Entwicklung von Verfahren zur Beurteilung und Verbesserung der Qualität von Navigationsdaten

Von der Fakultät Luft- und Raumfahrttechnik und Geodäsie
der Universität Stuttgart zur Erlangung der Würde eines Doktors der
Ingenieurwissenschaften (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von
Hainan Chen
aus Fujian, China

Hauptberichter:	Prof. Dr.-Ing. habil. Dieter Fritsch
Mitberichterin:	Frau Prof. Dr.-Ing. Liqiu Meng
Tag der mündlichen Prüfung:	1. April 2011

Institut für Photogrammetrie
der Universität Stuttgart

2011

Inhaltverzeichnis

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Entwicklung von Verfahren zur Überprüfung und Verbesserung der Qualität von raumbezogenen Daten, hier insbesondere Navigationsdaten, indem verschiedene Datenquellen integriert werden.

Zur Vorstellung der Grundlagen werden zunächst der Begriff der digitalen Karte und Arten der Mehrfachrepräsentationen von raumbezogenen Daten vorgestellt. Weiterhin erfolgen eine Gegenüberstellung der verschiedenen Qualitätsmodelle und eine detaillierte Vorstellung der einzelnen Qualitätsmerkmale. Der Begriff der Unsicherheit wird anschließend erklärt. Bestehende Verfahren zur Überprüfung der Qualität von raumbezogenen Daten werden im Rahmen der vorliegenden Arbeit untergliedert. Bei direkt interner Qualitätsprüfung werden Verfahren zur Konsistenzprüfung bzw. unter Nutzung von Integritätsbedingungen vorgestellt. Weiterhin werden vier Arten der Referenzinformation zur direkt externen Qualitätsprüfung präsentiert: Referenzstrecken, Daten aus der Photogrammetrie und Fernerkundung, nutzergenerierte Inhalte und andere Datenbestände. Danach werden indirekte Verfahren mit Metadaten und Techniken zum Qualitätsreport und zur Qualitätsvisualisierung vorgestellt. Schließlich erfolgt eine Ausarbeitung des Standes der Forschung zu Themen der Zuordnung und Verschmelzung von Datensätzen.

Basierend auf den Grundlagen widmet sich die vorliegende Arbeit der Integration sowohl zwischen kommerziellen digitalen Karten von unterschiedlichen Anbietern als auch zwischen kommerziellen und kostenfreien digitalen Karten, die aus nutzergenerierten Inhalten entstehen und eine immer wichtigere Rolle bei der Datenerfassung spielen. So werden eine kostenfreie (OpenStreetMap) und zwei kommerzielle digitale Karten (NavTeq und TeleAtlas) zur Untersuchung eingesetzt. Die Datenmodellierungen in den Datensätzen werden analysiert und verglichen. Um die Komplexität der Datenintegration aufgrund der Heterogenität der Datenmodellierung zu reduzieren, wird ein übergeordnetes Datenmodell entwickelt und die Datensätze werden mit unterschiedlichen Transformationsregeln in das übergeordnete Datenmodell abgebildet. Straßenobjekte werden mit der Relation *1:1*, *1:n* bzw. *n:1* in das übergeordnete Datenmodell abgebildet. Die semantische Heterogenität in den Datensätzen wird durch eine semantische Homogenisierung reduziert.

In der Folge werden die Zuordnung zur Ermittlung von Korrespondenzen in den Datensätzen, die Qualitätsanalyse mittels unterschiedlicher Ähnlichkeitsmaße und die Datenverschmelzung zur Verbesserung der Qualität der Daten erläutert.

Zur Bestimmung von Kantenkorrespondenzen in den Datensätzen wird das Zuordnungsmodell „Buffer Growing“ so erweitert, dass nicht nur Zuordnungen zwischen Kanten, sondern auch Zuordnungen zwischen Kanten und Knoten möglich sind. Der Ansatz wird jeweils mit Testdaten im Stadtgebiet und im ländlichen Raum untersucht. Korrespondenzen von Kanten werden unter Verwendung von einem ArcGIS-Tool manuell identifiziert. Zur Untersuchung der unterschiedlichen geometrischen Modellierungen in den Datensätzen werden acht grundlegende Formklassen definiert. Durch die Erkennung der Formklassen lassen sich die geometrischen Modellierungen in den Datensätzen vergleichen. Darüber hinaus werden Korrespondenzen von Knoten zur Berechnung der komplexen Objekte und Verschmelzung der Datensätze automatisch bestimmt.

Ähnlichkeitsmaße werden aus verschiedenen Aspekten zur Qualitätsprüfung entwickelt. Eine hohe Ähnlichkeit zwischen den Datensätzen bedeutet eine hohe relative Qualität. Auf der Ebene des Datensatzes werden die globale geometrische und topologische Ähnlichkeit sowie die

Vollständigkeit untersucht. Dabei werden vergleichbare Adjazenzmatrizen mit Hilfe von komplexen Objekten berechnet, welche anhand der Zuordnungsergebnisse ermittelt werden. Die globale geometrische Ähnlichkeit erfolgt durch die Ermittlung der durchschnittlichen und maximalen Abweichungen in den Adjazenzmatrizen. Die globale topologische Ähnlichkeit wird durch die Korrelation der Exzentrizitätsvektoren reflektiert. Für jedes Zuordnungspaar werden die Ähnlichkeit der Form als Maß der Komplexität der geometrischen Modellierung, die lokale geometrische Ähnlichkeit anhand der Hausdorff-Distanz und die lokale topologische Ähnlichkeit mittels der Erreichbarkeit untersucht. Darüber hinaus werden Attribute nach ihren Wertbereichen in verschiedene Kategorien aufgeteilt und mit unterschiedlichen Verfahren ausgewertet. Einzelne Attribute lassen sich direkt vergleichen oder über eine Konfusionsmatrix auswerten. Zusammengesetzte Attribute werden allerdings zunächst zerlegt und anschließend ausgewertet. Die Ergebnisse der Qualitätsprüfung werden diskutiert.

Zur Verbesserung der Qualität der Daten wird ein clusterbasierter Ansatz zur Verschmelzung der Datensätze unter Berücksichtigung der unterschiedlichen geometrischen Modellierungen vorgestellt. Anhand der Konnektivität der Kanten und der geometrischen Modellierungen wird eine gleiche Anzahl von Clustern für zugeordnete Kanten und Knoten ermittelt. Die Datenverschmelzung der zugeordneten Kanten und Knoten wird in drei Schritten durchgeführt. Im ersten Schritt werden die Verbindungsknoten ermittelt und die Zuordnungspaare zwischen Knoten und Kanten behandelt. Im weiteren Schritt werden die Mittellinien für die Cluster mit gleicher Form berechnet. Im letzten Schritt werden die Cluster aus den Zuordnungspaaren mit unterschiedlicher Form berechnet und in den Enddatensatz transformiert. Die Konflikte der unterschiedlichen geometrischen Modellierungen werden durch die Parametereinstellung der einfachen bzw. komplexen Form aufgehoben. Cluster für nicht zugeordnete Kanten und Knoten werden ebenfalls anhand der Konnektivität berechnet. Diese Cluster werden mit Hilfe von Verlinkungsknoten in den Enddatensatz transformiert. Konflikte, die nach der Datenverschmelzung entstanden sind, werden anhand von Beispielen diskutiert.

Die erzeilten Ergebnisse bei der Zuordnung, Qualitätsanalyse und Datenverschmelzung werden an den jeweiligen Stellen vorgestellt und diskutiert. Eine Zusammenfassung und ein Ausblick schließen die Arbeit ab.

Abstract

The present work deals with the development of methods for quality inspection and improvement of spatial data, especially navigation data, by integration of different data sources.

In order to impart the basic knowledge, the term of digital map and the types of multiple representations of spatial data are introduced at first. Then, a comparison of different quality models and a detailed introduction of individual quality characteristics take place. The term of uncertainty is explained subsequently. Existing methods for quality inspection of spatial data are subdivided within the work. In direct internal quality inspection, methods for consistency evaluation and subject to spatial constraints are introduced. Furthermore, four types of reference information for direct external quality inspection are presented: reference routes, data from photogrammetry and remote sensing, user generated contents and other data sources. Afterwards indirect methods with metadata and technologies for quality report and visualization are introduced. At last, current research on the topics of matching and conflation of datasets is presented.

The work at hand investigate the integration between commercial digital maps from different suppliers and between commercial and cost-free digital maps which result from user generated contents and play a more and more important role in the data collection. Therefore, one cost-free (OpenStreetMap) and two commercial digital maps (NavTeq and TeleAtlas) are applied for the research. The data modeling of the different datasets is analyzed and compared. In order to reduce the complexity of data integration due to differences of data modeling, a global data model is developed and the datasets are transformed into the global data model with different transformation rules. Street objects are transformed into the global data model with the relation *1:1*, *1:n* or *n:1*. The semantic heterogeneity in the datasets is reduced by a semantic homogenization.

In the following, the matching to identify correspondences in the datasets, the quality inspection with different similarity measures and the data conflation for quality improvement are introduced.

In order to determine correspondences between the datasets, the matching model “Buffer growing” is extended, so that not only matchings between edges, but also matchings between edges and nodes are possible. The approach is examined with test data in urban and rural areas. Correspondences of edges are identified manually with a tool developed under ArcGIS. To investigate the different geometric modeling in the datasets, eight basic form classes are defined. The geometric modeling of the datasets can be compared through recognition of the form classes. Furthermore, correspondences of nodes are computed automatically for calculation of complex objects and for conflation of the datasets.

