoten. Im dargestellten Ausschnitt des Straßennetzwerkes kann gesehen werden, daß eine der Straßen nur im ATKIS-Datensatz erfaßt wurde, jedoch

[2] In GDF könnte diese Situation auch mit Hilfe eines segmentierten Attributes modelliert werden (siehe Kapitel 3.3.2). In den Daten der Testgebiete wurde diese Technik jedoch nicht genutzt.

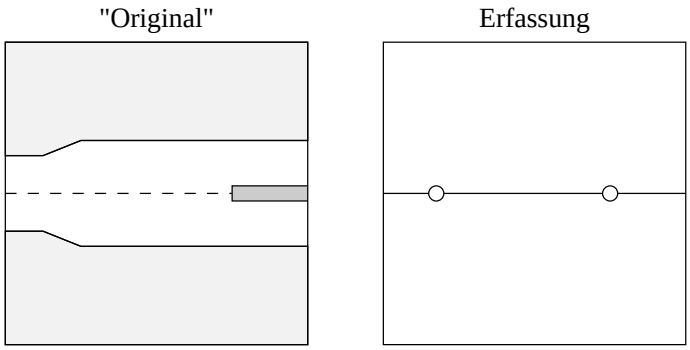


Abbildung 7.7: Erfassung von Knoten bei Attributänderungen

nicht im GDF-Datensatz. Trotzdem wurde an der Position der Einmündung im GDF-Datensatz ein Knoten (*N3*) digitalisiert. Hierbei handelt es sich entweder ebenfalls um einen Attributwechsel oder um die Bildung eines neuen Objektes, da sich beispielsweise der Straßenname an dieser Einmündung ändert. Aus den gleichen Gründen erfolgt die Bildung der Knoten *N2* bzw. *N4*.

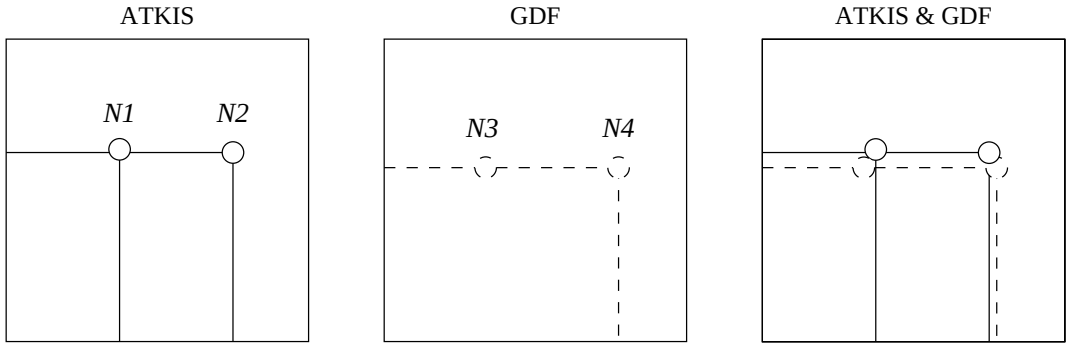


Abbildung 7.8: Bildung von Knoten mit zwei Kanten

Für den Zuordnungsprozeß sind jedoch nur solche Knoten interessant, bei denen tatsächlich eine topologische Änderung der Objekte des Straßennetzes vorliegt. Dies ist dann der Fall, wenn sich in einem Knoten mindestens drei Kanten treffen, oder wenn der Knoten aufgrund eines Objektwechsels gebildet wurde. Knoten, die nur einen Attributwechsels darstellen, sind für den Zuordnungsprozeß redundant und als Zwischenpunkte ohne topologische Bedeutung zu betrachten. Die Erkennung redundander Knoten kann mit Hilfe der Objektstrukturen erfolgen. Dies war bei den GDF-Daten jedoch nicht eindeutig möglich, da hier die Bildung komplexer Objekte (Ebene 2 Objekte, siehe Kapitel 3.3.4) nur in Teilbereichen durchgeführt war. Eine Erkennung redundanter Knoten wurde daher auf der Geometrieebene durchgeführt. Abbildung 7.9 zeigt die verwendeten Bedingungen zur Eliminierung redundanter Knoten.

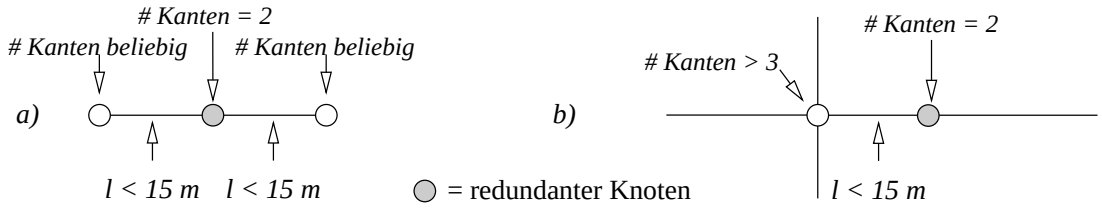


Abbildung 7.9: Eliminierung von redundanten Knoten

Redundante Knoten können daran erkannt werden, daß eine oder beide Kanten, die in diesem Knoten enden, sehr kurz sind. Als Erfahrungswert ergab sich, daß Kanten, welche kürzer als 15 m sind, auf redundante Knoten

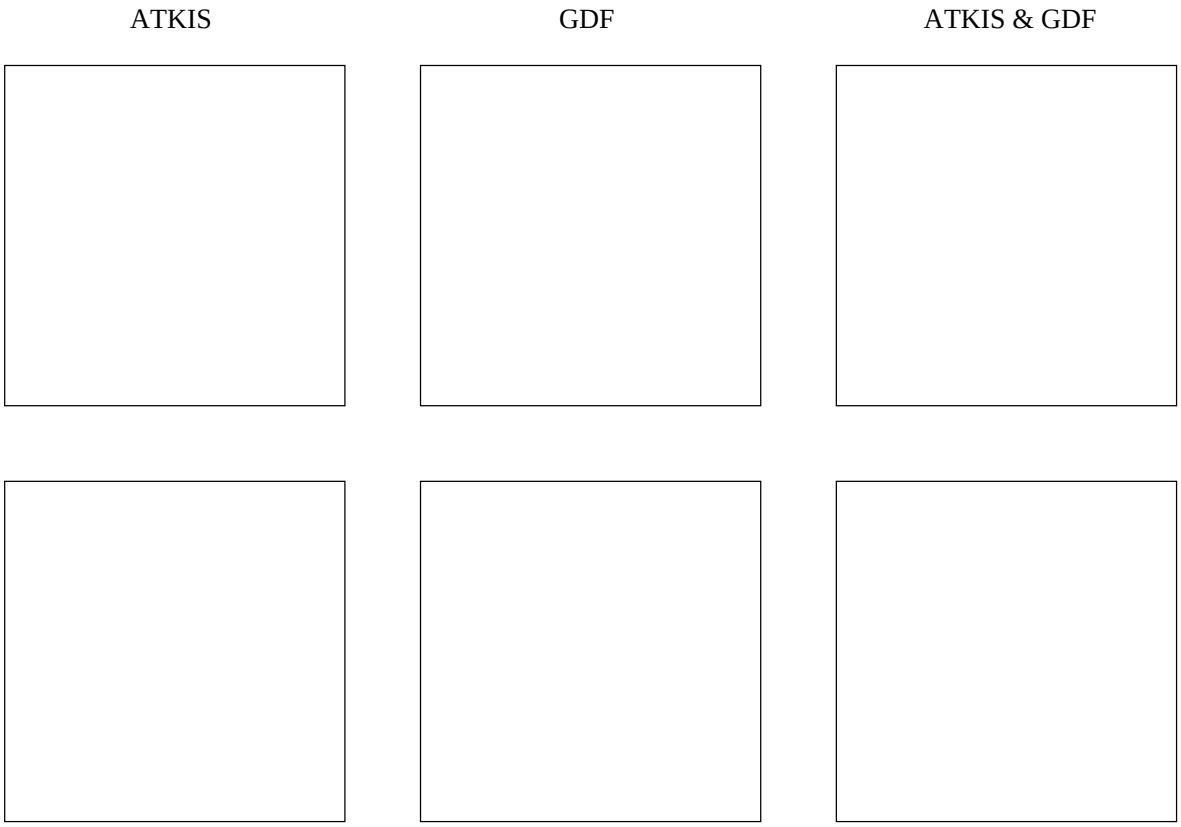


Abbildung 7.10: Datensätze mit (oben) und ohne (unten) redundanten Knoten

7.3 Aufstellen der potentiellen Zuordnungspaare

Beim Aufstellen der potentiellen Zuordnungspaare sind zwei unterschiedliche Randbedingungen zu beachten. Zum einen soll eine Minimierung der Rechenzeit erfolgen, und zum anderen soll garantiert sein, daß alle Zuordnungen gefunden werden, welche Bestandteil der optimalen Zuordnung sind.

7.3.1 Geometrische Beschränkungen

Zur Durchführung der automatischen Zuordnung, wird eine Liste mit potentiellen Zuordnungspaaren erstellt. Diese Liste soll alle Zuordnungspaare enthalten, welche in der manuellen Zuordnung vorkommen können. Da für die manuelle Zuordnung als Einschränkung gilt, daß sich Anfangs- und Endpunkte der Elemente einer Zuordnung in einem Fangkreis von 15 m befinden müssen, kann dies als geometrische Beschränkung für die automatische Zuordnung genutzt werden. Dies bedeutet, daß nur solche Zuordnungspaare in die Liste der potentiellen Zuordnungen aufgenommen werden, welche einen Längenunterschied von höchstens 30 m besitzen.

Als weitere geometrische Beschränkung der Zuordnungen kann der Winkel zwischen den zugeordneten Elementen betrachtet werden. Abbildung 7.11 zeigt die Häufigkeitsverteilung der Winkeldifferenz zwischen den Elementen der manuell erzeugten Zuordnungen. Der Winkel eines linienförmigen Elementes errechnet sich aus dem Anfangs- und Endpunkt, ohne das Zwischenpunkte betrachtet werden. In der Abbildung kann gesehen werden, daß die Mehrzahl der Zuordnungen eine Winkeldifferenz von weniger als 10 Grad besitzt. Im Bereich von 10 bis 30 Grad

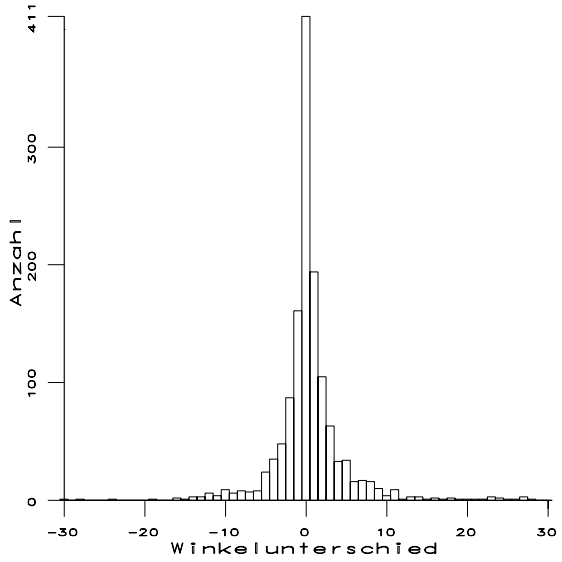


Abbildung 7.11: Winkelunterschiedsverteilung der manuellen Zuordnungen

sind nur noch sehr wenige Zuordnungen zu finden. Als weitere geometrische Beschränkung wird daher definiert, daß nur solche Zuordnungspaare in die Liste der potentiellen Zuordnungspaare aufgenommen werden, welche einen Winkeluntschied von weniger als 30 Grad besitzen.

7.3.2 Minimale Zuordnungen

Wie bereits in Kapitel 6.4 dargestellt, dürfen nur solche Zuordnungen in die Liste der potentiellen Zuordnungspaare eingetragen werden, welche sich nicht durch eine Kombination bereits bestehender Zuordnungen bilden lassen. Diese Zuordnungen werden im folgenden *minimale Zuordnungen* genannt. Die Bildung von minimalen Zuordnungen erfolgt zum einen zur Rechenzeitoptimierung und zum anderen wegen einer besseren Vergleichsmöglichkeit zwischen den manuell und automatisch erzeugten Zuordnungen. Mit Hilfe von minimalen Zuordnungen wird eine Gesamtzuordnung zwischen zwei Datensätzen eindeutig beschrieben.

Abbildung 7.12 zeigt ein Beispiel einer minimalen Zuordnung. Der zuzuordnende Datensatz besteht aus jeweils drei Elementen. Wäre es erlaubt, beliebige Paarungen zu bilden, sind insgesamt 28 Zuordnungspaare möglich. Dabei ist auch immer die Zuordnung zu einer "Wildcard" erlaubt, die aussagt, daß dieses Element *keinen* Partner im anderen Datensatz besitzt. Bei der minimalen Zuordnung gibt es nur noch 13 mögliche Zuordnungspaare. Das bedeutet, daß nur dann Elemente mit Hilfe des Buffer Growings zusammengefaßt werden dürfen, wenn es nicht möglich ist, die gleiche Zuordnung durch eine Kombination von kürzeren Elementen zu bilden.

7.3.3 Buffer Growing

Das Buffer Growing hat die Aufgabe einzelne Elemente der Datensätze zu logischen Elementen zusammenzufassen, um $n : m$ bzw. $1 : n$ und $n : 1$ Zuordnungen bilden zu können. Das Verfahren wurde in Kapitel 6.4 dargestellt. Um wiederum den Suchraum zu minimieren, werden auch für das Buffer Growing Einschränkungen definiert. Bei einer Betrachtung der manuellen Zuordnungen kann festgestellt werden, daß logische Elemente mit wenig Teilelementen wesentlich häufiger sind als logische Elemente mit vielen Teilelementen.

Tabelle 7.2 zeigt eine Auswertung der manuellen Zuordnungen. Das längste logische Element ist Teil des GDF-Datensatzes und besteht aus 9 Teilelementen. Es ist nicht nötig, das Buffer Growing beliebig lange durchzuführen. Für diese Arbeit wurde die Obergrenze auf maximal 9 Teilelemente gesetzt. Auch bei der Richtung, in die der Puffer wächst, können Beschränkungen definiert werden. Sind der Richtung des Puffers keine Schranken gesetzt, so können nicht sinnvolle logische Elemente entstehen, wie es in Abbildung 7.13 dargestellt ist.

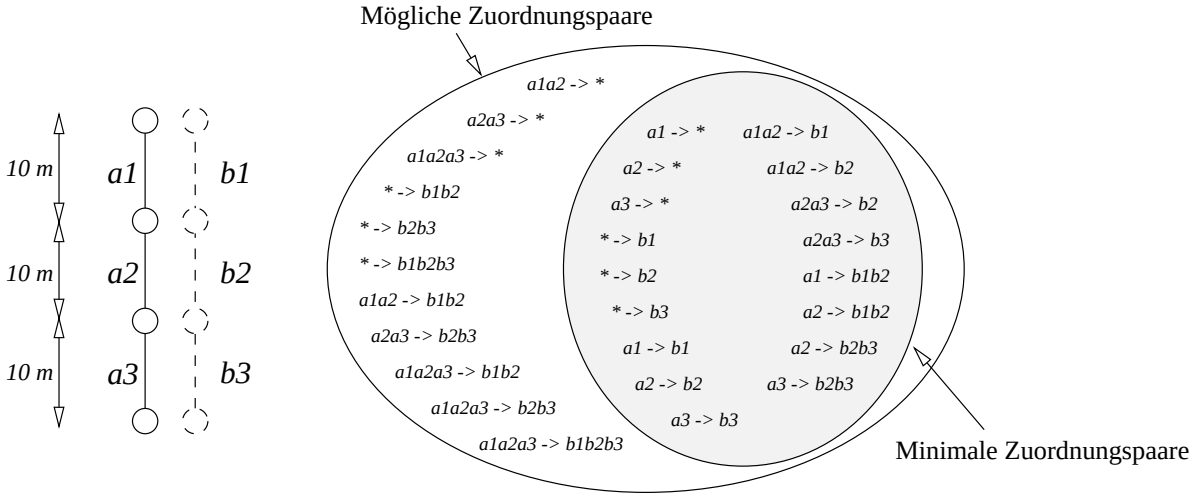


Abbildung 7.12: Beispiel für eine minimale Zuordnung

Anzahl der Teilelemente	Anzahl logische Elemente ATKIS	Anzahl logische Elemente GDF
1	1127	925
2	216	291
3	41	116
4	11	165
5	6	17
6	4	3
7	0	2
8	0	0
9	0	2

Tabelle 7.2: Anzahl der Einzelelemente in den logischen Elementen

Abbildung 7.14 zeigt die in dieser Arbeit verwendeten Beschränkungen des Buffer Growings. Abhängig von der Anzahl der abgehenden Kanten in einem Knoten wird entschieden, wann der Puffer weiter wachsen darf. Gehen von dem Knoten zwei Kanten ab, so gibt es keine Beschränkungen und der Puffer darf weiter wachsen. Treffen jedoch in dem Knoten drei oder mehr Kanten zusammen, so wird der Schnittwinkel der Liniensegmente, die in diesem Knoten enden, überprüft. Nur wenn dieser Schnittwinkel kleiner als 45 Grad ist, darf der Puffer weiter wachsen. Dies führt zu einer weiteren wesentlichen Einschränkung des Suchraumes.

7.3.4 Auswertung der potentiellen Zuordnungspaare

Für die manuelle Zuordnung der Daten wurde als einzige Beschränkung definiert, daß die Elemente in einem 15 m breiten Puffer/Fangkreis liegen müssen. Für die automatische Zuordnung wird zur Einschränkung des Suchraumes zusätzlich der Winkelunterschied der Zuordnungspartner überprüft und die Bildung logischer Elemente aus Teilelementen ist nur dann erlaubt, wenn der Winkel des Anfangs- und Endliniensegmentes kleiner als 45 Grad ist. Weiterhin wird versucht redundante Knoten automatisch zu identifizieren. Diese Einschränkungen können dazu führen, daß manuell erzeugte Zuordnungen nicht in der Liste der potentiellen Zuordnungspaare enthalten sind. Da die Beschränkungen für die automatische Zuordnung nicht sehr restriktiv sind, kommt diese Situation jedoch nur selten vor.