Similarity measures for quality inspection are developed from different points of view. A high similarity between the datasets is an indicator for a high relative quality. At the level of dataset, the global geometric and topologic similarity as well as the completeness are examined. For this purpose, comparable adjacency matrices are computed. The maximum and average differences in the adjacency matrices are calculated as global geometric similarity. The global topologic similarity is defined as correlation between the eccentricity vectors. For each matching pair, the similarity of form as degree of complexity of geometric modeling, the local geometric similarity with the Hausdorff distance and the local topologic similarity using accessibility are investigated. Attributes are subdivided into different categories according to their range of values and evaluated with different methods. Single attributes can be directly compared or analyzed by using a confusion

matrix. Combined attributes are separated and then evaluated. The results of the quality inspection are discussed.

In order to improve the quality of data, a cluster-based approach for conflation of the datasets considering the different geometric modeling is presented. Based on the connectivity of the edges and the geometric modeling, a same amount of clusters is calculated for the matched edges and nodes. The data conflation for the matched edges and nodes is performed in three steps. First, the connecting nodes are calculated and the matching pairs between nodes and edges are handled. Second, middle lines are built for clusters with the same form. Finally, the clusters of matching pairs with different forms are calculated and transformed into the final dataset. In the same way, clusters for unmatched edges and nodes are calculated according to their connectivity. These clusters are transformed into the final dataset with the aid of linking nodes. The conflicts which arise after the data conflation are discussed with examples.

The results of matching, quality inspection and data conflation are presented and discussed in the respective chapters. A summary and an outlook conclude the work.

1 Einleitung

1.1 Motivation

Die Qualität von raumbezogenen Daten stellt ein nach wie vor spannendes und anspruchsvolles Thema der Geoinformatik dar. Seit Mitte der 1990er Jahre beziehen sich zwei internationale Konferenzen spezifisch auf das Gebiet der Qualität von raumbezogenen Daten [Oort 2005]: International Symposium on Spatial Accuracy Accessement (1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010) und International Symposium on Spatial Data Quality (1999, 2003, 2004, 2005, 2007, 2009). Weiterhin wird die Datenqualität als ein bedeutendes Thema in vielen internationalen Konferenzen behandelt: z.B. Arbeitsgruppe „Quality of Spatio-Temporal Data and Models“ der International Society for Photogrammetry and Remote Sensing (ISPRS) und Arbeitsgruppe „Spatial Data Uncertainty and Map Quality“ der International Cartographic Association (ICA). Heutzutage ist die Datenqualität bereits ein wesentlicher Bestandteil in vielen Büchern der Geoinformatik und eine Reihe von Büchern wurde spezifisch zum Thema der Datenqualität herausgegeben (z.B. [Guptill & Morrison 1995; Devillers & Jeansoulin 2006]).

Unter der Datenqualität wird im weiteren Sinne die Eignung der Daten für spezifische Anwendungen („*fitness for use*“) verstanden [Chrisman 1983]. Mit der Entwicklung des World Wide Web und der zunehmenden Einsatzbereiche von raumbezogenen Daten in verschiedenen Applikationen gewinnt die Datenqualität immer mehr an Bedeutung [Devillers & Jeansoulin 2006]. „*Data quality is a problem we need to address if we in the geospatial industry expect to be a part of the enterprise IT picture. Our most pressing need is a simple, reliable way to answer: "Are these data fit for this purpose?" each time spatial data are merged or shared in an enterprise system*“ [Sonnen 2007].

Um den Bedarf der verschiedenen Einsatzbereiche zu befriedigen, werden raumbezogene Daten von unterschiedlichen staatlichen Institutionen und privaten Unternehmen aus verschiedenen Anwendungsaspekten mit verschiedenartigen Datenmodellen erfasst [Walter 1997]. Beispielsweise liegen digitale Straßenkarten weltweit in unterschiedlichen Arten (z.B. für Google Maps, Navigationssysteme oder Logistikunternehmen) vor. Allein für die Navigationsanwendung wird z.B. das vollständige Straßennetz in Europa von den Firmen NavTeq und TeleAtlas unabhängig voneinander digitalisiert. Zum Zwecke der topographischen Landaufnahme wird das Straßennetz in unterschiedlichen Maßstäben im Amtlichen Topographisch-Kartographischen Informationssystem (ATKIS) abgebildet. Für die Anwendung des Liegenschaftskatasters werden Straßen in der Automatisierten Liegenschaftskarte (ALK) als flächenförmige Objekte digitalisiert. Neben den kommerziellen bzw. amtlichen Datenbeständen stehen mittlerweile noch kostenfreie digitale Straßenkarten (z.B. OpenStreetMap) zur Verfügung, die von Freiwilligen erfasst werden.

Aus den Mehrfachrepräsentationen für dieselben Objekte der realen Welt ergeben sich redundante Informationen, die zur Qualitätsprüfung und -verbesserung genutzt werden können. In diesem Sinne werden heterogene Datenquellen auf verschiedene Arten integriert (siehe [Sester 2008]). Grundsätzlich sind dabei horizontale und vertikale Integration zu unterscheiden. Bei der horizontalen Integration werden die Daten von unterschiedlichen Gebieten zusammengeführt, während sich die vertikale Integration mit der Integration von Datensätzen für dasselbe Gebiet beschäftigt. Daher bietet die vertikale Integration Möglichkeiten zur Qualitätsprüfung von raumbezogenen Daten an. Auf dem Forschungsgebiet der Integration von raumbezogenen Daten existieren bereits viele Arbeiten. Beispielsweise betrachtet Walter [1997] in seiner Dissertation die

Integration von raumbezogenen Daten als ein Zuordnungsproblem, nämlich die Identifikation von korrespondierenden Objekten in verschiedenen Datenquellen. Später widmen sich einige Forscher der gemeinsamen Datenverarbeitung in heterogenen Ausgangsdaten (z.B. [Volz 2006a]). Allerdings fehlt es an Verfahren, um die Qualität von raumbezogenen Daten durch eine Datenintegration zu beurteilen. Aus diesem Grund besteht noch Forschungsbedarf auf dem Gebiet der Qualitätsprüfung durch die Integration von raumbezogenen Daten.

1.2 Zielsetzung der Arbeit

Die Qualität von Navigationsdaten spielt eine entscheidende Rolle bei Navigationsanwendungen. Die zukünftigen Navigationsanwendungen (insbesondere ADAS - Advanced Navigation and Driver Assistance Systems) verlangen mehr Informationen (Attribute) von digitalen Karten und stellen hohe Ansprüche an die geometrische und thematische Genauigkeit [Möhlenbrink et al. 2006]. Die Serienfreigabe neuer Kartenstände für das digitale Straßennetz erfordert in der Industrie umfangreiche und wiederholte Erprobungsfahrten. Trotz des hohen zeitlichen und finanziellen Aufwands ist es schwierig, zuverlässige Qualitätsaussagen zu machen, da es sich dabei um Stichprobenverfahren handelt. Weiterhin werden unterschiedliche Qualitätsregeln und Integritätsbedingungen zur Kontrolle der Datenqualität aufgestellt. Dabei wird häufig lediglich die Konsistenz geprüft (z.B. [Joos 2000]). Des Weiteren verwenden viele Forscher hochaufgelöste Luft- und Satellitenbilder zur Prüfung der Qualität des digitalen Straßennetzes (z.B. [Gerke et al. 2004; Mayer et al. 2008]). In diesem Fall können allerdings nur wenige Attribute überprüft werden. Die Qualitätsprüfung mittels Datenintegration eignet sich für eine flächendeckende Untersuchung der Datenqualität eines großen Gebiets. Dabei können nicht nur die Geometrie, sondern auch die Attribute der Daten überprüft werden.

Das grundlegende Ziel der vorliegenden Arbeit besteht in der Entwicklung von Verfahren zur Überprüfung und Verbesserung der Qualität von Navigationsdaten durch Kombination und Vergleich von unterschiedlichen Datenquellen, welche dieselben Objekte der realen Welt repräsentieren. So werden drei Datensätze im Rahmen der vorliegenden Arbeit eingesetzt: NavTeq, TeleAtlas und OpenStreetMap. Die drei Datensätze repräsentieren das gleiche Straßennetz und sind in sehr großen Gebieten verfügbar. Die Datensätze NavTeq und TeleAtlas basieren auf dem gleichen Datenmodel (GDF- Geographic Data File), während das Datenmodell von OpenStreetMap unterschiedlich ist. Allerdings ist die Modellierung für das Straßennetz in den drei Datensätzen in gewissem Maße ähnlich. Abbildung 1.1 gibt einen ersten Einblick auf die Unterschiede zwischen den Datensätzen NavTeq und TeleAtlas.

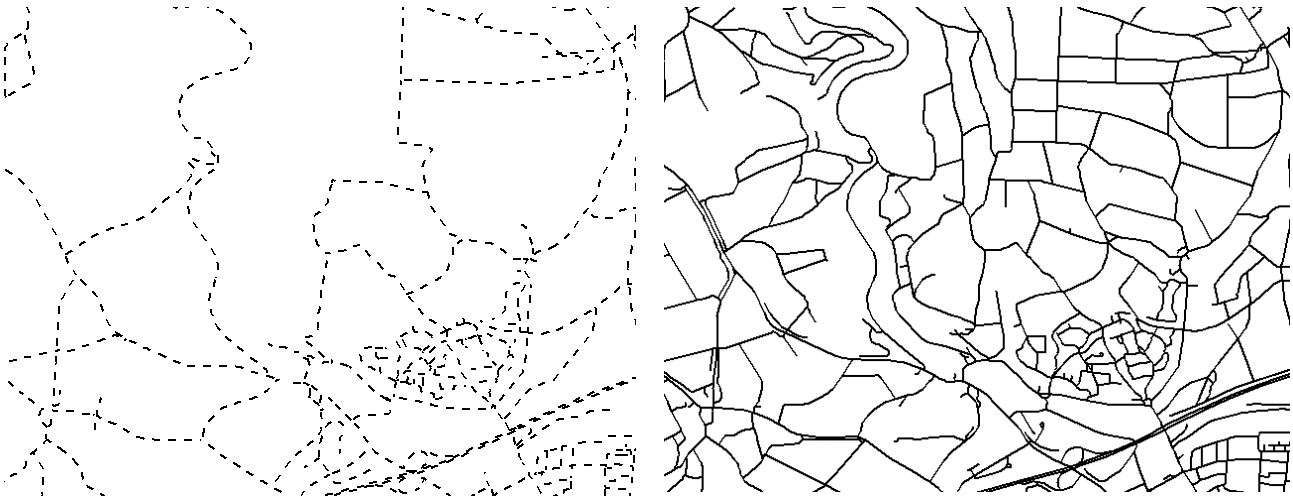


Abbildung 1.1: Unterschiedliche Erfassung zwischen NavTeq (links) und TeleAtlas (rechts)

Um die verschiedenen Datenquellen integrieren zu können, sind im ersten Schritt einander korrespondierende Objekte in den Datenquellen zu finden. Weiterhin ist die Qualität der korrespondierenden Objekte anhand von unterschiedlichen Ähnlichkeiten zu ermitteln. Dabei werden beispielsweise die Ähnlichkeiten der Straßengeometrie, Informationen über die Straßen (z.B. Straßenname) und Verbindungen zu anderen Straßen berücksichtigt. Im letzten Schritt werden die digitalen Karten mit verschiedenen Ansätzen verschmolzen, um eine neue digitale Karte mit einer verbesserten Qualität zu generieren.

1.3 Aufbau der Arbeit

Im weiteren Verlauf der Arbeit wird in Kapitel 2 zunächst auf den Begriff der digitalen Karte eingegangen. Daran schließen sich Erläuterungen zum Thema der Mehrfachrepräsentationen an, welche als ein Element der Interoperabilität betrachtet werden. Weiterhin werden die Definition der Qualität und ihre Merkmale ausführlich beschrieben. Im letzten Abschnitt des Kapitels wird die Unsicherheit der Daten im Zusammenhang mit der Qualität vorgestellt.

Kapitel 3 beschäftigt sich zunächst mit dem Stand der Forschung der Qualitätsprüfung von raumbezogenen Daten. Dabei werden Verfahren zur Qualitätsprüfung nach dem Einsatz von verschiedenartigen internen und externen Informationen untergliedert. Weiterhin werden Verfahren zur Qualitätsverbesserung durch die Integration von heterogenen Datensätzen präsentiert. Die aktuellen Techniken zur Zuordnung und Verschmelzung (Conflation) von Datensätzen werden ausführlich beschrieben.

In Kapitel 4 wird zunächst auf die Datenmodellierung der zu untersuchenden Datensätze eingegangen. Die Konzepte der Modellierung in den verschiedenen Datenmodellen werden miteinander verglichen. Um den Aufwand der Untersuchung zu minimieren, wird ein übergeordnetes Datenmodell entwickelt. Die Datensätze werden dann mit unterschiedlichen Transformationsregeln in das übergeordnete Datenmodell abgebildet.

Kapitel 5 widmet sich der Zuordnung der Datensätze. Zunächst wird das Zuordnungsmodell vorgestellt. Nach der Zuordnung von Kanten erfolgen die Erkennung der Form und die Ermittlung von Knotenkorrespondenzen. Kapitel 6 setzt sich mit der Qualitätsprüfung auseinander. Dabei wird zunächst eine globale Qualitätsauswertung auf der Ebene des Datensatzes vorgestellt. Weiterhin erfolgt eine lokale Qualitätsauswertung im Bezug auf einzelne Zuordnungspaare. Kapitel 7 beschäftigt sich mit der Qualitätsverbesserung durch Verschmelzung von den Datensätzen. Verfahren zur Verschmelzung von zugeordneten Kanten und Knoten sowie von nicht zugeordneten Kanten und Knoten werden detailliert erklärt.

Abschließend werden in Kapitel 8 die Ergebnisse diskutiert und ein Ausblick auf zukünftige Forschungsthemen gegeben.

2 Grundlagen

Das aktuelle Kapitel setzt sich zunächst mit den Grundbegriffen der digitalen Karte auseinander. Weiterhin erfolgt eine Einführung in das Problem der Mehrfachrepräsentationen. Ferner werden Definitionen der Qualität und einzelne Qualitätsmerkmale von ISO/TC 211 beschrieben. Der letzte Abschnitt des Kapitels beschäftigt sich mit der Unsicherheit von raumbezogenen Daten.

2.1 Digitale Karte

Laut International Cartograhic Association (ICA) ist eine Karte eine maßstäblich verkleinerte, generalisierte und erläuterte Grundrissdarstellung von Erscheinungen und Sachverhalten der Erde und ist nach gestellten Anforderungen generalisiert und inhaltlich festgelegt [Wilhelmy et al. 2002]. Karten werden in [Hake et al. 2002] nach verschiedenen Blickpunkten klassifiziert und sind nach unterschiedlichen Kartenarten (Karteninhalt, Anwendung) und Kartentypen (Merkmale der Kartengraphik, Maßstab) zu differenzieren. Wie eine traditionelle analoge Karte stellt eine digitale Karte Teile der Erdoberfläche über einen Signaturenkatalog dar [Czommer 2000]. Dabei handelt es sich eigentlich um eine Kartendatenbank, welche die reale Welt mit einem spezifischen Modell abbildet. Im Folgenden werden digitale Karten für Navigationsanwendungen als Beispiel diskutiert.

Fahrzeugnavigationssysteme gewinnen seit ihrer Markteinführung im Jahr 1994 stetig an Bedeutung [Schlott 1997]. Die Navigation im Fahrzeug erfolgt auf Grundlage einer digitalen Karte. Die Geometrie des Straßennetzes wird in Form von Vektoren (Knoten, Kanten) in einer Datenbank abgebildet. Zu den Vektoren werden verschiedene Informationen (Attribute) wie z.B. Straßennamen, Hausnummern und Abbiegeverbote gespeichert. Darüber hinaus werden POIs (Points of Interest) wie z.B. Sehenswürdigkeiten, Restaurants und Krankenhäuser oder flächenförmige Objekte wie z.B. Parkanlagen erfasst. Abbildung 2.1 stellt einen Ausschnitt einer digitalen Karte für die Navigation dar.

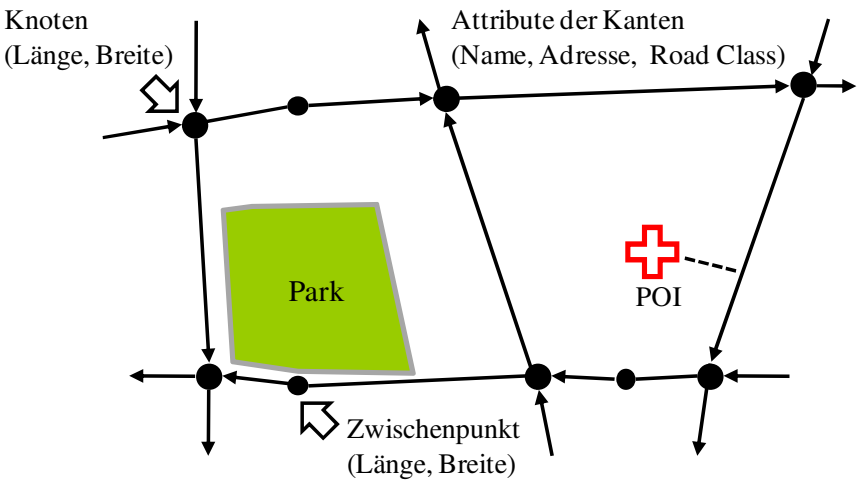


Abbildung 2.1: Ausschnitt einer digitalen Karte

In modernen Navigationssystemen werden immer mehr Datenquellen integriert, um eine bessere bzw. realitätsnahe Kartendarstellung zu erzielen und mehr Informationen akustisch oder visuell zu vermitteln. Dazu zählen z.B. Luftbilder, Satellitenbilder, 3D-Stadtmodelle, 3D-Landmarken,

Geländemodelle oder Bitmaps für Kreuzungen. Die Qualität der digitalen Karte ist ausschlagend für die Qualität der Navigation im Fahrzeug.