Tabelle 7.3 zeigt eine Auswertung der potentiellen Zuordnungspaare der verschiedenen Testgebiete. Es kann gesehen werden, daß ca. doppelt soviele Zuordnungspaare in der Liste der potentiellen Zuordnungspaare aufgenommen werden, als in den manuell vorgenommenen Zuordnungen enthalten sind. Trotz der verwendeten

Testgebiet	*1*	*2*	*3*	*4*	*Gesamt*
Anzahl potentielle Zuordnungen	431	548	863	923	2765
Anzahl manuell erstellte Zuordnungen	211	363	339	492	1405
Anzahl gefundene manuelle Zuordnungen	209	362	337	491	1399
Prozentsatz gefundene Zuordnungen	99,05	99,72	99,41	99,79	99,57
Nicht gefunden wegen Winkel > 30 Grad	1	1	0	0	2
Nicht gefunden wegen Buffer Growing	0	0	0	1	1
Nicht gefunden wg. Eliminierung red. Knoten	1	0	2	0	3

Tabelle 7.3: Auswertung der potentiellen Zuordnungen (ohne Wildcard-Zuordnungen)

geometrischen Einschränkungen werden 99,57 Prozent der manuell erzeugten Zuordnungen gefunden. Nach dem Aufstellen der potentiellen Zuordungspaare wird der eigentliche Zuordnungsprozeß durchgeführt, der die Aufgabe hat, aus der Liste der potentiellen Zuordnungspaare die Kombination von Zuordnungen herauszufinden, welche die beste Zuordnung darstellt.

7.4 Berechnung der gegenseitigen Information

Die Datensätze werden durch eine relationale Beschreibung $D = (P, R)$ dargestellt. Der Teil P enthält die Beschreibung der Elemente durch ihre Attribute und der Teil R die Relationen zwischen den Elementen. Bei der Zuordnung der ATKIS- und GDF-Daten werden die linienhaften Elemente der Datensätze betrachtet. Als Attribute werden die Eigenschaften *Länge*, *Form* und *Position* verwendet. Der relationale Teil wird durch die topologische Relation *verbunden* beschrieben. Die gegenseitige Information $I_h(D_1; D_2)$ einer Zuordnung h berechnet sich daher aus vier Teilen:

$$\begin{aligned} I_h(D_1; D_2) &= I_h(P_1; P_2) + I_h(R_1; R_2) \\ &= \sum_{p \in h} I(\text{Länge}(p); \text{Länge}(h(p))) + \\ &\quad \sum_{p \in h} I(\text{Form}(p); \text{Form}(h(p))) + \\ &\quad \sum_{p \in h} I(\text{Position}(p); \text{Position}(h(p))) + \\ &\quad \sum_{r_i \in h} \sum_{r_j \in h} I(\text{Verbunden}(r_i, r_j); \text{Verbunden}(h(r_i), h(r_j))) \end{aligned} \tag{7.1}$$

Die Berechnung der gegenseitigen Information erfolgt mit Hilfe der bedingten Wahrscheinlichkeiten der Attribute und Relationen der Datensätze. Um diese Wahrscheinlichkeiten zu erhalten, gibt es grundsätzlich drei verschie-

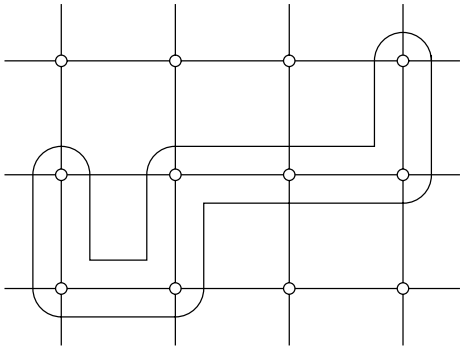


Abbildung 7.13: Wachstum des Puffers ohne Beschränkungen

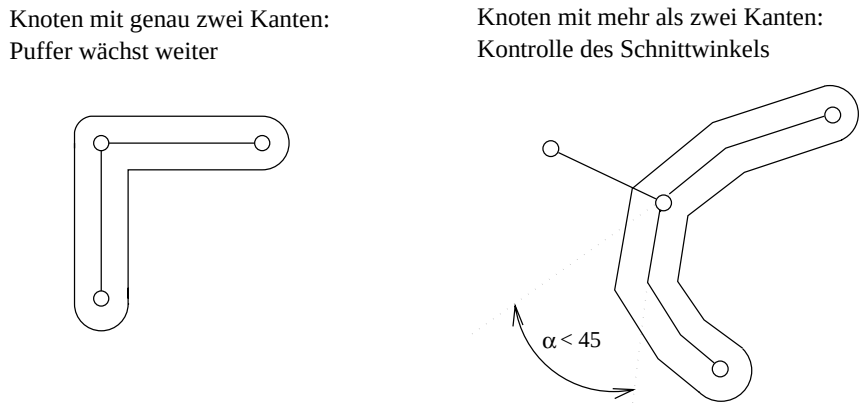


Abbildung 7.14: Beschränkungen beim Buffer Growing

dene Möglichkeiten [Vosselman 1992]: (1) durch analytische Berechnung, (2) durch numerische Simulation der Zuordnungen und (3) durch Trainingszuordnungen. Da der Zusammenhang zwischen ATKIS- und GDF-Daten weder analytisch beschreibbar noch numerisch simulierbar ist, müssen die statistischen Untersuchungen mit Hilfe von manuellen Trainingszuordnungen durchgeführt werden. Im folgenden wird die Berechnung der gegenseitigen Information anhand der verschiedenen betrachteten Attribute und Relationen in einzelnen Unterkapiteln beschrieben. Die statistischen Auswertungen werden mit Hilfe von Häufigkeitsverteilungen dargestellt.

7.4.1 Gegenseitige Information der Form

Die Form der Linien wird in drei Klassen eingeteilt. Um die Form zu berechnen, wird zu einer gedachten Hilfslinie, welche vom Anfangs- zum Endpunkt verläuft, die maximale Entfernung des Linienverlaufs zur Hilfslinie berechnet (siehe Abbildung 7.15). Abhängig von dieser Entfernung erfolgt die Einteilung in die verschiedenen Klassen. Dies ist ein sehr einfaches Maß zur Bestimmung der Form einer Linie. Komplexere Klassifizierungen finden sich z.B. in [Frydrychowicz 1990]. Für diese Arbeit ist jedoch ein einfaches Maß der Form ausreichend, da durch die Prüfung, ob eine zuzuordnende Linie vollständig im Puffer der Ausgangslinie liegt, die Form nicht sehr gegenüber der Ausgangslinie variieren kann.

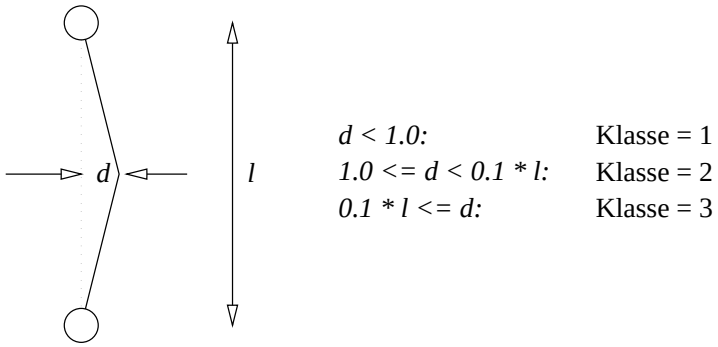


Abbildung 7.15: Berechnung der Form der Linien

Tabelle 7.4 zeigt die Auftretenswahrscheinlichkeiten der einzelnen Klassen im ATKIS-Datensatz. Eine weitere Verfeinerung kann durchgeführt werden, in dem die Auftretenswahrscheinlichkeiten für die Form für verschiedene Linienlängen getrennt voneinander untersucht werden. Je länger ein Element ist, desto wahrscheinlicher ist es, daß es einer höheren Klasse angehört als eine kurzes Element [Muratoglu 1996].

In Tabelle 7.5 sind die bedingten Wahrscheinlichkeiten für die verschiedenen Formklassen angegeben. Für eine gegebene Klasse ist die bedingte Wahrscheinlichkeit der Zuordnung in dieselbe Klasse immer am höchsten. Der Wechsel in eine andere Klasse ist umso unwahrscheinlicher, je weiter die Klasse von der ursprünglichen Klasse “entfernt“ ist.

	$P_{Form}(a_i)$
$a_i \in$ Klasse 1	0,59
$a_i \in$ Klasse 2	0,27
$a_i \in$ Klasse 3	0,14

Tabelle 7.4: Wahrscheinlichkeitsverteilung der Form

$P_{Form}(a_i \mid b_j)$	$b_j \in$ Klasse 1	$b_j \in$ Klasse 2	$b_j \in$ Klasse 3
$a_i \in$ Klasse 1	0,87	0,11	0,02
$a_i \in$ Klasse 2	0,28	0,61	0,11
$a_i \in$ Klasse 3	0,09	0,14	0,77

Tabelle 7.5: Bedingte Wahrscheinlichkeit der Form

7.4.2 Gegenseitige Information des Winkels

Der Winkel wird durch den Anfangs- und Endpunkt der Elemente berechnet. Während bei dem Attribut Form nur drei verschiedene Klassen definiert wurden, handelt es sich beim Winkel um ein kontinuierliches Maß. Im Anhang B wird gezeigt, daß die differentiellen Informationsmaße für kontinuierliche Signale analog zu den Maßen für endliche Eingabealphabete definiert sind. Für eine numerische Verarbeitung der Signale muß jedoch immer eine Diskretisierung erfolgen. Abhängig von der Größe des Diskretisierungsintervall entsteht ein Informationsverlust im Kanal. Im Anhang B.2 wird gezeigt, daß, wenn das Diskretisierungsintervall wesentlich kleiner als die Bandbreite der Rauschfunktion gewählt wird, der Informationsverlust vernachlässigbar klein wird.

Um die gegenseitige Information eines Attributes zu berechnen, wird die Wahrscheinlichkeit des Auftretens des Attributes sowie die bedingte Wahrscheinlichkeit zwischen den zugeordneten Attributen benötigt. Als Verteilungsfunktion des Attributes Winkel wird eine Gleichverteilung im Intervall [0, 360 Grad] angenommen, da Straßen in Stadtgebieten (zumindest in Deutschland[3]) keine bevorzugte Richtung aufweisen. Somit ist die Auftretenswahrscheinlichkeit bei einem Diskretisierungsintervall von einem Grad:

$$P_{Winkel}(a_i) = \frac{1}{360} \tag{7.2}$$

Die bedingte Wahrscheinlichkeit des Attributes Winkel kann direkt aus der Häufigkeitsverteilung in Abbildung 7.11 errechnet werden. In der Abbildung kann gesehen werden, daß Zuordnungen, deren Elemente den gleichen Winkel besitzen, am häufigsten vorkommen und mit zunehmenden Winkelunterschied die Anzahl der Zuordnungen abnimmt. Dies bedeutet, daß die gegenseitige Information umso größer wird, je kleiner der Winkelunterschied der zuzuordnenden Elemente ist.

7.4.3 Gegenseitige Information der Länge

Bei der Wahrscheinlichkeitsfunktion der Länge handelt es sich im Gegensatz zum Winkel nicht um eine Gleichverteilung. Abbildung 7.16 zeigt die Häufigkeitsverteilung des Attributes Länge der ATKIS-Elemente. Die Mehrzahl der Elemente haben eine Länge bis 180 m. Danach ist ein deutlicher Rückgang der Häufigkeit festzustellen. Dies bedeutet, daß sehr lange Elemente einen deutlich höhereren Selbstinformationsgehalt besitzen als kurze Elemente. Dadurch werden Zuordnungen mit langen Elementen im Zuordnungsprozeß gegenüber Zuordnungen mit kurzen Elementen bevorzugt.

Die Häufigkeitsverteilung der Längendifferenz der zugeordneten Elemente ist in Abbildung 7.17 dargestellt. Das Intervall der Rauschfunktion ist [-30 m, 30 m], da durch die Anwendung der geometrischen Beschränkungen keine größeren Längendifferenzen möglich sind. Als Breite des Diskretisierungsintervall wurde 1 m gewählt. Das Kanalverhalten wird mit einer Rauschfunktion pro betrachtetem Attribut modelliert. Mit der Modellierung von unterschiedlichem Kanalverhalten abhängig von den Attributwerten kann eine Verfeinerung der Ergebnisse

[3] Bei einer Analyse des Straßennetzes von New York wäre intuitiv z.B. eine höhere Auftretenswahrscheinlichkeit von Straßen in Nord-Süd- bzw. Ost-West-Richtung gegenüber Straßen anderer Richtungen zu erwarten.

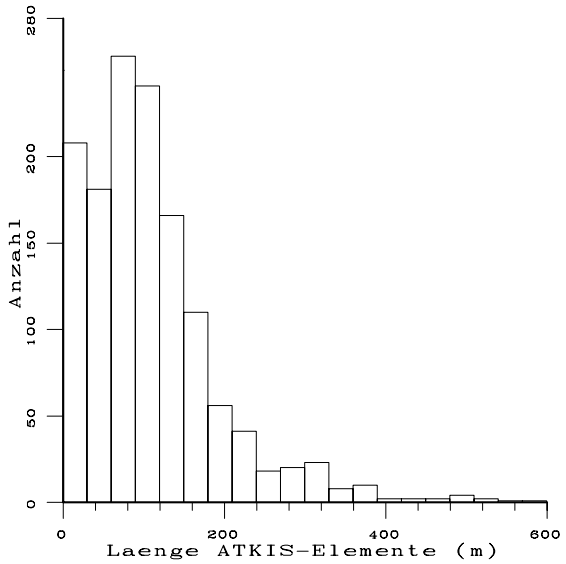


Abbildung 7.16: Häufigkeitsverteilung der Länge

erzielt werden. Für das Attribut Länge ist bei kurzen Linien z.B. eine steilere Kurve der bedingten Wahrscheinlichkeit der Längendifferenz zu erwarten als bei langen Linien, da der prozentuale Fehler mit zunehmender Länge bei gleicher Längendifferenz abnimmt (dieser Zusammenhang wird in [Muratoglu 1996] aufgezeigt).

7.4.4 Gegenseitige Information der Position

Zur Bestimmung der Position der Elemente werden die Koordinaten des Anfangs- und des Endpunktes betrachtet. Hierzu wird ein zweidimensionales Gitter über das Testgebiet gelegt. Als Verteilungsfunktion wird eine Gleichverteilung innerhalb des Testgebietes angenommen. Die Wahrscheinlichkeit, daß eine Koordinate in einer bestimmten Gitterzelle liegt, ist dann:

$$P_{Gitterzelle}(a_i) = \frac{1}{n * m} \tag{7.3}$$

wobei n und m die Anzahl der Gitterzellen in x- bzw. y-Richtung sind. Daraus folgt, daß der Informationsgehalt eines Punktes abhängig von der Größe des Testgebietes ist. Dies ist aus der Sicht der Informationstheorie damit zu erklären, daß mit zunehmender Größe des Zeichenvorrates des Eingabealphabetes (in diesem Fall die Anzahl der Gitterzellen) der Selbstinformationsgehalt eines Zeichen, unter der Annahme einer Gleichverteilung, zunimmt.

Abbildung 7.18 zeigt ein zweidimensionales Impulsdiagramm der Häufigkeitsverteilung der zugeordneten Positionen im 1 Meter-Raster. Die Zuordnung zweier Punkte in die gleiche Gitterzelle kommt am häufigsten vor. Aufgrund der geometrischen Beschränkung, daß zwei zugeordnete Punkte in einem Fangkreis von 15 m liegen müssen, ergibt sich die kreisförmige Abgrenzung der Verteilung im Diagramm.

Aus der zweidimensionalen Häufigkeitsverteilung aus Abbildung 7.18 läßt sich eine eindimensionale Verteilung der Entfernung der zugeordneten Punkte berechnen (Abbildung 7.19). Im Diagramm kann gesehen werden, daß Zuordnungen deren Anfangs- bzw. Endpunkte zwischen zwei und drei Meter voneinander entfernt sind, am häufigsten vorkommen. Je größer die Entfernung ist, desto weniger Elemente werden zugeordnet. Auffällig ist, daß Zuordnungen von Elementen, deren Anfangs- bzw. Endpunkte nicht weiter als ein Meter voneinander entfernt sind, gegenüber anderen Zuordnungen nur selten vorkommen.