2.2 Mehrfachrepräsentationen

Informationen mit Raumbezug werden als *raumbezogene Daten* bzw. *Geodaten* bezeichnet, die zu Bestandteilen eines Geoinformationssystems (GIS) gehören. Der Begriff *Geoinformationssystem* wird in [Möser et al. 2004] wie folgend definiert:

„*Ein Geoinformationssystem (GIS) ist ein System bestehend aus Hardware, Software und Daten zur Erfassung, Verwaltung, Analyse, Präsentation aller Daten, die einen Teil der Erdoberfläche und die darauf befindlichen technischen und administrativen Einrichtungen sowie ökonomische und ökologische Gegebenheiten beschreiben.*“

Geoinformationssysteme (GIS) werden in vielen Anwendungsbereichen wie z.B. Liegenschaftskataster, topographische Landaufnahme und Umweltschutz eingesetzt [Hake et al. 2002]. Welche Objekte der realen Welt in einem Geoinformationssystem erscheinen und welche Informationen (z.B. Attribute) sie enthalten, ist von der Art der Anwendung abhängig [Möser et al. 2004]. Zur Auswahl und Gestaltung der Objekte werden verschiedenartige Modelle bzw. Schemata nach Anwendungen eingeführt, welche Phänomene der realen Welt in gewisser Weise abbilden. Raumbezogene Daten sind infolgedessen eine subjektive Selektion der realen Welt [Stoter & Zlatanova 2004]. Aus diesem Grund können Mehrfachrepräsentationen (MR) für die gleichen Objekte der realen Welt entstehen. Weiterhin entstehen heterogene Repräsentationen aufgrund unterschiedlicher Auflösungen [Balley et al. 2004].

Abbildung 2.2 illustriert ein Beispiel für unterschiedliche Repräsentationen eines Kreisverkehrs. In Repräsentation 1 wird der Kreisverkehr mit flächenförmigen Objekten erfasst, während er in Repräsentation 2 als ein punktförmiges Objekt und in Repräsentation 3 mit linienförmigen Objekten erfasst ist.

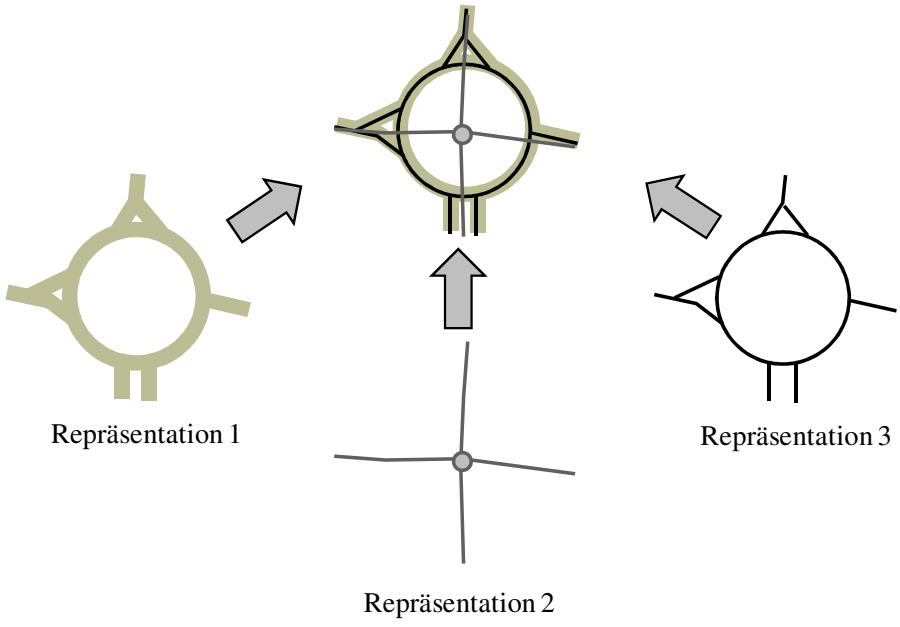


Abbildung 2.2: Unterschiedliche Repräsentationen für denselben Kreisverkehr (nach [Balley et al. 2004])

2.2.1 Arten der Mehrfachrepräsentationen

Mehrfachrepräsentationen werden in [Volz 2006a] nach Ursachen ihrer Entstehung klassifiziert (siehe Abbildung 2.3). Schemabedingte Mehrfachrepräsentationen befassen sich mit den oben beschriebenen multiplen Repräsentationen angesichts unterschiedlicher Anwendungszwecke. Auch bei Verwendung eines identischen Datenmodells können heterogene Repräsentationen aufgrund der unterschiedlichen Operateure oder Zeitpunkte der Datenerfassung entstehen. Des Weiteren werden häufig temporäre Objektkopien während der Bereitstellung von raumbezogenden Daten erzeugt (verwaltungsbedingte Mehrfachrepräsentationen). Die Veränderung und Erweiterung von raumbezogenen Daten führen zu Mehrfachrepräsentationen durch Veredelung. Formatbedingte Mehrfachrepräsentationen treten aufgrund der Konvertierung von raumbezogenen Daten in unterschiedlichen physikalischen Formaten auf.

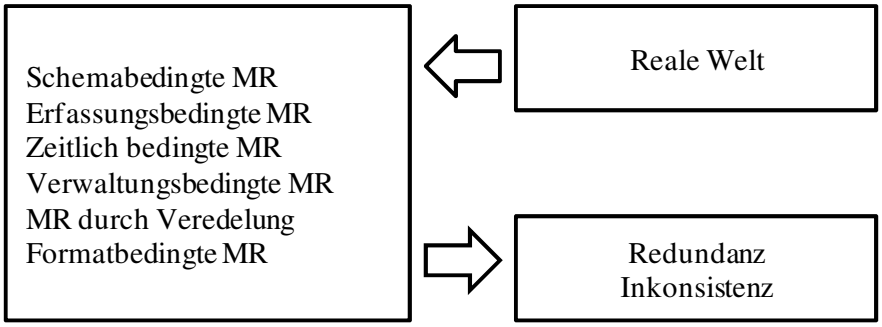


Abbildung 2.3: Klassifizierung der Mehrfachrepräsentationen (MR)

Mehrfachrepräsentationen führen zu Inkonsistenzen und Redundanzen. In [Rodríguez 2005] werden interne Konsistenz, Konsistenz in den unterschiedlichen Detailierungsstufen (LODs) und externe Konsistenz zwischen unterschiedlichen Datensätzen unterschieden (siehe Abbildung 2.4). Um die Konsistenz in unterschiedlichen Detailierungsstufen zu erzielen, sind idealerweise Objekte in einer hochdetailliertesten Stufe zu erfassen. Abstrahierte Stufen sind anschließend über die Generalisierung automatisch abzuleiten. Ein solches Generalisierungsverfahren existiert allerdings noch nicht [Sester 2000; Volz 2006a]. Die relevanten Untersuchungen zur Prüfung der Konsistenz werden in Kapitel 3 ausführlich beschrieben.

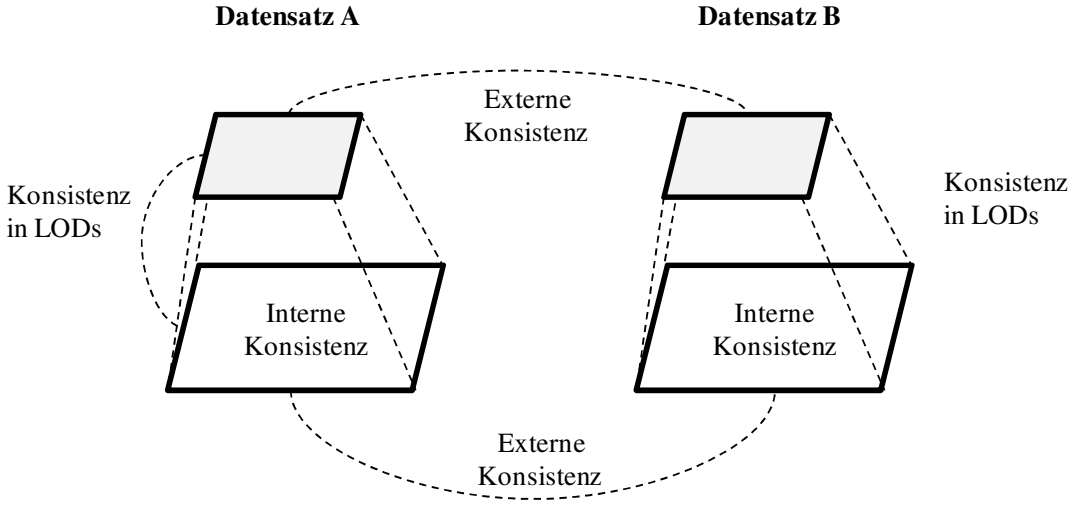


Abbildung 2.4: Konsistenz in Mehrfachrepräsentationen

Intentionale und extentionale Redundanzen sind voneinander zu differenzieren. Die Intension umfasst die Menge der Schemainformationen und deren Semantik (Bedeutung) und die Extension ist die Menge aller zur Intension gehörigen und zugreifbaren Daten. Zusammenfassend bezieht sich die Intension auf die Schemaebene und die Extension auf die Instanzebene. Redundanzen dienen zur Verifikation und zur Komplementierung. Durch eine Integration von Mehrfachrepräsentationen lässt sich die Qualität der Daten (z.B. Vollständigkeit) verbessern [Leser & Naumann 2007].

- *Intentionale Redundanz* ermöglicht extentionale Komplementierung. Das heißt, dass zwei Datensätze mit gleichem Schema zu einem überdeckenderen Datensatz integriert werden können.

- *Extentionale Redundanz* gewährt intentionale Komplementierung. Zwei Datensätze, die dasselbe Objekt repräsentieren, können zu einem dichteren Datensatz integriert werden.

2.2.2 Interoperabilität und Integration

Das Problem der Mehrfachrepräsentationen wird in [Volz 2006a] als ein Bestandteil der Interoperabilität betrachtet. ISO 19118 [2005] definiert die Interoperabilität folgendermaßen:

„*capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units*."

In [Volz 2006a] werden technische und semantische Interoperabilität unterschieden. Die technische Interoperabilität stellt Daten- bzw. Funktionsschnittstellen zur Verfügung und die semantische Interoperabilität befasst sich mit der Übersetzbarkeit des Dateninhalts bzw. der heterogenen semantischen Bedeutungen. In [Parent & Spaccapietra 2000] wird die Interoperabilität nach drei Stufen der Integration unterteilt:

1. *Niedrigste Stufe*: Die heterogenen Datensätze in unterschiedlichen Formaten sind durch eine gemeinsame Schnittstelle (z.B. ODBC - Open DataBase Connectivity) austauschbar. In dieser Stufe findet keine eigentliche Integration statt.

2. *Mittlere Stufe*: Hierzu wird ein benutzergesteuerter Zugriff bzw. Integration von heterogenen Datensätzen ermöglicht. Allerdings sind die Benutzer verantwortlich für die Behandlung der Inkonsistenzen in den Datensätzen. Funktionen für eine Transformation sind zu diesem Zweck zu entwickeln.

3. *Höchste Stufe*: Ein globales System zur Abdeckung der heterogenen Datensätze wird in diesem Fall entwickelt und Inkonsistenzen werden bei der Integration automatisch beseitigt.

Zur Verbesserung der Interoperabilität widmen sich ISO/TC 211 (Technical Committee 211 of the International Standardisation Organisation) und OGC (Open Geospatial Consortium) der Spezifikation von Normen und Standards zur Erleichterung des Zugriffs auf heterogene Datensätze. Eine Übersicht der Normen und Standards von ISO/TC 211 wird in [ISO/TC211 2009] gegeben. Zusammenfassend lassen sich die ISO-Normen in folgende Bereiche untergliedern: Infrastruktur, Datenmodell, GIS-Management, GIS-Service, GIS-Codierung und spezifische Telematikbereiche [ISO/TC211 2009].

Um heterogene Datensätze gemeinsam verarbeiten zu können, ist eine Integration erforderlich [Volz 2006a]. Hierzu sind die Korrespondenzen in den heterogenen Repräsentationen zu finden. Modelle zur Integration der Mehrfachrepräsentationen [Balley et al. 2004] und Multiple Repräsentationsdatenbanken (MRDB) werden aus diesem Grund entwickelt. Durch die Integration von Mehrfachrepräsentationen werden folgende Vorteile erzielt:

- Durch eine Integration werden Inkonsistenzen und Redundanzen beseitigt. Die Integration von Mehrfachrepräsentationen ermöglicht eine Qualitätsprüfung der integrierten Daten und Identifizierung von potentiellen Fehlern [Sester 2008]. Dieser Aspekt ist für die vorliegende Arbeit von besonderer Bedeutung.

- Informationen bzw. Attribute lassen sich durch eine Integration von einem Datensatz zu einem anderen Datensatz übertragen. Darüber hinaus können neue Informationen abgeleitet werden, welche bei der Verwendung von nur einem Datensatz nicht vorhanden sind [Butenuth et al. 2007]. Weiterhin ist die Fortführung nur einmal durchzuführen und kann anschließend auf einen anderen Datensatz übertragen werden, um den zeitlichen und finanziellen Aufwand bei der Aktualisierung zu reduzieren.

- Aus Mehrfachrepräsentationen ergeben sich neue Möglichkeiten der Analyse. Beispielsweise implementiert Volz [2006a] eine Wegesuche in heterogenen Datensätzen.

- Die Integration von Repräsentation in unterschiedlichen Detailierungsstufen ermöglicht eine Darstellung von raumbezogenen Daten in unterschiedlichen Maßstäben [Hampe 2007].

2.3 Qualität

Der Begriff „Qualität“ stammt aus dem Lateinischen „qualitas“. Der erste Beitrag im Bereich der Qualität wurde von Taylor [1911] durch die Definition von Prinzipen zur Unterstützung des Arbeitsmanagements geleistet, um die Qualität eines hergestellten Produkts zu verbessern [Devillers & Jeansoulin 2006]. Eine Vielzahl von Definitionen des Begriffs der Qualität wurde vorgeschlagen. In ISO 9000 [2005] wird der Begriff *Qualität* wie folgend definiert:

"Grad, in dem ein Satz inhärenter Merkmale Anforderungen erfüllt."

2.3.1 Qualitätsmodell

Zur Beschreibung der Qualität von raumbezogenen Daten werden eine Reihe von Qualitätsmodellen mit unterschiedlichen Charakteristiken (Qualitätsmerkmale) vorgestellt. Eine Gegenüberstellung der drei Qualitätsmodelle ICA 1995 [Guptill & Morrison 1995], Joos 2000 [Joos 2000] und ISO/TC211 2002 [ISO19113 2002] ist in Abbildung 2.5 dargestellt.

Abbildung 2.5: Gegenüberstellung der Qualitätsmodelle

Im Namen der International Cartograhic Association wurde ein Buch mit dem Titel „Elements of Spatial Data Quality“ von Guptill und Morrison [1995] veröffentlicht. Gegenüber den anderen zwei Qualitätsmodellen wird die *semantische Genauigkeit* im Qualitätsmodell ICA 1995 als ein zusätzliches Qualitätsmerkmal definiert. Weiterhin wird die *temporäre Genauigkeit* im ICA 1995 durch *temporäre Information* nicht explizit spezifiziert.

In [Joos 2000] werden die *geometrische* und *thematische Genauigkeit* als ein Qualitätsmerkmal *Genauigkeit* zusammengefasst. Die *temporäre Genauigkeit* wird dann nicht explizit im Qualitätsmodell Joos 2000 definiert. Weiterhin werden Qualitätsinformationen (z.B. Herkunft, Aktualität) in Metadaten dokumentiert. Ferner wird die *Richtigkeit* (Korrektheit) als ein Qualitätsmerkmal definiert, um den Grad der Übereinstimmung der Information mit der konzeptionellen Realität anzugeben [Wiltschko 2004].

Die Qualitätsmerkmale von ISO/TC 211 werden in quantitative und nichtquantitative Qualitätsmerkmale eingeteilt [ISO19113 2002]. Zu quantitativen Qualitätsmerkmalen gehören die *Vollständigkeit*, *logische Konsistenz*, *geometrische*, *thematische* sowie *temporäre Genauigkeit*. Eine weitere Klassifikation der Qualitätsmerkmale findet sich in [Devillers & Jeansoulin 2006]: interne und externe Qualitätsmerkmale. Die internen Qualitätsmerkmale beinhalten die gleichen Merkmale wie quantitative Qualitätsmerkmale und bezeichnen den Grad der Ähnlichkeit zwischen den hergestellten und den perfekten Daten, die hergestellt werden sollen. Die externen Qualitätsmerkmale befassen sich mit der Eignung der Nutzung und kennzeichnen den Grad der Übereinstimmung zwischen einem Produkt und Benutzeranforderungen.

2.3.2 Quantitative Qualitätsmerkmale

In der Folge werden die einzelnen quantitativen Qualitätsmerkmale des Qualitätsmodells ISO/TC211 2002 ausführlich beschrieben.

Vollständigkeit: Raumbezogene Daten (Ein Datenbestand) bestehen aus unterschiedlichen Einheiten. Ein Datenbestand ist vollständig, wenn alle Einheiten (z.B. Objekte, Attribute), die in der Spezifikation oder durch Anforderungen festgelegt sind, vollständig erfasst werden [PAS1071 2007]. In [ISO19113 2002] werden zwei Subelemente für die Vollständigkeit, nämlich Datenüberschuss (Übervollständigkeit) und Datenmangel (Untervollständigkeit) definiert. In [Brassel et al. 1995] wird die Vollständigkeit hierarchisch auf unterschiedliche Ebenen unterteilt (siehe Abbildung 2.6).

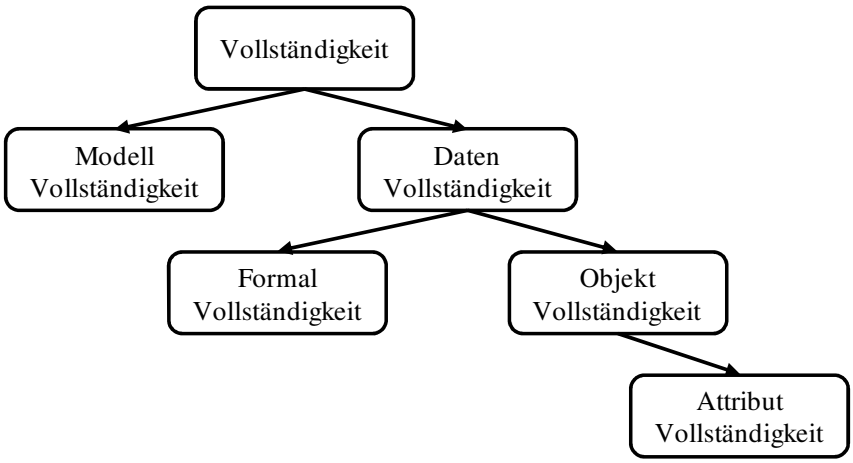


Abbildung 2.6: Hierarchie der Vollständigkeit (aus [Brassel et al. 1995])

Logische Konsistenz repräsentiert das Ausmaß der Einhaltung der logischen Regeln der Datenstruktur, Attribute und Relationen [Devillers & Jeansoulin 2006]. Die logische Konsistenz bezieht sich prinzipiell auf einen Datensatz und wird als interne Konsistenz bezeichnet. Die Datenstruktur kann sich allerdings auf die konzeptionelle, logische und physikalische Ebene beziehen. Die logische Konsistenz besteht aus folgenden Subelementen [Steyer et al. 2004; PAS1071 2007]: konzeptionelle Konsistenz (Schema), Bereichskonsistenz (Weltbereich), Formatkonsistenz, topologische Konsistenz, geometrische Konsistenz und thematische Konsistenz.