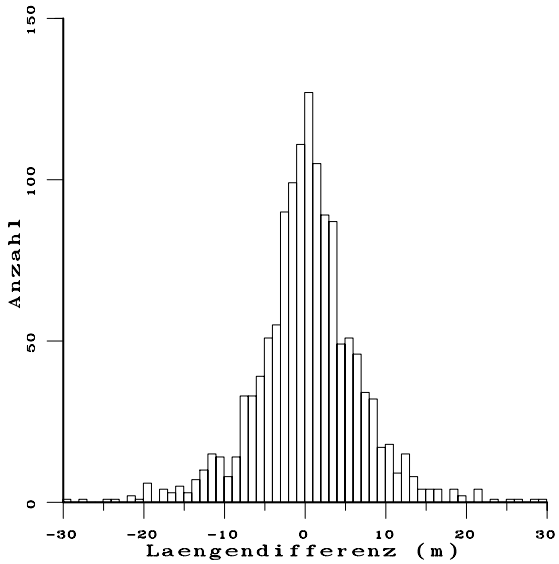


Abbildung 7.17: Häufigkeitsverteilung der Längendifferenz

7.4.5 Gegenseitige Information des relationalen Teils

Zur Leistungsberechnung des relationalen Teils R wird die gegenseitige Information aller möglichen Paare von Zuordnungen berechnet:

$$\begin{aligned} I_h(R_1;R_2) &= \sum_{r_i \in h} \sum_{r_j \in h} I(\text{Verbunden}(r_i, r_j); \text{Verbunden}(h(r_i), h(r_j))) \\ &= \sum_{r_i \in h} \sum_{r_j \in h} -\log(P(r_i, r_j)) + \log(P(r_i, r_j | h(r_i, r_j))) \end{aligned} \tag{7.4}$$

Für eine Abschätzung der Selbstinformation einer Zuordnung wird die durchschnittliche Anzahl von Elementen ermittelt, welche mit einem betrachteten Element verbunden sind. Eine Auswertung der manuellen Zuordnung ergibt, daß jedes Element mit durchschnittlich 2,7 anderen Elementen verbunden ist. Es sei $P(W)$ die Wahrscheinlichkeit des Ereignisses, daß zwei Elemente miteinander verbunden sind und $P(F)$ das zu $P(W)$ entgegengesetzte Ereignis. Dann gibt es für einen Datensatz mit n Zuordnungspaaren $n*(n-1)$ Relationen des Typs *verbunden* und es gilt[4]:

$$P(W) = \frac{n * 2,7}{n * (n-1)} = \frac{2,7}{n-1} \tag{7.5}$$

und

$$P(F) = 1 - P(W) = \frac{(n * (n-1)) - (n * 2,7)}{n * (n-1)} = \frac{n-3,7}{n-1} \tag{7.6}$$

Da im voraus nicht bekannt ist, wieviele Zuordnungspaare in der Endzuordnung vorkommen, muß n ebenfalls abgeschätzt werden. Bei einer Untersuchung der Testgebiete kann gesehen werden, daß die Anzahl der gefundenen Zuordnungen im Bereich von 57 Prozent bis 73 Prozent der Anzahl der ATKIS-Elemente liegt (siehe Tabelle 7.7). Aus dem gewichteten arithmetischen Mittel ergibt sich n zu 65,2 Prozent der Anzahl der ATKIS-Elemente. Das bedeutet, daß die Information der Relation *verbunden* von der Größe der Datensätze abhängt. Dies ist jedoch beabsichtigt, da das Maß der Überraschung, daß zwei Elemente miteinander verbunden sind, in einem Datensatz mit vielen Elementen höher ist, als in einem Datensatz mit nur wenigen Elementen.

[4] Die Relation, daß ein Element mit sich selbst verbunden ist, ist nicht erlaubt.

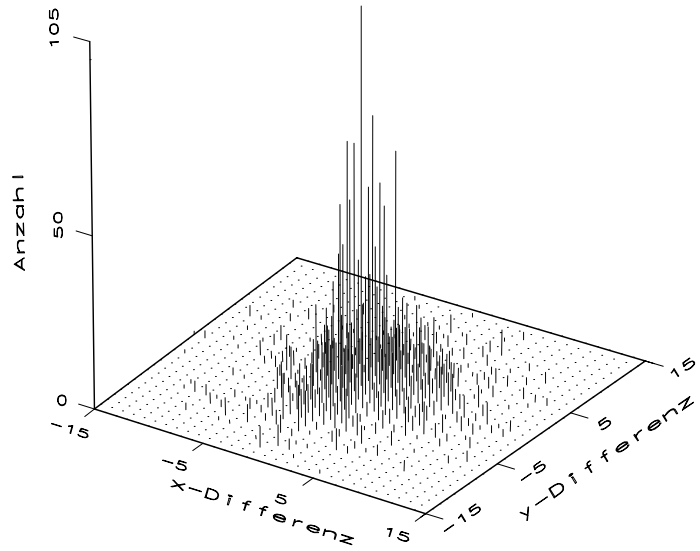


Abbildung 7.18: Bedingte Häufigkeitsverteilung der Position

$P(r_i\|r_j)$	$r_j = W$	$r_j = F$
$r_i = W$	0,96	0,04
$r_i = F$	0,00043	0,99957

Tabelle 7.6: Bedingte Wahrscheinlichkeiten der Relation *Verbunden*

Die bedingte Wahrscheinlichkeit, daß, falls zwei Elemente im Ausgangsdatensatz miteinander verbunden sind, auch die zugeordneten Elemente des anderen Datensatzes verbunden sind, ist dagegen nicht von der Anzahl der Elemente abhängig, sondern nur von den zugrundeliegenden Datenmodellen. Tabelle 7.6 zeigt die bedingten Wahrscheinlichkeiten, welche aus den manuellen Zuordnungen bestimmt wurden.

7.5 Dynamische Berechnung der relationalen Leistung

Da der attributive Anteil der Leistung einer Zuordnung nur von den beteiligten Elementen dieser einen Zuordnung abhängig ist, kann er bereits vor dem Aufstellen des Suchbaumes vollständig berechnet werden. Der relationale Anteil ist dagegen eine Funktion aller beteiligten Zuordnungen. Da jeder Knoten im Baum für eine andere Kombination von Zuordnungen steht, wird der relationale Anteil dynamisch während des Aufbaus des Suchbaumes für jeden Knoten neu berechnet. Daher ist er bei Beginn der Baumsuche Null und kann erst bei Erreichen eines Blattes vollständig bestimmt werden. Zur Berechnung ist für die aktuelle Kombination von Zuordnungen je eine Adjazenzmatrix für die ATKIS- und die GDF-Elemente zu erstellen. In diesen beiden Matrizen sind alle Informationen enthalten, um den relationalen Anteil zu berechnen. Bei der Lösung des Zuordnungsproblem werden sehr viele Knoten expandiert. Daher muß versucht werden, den Speicher- und Zeitbedarf für die Auswertung der Matrizen zu minimieren.

Bei der Durchführung des Algorithmus aus Kapitel 6.11 gibt es zwei Möglichkeiten, um von einem Knoten im Baum zum nächsten zu kommen. Aus der Liste der noch nicht verwendeten Zuordnungen wird die Zuordnung

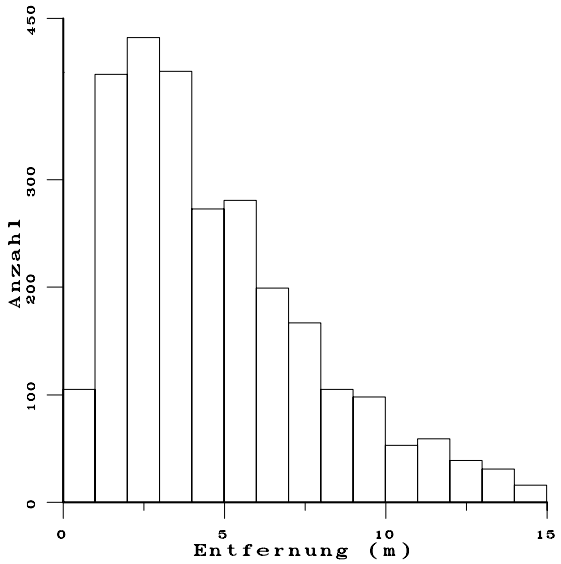


Abbildung 7.19: Häufigkeitsverteilung der Entfernung

mit der höchsten Leistungsabschätzung herausgegriffen und entweder in die Lösungsmenge aufgenommen oder verworfen. Wird sie verworfen, so ändert sich an der Leistung des relationalen Anteil nichts, da weiterhin die gleiche Kombination von Zuordnungen betrachtet wird. Wird sie dagegen in die Lösungsmenge aufgenommen, muß der relationale Anteil neu berechnet werden. Hierzu ist es nicht notwendig die Adjazenzmatrizen vollständig neu zu berechnen, sondern es ist ausreichend, nur die hinzugekommene Zuordnung mit den bereits bestehenden Zuordnungen zu vergleichen.

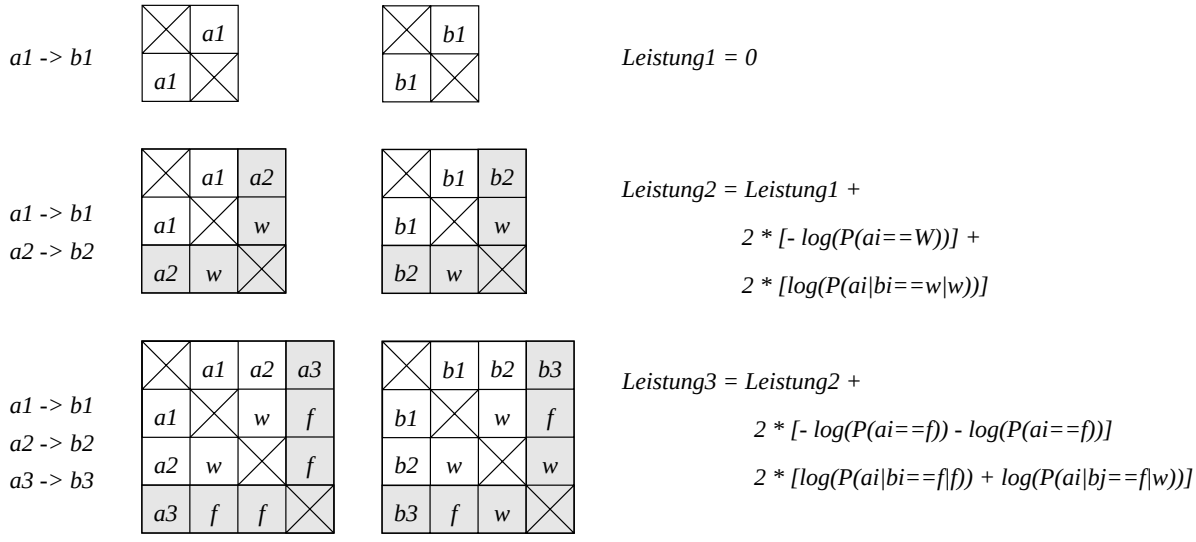


Abbildung 7.20: Dynamische Berechnung der relationalen Information

Abbildung 7.20 zeigt die dynamische Berechnung der relationalen Information. Es sind die Adjazenzmatrizen für drei aufeinanderfolgende Knoten im Suchbaum dargestellt. Die ersten beiden Matrizen stellen die Situation dar, in der erst eine Zuordnung in die Lösungsmenge aufgenommen wurde. In diesem Fall ist der relationale Anteil der Leistung gleich Null, da bei nur einem Element noch keine Relationen innerhalb der Datensätze bestehen. Bei Hinzunahme einer zweiten Zuordnung kann ein relationaler Anteil berechnet werden. Die neue Leistung errechnet sich durch die Leistung des Vorgängerknotens plus die Leistung durch die neu hinzugekommene Spalte bzw.

Zeile (da die Adjazenzmatrizen symmetrisch sind, reicht es aus, nur eine Zeile oder Spalte zu betrachten). Aus der Adjazenzmatrix der ATKIS-Elemente wird der Selbstinformationsgehalt und durch einen Vergleich mit der GDF-Adjazenzmatrix die bedingte Information berechnet. Da zur Berechnung der neuen Leistung nur der Wert der Leistung des Vorgängerknotens benötigt wird, ist es nicht nötig die Matrizen in den Knoten vollständig zu speichern, sondern es reicht aus, den Wert der Leistung von einem Knoten zum nächsten zu übergeben. Allgemein berechnet sie die Leistung I in einem Knoten j:

$$I(Knoten_j) = I(Knoten_{j-1}) + 2 * \sum_{n=1}^{j-1} -\log(P(A_{jn})) + \log(P(A_{jn}|B_{jn})) \tag{7.7}$$

wobei A für die ATKIS-Adjazenzmatrix steht und B für die GDF-Adjazenzmatrix.

7.6 Abschätzung der Leistung

Bevor ein Knoten beim Aufbau des Suchbaums expandiert wird, erfolgt eine Abschätzung der maximal erreichbaren Leistung des unter diesen Knoten liegenden Teilbaumes. Nur wenn die Abschätzung höher ist, als die bisher erzielte Höchstleistung muß der Teilbaum evaluiert werden. Wie in Kapitel 6.11 aufgezeigt, besteht die Abschätzung $f^*(n)$ in einem Knoten n aus zwei Teilen:

$$f^*(n) = g(n) + h^*(n) \tag{7.8}$$

Dabei steht $g(n)$ für die bisher erzielte Leistung (die Berechnung von $g(n)$ wurde im vorigen Abschnitt diskutiert) und $h^*(n)$ für die Abschätzung der maximalen Leistung $h(n)$, die durch den Teilbaum unterhalb des Knotens n erreicht wird. Als Randbedingung gilt $h^*(n) \geq h(n)$ für alle n, damit die optimale Lösung auf jeden Fall gefunden wird.

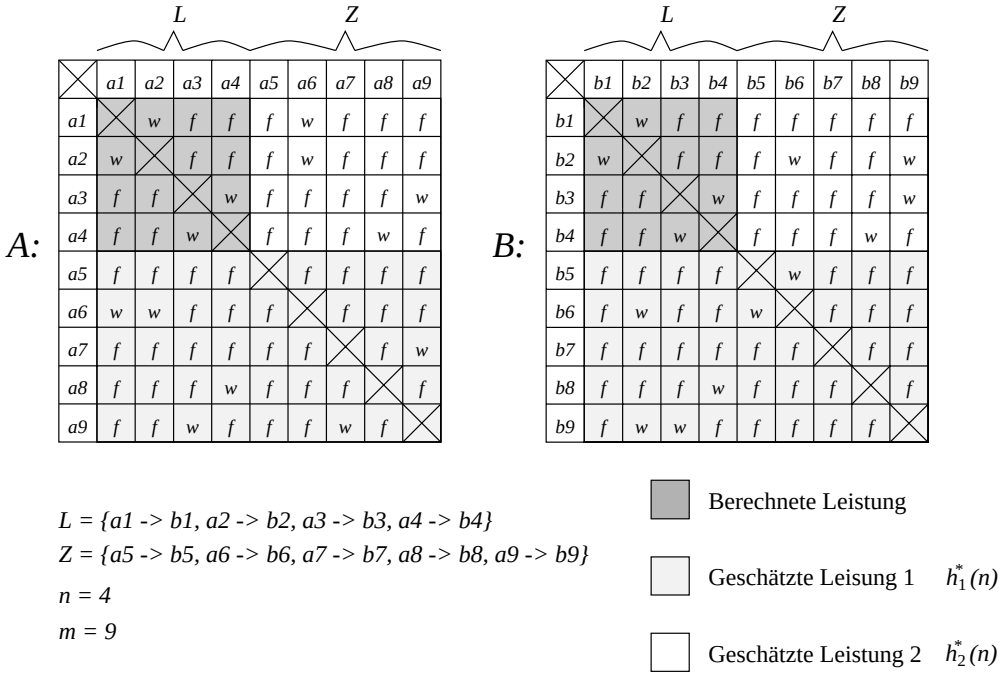


Abbildung 7.21: Matrix zur Abschätzung der Information

Abbildung 7.21 zeigt an einem Beispiel die zur Abschätzung der Leistung benötigten Adjazenzmatrizen. Die linke Matrix A wird aus den Elementen des ATKIS-Datensatzes berechnet, die rechte Matrix B aus den Elementen des GDF-Datensatzes. Die akzeptierten Zuordnungen sind in der Menge $L = \{z_1, z_2, \ldots, z_n\}$ abgespeichert. Diese Menge ist eindeutig und darf keine Widersprüche mehr enthalten (siehe Kapitel 6.11). Die Menge $Z = \{z_{n+1}, z_{n+2}, \ldots, z_m\}$ enthält alle Zuordnungen, die noch nicht verwendet oder durch das Aufdecken

Testgebiet	1	2	3	4	Gesamt
Anzahl gefundene Zuordnungen	208	349	332	469	1358
Anzahl ATKIS-Elemente im Testgebiet	363	530	530	640	2063
Anzahl richtig zugeordnete ATKIS-Elemente	328	523	515	620	1986
In Prozent	90,35	98,67	97,16	96,87	96,26

Tabelle 7.7: Ergebnisse der automatischen Zuordnung

von Widersprüchen eliminiert wurden. Aus den linken oberen $n*n$ Elementen wird die bereits erreichte Leistung $g(n)$ berechnet. Die Abschätzung der Leistung $h^*(n)$ erfolgt in zwei Schritten aus den restlichen Elementen der Matrizen:

$$h^*(n) = h_1^*(n) + h_2^*(n) \tag{7.9}$$

Eine Endzuordnung, die unterhalb eines betrachteten Knoten liegt, besteht aus den Zuordnungen der Menge L zuzüglich einer Kombination von beliebig vielen Elementen der Menge Z. Zuerst werden nur die Elemente aus der Menge Z betrachtet. Da im voraus nicht bekannt ist, welche Elemente aus Z in der Endzuordnung vorkommen (dies ist ja genau die Aufgabe des Zuordnungsprozesses), wird für jede Zuordnung aus Z der maximal erreichbare Betrag berechnet, den diese Zuordnung zur Gesamtzuordnung beitragen kann. Die Summe der Beträge ergibt den Teil 1 der Abschätzung der Gesamtleistung:

$$h_1^*(n) = \sum_{i=n+1}^{m} Max\_Leistung(z_i) \text{ für alle } Max\_Leistung(z_i) \geq 0 \tag{7.10}$$

Die Leistung der Attribute ist konstant, wogegen die Leistung der Relationen abhängig von der gerade betrachteten Kombination der Zuordnungen ist. Daher errechnet sich die maximale erreichbare Leistung einer Zuordnung durch die Attributleistung zuzüglich der Summe der Leistungen der Relationen, falls diese positiv sind:

$$Max\_Leistung(z_i) = I_{attr}(z_i) + \sum_{j=1}^{m} I_{rel}(z_i, z_j) \text{ für alle } I_{rel}(z_i, z_j)) \geq 0 \tag{7.11}$$

Der zweite Teil der Abschätzung errechnet sich analog zu Teil 1 aus den Zuordnungen der Menge L, jedoch muß hier die bereits erzielte Leistung $g(n)$ von der Abschätzung abgezogen werden:

$$h_2^*(n) = \sum_{i=1}^{n} Max\_Leistung(z_i) - g(n) \text{ für alle } Max\_Leistung(z_i) \geq 0 \tag{7.12}$$

7.7 Endgültige Zuordnungen

Zur Bewertung des Verfahrens werden die berechneten Zuordnungen mit den manuell erzeugten Zuordnungen verglichen. Tabelle 7.7 zeigt die Auswertung der vier Testgebiete. Als Maß zur Beurteilung der automatisch erzeugten Zuordnungen wird der Prozentsatz der richtig zugeordneten ATKIS-Elemente verwendet. Ein ATKIS-Element gilt dann als richtig zugeordnet wenn es exakt den gleichen Elementen zugeordnet wurde, wie bei der manuellen Zuordnung.