Geometrische Genauigkeit wird ebenfalls in ICA 1995 als Positionsgenauigkeit bezeichnet und beschreibt die Genauigkeit der Werte der Koordinaten [Oort 2005]. Relative (innere) und absolute (äußere) Positionsgenauigkeit und Rasterdatengenauigkeit sind in [ISO19113 2002] zu unterscheiden. Für dreidimensionale Daten sind vertikale und horizontale Positionsgenauigkeit in [PAS1071 2007] zu differenzieren.

Absolute Positionsgenauigkeit	kennzeichnet das Maß der Übereinstimmung zwischen den ermittelten Koordinatenwerten und wahren bzw. angenommenen wahren Werten.
Relative Positionsgenauigkeit	kennzeichnet das Maß der Übereinstimmung zwischen der ermittelten relativen Position durch Merkmale im Datensatz und wahren bzw. angenommenen wahren relativen Position.
Rasterdatengenauigkeit	kennzeichnet das Maß der Übereinstimmung zwischen der ermittelten und wahren bzw. angenommenen wahren Rasterposition.
Horizontale Positionsgenauigkeit	entspricht der Lagegenauigkeit.
Vertikale Positionsgenauigkeit	entspricht der Höhengenauigkeit und kann auch als thematische Genauigkeit behandelt werden (Höhe ist oft als Attribut modelliert).

Thematische Genauigkeit gibt, ähnlich wie die geometrische Genauigkeit, den Grad der Übereinstimmung zwischen dem ermittelten und dem wahren bzw. angenommen wahren Wert an [Devillers & Jeansoulin 2006]. Attribute werden nach vier Skalen ermittelt: Nominal, Ordinal, Intervall und Quotient [Goodchild 1995]. Die Nominalskala dient zum Charakterisieren und Differenzieren (z.B. Landnutzung) und kann als Menge von nicht geordneten numerischen Werten

dargestellt werden. Die Ordinalskala wird zur Klassifizierung bzw. Sortierung genutzt und enthält numerische bzw. qualitative Attribute (wie z.B. niedriges, mittleres und hohes Niveau) [Devillers & Jeansoulin 2006]. Intervall- und Verhältnisskala sind streng numerisch dargestellt [Goodchild 1995].

Die thematische Genauigkeit wird durch folgende Subelemente gekennzeichnet: Richtigkeit der Klassifizierung, Genauigkeit der qualitativen Attribute und Richtigkeit der nicht qualitativen Attribute [ISO19113 2002; PAS1071 2007].

Temporäre (Zeitliche) Genauigkeit beschreibt die Genauigkeit der temporären Attribute und temporären Relationen zwischen Objekten [Devillers & Jeansoulin 2006]. Raumbezogene Daten sind nicht nur rauminvariabel, sondern auch zeitbezogen [PAS1071 2007]. Während der Herstellungsperiode von raumbezogenen Daten nimmt die Aktualität der Daten bzw. deren Referenz kontinuierlich ab. In [ISO19113 2002] wird die zeitliche Genauigkeit durch folgende drei Subelemente repräsentiert:

Genauigkeit der Zeitmessung	Präzision der zeitlichen Angabe zu einem Datenbestand
Zeitliche Konsistenz	Korrektheit der Reihenfolge der Ereignisse
Zeitliche Gültigkeit	Gültigkeit der Daten bezüglich der Zeit

2.3.3 Nichtquantitative Qualitätsmerkmale

Abgesehen von den oben beschriebenen quantitativen Qualitätsmerkmalen werden noch drei nichtquantitative Qualitätsmerkmale in [ISO19113 2002] definiert, die in [Steyer et al. 2004] als Datenüberblicksmerkmale bezeichnet werden. Weitere nichtqualitative Qualitätsmerkmale finden sich in [PAS1071 2007].

Zweck dokumentiert die geplante Nutzung bzw. die ursprüngliche Intention für die Erstellung des Datensatzes [Steyer et al. 2004].

Nutzen beschreibt die tatsächliche Nutzung [ISO19113 2002].

Herkunft dokumentiert die Geschichte des Datensatzes und ermöglicht die Ableitung der Abstammungsinformation und die Nachvollziehung der sukzessiven Entwicklung und derzeitiger Form des Datensatzes [Steyer et al. 2004].

2.4 Unsicherheit

Die Abstraktion der realen Welt im Geoinformationssystem führt zu Informationsverlusten und ruft aus diesem Grund die Unsicherheit hervor [Devillers & Jeansoulin 2006]. Unter *Unsicherheit* wird die Abweichung eines digitalisierten Elements von der Wirklichkeit verstanden und raumbezogene Daten sind stets zu einem gewissen Grad mit Unsicherheit behaftet [Glemser 2001].

Abbildung 2.7 stellt den Zusammenhang der relevanten Begriffe bezüglich der Unsicherheit dar [Glemser 2001]. Je nach der Verfügbarkeit von wahren Werten lässt sich die Unsicherheit *absolut* und *relativ* bestimmen. Der Begriff *Unschärfe* repräsentiert Abweichungen eines Modells von der Wirklichkeit. *Fehler* kennzeichnen „*Abweichungen erfasster Elemente vom Modell*“ [Glemser 2001] und unterteilen sich weiterhin in *grobe*, *systematische* und *zufällige* Fehler. *Grobe Fehler* sind falsch erfasste Einheiten und entstammen aufgrund der nicht ausreichenden Kontrolle.

Systematische Fehler geben Abweichungen an, die permanent in gleicher Weise wirken. *Zufällige Fehler* sind in einem Zufallsprozess auftretende statistische Abweichungen. Durch angemessene Kontrolle lassen sich die systematischen und groben Fehler identifizieren und eliminieren. Idealerweise werden zufällige Fehler ebenfalls bestimmt.

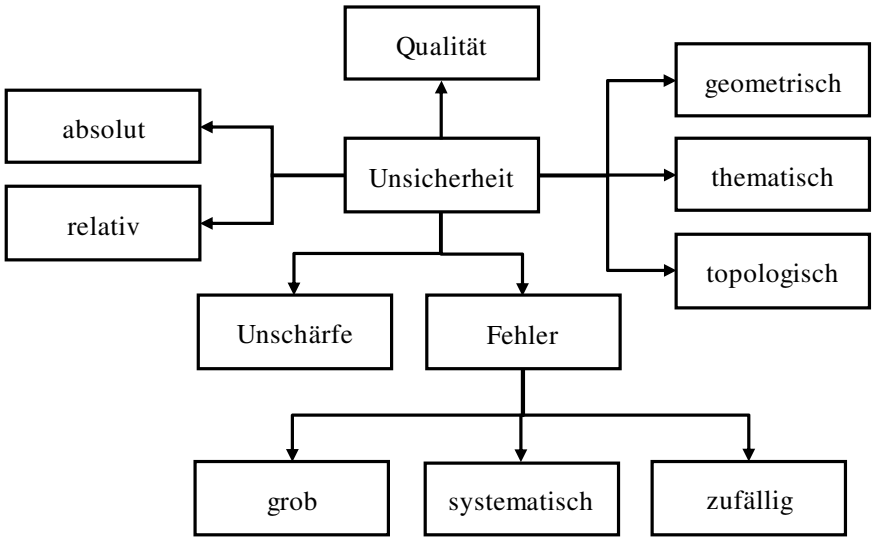


Abbildung 2.7: Zusammenhang der Begriffe im Bezug auf Unsicherheit (nach [Glemser 2001])

Im Grunde sind geometrische, thematische und topologische Unsicherheit voneinander zu differenzieren.

- *Geometrische Unsicherheit*: Die Ursachen zum Hervorrufen einer geometrischen Unsicherheit sind vielfältig [Glemser 2001]. Wie bereits in Abbildung 2.2 dargestellt, ist die geometrische Modellierung für ein Objekt der realen Welt aufgrund verschiedenartiger Modellierungsvorschriften unterschiedlich. Besonders zu nennen ist der Erfassungsvorgang, welcher die reale Welt in einem GIS-Modell abbildet [Fritsch et al. 1998]. Durch Diskretisierung und Messung wird die kontinuierliche Geometrie mit diskreten Koordinaten approximiert. Alle Faktoren führen zur geometrischen Unsicherheit. Glemser [2001] setzt sich in seiner Dissertation vertiefend mit der geometrischen Unsicherheit auseinander und entwickelt drei verschiedene Modelle, nämlich das stochastische Modell, das Minimum-Maximum Modell und das Fuzzy-Modell zur Modellierung der Unsicherheit.

- *Thematische Unsicherheit*: Die geometrische und thematische Unsicherheit sind in manchen Fällen zusammenhängend. Beispielsweise verursacht die geometrische Unsicherheit einer Grenze Unsicherheit der thematischen Klassifizierung (siehe Abbildung 2.8). Die Thematik von Objekten wird durch Attribute repräsentiert, die diskrete oder kontinuierliche Werte besitzen. Die Unsicherheit der Attribute ist vor allem ein Thema der Statistik. Goodchild [1995] stellt unterschiedliche Modelle zur Beschreibung der thematischen Unsicherheit (Normalverteilung, Konfusionsmatrix usw.) vor.