Im Durchschnitt wurden 96,26 Prozent der ATKIS-Elemente richtig zugeordnet. Es kann gesehen werden, daß Testgebiet 1 einen signifikant schlechteren Prozentsatz an richtigen Zuordnungen gegenüber den anderen Testgebieten aufweist. Der Grund hierfür ist, daß sich die im Rahmen dieser Arbeit verwendeteten ATKIS- und GDF-Daten in ihrer Aktualität unterscheiden. Im Testgebiet 1 wurde eine große Anzahl von Kreuzungsbereichen aufgrund einer Verkehrsberuhigung stark geändert. Im GDF-Datensatz sind die Kreuzungsbereiche im alten Zustand und im ATKIS-Datensatz im geänderten Zustand erfaßt. Abbildung 7.22 zeigt ein Beispiel dieser Situation. Während die Straßen, die zum Kreuzungsbereich führen, eindeutig zugeordnet werden können, ist es nicht möglich, sinnvolle Zuordnungen im Kreuzungsbereich zu bilden. Beim automatischen Zuordnungsprozeß

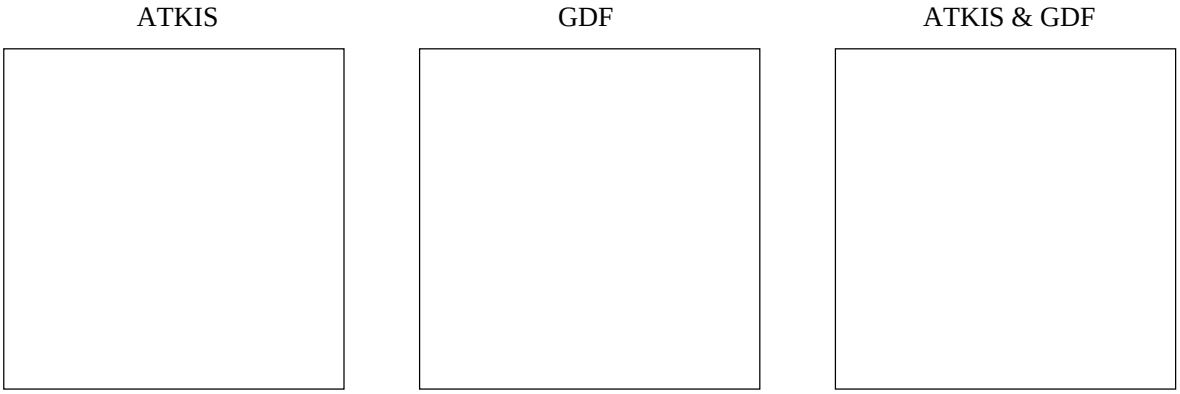


Abbildung 7.23: Beispiel für Problembereich beim Vergleich manueller und automatischer Zuordnung

7.8 Qualitätsmaße der Zuordnungen

Nach der automatischen Zuordnung der Daten, kann eine manuelle Nachbearbeitung erfolgen. Zur Reduzierung des Aufwands der Nachbearbeitung werden Qualitätsmaße benötigt, die angeben, wie sicher eine einzelne Zuordnung bzw. wie sicher die Gesamtzuordnung zwischen zwei Datensätzen ist. Zur Bewertung von einzelnen Zuordnungen können die Attribute und die Relationen der Elemente der Zuordnungen betrachtet werden. Hierbei kann abhängig von der zugrunde liegenden Anwendung definiert werden, ab wann eine Zuordnung nicht mehr als sicher einzustufen ist. Wird z.B. eine Zuordnung durchgeführt, um die geometrische Qualität eines Datensatzes zu verbessern, können alle Zuordnungen als unsicher markiert werden, deren Elemente einen bestimmten Abstand überschreiten. Eine Anwendung, die Daten zur Verkehrsnavigation nutzt, kann z.B. alle Zuordnungen als unsicher markieren, deren Elemente nicht exakt die gleichen topologischen Eigenschaften besitzen.

Testgebiet	1	2	3	4
Anzahl Zuordnungen (ohne Wildcard-Zuordnungen)	208	349	332	469
Summe bedingte Information	2200	3278	3225	4579
Durchschnittliche bedingte Information pro Zuordnung	10.57	9.36	9.71	9.76
Prozentsatz richtig zugeordneter ATKIS-Elemente	90.35	98.67	97.16	96.87

Tabelle 7.8: Durchschnittliche gegenseitige Information als Qualitätsmaß

7.8.1 Qualität der Gesamtzuordnung

Im folgenden wird untersucht, ob die Informationsmaße, welche während des Zuordnungsprozesses berechnet werden, ebenfalls zur Bestimmung der Qualität der Zuordnungen genutzt werden können. Als erstes wird die Qualität einer Gesamtzuordnung zweier Datensätze betrachtet. Hierbei ist die bedingte Information zur Beurteilung der Qualität geeignet, die ein Maß der Überraschung ist, daß ein Symbol b_j empfangen wurde, wenn ein Symbol a_i gesendet wurde. Je niedriger die durchschnittliche bedingte Information pro Zuordnung ist, desto mehr entspricht der Übertragungskanal einem idealen Kanal[5]. Es wird daher die durchschnittliche bedingte Information pro Zuordnung zwischen zwei Datensätzen als Qualitätsmerkmal der Gesamtzuordnung definiert. Tabelle 7.8 zeigt die Auswertungen für die einzelnen Testgebiete. Die Auswertungen in der Tabelle geben an, daß die durchschnittliche bedingte Information pro Zuordnung ein sehr guter Indikator für die Prozentzahl der richtig zugeordneten Elemente ist. Je höher die durchschnittliche bedingte Information pro Zuordnung ist, desto weiter ist der Übertragungskanal von einem idealen Kanal entfernt und desto niedriger ist der Prozentsatz richtig zugeordneter Elemente.

7.8.2 Qualität eines Zuordnungspaares

Um die Qualität eines einzelnen Zuordnungspaares zu bewerten, kann ebenfalls die bedingte Information genutzt werden. Da die bedingte Information den Informationsverlust im Kanal beschreibt, ist sie ein Maß für die Ähnlichkeit der Zuordnungspartner. Je höher die bedingte Information ist, desto größer ist der Unterschied der Attribute und Relationen der Zuordnungspartner. Für eine interaktive Nachbearbeitung einer automatischen Zuordnung müssen insbesondere die Fehlzuordnungen identifiziert werden. Abbildung 7.24 zeigt einen Vergleich der Häufigkeiten der bedingten Information zwischen allen Zuordnungen und den Fehlzuordnungen. In der Abbildung kann gesehen werden, daß das Maximum der Häufigkeiten der bedingten Information der Fehlzuordnungen auf der X-Achse nach rechts, gegenüber der Auswertung aller Zuordnungen, verschoben ist.

Die Aussagekraft der bedingten Information alleine ist jedoch zu schwach um Fehlzuordnungen automatisch zu identifizieren. Als weiteres Maß wird die gegenseitige Information betrachtet. Abbildung 7.25 zeigt die Häufigkeitsauswertung der gegenseitigen Information. Bei einer niedrigen gegenseitigen Information ist die Wahrscheinlichkeit höher, daß es sich um eine Fehlzuordnung handelt als bei einer hohen gegenseitigen Information.

Um die Zuordnungen zu bewerten werden drei Klassen definiert. Hierzu wird die durchschnittliche bedingte Information pro Zuordnung $Info^{\emptyset}_{bed}$ sowie die durchschnittliche gegenseitige Information pro Zuordnung $Info^{\emptyset}_{geg}$ berechnet. Die Einteilung einer Zuordnung Z_i in die unterschiedlichen Klassen wird dann wie folgt durchgeführt:

$$Klasse(Z_i) = \begin{cases} 1 \quad \text{für} \quad Info_{geg}(Z_i) < Info^{\emptyset}_{geg} \wedge Info_{bed}(Z_i) > Info^{\emptyset}_{bed} \\ 2 \quad \text{für} \quad (Info_{geg}(Z_i) < Info^{\emptyset}_{geg} \wedge Info_{bed}(Z_i) < Info^{\emptyset}_{bed}) \quad \vee \\ \qquad\qquad (Info_{geg}(Z_i) > Info^{\emptyset}_{geg} \wedge Info_{bed}(Z_i) > Info^{\emptyset}_{bed}) \\ 3 \quad \text{für} \quad \text{sonst} \end{cases} \tag{7.13}$$

Tabelle 7.9 zeigt die Auswertung der Zuordnungen gemäß dieser Einteilung.

Klasse 1 beschreibt alle Zuordnungen, deren bedingte Information höher als die durchschnittliche bedingte Information und deren gegenseitige Information niedriger als die durchschnittliche bedingte Information ist. Diese Klasse enthält ca. 27 Prozent aller Zuordnungen der Testgebiete und ca. 64,5 Prozent aller Fehlzuordnungen. Zuordnungen, die in dieser Klasse liegen, besitzen stark unterschiedliche Attribute und Relationen und werden daher als *unsicher* eingestuft.

[5] Ein idealer Kanal liegt dann vor, wenn die durchschnittliche bedingte Information pro Zuordnung gleich Null ist.

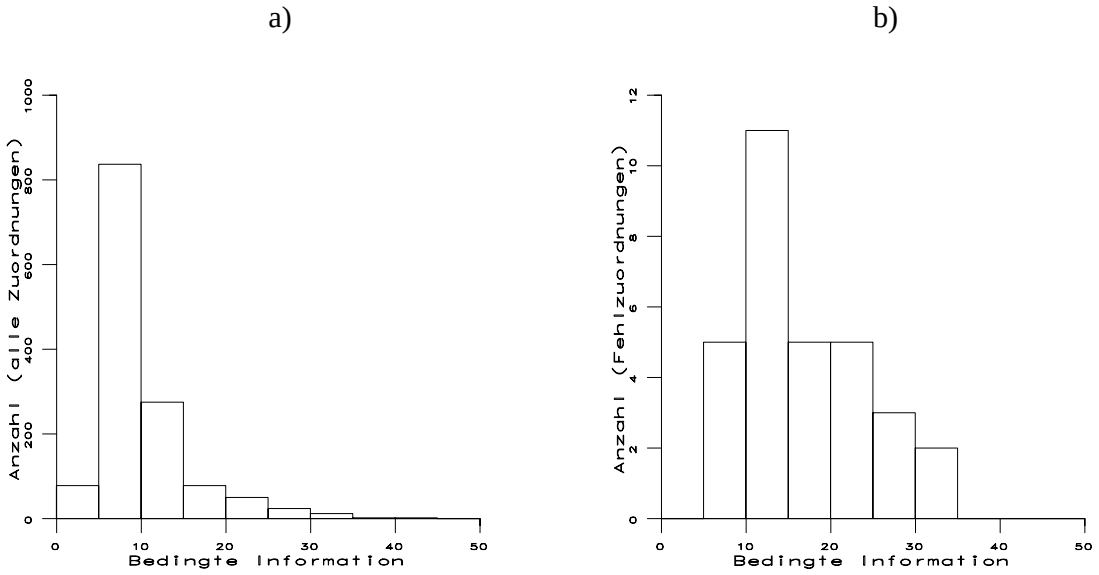


Abbildung 7.24: Auswertung der bedingten Information

Testgebiet	1			2			3			4		
Klasse	1	2	3	1	2	3	1	2	3	1	2	3
Anz. Zuordnungen	68	68	72	102	118	129	84	132	116	113	176	180
Anz. Fehlzuordnungen	9	3	1	3	1	0	5	1	0	3	4	1

Tabelle 7.9: Einteilung der Zuordnungen in die verschiedenen Klassen

In der Klasse 2 befinden sich alle Zuordnungen, die nur eines der Kriterien von Klasse 1 erfüllen. In dieser Klasse sind ca. 36 Prozent aller Zuordnungen sowie ca. 29 Prozent aller Fehlzuordnungen zu finden. Zuordnungen dieser Klasse werden als *bedingt sicher* eingestuft.

Klasse 3 enthält alle Zuordnungen, die weder in Klasse 1 noch in Klasse 2 liegen. Ca. 36,6 Prozent der Zuordnungen der vier Testgebiete befinden sich in dieser Klasse, jedoch nur ca. 6 Prozent der Fehlzuordnungen. Die Zuordnungen in dieser Klasse sind als *sicher* einzuordnen. Fehlzuordnungen, die trotzdem in dieser Klasse liegen, entstehen in Bereichen, in denen keine eindeutige Zuordnung vom Operateur gefunden werden kann, da mehrere Kombinationen von Zuordnungen gleich gut zu bewerten sind (siehe auch Kapitel 7.1.3).

7.9 Zeitverhalten

Zur Realisierung des Verfahrens wurden zwei Teilprogramme entwickelt. Im ersten Teil werden die Daten eingelesen und die Liste der potentiellen Zuordnungspaare berechnet, wogegen die eigentliche Zuordnung durch das Baumsuchverfahren in einem zweiten Teil durchgeführt wird. Die Implementierung des ersten Teils erfolgte in der Rapid Prototyping Sprache *Python* [Rossum 1994*b*]. Hierbei handelt es sich um eine objektorientierte Interpretersprache, die es erlaubt, sehr kompakte und leicht lesbare Programme in kurzer Entwicklungszeit zu erstellen. Der Nachteil der Sprache ist, daß die kürzere Entwicklungszeit oftmals mit längeren Programmlaufzeiten erkauft werden muß. Da zum Zuordnen der Daten sehr viele Knoten im Suchbaum expandiert werden müssen, wurde das Baumsuchverfahren selbst in der Programmiersprache C geschrieben, um bessere Programmlaufzeiten zu erzielen. Die angegebenen Zeiten wurden auf einer Silicon Graphics INDY mit einem R5000 180 Mhz Prozessor ermittelt.

Tabelle 7.10 zeigt die benötigten Rechenzeiten der einzelnen Testgebiete. Die Rechenzeit hängt von mehreren Faktoren ab. Den Hauptanteil macht dabei die Anzahl der beteiligten Elemente aus. Weiter spielt die Ähnlichkeit der Datensätze und Dichteverteilung des Straßennetzes eine Rolle. In Bereichen in denen die Elemente viele topologische Partnerelemente besitzen, ergeben sich große Cluster, was zu großen Suchbäumen führt und dadurch

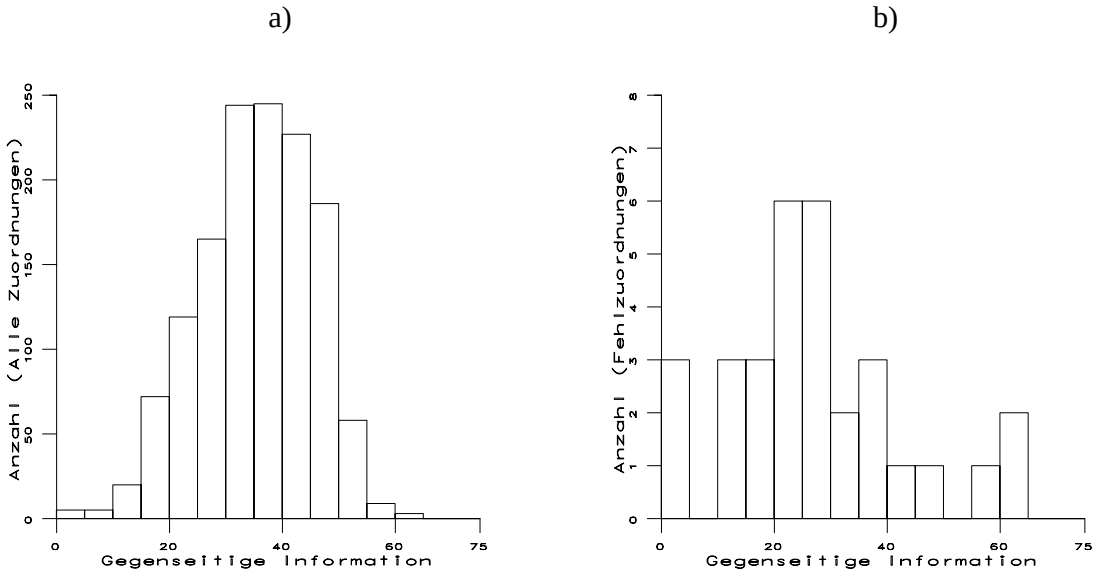


Abbildung 7.25: Auswertung der gegenseitigen Information

Testgebiet	1	2	3	4
Anzahl ATKIS-Elemente	363	530	530	640
Anzahl GDF-Elemente	435	668	853	963
Gesamt	798	1198	1383	1603
Anzahl expandierte Knoten	1.145.386	917.545	2.196.406	1.562.566
Rechenzeit Vorverarbeitung	54 min	1 h 31 min	2 h 21 min	2 h 34 min
Rechenzeit Baumsuche	1 h 14 min	1 h 47 min	4 h 08 min	4 h 13 min
Gesamtrechenzeit	2 h 07 min	3 h 18 min	6 h 29 min	6 h 47 min

Tabelle 7.10: Rechenzeit der Testgebiete

die Rechenzeit erhöht. Der Grund für die überdurchschnittliche Zahl von expandierten Knoten des Testgebietes 2 ist eine hohe Anzahl von nah beieinander stehenden parallelen Linienstücken. Dadurch enstehen sehr viele mögliche Kombinationen von Zuordnungspaaren.

Kapitel 8

Zusammenfassung und Diskussion

Objekte aus dem Bereich des Straßenverkehrs werden in Deutschland sowohl im Amtlichen Topographischen Kartographischen Informationssystem (ATKIS) als auch im Geographic Data File (GDF) erfaßt. Eine steigende Nachfrage nach diesen Daten erfolgt unter anderem von Betreibern von Verkehrsmanagement- und Leitsystemen, welche hohe Ansprüche an die Qualität und Aktualität stellen. Um Ansprüche dieser Art befriedigen zu können, ist ein großer Kosten- und Zeitaufwand bei der Erfassung und Fortführung notwendig. Eine Integration von ATKIS- und GDF-Daten könnte diesen Aufwand reduzieren und gleichzeitig das Anwendungspotential steigern. Um eine Integration durchführen zu können, müssen Elemente, welche die gleichen Objekte beschreiben, in den Datensätzen identifiziert werden. Da die Objektstrukturen aufgrund der verschiedenen Weltsichten sehr unterschiedlich sind, muß die Identifikation gleicher Objekte mit Hilfe von Zuordnungsalgorithmen durchgeführt werden.

In dieser Arbeit werden Zuordnungsverfahren für raumbezogene Daten diskutiert. Um in die Problematik einzuführen, erfolgt zuerst eine Darstellung der Mehrfacherfassung von Straßenverkehrsdaten in Deutschland. Dies ermöglicht eine Einteilung in verschiedene Stufen der Integration sowie eine Klassifizierung der grundsätzlichen Vorgehensweisen. Danach werden die Datenmodelle von ATKIS und GDF eingehend untersucht. Diese Untersuchungen sind die Grundlage für einen anschließenden Vergleich von ATKIS und GDF. Die Zuordnung von raumbezogenen Daten wird zuerst von einem allgemeinen Standpunkt aus betrachtet, ohne dabei auf anwendungsspezifische Fragestellungen einzugehen. Es wird ein Verfahren vorgestellt, welches sich in mehrere Teilschritte untergliedern läßt. Zuerst werden in einem Vorverarbeitungsschritt globale Fehler in den Datensätzen minimiert. Anschließend wird eine Liste mit potentiellen Zuordnungspaaren erstellt. Diese Liste enthält nur Zuordnungen die bestimmte geometrische und semantische Vorgaben erfüllen. Es erfolgt eine Bewertung der Zuordnungspaare mit Hilfe einer Leistungsfunktion. Anschließend wird mit einem Baumsuchverfahren die optimale Zuordnung zwischen den beiden Datensätzen bestimmt. Das Verfahren wird mit ATKIS- und GDF-Daten aus dem Stadtgebiet Stuttgart getestet. Die zur Bestimmung der Leistungsfunktion notwendigen statistischen Auswertungen zwischen den Datensätzen werden ausführlich diskutiert. Eine abschließende Diskussion der Ergebnisse sowie ein Ausblick auf weitere Forschungstätigkeiten erfolgen in diesem Kapitel.

8.1 Diskussion der Ergebnisse

Die zwei Hauptprobleme bei der automatischen Zuordnung von raumbezogenen Daten sind das exponentielle Wachstum des Suchraumes sowie die Definition eines geeigneten Ähnlichkeitsmaßes zur Bewertung der Zuordnungen. Eine Vollraumsuche kann aufgrund der großen Anzahl der Elemente nicht durchgeführt werden. Es werden daher Heuristiken verwendet, um den Suchraum zu verkleinern. In dieser Arbeit wurden verschiedene Techniken untersucht, um die Anzahl der potentiellen Zuordnungspartner zu minimieren. Von großer Bedeutung ist eine Vorverarbeitung, bei der die Daten topologisch bereinigt und geometrische globale Fehler zwischen den Datensätzen minimiert werden. Hierzu werden interaktiv Passpunkte gemessen und daraus die Parameter einer globalen Transformation bestimmt. Bei einer Überlagerung der Datensätze kann festgestellt werden, daß trotz der Minimierung des globalen Fehlers, aufgrund lokaler Inhomogenitäten eine Pufferbreite von 15 m notwendig ist, um alle Elemente einander zuzuordnen, welche dieselben Objekte der Landschaft beschreiben.

In den Puffern werden viele Elemente gefunden, die aufgrund ihrer geometrischen Ausprägung nicht als Zuordnungspartner in Betracht kommen. Daher werden zur Einschränkung des Suchraumes nur solche Zuordnungspaare in die Liste der potentiellen Zuordnungen aufgenommen, deren Winkel sich um weniger als 30 Grad unterscheidet. Diese Einschränkung gilt nur für die automatische Zuordnung der Daten. Bei der manuellen Zuordnung wird kein Maximum für die Winkeldifferenz definiert. Ein Vergleich der manuellen Zuordnungen mit der Liste der potentiellen Zuordnungen zeigt jedoch, daß durch diese Einschränkung lediglich zwei von 1405 manuellen Zuordnungen nicht gefunden werden.

Da bei der Modellierung von ATKIS und GDF verschiedene Weltsichten zugrunde liegen, gibt es teilweise starke Unterschiede in der geometrischen Ausprägung der Daten. Daher reicht es nicht aus, nur 1 : 1 Zuordnungen zu

betrachten, sondern es müssen $n : m$ Zuordnungen mit $n, m \geq 0$ zugelassen werden. Insbesondere der Fall n bzw. m gleich Null ist von großer Bedeutung, da er die Situation beschreibt, in der ein Element in einem Datensatz erfaßt wurde, jedoch nicht im anderen. Um die $n : m$ Zuordnungen bilden zu können, wurde ein Verfahren entwickelt, welches mit wachsenden Puffern einzelne Elemente zu logischen Elementmengen gruppiert. Auch beim Wachstum des Puffers müssen Einschränkungen verwendet werden, damit die Liste der potentiellen Zuordnungspaare nicht zu stark anwächst. Elemente dürfen nur dann gruppiert werden, wenn der Winkel zwischen den Liniensegmenten des gemeinsamen Knotens kleiner als 45 Grad ist. Bei Knoten, in denen genau zwei Linien enden, darf der Puffer unabhängig vom Schnittwinkel der angrenzenden Liniensegmente wachsen. Die maximale Anzahl von Teilelementen einer logischen Elementmenge wurde in dieser Arbeit auf neun beschränkt.

Eine weitere Einschränkung von potentiellen Zuordnungspaaren wird durch die Eliminierung redundanter Knoten erreicht. Hierbei handelt es sich um Knoten, die nur aufgrund eines Attributwechsel erzeugt wurden und keine topologische Änderung oder einen Objektwechsel darstellen. Knoten dieser Art werden vor dem eigentlichen Zuordnungsprozess entfernt. Trotz dieser Vielzahl von Einschränkungen wurden lediglich 0.43 Prozent aller manuellen Zuordnungen nicht in die Liste der potentiellen Zuordnungspaare aufgenommen.

Um die Zuordnungen zu bewerten, wird ein Maß benötigt, welches die Ähnlichkeit der Attribute und Relationen wiederspiegelt. Die gegenseitige Information ist ein Maß dafür, wieviel Information ein Element über ein anderes Element besitzt. Von großer Wichtigkeit ist hierbei, daß Attributwerte oder Relationen, die nur selten vorkommen, bei der Zuordnung ein höheres Gewicht haben als solche, die sehr oft vorkommen. Die gegenseitige Information ist eine Leistungsfunktion und hat gegenüber einer Kostenfunktion den Vorteil, daß die Zuordnung eines Elementes zu einer "Wildcard" ($n : m$ Zuordnung mit n oder $m = 0$) mit keinen Strafkosten verbunden ist. Um die Parameter der Leistungsfunktion zu bestimmen, werden statistische Auswertungen zwischen den Datensätzen durchgeführt. Es ist jedoch nicht notwendig, empirisch Gewichtsfaktoren oder Startwerte durch Testläufe zu ermitteln. Die statistischen Auswertungen werden zusätzlich zur Bestimmung der geometrischen Beschränkungen genutzt.

Die Informationsmaße sind sowohl für diskrete als auch für kontinuierliche Signale definiert. Es wurde gezeigt, daß die Verluste durch die Diskretisierung eines kontinuierlichen Signal vernachlässigbar sind, wenn das Diskretisierungsintervall klein genug gewählt wird. Dadurch können Informationsmaße von diskreten und kontinuierlichen Signalen kombiniert werden und gemeinsam in den Zuordnungsprozess einfließen. Dieses gilt ebenso für symbolische (nicht numerische) Attribute, wie z.B. die Darstellungsfarbe von Elementen.

Die Summe der gegenseitige Information aller Zuordnungen wird mit einem Baumsuchverfahren maximiert. Es wird ein Bergsteige-Algorithmus verwendet, der immer den Knoten im Baum expandiert, bei dem der steilste Anstieg der Zielfunktion erfolgt. Vor der Expansion eines Knotens wird abgeschätzt, ob der unterhalb dieses Knotens liegende Teilbaum das bisher erreichte Leistungsmaxima überschreiten kann. Weiterhin werden im Baum alle Knoten eliminiert, die zu Widersprüchen in den Zuordnungen führen. Trotz dieser Techniken ist der verbleibende Suchraum immer noch zu groß, um damit räumliche Gebiete großer Ausdehnung in akzeptabler Zeit zu optimieren. Daher erfolgt eine Aufteilung des Suchraums in Cluster.

Ein Cluster besteht aus einer einzelnen Zuordnung plus den topologischen Nachbarn der Elemente der Zuordnung sowie deren topologischen Nachbarn. Die Cluster werden sukzessive, der Größe nach aufsteigend, optimiert. Die Bewertung der Leistungsfunktion erfolgt jedoch immer global für alle Zuordnungen. Diese Strategie zeigt Analogien zur Vorgehensweise bei der manuellen Zuordnung. Kann ein Operateur mit Hilfe der Attribute nicht entscheiden, ob zwei Elemente zuzuordnen sind, wird er die Nachbarelemente betrachten und unter Umständen zusätzlich deren Nachbarelemente. Eine Abhängigkeit der Zuordnung zu weiter entfernten Zuordnungen ist nur selten zu erwarten. Weiterhin wird er versuchen, diejenigen Elemente zuerst zuzuordnen, bei denen nur wenige Alternativen möglich sind und diese Zuordnungen dann als "Anker" für komplexere Bereiche zu nutzen. Dies entspricht der Vorgehensweise, die Cluster in der Reihenfolge ihrer Größe abzuarbeiten.

Um die Ergebnisse der automatischen Zuordnung bewerten zu können werden sie mit den manuell erzeugten Zuordnungen verglichen. Insgesamt werden 96,26 Prozent der ATKIS-Elemente, durch das automatische Verfahren genauso wie bei der manuellen Zuordnung einander zugeordnet. Bei der automatischen Zuordnung können Fehlzuordnungen innerhalb von stark unterschiedlich erfaßten Kreuzungsbereichen entstehen, bei denen oftmals nur eine Teilmenge der Daten zugeordnet werden kann. Werden Kreuzungsbereiche in einem Datensatz punktförmig und im anderen linienförmig erfaßt, können ebenfalls Fehlzuordnungen entstehen, da in diesen Fällen häufig keine eindeutige Zuordnung existiert. Hierbei handelt es sich jedoch oft nur um sehr kleine Inhomogenitäten, bei denen ein sehr kurzes Straßenstück zur topologisch benachbarten logischen Elementmenge gruppiert wird. Da Zuordnungspartner nur innerhalb eines Puffers um die Elemente gesucht werden, kann es zu keinen sehr groben Zuordnungsfehlern kommen.

Zur nachträglichen automatischen Identifizierung von unsicheren Zuordnungen oder Fehlzuordnungen können die Attribute und Relationen der Elemente untersucht werden. Hierzu können, abhängig von der zugrunde-

liegenden Anwendung, Schranken definiert werden, ab denen eine Zuordnung nochmals manuell zu überprüfen oder ganz zu entfernen ist. Als weiteres Qualitätsmaß einer Zuordnung können die zur Optimierung berechneten Informationsmaße genutzt werden. Zur Identifizierung von Fehlzuordnungen werden alle Zuordnungspaare in die drei Klassen *sicher*, *bedingt sicher* und *unsicher* eingeteilt. Die Klasseneinteilung erfolgt mit Hilfe der bedingten und der gegenseitigen Information. Die Klasse *unsicher* enthält ca. 27 Prozent aller Zuordnungen der Testgebiete sowie ca. 64,5 Prozent aller Fehlzuordnungen. Als weiteres Qualitätsmaß wurde die durchschnittliche bedingte Information definiert, die ein guter Indikator für den Prozentsatz richtiger Zuordnungen eines Datensatzes ist.

Das Verfahren wurde an Gebieten mit unterschiedlicher Straßendichte getestet. Es ist sehr robust und findet in allen Gebieten ungefähr den gleichen Prozentsatz an richtigen Zuordnungen. Ausnahme davon ist ein Testgebiet, in dem sich die ATKIS- und GDF-Daten sehr stark unterscheiden. Hier wurde mit 9.65 Prozent ein signifikant höhererer Prozentsatz an Fehlzuordnungen gefunden als in den anderen Testgebieten. Durch die Bewertung der Qualität der Einzelzuordnungen und der Gesamtzuordnung können diese Fehlzuordnungen jedoch automatisch identifiziert werden. Die Rechenzeit des Programms liegt für alle vier Testgebiete im Stundenbereich.

Im Bereich der Zuordnung bzw. Integration von raumbezogenen Daten gibt es zum jetzigen Zeitpunkt nur wenige Untersuchungen. Gerade bei der Durchführung interdisziplinärer Projekte ist es jedoch oft notwendig, Daten aus verschiedensten Quellen zu nutzen. Die vorgestellten Verfahren zeigen auf, wie ein interoperables GIS (IOGIS) realisiert werden kann. Neben der eigentlichen Zuordnung der Daten ergeben sich durch die informationstheoretischen Untersuchungen neue Maße um zwei Datensätze qualitativ miteinander vergleichen zu können. In der vorliegenden Arbeit wird erstmals eine Integration von ATKIS- und GDF-Daten eingehend untersucht. Erste Anwendungen lassen sich direkt aus der Zuordnung der Datensätze ableiten. Durch die automatische Zuordnung wird es z.B. möglich die Geometrien der beiden Datensätze einem Vergleich zu unterziehen. Hierdurch können Erfassungsfehler bzw. Änderungen durch Fortführungen automatisch erkannt werden. Eine weitere Anwendung mit hohem Automatisierungspotential ist z.B. der Import von GDF-Straßennamen in den ATKIS-Datensatz, da diese in der derzeitigen Ausbaustufe DLM 25/1 nicht erfaßt werden.

8.2 Ausblick auf zukünftige Arbeiten

In diesem Abschnitt werden Fragestellungen diskutiert, welche im Zusammenhang mit der automatischen Zuordnung von raumbezogenen Daten von Bedeutung sind und Schwerpunkt zukünftiger Forschungen sein sollten.

Sowohl bei der manuellen als auch bei der automatischen Zuordnung ergeben sich in komplexen Kreuzungsbereichen Probleme beim Aufstellen der Zuordnungspaare. Diese Probleme können umgangen werden, indem diese Bereiche vor der Zuordnung maskiert werden. Dies kann manuell durch einen Operateur erfolgen oder mit Hilfe eines automatischen Verfahrens, welches in der Lage ist, in vektoriellen Straßendaten komplexe Kreuzungsbereiche zu erkennen. Hierzu müssen implizite Informationen aus den Daten abgeleitet werden. Die Entwicklung von Techniken dieser Art ist ein sehr aktueller Forschungszweig, welcher auch mit dem Begriff “Data Mining“ bezeichnet wird (siehe z.B. [Holsheimer und Siebes 1994]). Mit Hilfe einer automatischen Erkennung von Kreuzungsbereichen können diese nicht nur ausmaskiert werden, sondern es können in diesen Bereichen gezielt andere Techniken für die Zuordung der Daten verwendet werden. Weitere Informationen, welche den Zuordnungsprozess unterstützen können, sind z.B. Siedlungsformen oder Flächennutzung.

Die Qualität der Ergebnisse der automatischen Zuordnung hängt wesentlich von den statistischen Auswertungen der Datensätze ab. Die in dieser Arbeit untersuchten Testgebiete enthalten Daten aus unterschiedlich dicht besiedelten Stadtgebieten. Hier ist eine Verfeinerung der Auswertungen durchzuführen. Es ist zu untersuchen, ob sich die Auswertungen für dicht und dünn besiedelte Gebiete unterscheiden, ob es regionale Unterschiede gibt und welche Auswirkungen die betrachteten Objektklassen haben. Weiter ist eine Auswertung von Datensätzen aus verschiedenen Bundesländern durchzuführen.

Die Durchführung statistischer Auswertungen für viele unterschiedliche Gebiete ist zeitintensiv. Der hierbei einzubringende Aufwand lohnt sich nur, wenn anschließend Daten in großem Umfang automatisch zugeordnet werden sollen. Mit Hilfe eines lernenden Verfahren kann der Aufwand für die manuelle Zuordnung reduziert werden. Um Daten zuzuordnen, für die keine statistischen Auswertungen vorliegen, kann mit Defaultparametern eine automatische Zuordnung durchgeführt werden. Anschließend werden die Ergebnisse einem Operateur interaktiv zur Bewertung dargestellt. Mit Hilfe dieser Bewertungen kann die Berechnung der Leistungsfunktion solange sukzessive verfeinert werden, bis die Ergebnisse den Anforderungen entsprechen.