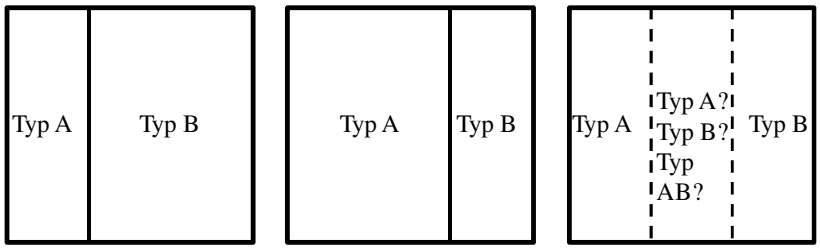


Abbildung 2.8: Unsicherheit der thematischen Klassifizierung aufgrund der geometrischen Unsicherheit (nach [Devillers & Jeansoulin 2006])

- *Topologische Unsicherheit*: Bei der topologischen Unsicherheit handelt es sich um unsichere topologische Beziehungen bzw. Relationen von raumbezogenen Objekten. Grundsätzlich werden acht topologische Relationen unterschieden [Devillers & Jeansoulin 2006]. Zur Formulierung der topologischen Relationen zwischen räumlichen Objekten wurde zunächst ein 4-Intersection Modell von Egenhofer und Franzosa [1991] und anschließend ein 9-Intersection Modell von Egenhofer et al. [1994a] vorgestellt. Die topologischen Relationen sind zwar von der Geometrie unabhängig [Winter 1994], aber sie werden häufig von der geometrischen Repräsentationen abgeleitet. So können geometrische Änderungen zu Änderung der topologischen Relation führen. Unterschiedliche statistische Ansätze zur Auswertung der topologischen Unsicherheit werden vorgestellt (z.B. [Winter 1994; Krauß 1998]).

3 Stand der Forschung

In diesem Kapitel erfolgt zunächst eine Sammlung und Beschreibung bisheriger relevanter Forschungen zum Thema der Qualitätsprüfung von raumbezogenen Daten. Im Anschluss daran wird auf aktuelle Verfahren der Qualitätsverbesserung von raumbezogenen Daten durch Datenintegration eingegangen. Techniken der Zuordnung und Conflation (Verschmelzung), die wesentlicher Bestandteil der Datenintegration sind, werden vorgestellt.

3.1 Qualitätsprüfung

Verfahren zur Ermittlung der Datenqualität lassen sich nach [ISO19114 2003] in *direkte* und *indirekte* Verfahren unterteilen (siehe Abbildung 3.1). Die *direkten* Qualitätsprüfungsverfahren beurteilen die Qualität durch Vergleich von raumbezogenen Daten mit *internen* bzw. *externen* Referenzinformationen. Nach Untersuchung mittels Fragebogen ziehen Kaufmann und Wiltschko [2005] das Fazit, dass die logische Konsistenz mit direkt internen Verfahren prüfbar ist, während die anderen quantitativen Qualitätsmerkmale (z.B. Genauigkeit, Vollständigkeit) mit direkt externen Verfahren zu kontrollieren sind. Im Gegensatz zu direkten Verfahren beurteilen die *indirekten* Verfahren die Datenqualität, indem nicht raumbezogene Daten an sich, sondern Informationen über die Daten (Metadaten) analysiert werden.

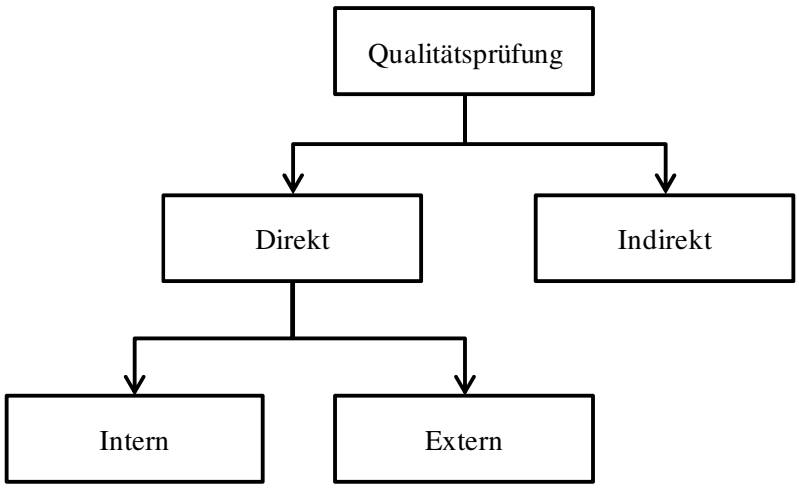


Abbildung 3.1: Verfahren zur Qualitätsprüfung (nach [ISO19114 2003])

Weiterhin lassen sich manuelle und automatische Verfahren zur Qualitätsprüfung unterscheiden. Die Beurteilung der Qualität kann vollständig oder stichprobenweise durchgeführt werden. Eine Stichprobe lässt sich in vier Schritte aufteilen [ISO14825 2004]. Zunächst ist ein Konfidenzniveau zu definieren. Anschließend ist die Anzahl der Stichproben auf Basis des Konfidenzniveaus und der gesamten Anzahl des Items (Feature) im Lot (Datensatz) festzulegen. Danach werden die Charakteristiken in den Items geprüft. Letztendlich ist eine Entscheidung für eine Annahme oder Ablehnung auf Basis der Fehlerquote zu treffen.

3.1.1 Interne Qualitätsprüfung

Die internen Verfahren überprüfen die Qualität der Daten ohne Verwendung von externen Referenzinformationen. In den meisten Fällen wird die Konsistenz der Daten kontrolliert. Nach den Ebenen der Modellierung unterscheiden sich physikalische, logische und konzeptionelle Konsistenz

[Joos 2000]. Zur Überprüfung der Konsistenz sind Regeln zu definieren, die sich auf die Datenmodellierung beziehen. Beispielsweise werden über tausend Regeln von TeleAtlas zur Sicherung der Datenqualität festgelegt [TeleAtlas 2009a]. Im Folgenden wird zunächst die Konsistenzprüfung auf der logischen Ebene diskutiert. Daran schließt sich die Vorstellung der Konsistenzprüfung mittels Integritätsbedingungen an. Wird der zu prüfende Datenbestand in unterschiedlichen Detailierungsstufen (LODs) abgebildet, dann ist die Konsistenz in den verschiedenen LODs zu überprüfen.

Logische Konsistenz

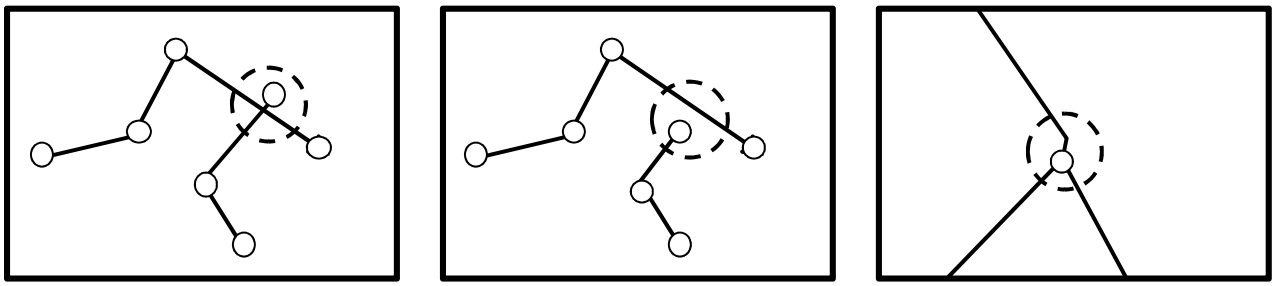
Daten, die in spezifischen Formaten (z.B. GDF - Geographic Data File) gespeichert sind, müssen mit einer Formatprüfung kontrolliert werden [Claussen 1996]. Dies kann automatisch durchgeführt werden. Die Formatprüfung für Navigationsdaten in GDF Format umfasst die Prüfung von Syntax, Wertbereichen, Datenbankintegrität, Topologie und Wertintegrität [ISO14825 2004]:

Syntaxfehler:	Nutzung von nicht spezifizierten Zeichen Fehlerhafte Feldlänge
Wertfehler:	Fehlerhafte Feature-Schlüssel Fehlerhafte Attribut-Schlüssel Fehlerhafte Relationen-Schlüssel
Datenbankintegritätsfehler:	Fehlerhafte Zeiger Fehlerhafte Feldzählerwerte
Topologische Fehler:	Fehlerhafte Umrissdefinition Unterbrochene Linienobjekte
Wertintegritätsfehler:	Fehlerhafte Zuordnung zwischen Feature und Attribut Fehlerhafte Zuordnung zwischen Feature und Relation Fehlerhafte Zuordnung zwischen Attribut und Relation Koordinaten außerhalb der Bereichsgrenze