Ein großer Vorteil des in dieser Arbeit vorgestellten Ansatz ist, daß unsichere Zuordnungen und Fehlzuordnungen aufgrund ihrer Geometrie und Topologie sowie mit Hilfe der bedingten Information identifiziert werden können. Der nächste Schritt wäre nun eine automatische Konfliktlösungsstrategie. Hierzu können wissensbasierte Methoden verwendet werden, welche die Konfliktstelle untersuchen, um zu klären, was die Ursache des

Fehlers ist. Eine Fehlzuordnung kann beispielsweise dadurch entstehen, daß in einem der Datensätze ein Kreuzungsbereich punktförmig und im anderen linienförmig erfaßt wurde. Kann diese Situation automatisch erkannt werden, besteht die Möglichkeit in diesem Fall eine Zuordnung zwischen einem oder mehreren linienförmigen Elementen zu einem punktförmigen Element durchzuführen. Kann ein Element nicht zugeordnet werden, liegt es häufig daran, daß dieses Element im anderen Datensatz überhaupt nicht erfaßt wurde. In diesem Fall kann eine Konfliktlösungsstrategie das Erzeugen eines neuen Elementes vorschlagen. Es werden daher automatisierte Verfahren benötigt, welche die manuelle Nachbearbeitung der zugeordneten Datensätze minimieren.

Die automatische Zuordnung von raumbezogenen Daten ist ein erster Schritt einer Integration verschiedener Datenmodelle. Nachdem die Daten zugeordnet wurden, stellt sich die Frage der Weiterverarbeitung der Zuordnungen. Es sind zwei unterschiedliche Möglichkeiten denkbar. Die erste Möglichkeit ist, die Daten in beiden Datenmodellen weiterhin getrennt zu speichern. In diesem Fall ist zu untersuchen, wie bidirektionale Links zwischen den Datensätzen aufgebaut werden können, die es ermöglichen, Änderungen, die in einem der Datensätze durchgeführt wurden, im anderen Datensatz nachzuvollziehen. Beim Löschen oder Erzeugen von neuen Daten kann sich dies jedoch auch auf die Zuordnungen auswirken, was zu Problemen führen kann. Die andere Möglichkeit ist, die Daten vollständig zu einem Datensatz zusammenzuführen, um dadurch die Qualität zu verbessern und das Anwendungspotential, durch eine größere Anzahl verschiedener Objektklassen, zu erhöhen. Hier ist zu untersuchen, wie die Geometrien mit Hilfe der Zuordnungen homogenisiert werden können. Weiterhin sind die verschiedenen Objektstrukturen zu integrieren. Probleme dieser Art sind auch unter dem Begriff "Multiple Representation" bekannt (siehe z.B. [Buttenfield 1989]). In diesem Forschungszweig wird untersucht, wie verschiedene Daten, welche dieselben Objekte beschreiben aber aus unterschiedlicher Herkunft stammen, zu speichern und zu verarbeiten sind.

Die in diesem Abschnitt dargestellten Problemstellungen zeigen auf, daß die automatische Zuordnung als ein Baustein in einer Prozeßkette zu sehen ist, welche sich allgemein mit der Integration von raumbezogenen Daten befaßt. Die erzielten Ergebnisse sind vielversprechend und sollen daher die Forschung in diesem Themengebiet stimulieren. Da die Erfassung und Fortführung raumbezogener Daten sehr kosten- und zeitintensiv ist, sind Techniken, die die Wiederverwendbarkeit der Daten erhöhen, von großem volkswirtschaftlichen Nutzen und werden von Anwendern raumbezogener Daten gefordert.

Literaturverzeichnis

AdV [1988], 'Amtlich Topographisches-Kartographisches Informationssystem (ATKIS)', Arbeitsgemeinschaft der Länder der Vermessungsverwaltungen der Bundesrepublik Deutschland (AdV), Bonn.

ALK [1986], 'ALK-Dokumentation 2.1: Einheitliche Datenbankschnittstelle (EDBS)', AG Hannover, Niedersächsisches Landesverwaltungsamt - Landesvermessung - Hannover.

ALK [1993], 'Dokumentation zum ALK/ATKIS-Datenaustausch', AG Hannover, Niedersächsisches Landesverwaltungsamt - Landesvermessung - Hannover.

Ament, R. [1993], 'ATKIS - Datenbank und Datenaustausch', *Deutscher Verein für Vermessungswesen, Landesverein Baden Württemberg e. V. Mitteilungen.*

Ballard, D. und Brown, C. [1982], *Computer Vision*, Prentice Hall, Englewood cliffs, N.J.

Bartelme, N., Hrsg. [1995], *Grazer Geoinformatiktage '95 - GIS in Transport und Verkehr*, number 80 *in:* 'Mitteilungen der geodätischen Institute der Technischen Universität Graz'.

Barwinski, K. [1988], 'European Road Database - Aktivitäten innerhalb der CERCO', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Nummer 103, Verlag des Instituts für angewandte Geodäsie Frankfurt am Main.*

Barwinski, K. [1994], Begrüßung des AdV Symposiums, *in:* R. Harbeck, Hrsg., 'Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung, Vorträge anläßlich des AdV-Symposiums ATKIS am 15. und 16. Juni 1994', Landesvermessungsamt Nordrhein-Westfalen, 11–12.

Bauer, F. und Goos, G. [1982], *Informatik - Eine einführende Übersicht - Erster Teil/ 3. Auflage*, Heidelberger Taschenbücher Sammlung Informatik, Springer Verlag Berlin, Heidelberg, New York u.a.

Bill, R. und Fritsch, D. [1991], *Grundlagen der Geo-Informationssysteme*, Vol. 1, Wichmann Verlag Karlsruhe.

Bosch [1995], 'MultiMap', Robert Bosch GmbH, Mobile Communication Division, Product Division Digital Road Map, Postfach 77 77 77, D-31132 Hildesheim.

Boyer, K. und Kak, A. [1986], Symbolic Stereo from Structural Descriptions, Technical Report TR-EE 86-12, School of Electrical Engineering, Purdue University, West Lafayette, Indiana.

Boyer, K. und Kak, A. [1988], Structural Stereopsis for 3-D Vision, *in:* 'IEEE Transactions on Pattern Analysis and Machine Intelligence', Vol. 10/2, 144–166.

Brown, J., Rao, A. und Baran, J. [1995], Are you Conflated? Integrating TIGER and other Datasets Trough Automated Network Conflation, *in:* 'GIS-T 95', GIS/Trans, Ltd, Cambridge.

Brüggemann, H. [1990], 'Das ATKIS-Datenmodell', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Nummer 105, Verlag des Instituts für angewandte Geodäsie Frankfurt am Main.*

Brüggemann, H. [1992], Normen und Standards für Geodaten - Stand und Perspektiven der nationalen und internationalen Entwicklung, *in:* O. Günther and K. Schulz and J. Seggelke, Hrsg., 'Umweltanwendungen geographischer Informationssysteme', Wichmann Verlag, Karlsruhe.

Brüggemann, H. [1994], Zukunftsaspekte des Geoinformationssystems ATKIS, *in:* R. Harbeck, Hrsg., 'Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung, Vorträge anläßlich des AdV-Symposiums ATKIS am 15. und 16. Juni 1994', Landesvermessungsamt Nordrhein-Westfalen, 171–188.

Brüggemann, H. [1995], Koordinierung, Standardisierung und Normung auf dem Gebiet Geoinformation in Europa, Deutschland und den Bundesländern, *in* Kophstahl und Sellge [1995], 125 –139.

Buttenfield, B [1989], Multiple Representations: Initiative 3 Specialist Meeting Report, Technical Report Technical Report 89-3, NCGIA, Santa Barbara.

Charniak, E. und McDermott, D. [1985], *Introduction to Artificial Intelligence*, Addison-Wesley Publishing Company, Inc.

Claussen, H. [1989], 'GDF - Ein Austauschformat für geographische Daten', *Nachrichten aus dem Karten und Vermessungswesen* 37–44.

Claussen, H. [1992], GDF 2.0 - Towards a pan European Standard for Geographic Information, *in:* 'International Workshop European Digital Road Map'.

Claussen, H. [1995], Qualitätsanforderungen an die digitale Karte aus Anwendersicht, *in* Bartelme [1995].

Daimler [1994], 'GDF 2.1 Extension for Airports', Daimler Benz AG, Document VR017-D2 (1994).

DDGI [1996], 'Ergebnisprotokoll Aktualisierung von Geodaten am 16. Januar 1996', *DDGI aktuell* in der Zeitschrift Geo-Informationssyteme, Vol. 9 , Nr. 1.

EURONAV [1992], *EURONAV 92 - Digital Mapping and Navigation - The 1992 International Conference of The Royal Institute of Navigation and The German Institute of Navigation (DGON)*, London.

Förstner, W. [1986], A Feature Based Correspondence Algorithm for Image Matching, *in:* 'IAP Rovaniemi', Vol. 26/3, 150–166.

Fritsch, D. und Anders, K. [1996], 'Objektorientierte Konzepte in Geo-Informationssystemen', *Geo-Informations-Systeme (GIS)* **9**(2), 2–14.

Frydrychowicz, S. [1990], Strukturvergleich ebener Kurven mit lokalen Formelementen, *in:* 'Mustererkennung 1990, 12. DAGM-Symposium', Informatik-Fachberichte 254, Springer Verlag, Berlin, Heidelber, New York u.a., 332 – 339.

Gabay, Y. und Doytsher, Y. [1994], Adjustment of Line Maps, *in:* 'GIS/LIS '94, Phoenix, Arizona', 191 – 199.

Gabay, Y. und Doytsher, Y. [1995], Automatic Feature correction in Mergin Line Maps, *in:* '1995 ACSM/ASPRS Annual Convention & Exposition Technical Papers - Charlotte, North Carolina', Vol. 2.

Gillmann, D. [1985], Triangulations for Rubber-Sheeting, *in:* 'Auto-Carto 7 Proceedings', American Society of Photogrammetry, 191 – 199.

Gran, H. [1988], 'Der Katalog der Landschaftsobjekte - ein Modellierungsinstrument des Informationssystem ATKIS', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Nummer 103, Verlag des Instituts für angewandte Geodäsie Frankfurt am Main.*

Grimson, W. E. L. [1990], *Object recognition by computer*, MIT Press.

Grünreich, D. [1990], 'ATKIS - Amtliches Topographisches-Kartographisches Informationssystem der Landesvermessung', *Geo-Informationssysteme, Nummer 4.*

Grünreich, D. [1995], Anforderungen an die GIS-Technologie, *in* Kophstahl und Sellge [1995].

Haala, N. und Vosselman, G. [1992], Recognition of Road and River Patterns by Relational Matching, *in* ISPRS [1992], 969–975.

Hake, G. [1982], *Kartographie I*, Verlag de Gruyter, Berlin/New York.

Harbeck, R. [1994], Überblick über Konzeption, Aufbau und Datenangebot des Geoinformationssystems ATKIS, *in:* R. Harbeck, Hrsg., 'Das Geoinformationssystem ATKIS und seine Nutzung in Wirtscahft und Verwaltung', Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 33–46.

Harbeck, R. [1995], Überblick über Konzeption, Aufbau und Datenangebot des Geoinformationssystem ATKIS, *in* Kophstahl und Sellge [1995].

Helmle, R. [1995], Die Bedeutung von Partnern beim Aufbau und Pflege einer geographischen Datenbank, *in* Bartelme [1995], 47 –54.

Herdeg, E. [1994], Rechtliche Bedingungen und Entgelte bei der Bereitstellung von ATKIS-Daten und anderen geotopographischen Daten, *in:* R. Harbeck, Hrsg., 'Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung, Vorträge anläßlich des AdV-Symposiums ATKIS am 15. und 16. Juni 1994', Landesvermessungsamt Nordrhein-Westfalen, 121–130.

Heres, L., Berthet, P., Claussen, H. und Hiestermann, V. [1991], *GDF-Documentation Vol.1 - Vol. 8*, Task Force EDRM.

Heres, L. und Wood, T. F. [1992], GDF, a Lingua Franca for Geographic Information, *in* EURONAV [1992].

Hiestermann, V. [1992], 'GDF 2.1 Extension for STORM', DRIVE II project V2052.

Holsheimer, M. und Siebes, A. [1994], Data Mining: the search for knowledge in databases, Technical Report CS-R9406, Centrum voor Wiskunde en Informatica, Computer Science Department of Algorithms and Architecture, Amsterdam.

Illert, A. [1995], Aspekte der Zusammenführung digitaler Datensätze unterschiedlicher Quellen, *in:* 'Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Nummer 113', Verlag des Institutes für Angewandte Geodäsie, Frankfurt am Main, 105 – 115.

Ingels, F. [1971], *Information and Coding Theory*, Intext Educational Publishers, Scranton, Pennsylvania.

Ireland, P. [1994*a*], 'Europe at the digital crossroads', *GIS Europe* **3**(2), 42–45.

Ireland, P. [1994*b*], 'Europe at the digital crossroads - Part2', *GIS Europe* **3**(3), 28–31.

ISO [1985], 'International Standard ISO 8211 - Information processing - Specification descriptive file for information interchange'.

ISPRS [1992], *International Archives of Photogrammetry and Remote Sensing (ISPRS) - Comission III*, Vol. 24.

Killick, J. H. [1992], Data Representation using a Standard Data Transfer Specification, *in* EURONAV [1992].

Kophstahl, E. [1988], 'ATKIS - Raumbezogene Basisinformationen der Bundesrepublik Deutschland - Realisierung und Anwendung in Niedersachsen', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Verlag des Instituts für Angewandte Geodäsie Frankfurt am Main.*

Kophstahl, E. [1994], Überblick über Anwendungen des Geoinformationssytem ATKIS - Datenintegrationskonzept, *in:* R. Harbeck, Hrsg., 'Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung, Vorträge anläßlich des AdV-Symposiums ATKIS am 15. und 16. Juni 1994', Landesvermessungsamt Nordrhein-Westfalen, 33–46.

Kophstahl, E. [1995], Überblick über Anwendungen des Geoinformationssystem ATKIS, *in* Kophstahl und Sellge [1995], 39–51.

Kophstahl, E. und Sellge, H., Hrsg. [1995], *Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung - Vorträge anläßlich des 2. ADV-Symposiums ATKIS am 276. und 28. Juni 1995 in Hannover*, Niedersächsisches Landesvermessungsamt - Landesvermessung.

Kraft, W. [1995], Entwurf von Zuordnungs-Algorithmen zur Fortführung und Überprüfung von raumbezogenen Datenbeständen, Diplomarbeit (nicht veröffentlicht), Institut für Photogrammetrie, Universität Stuttgart.

Muratoglu, S. [1996], Entwurf einer Applikation zur semiautomatischen Überprüfung und Erstellung von Zuordnungen von raumbezogenen Daten aus verschiedenen Datenmodellen, Diplomarbeit (nicht veröffentlicht), Institut für Photogrammetrie, Universität Stuttgart.

Nijsen, G. M. und Halpin, T. [1989], *Conceptual Scheme and Relational Database Design - a fact oriented Approach*, Prentice Hall.

Portele, C. [1993], Specification Implementing Records Using ISO 8211, Technical report, EDRM-2 DRIVE projekt V 2052 Standardization Work Package 3100.

Portier, M. [1994], Optimized Network Modelling for Route Planning, *in:* J. Harts, H. Ottens und H. Scholten, Hrsg., 'EGIS/MARI '94', Vol. 2, 1807 – 1816.

Rappe, B. [1995], Erfassung und Integration von Geo-Daten aus unterschiedlichen Quellen, *in:* G. Buziek, Hrsg., 'GIS in Forschung und Praxis', Verlag Konrad Wittwer Stuttgart, 123–140.

Rautenstrauch, C. und Moazzami, M. [1990], *Effiziente Systementwicklung mit ORACLE: ein Handbuch für die Praxis des Anwendungsentwicklers*, Addison Wesley.

Rosen, B. und Saalfeld, A. [1985], Match Criteria for Automatic Alignment, *in:* 'Auto-Carto 7, Washington D.C.'.

Rossol, G. [1988], 'Die Einheitliche Datenbankschnittstelle (EDBS) als Schnittstelle für das Amtliche Topographisch-Kartographische Informationssystem (ATKIS)', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Verlag des Instituts für angewandte Geodäsie Frankfurt am Main.*

Rossum, G. [1994*a*], *Python Library Reference*, Dept. CST, CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. Release 1.1.

Rossum, G. [1994*b*], *Python Reference Manual*, Dept. CST, CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. Release 1.0.3.

Rossum, G. [1994*c*], *Python Tutorial*, Dept. CST, CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands. Release 1.1.

Saalfeld, A. [1988], 'Automated Map Compilation', *International Journal of Geographical Information Systems* **2**(3), 217 – 228.

Salgé, F. und Brüggemann, H. [1992], Conditions for the Development of European-Wide Geographical Reference Information, *in:* 'EURNAV 92 - Digital Mapping and Navigation', London.

Scholl, P. [1995], ATKIS-DLM 25: Erweiterung und Aktualisierung der vorhandenen Datenbestände, *in* Kophstahl und Sellge [1995], 221 – 230.

Sellge, H. [1992], Basisinformationen der Vermessungsverwaltungen (ALK und ATKIS), *in:* G. Buziek, Hrsg., 'Gewinnung von Basisdaten für GIS', Schriftenreihe des DVW im Verlag Konrad Wittwer, Stuttgart, 87–95.

Shannon C. E. and Weaver W. [1949], *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana (Deutsche Ausgabe: Oldenbourg, München 1976.

Shapiro, L. G. und Haralick, R. [1981], Structural Description and Inexact Matching, *in:* 'IEEE Transactions on Pattern Annalysis and Machine Intelligence', Vol. 3, 505–519.

Siemens [1993*a*], 'SICAD-ALK/ATKIS', Siemens Nixdorf Informationssysteme AG, AP Internationales Druckzentrum, München.

Siemens [1993*b*], 'SICAD-GDB-X V1.0', Siemens Nixdorf Informationssysteme AG, AP Internationales Druckzentrum, München.

Siemens [1993*c*], 'SICAD-GR V6.0', Siemens Nixdorf Informationssysteme AG, AP Internationales Druckzentrum, München.

Siemens [1993*d*], 'SICAD-KRT1 V3.0', Siemens Nixdorf Informationssysteme AG, AP Internationales Druckzentrum, München.

van Essen, R. [1994], 'CDPA Data Model & Content Specification', CDPA TELE ATLAS Begium NV.

Vickus, G. [1991], 'Objektgeneralisierung vom DLM zum DKM und ihre Anforderungen an die ATKIS-Modellierung', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe 1, Verlag des Instituts für angewandte Geodäsie Frankfurt am Main.*

Vosselman, G. [1992], *Relational Matching*, Springer-Verlag, Berlin, Heidelberg, New York u.a.

Vosselman, G. und Haala, N. [1992], 'Erkennung topographischer Paßpunkte durch relationale Zuordnung', *Zeitschrift für Photogrammetrie und Fernerkundung (ZPF)* **60**(6), 170–176.

Wagner, G. [1995], Geographische Information im Straßenverkehr, *in* Bartelme [1995], 55–60.

Walter, V. [1995*a*], Abbildung des GDF-Datenmodells auf SICAD-ALK/ATKIS - Interner Bericht -, Institut für Photogrammetrie, Universität Stuttgart.

Walter, V. [1995*b*], Modellierung und Integration von Verkehrsdaten in Geo-Informationssystemen, *in:* '4. Internationales Anwenderforum 1995 - Geoinformationssysteme im Wandel', Siemens Nixdorf Informationssysteme AG, 269–276.

Walter, V. und Fritsch, D. [1994*a*], GIS Data Structures for Vehicle Navigation Systems, *in:* 'Procedings of the 6th Canadian Conference on GIS, Ottawa', Vol. 1, 489 – 501.

Walter, V. und Fritsch, D. [1994*b*], Modelling and Storage of Road Network Data, *in:* 'AGDM '94'.

Walter, V. und Fritsch, D. [1995], Matching Techniques for Road Network Data in different Data Models, *in:* J. Soliman und D. Roller, Hrsg., '28th International Symposium on Automotive Technology and Automation', Automotive Automation Limited, Croydon, England, 633–640.

Walter, V. und Fritsch, D. [1996], 'Integration von Straßenverkehrsdaten aus unterschiedlichen Datenmodellen', *Nachrichten aus dem Karten- und Vermessungswesen, Reihe I* (115), 179–192.

Winston, P. H. [1987], *Künstliche Intelligenz*, Addison-Wesley-Publishing Company.

Zilberstein, O. [1992], Relational matching for Stereopsis, *in* ISPRS [1992], 711–719.

Anhang A

Grundlagen Informationstheorie

Die Informationstheorie befaßt sich mit der Übertragung und Optimierung von Nachrichten über Kanäle. Ein Kommunikationsystem besteht aus einem Sender, einem oder mehreren Übertragungskanälen und einem Empfänger. Der Sender kodiert eine Nachricht und sendet sie über einen Übertragungskanal zum Empfänger. Dort wird die Nachricht wieder dekodiert. Ist der Übertragungskanal optimal, so entspricht die empfangene Nachricht exakt der gesendeten Nachricht. Jedoch kann die Nachricht im Übertragungskanal mit Rauschen überlagert sein. Dadurch entstehen Veränderungen und die empfangene Nachricht entspricht nicht mehr der gesendeten Nachricht.

Die Informationstheorie stellt die Grundlagen zur Verfügung, um optimale Übertragungscodes zu berechnen. Optimal kann sich dabei auf die Effizienz beziehen, und damit den Code bezeichnen, welcher am kürzesten ist und damit besonders schnell übertragen werden kann, oder es kann beispielsweise auch der Code gesucht werden, welcher mit hoher Wahrscheinlichkeit ohne Fehler übertragen wird, aber dadurch Redundanzen aufweist und somit nicht der kürzeste Code ist. Um Fragestellungen dieser Art entscheiden zu können, werden Maße benötigt, die es erlauben, Aussagen über z.B. Informationsinhalt, Länge oder Redundanz eines Codes zu machen.

Der Informationsinhalt einer Nachricht ist abhängig von ihrer Auftretenswahrscheinlichkeit. Nachrichten, die sehr häufig gesendet werden, haben einen geringen Informationsgehalt und Nachrichten, die selten gesendet werden, haben eine hohen Informationsgehalt. Der Informationsgehalt entspricht dem Maß für die Überraschung, die auftritt, wenn eine Nachricht empfangen wird. So hat die Nachricht *es ist 5 Grad unter Null* in Brasilien einen höheren Informationsinhalt als in Alaska. Claude Shannon hat 1949 in der Shannonschen Informationstheorie Maße für die Information mathematisch definiert [Shannon C. E. and Weaver W. 1949].

Ein Kanal wird als diskret bezeichnet, wenn er nur eine endliche Anzahl von Symbolen überträgt. Sei a_i ein Symbol aus einem Alphabet $A = \{a_1, a_2, \ldots, a_n\}$ und $P(a_i)$ die Auftretenswahrscheinlichkeit des Symbols, dann ist der Informationsgehalt $I(a_i)$ dieses Symbols:

$$I(a_i) = \log_b \frac{1}{P(a_i)} = -\log_b P(a_i) \tag{A.1}$$

wobei b die Basis des Logarithmus ist und die Maßeinheit der Information bezeichnet. Im folgenden soll als Basis 2 gewählt werden, was der Maßeinheit bit entspricht. Aus der Definition der Information kann man sehen, daß, falls das Alphabet nur aus einem einzigen Zeichen besteht (also die Auftretenswahrscheinlichkeit dieses Zeichens Eins ist) der Informationsgehalt Null ist. D.h. das Maß der Überraschung, daß ein Zeichen aus einem Alphabet mit genau einem Zeichen gewählt wurde, ist Null.

Summiert man nun den Informationsgehalt $I(a_i)$ aller Zeichen des Alphabets A gewichtet mit ihrer Auftretenswahrscheinlichkeit $P(a_i)$, so erhält man den mittleren Informationsgehalt H(A) pro Zeichen:

$$H(A) = \sum_{i=1}^{n} P(a_i) * I(a_i) \tag{A.2}$$

Dieser mittlere Informationsgehalt $H(A)$ wird als Entropie der Nachrichtenquelle bezeichnet [Bauer und Goos 1982]. Ist nur ein Symbol bei der Übertragung möglich, d.h. $P(a_k) = 1$ und alle anderen Wahrscheinlichkeiten sind Null, dann ist auch die Entropie Null. Das Maxium der Entropie ergibt sich bei einer Gleichverteilung aller Zeichen.

Die Zeichen, die vom Empfänger empfangen werden, sollen mit $\{b_1, b_2, \ldots, b_m\} = B$ bezeichnet werden. Die Wahrscheinlichkeit, daß ein Symbol b_j empfangen wurde, wenn das Symbol a_i gesendet wurde, wird als bedingte Wahrscheinlichkeit $P(a_i|b_j)$ bezeichnet. Falls es sich bei dem Übertragungskanal um einen idealen Kanal handelt, so gibt es eine 1 : 1 Entsprechung zwischen dem Sender-Alphabet A und dem Empfänger-Alphabet B, d.h.

$P(a_i|b_j) = 1$ für genau ein i und j und sonst Null. Falls der Kanal mit Rauschen überlagert ist, können die Wahrscheinlichkeiten alle Werte zwischen Null und Eins annehmen. Die Überraschung, daß ein Symbol b_j empfangen wurde, wenn das Symbol a_i gesendet wurde, wird durch den bedingten Informationsgehalt ausgedrückt:

$$I(a_i|b_j) = -\log P(a_i|b_j) \tag{A.3}$$

Die Überraschung, daß ein Symbol b_j empfangen wurde, wenn das Symbol a_i gesendet wurde ist gleich Null, falls es sich um einen idealen Kanal handelt, da in diesem Fall $P(a_i|b_j) = 1$ ist. Umso niedriger die bedingte Wahrscheinlichkeit ist, desto höher ist die Überraschung.

Die Unsicherheit darüber, welches Symbol gesendet wurde, wenn das Symbol b_j empfangen wird, errechnet sich aus der Aufsummierung über alle Symbole der bedingten Information gewichtet mit der Wahrscheinlichkeit, daß ein bestimmtes Symbol gesendet wurde:

$$H(A|b_j) = \sum_{i=1}^{N} P(a_i) I(a_i|b_j) = \sum_{i=1}^{N} -P(a_i) \log P(a_i|b_j) \tag{A.4}$$

Die bedingte Entropie errechnet sich aus der gewichteten Summe aller bedingten Informationen über alle Kombinationen von Ein- und Ausgabezeichen:

$$\begin{aligned} H(A|B) &= \sum_{i=1}^{N} \sum_{j=1}^{M} P(a_i, b_j) I(a_i|b_j) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} -P(a_i, b_j) \log P(a_i|b_j) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} -P(b_j) P(a_i|b_j) \log P(a_i|b_j) \\ &= \sum_{j=1}^{M} P(b_j) \sum_{i=1}^{N} -P(a_i|b_j) \log P(a_i|b_j) \end{aligned} \tag{A.5}$$

Die bedingte Entropie gibt die durchschnittliche Unsicherheit für ein übertragenes Symbol als Durchschnitt über alle empfangenen Symbole an. Sie ist ein Maß für den durchschnittlichen Verlust eines Übertragungskanal [Vosselman 1992].

In dieser Arbeit wird als Abstandsmaß zwischen zwei relationalen Beschreibungen die gegenseitige Information verwendet. Sie ergibt sich durch die Differenz des Selbstinformationsgehalt und der bedingten Information:

$$I(a_i; b_j) = I(a_i) - I(a_i|b_j) \tag{A.6}$$

Die gegenseitige Information ist ein Maß dafür, wieviel Information ein Symbol über ein anderes aussagt [Vosselman 1992]. Ist die bedingte Information (die Überraschung, daß ein Symbol b_j empfangen wurde, wenn das Symbol a_i gesendet wurde) sehr groß oder ist der Informationsgehalt von a_i sowieso klein, dann ist die gegenseitige Information auch klein. Dementgegen ist die gegenseitige Information dann hoch, wenn der Informationgehalt von a_i hoch ist und die Überraschung, daß das Symbol b_j empfangen wurde, wenn das Symbol a_i gesendet wurde, niedrig ist. Analog dazu kann die gegenseitige Entropie durch die Differenz der Selbst-Entropie und der bedingten Entropie errechnet werden:

$$H(A; B) = H(A) - H(A|B) \tag{A.7}$$

Die gegenseitige Entropie ist ein Maß über den durchschnittlichen Wert der Information, der pro übertragenes Symbol erhalten wird. Sie ist ein wichtiges Maß für das Design und die Performance von Kommunikationsystemen [Vosselman 1992].

Anhang B

Informationsmaße für kontinuierliche Signale

B.1 Definitionen

Die bisher definierten Informationsmaße wurden für ein endliches Alphabet von Eingabezeichen $A = \{a_1, a_2, \ldots a_n\}$ definiert. Um den Informationsgehalt kontinuierlicher Signale messen zu können, müssen die bisher verwendeten Definitionen modifiziert werden.

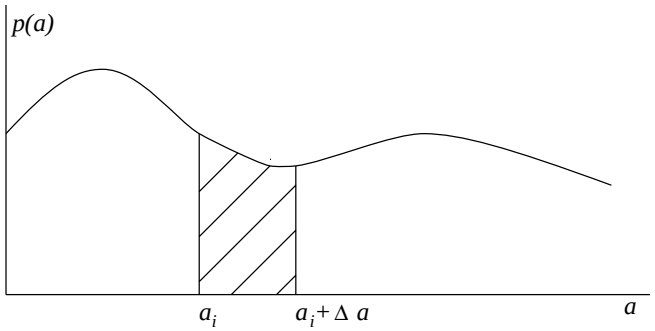


Abbildung B.1: Wahrscheinlichkeitsdichtefunktion p(a)

In Abbildung B.1 ist eine kontinuierliche Wahrscheinlichkeitsdichtefunktion $p(a)$ dargestellt. Die Wahrscheinlichkeit, daß ein Ereignis im Intervall a_i und $a_i + \Delta a$ eintritt, kann mit $p(a_i) * \Delta a$ für hinreichend kleine Δa approximiert werden. Wird jedoch Δa unendlich klein (keine Diskretisierung), so wird die Wahrscheinlichkeit dieses Ereignisses Null und der zugehörige Informationsgehalt unendlich. Um den Informationsgehalt kontinuierlicher Signale vergleichen zu können, wird daher der differentielle Informationsgehalt [Ingels 1971] eingeführt:

$$I_{dif}(a_i) = -\log p(a_i) \tag{B.1}$$

Hierbei gilt zu beachten, daß Informationsmaße von diskreten Signalen nicht mit differentiellen Informationsmaßen verglichen werden können. Die differentielle Information kann auch negative Werte annehmen, da der Wert einer Wahrscheinlichkeitsdichtefunktion größer Eins werden kann. Daher kann die differentielle Information nur zum Sortieren der Informationgehalte von kontinuierlichen Signalen verwendet werden, aber nicht zum Berechnen des Informationsunterschiedes [Vosselman 1992].

Bei der differentiellen Entropie eines kontinuierlichen Signals wird nicht mehr über die Elemente des Eingabealphabets aufsummiert, sondern über den gesamten Wertebereich integriert:

$$\begin{aligned} H_{dif}(A) &= \int_{-\infty}^{\infty} p(a) I_{dif}(a)\ da \\ &= -\int_{-\infty}^{\infty} p(a) \log p(a)\ da \end{aligned} \tag{B.2}$$

Die differentielle bedingte Information definiert sich analog zur bedingten Wahrscheinlichkeitsdichtefunktion zu [Vosselman 1992]:

$$I_{dif}(a_i|b_j) = -\log p(a_i|b_j) \tag{B.3}$$

und die differentielle bedingte Entropie durch:

$$\begin{aligned} H_{dif}(A|B) &= \int_{-\infty}^{\infty} p(b) \int_{-\infty}^{\infty} p(a|b) I_{dif}(a|b) \\ &= -\int_{\infty}^{\infty} -p(b) \int_{-\infty}^{\infty} p(a|b) \log p(a|b)\ da\ db \end{aligned} \tag{B.4}$$

Die gegenseitige Information hat ebenfalls die gleiche Form wie im diskreten Fall:

$$\begin{aligned} I_{dif}(a_i; b_j) &= I_{dif}(a_i) - I_{dif}(a_i|b_j) \\ &= -\log p(a_i) + \log p(a_i|b_j) \end{aligned} \tag{B.5}$$

Die gegenseitige Entropie errechnet sich aus:

$$\begin{aligned} H_{dif}(A; B) &= H_{dif}(A) - H_{dif}(A|B) \\ &= -\int_{-\infty}^{\infty} p(a) \log p(a)\ da + \int_{-\infty}^{\infty} p(b) \int_{-\infty}^{\infty} p(a|b) \log p(a|b)\ da\ db \end{aligned} \tag{B.6}$$

Mit Hilfe der gegenseitigen Entropie (das durchschnittliche Maß an Information, das pro Zeichen übertragen wird) läßt sich zeigen, daß bei Vorliegen einer Rauschquelle N die Information, die über einen Kanal übertragen wird, nur endlich groß sein kann [Vosselman 1992]. Hierzu wird eine Quelle A eingeführt, deren Eingabezeichen für den Wertebereich gleichverteilt sind (z.B. der Winkel von linienförmigen Straßenelementen zur X-Achse):

$$p(a) = \frac{1}{x_2 - x_1} \quad \text{für} \quad [x_1, x_2] \tag{B.7}$$

Die Rauschquelle N wird durch eine Gleichverteilung über ein Intervall der Breite Δn modelliert:

$$p(n) = \begin{cases} \frac{1}{\Delta n} & : \text{ für } [-\frac{1}{2}\Delta n, \frac{1}{2}\Delta n] \\ 0 & : \text{ sonst} \end{cases} \tag{B.8}$$

Die Wahrscheinlichkeitsdichtefunktion des Signals des Empfängers B setzt sich aus dem Eingabesignal und dem Rauschsignal zusammen und ergibt wieder eine Gleichverteilung im Intervall $[x_1, x_2]$, solange man Effekte an den Intervallgrenzen außer acht läßt:

$$p(b) = \frac{1}{x_2 - x_1} \quad \text{für} \quad [x_1, x_2] \tag{B.9}$$

Die bedingte Wahrscheinlichkeitsdichtefunktion ergibt sich dann zu:

$$p(a|b) = \begin{cases} \frac{1}{\Delta n} & : \text{ für } [b - \frac{1}{2}\Delta n \le a \le b + \frac{1}{2}\Delta n] \\ 0 & : \text{ sonst} \end{cases} \tag{B.10}$$

Die durchschnittliche Information, die ein empfangenes Signal besitzt, ergibt sich durch die gegenseitige Entropie:

$$\begin{aligned} H(A; B) &= -\int_{x_1}^{x_2} p(a) \log p(a)\ da + \int_{x_1}^{x_2} p(b) \int_{b-\frac{1}{2}\Delta n}^{b+\frac{1}{2}\Delta n} p(a|b) \log p(a|b)\ da\ db \\ &= -\int_{x_1}^{x_2} \frac{1}{x_2 - x_1} \log\left(\frac{1}{x_2 - x_1}\right) da \\ &\quad + \int_{x_1}^{x_2} \frac{1}{x_2 - x_1} \int_{b-\frac{1}{2}\Delta n}^{b+\frac{1}{2}\Delta n} \frac{1}{\Delta n} \log(\frac{1}{\Delta n})\ da\ db \\ &= -\log\left(\frac{1}{x_2 - x_1}\right) + \int_{x_1}^{x_2} \frac{1}{\Delta n} \log\left(\frac{1}{\Delta n}\right)\ db \\ &= -\log\left(\frac{1}{x_2 - x_1}\right) + \log\left(\frac{1}{\Delta n}\right) \\ &= \log\left(\frac{x_2 - x_1}{\Delta n}\right) \end{aligned} \tag{B.11}$$

Die gegenseitige Entropie ergibt sich also aus dem Logarithmus des Quotienten des Wertebereiches und der Breite des Rauschens. Dies bedeutet jedoch, daß nur dann unendlich viel Information über einen Kanal übertragen werden kann, wenn die Breite des Rauschsignals gleich Null ist (idealer Kanal). Beim Vorliegen einer Rauschquelle hat der Kanal nur eine endliche Übertragungskapazität und der durchschnittliche Informationsgehalt des empfangenen Signals kann nur endlich sein.

B.2 Diskretisierung kontinuierlicher Signale

Es stellt sich nun die Frage, ob der Informationsverlust, der durch eine Diskretisierung der Rauschfunktion entsteht, durch eine geignete Wahl eines hinreichend kleinen Diskretisierungsintervall so minimiert werden kann, daß er gegenüber dem Informationsverlust im Kanal vernachläßigbar ist. C sei das diskretisierte Signal von B im Intervall $[x_1, x_2]$, welches in n Teilintervalle der Breite Δx aufgeteilt ist. Da B gleichverteilt ist, ergibt sich die Wahrscheinlichkeitsdichtefunktion von C zu:

$$P(c_i) = \frac{\Delta x}{x_2 - x_1} \quad \text{für } i = 1, 2, \ldots, n \tag{B.12}$$

Die bedingte Wahrscheinlichkeitsdichtefunktion der diskretisierten Werte ist:

$$p(b|c_i) = \begin{cases} \frac{1}{\Delta x} & : \text{ für } \quad [c_i - \frac{1}{2}\Delta x \leq b \leq c_i + \frac{1}{2}\Delta x] \\ 0 & : \text{ sonst} \end{cases} \tag{B.13}$$

Die bedingte Wahrscheinlichkeitsdichtefunktion von der Quelle A zum beobachteten diskretisierten Signal C ergibt sich dann zu:

$$p(a|c_i) = \int_{x_1}^{x_2} p(a|b)p(b|c_i)\ db \tag{B.14}$$

Mit Hilfe dieser Definitionen läßt sich der durchschnittliche Verlust berechnen, der durch die Diskretisierung auftritt (Beweis siehe [Vosselman 1992]):

$$\begin{aligned} H(A;C) &= \sum_{i=1}^{n} P(c_i) \int_{x_1}^{x_2} p(a|c_i) \log\left(\frac{p(a|c_i)}{p(a)}\right)\ da \\ &= \begin{cases} \log\left(\frac{x_2 - x_1}{\Delta n}\right) - \frac{1}{2}\frac{\Delta x}{\Delta n} & \text{für } \Delta x \leq \Delta n \\ \log\left(\frac{x_2 - x_1}{\Delta n}\right) - \frac{1}{2}\frac{\Delta n}{\Delta x} & \text{für } \Delta n > \Delta x \end{cases} \\ &= \begin{cases} H(A;B) - \frac{1}{2}\frac{\Delta x}{\Delta n} & \text{für } \Delta x \leq \Delta n \\ H(A;B) - \log\frac{\Delta x}{\Delta n} - \frac{1}{2}\frac{\Delta n}{\Delta x} & \text{für } \Delta n > \Delta x \end{cases} \end{aligned} \tag{B.15}$$

Interressant ist hier der Fall $\Delta x \leq \Delta n$, d.h. das Diskretisierungsintervall ist kleiner als die Breite der Rauschfunktion. Von der obigen Gleichung wird deutlich, daß durch die Diskretisierung ein Informationsverlust entsteht. Dieser Verlust wird jedoch vernachlässigbar klein, wenn das Diskretisierungsintervall Δx viel kleiner als die Bandbreite der Rauschquelle gewählt wird.

Anhang C

Testgebiete

Im folgenden werden die Daten der Testgebiete dargestellt. In Tabelle C.1 ist eine Übersicht über den Inhalt der Testgebiete sowie eine Auswertung der manuellen Zuordnungen präsentiert. Anschließend erfolgt eine Darstellung der einzelnen Testgebiete.

Testgebiet	*1*	*2*	*3*	*4*	Gesamt
Anzahl Elemente ATKIS	363	530	530	640	2057
Anzahl Elemente GDF	435	668	853	963	2919
$1 : *$ Zuordnungen	83	52	83	65	283
$* : 1$ Zuordnungen	123	125	261	224	773
$1 : 1$ Zuordnungen	115	211	160	295	781
$1 : n$ Zuordnungen	43	83	96	124	346
$n : 1$ Zuordnungen	27	38	38	41	144
$n : m$ Zuordnungen	26	31	45	32	134

Tabelle C.1: Inhalt der Testgebiete und Auswertung der manuellen Zuordnungen

Abbildung C.1: Testgebiet 1 (2 x 2 km^2)

Abbildung C.2: Testgebiet 2 (2 x 2 km^2)

Abbildung C.3: Testgebiet 3 (2 x 2 km^2)

Abbildung C.4: Testgebiet 4 (2 x 2 km^2)

Anhang D

Übersicht über GDF-Feature-Arten

Die nachfolgende Tabelle gibt eine Übersicht über die bei Bosch/Teleatlas derzeit erfaßten Feature-Arten [van Essen 1994].

Feature Theme	*Feature Class*
Administrative Area	Country Order 1 Area - Order 9 Area Centre of Administrative Area Boundary Junction Boundary Element
Settlements	Built up Area Centre of Settlement
Roads and Ferries	Address Area Boundary Element Adress Area Road Element Junction Enclosed Traffic Area Ferry Element Road Intersection Freeway Exit Centre of Freeway Exit
Railways	Railway Element Railway Element Junction
Waterways	Waterway Junction Water Area Water Line Element
Road Furniture	Signpost
Services	Airport Hotel or Motel Parking Area Petrol Station Railway Station Rest Area Restaurant
Regions	Free Port Park Island
Brunnels	Brunnel

Tabelle D.1: Erfaßte Feature-Arten bei Bosch/Teleatlas

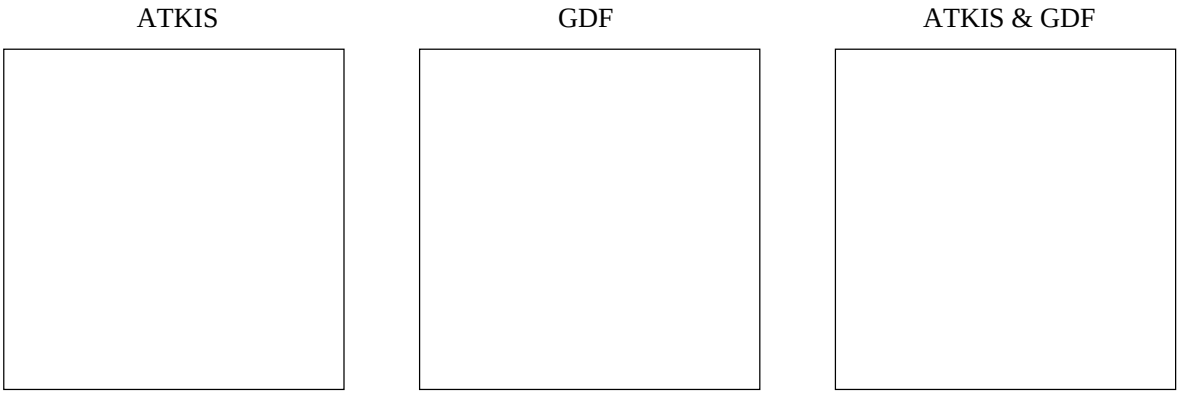


Abbildung E.1: Zuordnungsbeispiel 1

ATKIS-Elemente	$\rightarrow$	GDF-Elemente	ATKIS-Elemente	$\rightarrow$	GDF-Elemente
$a_1 a_5$	$\rightarrow$	$b_1 b_3$	a_9	$\rightarrow$	b_4
a_2	$\rightarrow$	b_2	a_{10}	$\rightarrow$	$b_{11} b_{12}$
a_3	$\rightarrow$	$b_5 b_6 b_7$	a_{13}	$\rightarrow$	$b_{14} b_{16}$
a_4	$\rightarrow$	b_8	a_7	$\rightarrow$	$*$
$a_6 a_{12}$	$\rightarrow$	b_{13}	a_8	$\rightarrow$	$b_9 b_{10}$

Tabelle E.1: Zuordnungsliste für Beispiel 1

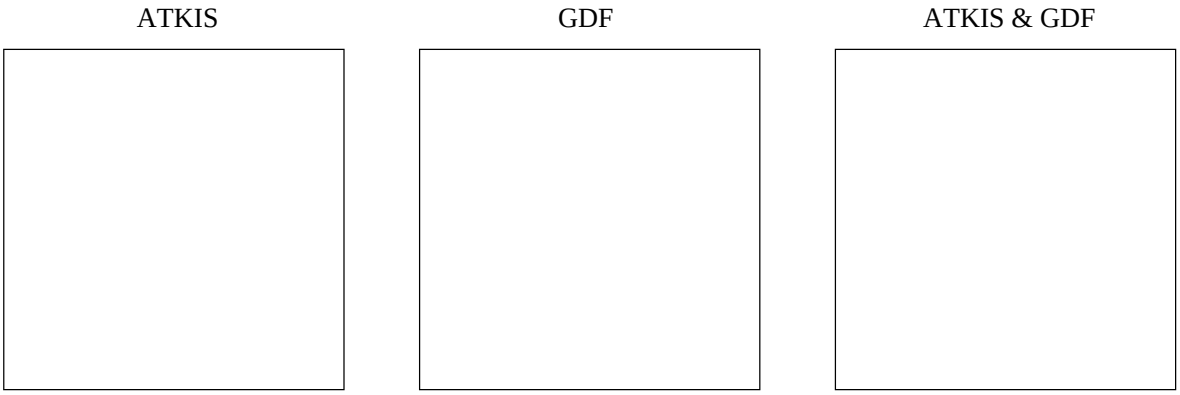


Abbildung E.2: Zuordnungsbeispiel 2

ATKIS-Elemente	$\rightarrow$	GDF-Elemente	ATKIS-Elemente	$\rightarrow$	GDF-Elemente
a_1	$\rightarrow$	$b_1 b_3$	a_4	$\rightarrow$	b_{11}
a_2	$\rightarrow$	$b_2 b_4$	$a_5 a_6$	$\rightarrow$	b_7
a_3	$\rightarrow$	b_6	$a_7 a_8$	$\rightarrow$	b_8
a_3	$\rightarrow$	$b_9 b_{10}$	$a_5 a_6 a_7 a_8$	$\rightarrow$	b_{12}
a_4	$\rightarrow$	b_5			

Tabelle E.2: Zuordnungsliste für Beispiel 2

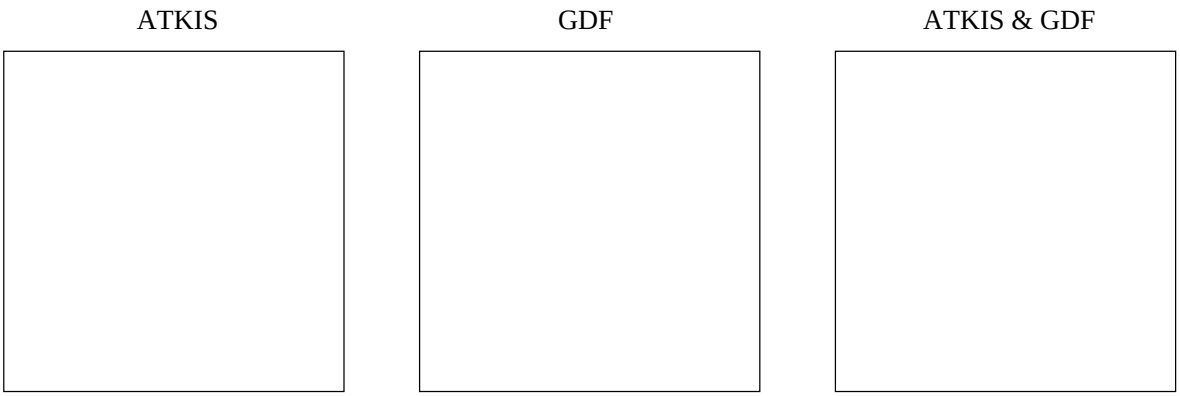


Abbildung E.3: Zuordnungsbeispiel 3

ATKIS-Elemente	→	GDF-Elemente	ATKIS-Elemente	→	GDF-Elemente
a_1	→	b_1	a_{10}	→	$b_9 b_{10}$
a_2	→	b_2	a_4	→	b_{14}
$a_5 a_6$	→	b_0	a_{12}	→	$b_6 b_7$
a_7	→	b_{13}	a_{13}	→	$b_{11} b_{12} b_7$
a_8	→	$b_4 b_5$	a_{11}	→	b_3
a_9	→	$*$	a_{11}	→	b_8

Tabelle E.3: Zuordnungsliste für Beispiel 3

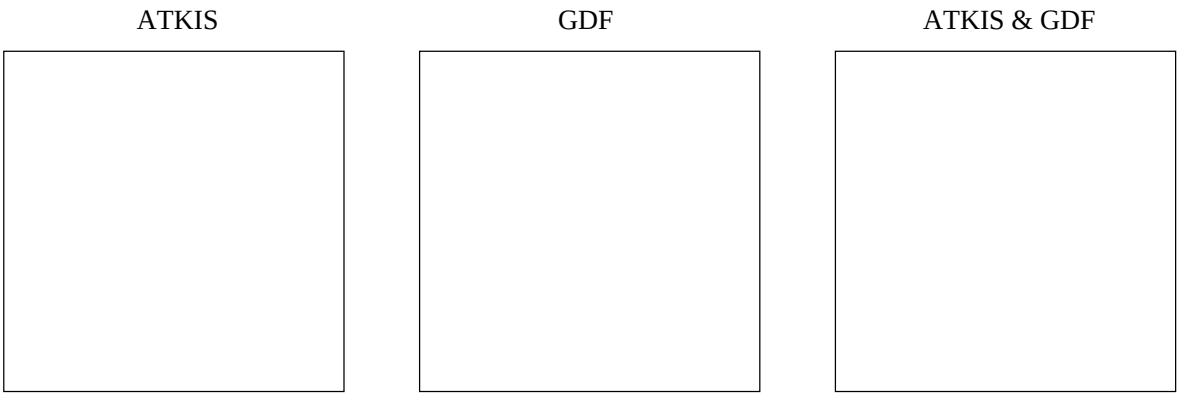


Abbildung E.4: Zuordnungsbeispiel 4

ATKIS-Elemente	→	GDF-Elemente	ATKIS-Elemente	→	GDF-Elemente
a_1	→	b_1	a_7	→	b_{11}
a_2	→	$b_2 b_4$	a_{10}	→	b_{18}
a_3	→	b_7	a_8	→	b_{12}
a_5	→	b_6	a_9	→	b_{13}
a_4	→	$b_8 b_9$	a_{12}	→	$b_{14} b_{15}$
a_{13}	→	b_{10}	a_{11}	→	b_{16}
a_6	→	b_5			

Tabelle E.4: Zuordnungsliste für Beispiel 4

Danksagung

Diese Arbeit entstand im Rahmen eines Promotionsstipendiums der Firma Siemens Nixdorf, München. Für die finanzielle Unterstützung möchte ich mich sehr herzlich bedanken. Ich bedanke mich vor allem bei Herrn Alt, Herrn Dr. Reinhardt und Herrn Singer von Siemens Nixdorf, München sowie Herrn Kutzschmar von Siemens Nixdorf, Stuttgart für die Diskussionsbereitschaft und die tatkräftige Unterstützung bei SICAD-Fragen. Auch allen anderen nicht genannten Mitarbeitern von Siemens Nixdorf, die mir oft weitergeholfen haben, möchte ich meinen Dank aussprechen.

Mein ganz besonderer Dank gilt Herrn Prof. Dieter Fritsch und Herrn Prof. Matthäus Schilcher, die diese Arbeit ins Leben gerufen und betreut haben. Sie gaben mir wertvolle Anregungen und standen mit fachkundigem Rat zur Verfügung.

Ich bedanke mich beim Landesvermessungsamt Stuttgart, beim Niedersächsischen Landesverwaltungsamt und bei der Firma Bosch, Hildesheim für die unbürokratische Bereitstellung von ATKIS- und GDF-Daten.

Bedanken möchte ich mich auch bei meinen Kolleginnen und Kollegen des Institutes für Photogrammetrie der Universität Suttgart, die mich während der Entstehung dieser Arbeit begleitet haben. Durch ihre freundschaftliche Unterstützung und ihre stete Diskussions- und Hilfsbereitschaft in allen Fragen haben sie wesentlich zum Gelingen dieser Arbeit beigetragen. Besonderer Dank gebührt denjenigen, die durch aufmerksame Durchsicht und Korrektur bei der Entstehung dieser Ausarbeitung mitgeholfen haben.

Lebenslauf

von

Volker Walter

25. September 1962	Geboren in Stuttgart
1969 - 1973	Grundschule
1973 - 1979	Realschule
1979 - 1982	Schreinerlehre Abendschule Technisches Gymnasium 11. Klasse
1982 - 1984	Technisches Gymnasium 12. und 13. Klasse, Abschluß Abitur
1984 - 1985	Wehrdienst
1985 - 1992	Studium der Informatik an der Universität Stuttgart
1993 - 1996	Institut für Photogrammetrie, Universität Stuttgart Promotionsstipendium der Firma Siemens Nixdorf Informationssysteme AG, München