In [Joos 2000] werden Verfahren zur Prüfung der topologischen Konsistenz vorgestellt. Planare Graphen lassen sich als Regeln formulieren und sind aus diesem Grund für eine logische Prüfung geeignet. Für einen planaren zusammenhängenden Graphen *G* gilt

$$n(\mathrm{G}) + f(\mathrm{G}) - m(\mathrm{G}) = 2,$$

wobei $n(\mathrm{G})$ die Anzahl der Knoten, $f(\mathrm{G})$ die Anzahl der Maschen und $m(\mathrm{G})$ die Anzahl der Kanten des Graphen kennzeichnen. Darüber hinaus bietet es sich an, typische geometrische bzw. topologische Stellen mit hoher Wahrscheinlichkeit eines Fehlers bei der Datenerfassung zu untersuchen (siehe Abbildung 3.2).



a) Überschneidung b) Unterschneidung c) Versatzstelle

Abbildung 3.2: Typische Stellen mit hoher Wahrscheinlichkeit von typischen geometrischen bzw. topologischen Inkonsistenzen (a und b aus [Joos 2000])

Integritätsbedingungen

Integritätsbedingungen spielen eine bedeutende Rolle bei der automatischen Qualitätsprüfung von raumbezogenen Daten [Mäs 2008]. Inkonsistenzen entstehen, wenn Integritätsbedingungen verletzt sind. In [Cockcroft 1997] werden räumliche und nicht räumliche bzw. traditionelle Integritätsbedingungen unterschieden. Die räumlichen Integritätsbedingungen werden in [Cockcroft 1997; Rodríguez 2005] wie folgt klassifiziert:

- *Topologische Integritätsbedingungen*: Diese Integritätsbedingungen betrachten geometrische Eigenschaften und topologische Relationen der räumlichen Objekte. Beispiele sind, dass sich Kanten an Knoten schließen müssen oder die Länge einer Kante größer als zehn Meter sein muss.

- *Semantische Integritätsbedingungen*: Hierbei wird die Bedeutung der Objekte betrachtet. Beispielsweise müssen sich die Mittellinien von Straßen an Kreuzungen schließen. Die Semantik „Straße“ wird in diesem Fall berücksichtigt.

- *Benutzerdefinierte Integritätsbedingungen*: Diese Arten von Integritätsbedingungen sind äquivalent wie Geschäftsregeln in einem nicht räumlichen Datenbankmanagementsystem (DBMS) (siehe [Cockcroft 1997]). Zum Beispiel muss die Lage eines Atomkraftwerkes außerhalb einer vorgegebenen Entfernung zu Siedlungsgebieten sein.

Integritätsbedingungen lassen sich unterschiedlich formulieren und verwalten. Hadzilacos und Tryfona [1992] präsentieren ein logisches Modell zur Formulierung von topologischen Integritätsbedingungen. Servigne et al. [2000] entwickeln einen Ansatz zur Überprüfung und Verbesserung der topologisch-semantischen Konsistenz von raumbezogenen Daten. Die topologisch-semantischen Regeln sind interaktiv einzugeben. Zunächst sind zwei Objektklassen auszuwählen und anschließend sind die topologischen Relationen sowie die Beziehung zwischen diesen zu definieren. Integritätsbedingungen lassen sich auf unterschiedliche Art und Weise darstellen. Cockcroft [2001] betrachtet die Integritätsbedingungen auf der Metadatenebene. Mostafavi et al. [2004] benutzen die Sprache *Prolog* (Programming in Logic) zur Detektion von Inkonsistenzen mit vorgegebenen Regeln. Ein ontologiebasierter Ansatz wird dafür entwickelt. Die Ontologie der raumbezogenen Datenbank wird in eine Kenntnisdatenbank in *Prolog* konvertiert. Regeln zur Detektion von Inkonsistenzen werden anschließend definiert. Der Ansatz wurde mit der NTDB (National Topographic Data Base) von Kanada getestet. Mäs [2008] definiert semantische Integritätsbedingungen von Objektklassen in einem sogenannten Integritätsnetzwerk.

Weitere Regeln zur Beurteilung der Datenqualität ohne Nutzung von externen Referenzinformationen, die nicht zu topologischen und semantischen Integritätsbedingungen gehören, werden in der vorliegenden Arbeit als benutzerdefinierte Integritätsbedingungen bezeichnet. Fang [2008] untersucht die automatische Identifikation von Schwachstellen der Verteilung von POIs (Point of Interest) in einer Karte. Die Abhängigkeiten der POI-Verteilung von POI-Kategorien werden untersucht und durch ein lineares Modell mit der Methode der kleinsten Quadrate geschätzt. Mit einem Schwellwert werden die Schwachstellen aufgedeckt.

Die aus einer Konsistenzprüfung mittels Integritätsbedingungen bestimmten Inkonsistenzen können für eine automatische Fehlerdetektion verwendet werden [Gong & Mu 2000]. Die Inkonsistenzen werden in räumliche Inkonsistenzen, temporäre Inkonsistenzen, Inkonsistenzen von Attributen sowie Inkonsistenzen zwischen beliebigen Kombinationen von Raum, Zeit und Attributen unterteilt. Die Inkonsistenzen sind manuell bzw. interaktiv zu verifizieren. Eine detaillierte Untersuchung zum Thema der Qualitätsverbesserung von raumbezogenen Daten unter Nutzung von räumlichen Integritätsbedingungen findet sich in [Devillers & Jeansoulin 2006].

Konsistenz in multiskaligen Datenbanken

Für Geoinformationssysteme mit Repräsentationen in verschiedenen Detailierungsstufen (LODs) ist es erforderlich, die Konsistenz in den unterschiedlichen Repräsentationsstufen zu kontrollieren. Nach [Egenhofer 1994b] sind topologische Relationen zwischen räumlichen Objekten die herausragenden Informationen für die Auswertung der Konsistenz. Egenhofer [1994b] beschreibt ein Konzept zur Ermittlung der topologischen Inkonsistenzen in heterogenen Repräsentationen. Die Topologie eines Objekts und die topologische Relation zwischen Objekten müssen in fortlaufenden Detailierungsstufen identisch sein oder kontinuierlich nach der Komplexität und Detaillierung verringert werden.

In heutigen Navigationssystemen werden digitale Karten in verschiedenen Detailierungsstufen (Maßstäben) dargestellt. Die Darstellung einer Karte in verschiedenen Maßstäben ist i.d.R. unterschiedlich [Spaccapietra et al. 2000]. Die Umstellung des Kartenmaßstabs führt möglicherweise zur Änderung einer topologischen Relation [Ulugtekin et al. 2004]. Abbildung 3.3 visualisiert ein Beispiel für topologische Inkonsistenz in zwei fortlaufenden Maßstäben. Das eingegebene Ziel („G“) befindet sich auf dem Land im Maßstab *1km* (Abbildung 3.3 links), während es im Wasserbereich im Maßstab *2km* (Abbildung 3.3 rechts) liegt. Ein möglicher Grund für die Änderung der topologischen Relation ist die geometrische Änderung des Landpolygons in den unterschiedlichen Maßstäben.

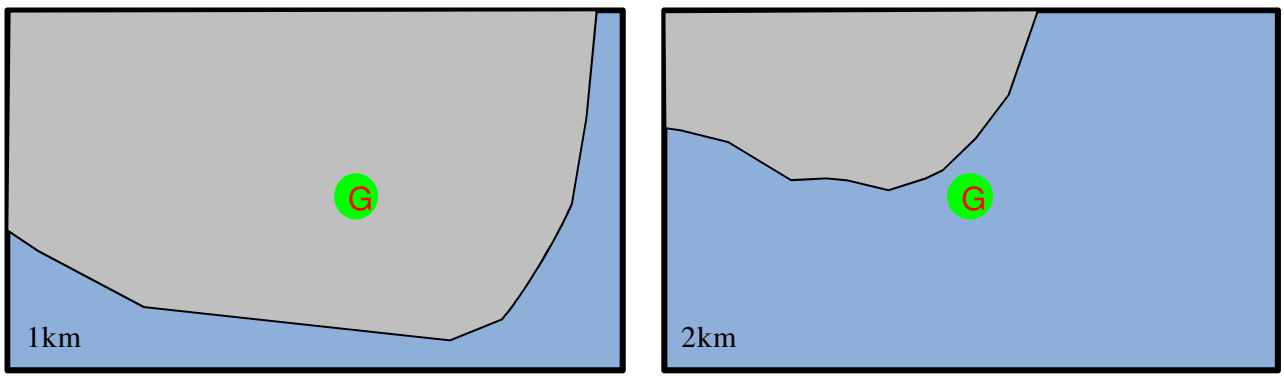


Abbildung 3.3: Topologische Inkonsistenz in verschiedenen Maßstäben

Paiva [1998] entwickelt in seiner Dissertation ein Modell zur Analyse der topologischen Äquivalenz und Ähnlichkeit zwischen räumlichen Objekten in multiskaligen Datenbanken. Das Model fokussiert auf räumliche Relationen zwischen Objekten und abstrahiert die geometrischen Repräsentationen der Objekte. Bedingungen der topologischen Konsistenz müssen in den verschiedenen Repräsentationen eingehalten werden. Eine Vektor-Repräsentation wird in eine symbolische Repräsentation konvertiert. Eine räumliche Szene wird dann durch einen Graph dargestellt, dessen Knoten die Objektrepräsentationen und dessen Kanten die topologischen Relationen zwischen den Objektrepräsentationen darstellen. Auf Basis dieses Modelles werden Verfahren zur Analyse der topologischen Äquivalenz und Ähnlichkeit zwischen räumlichen Szenen entwickelt.

Der Prozess zur Ableitung einer Karte mit kleinerem Maßstab aus einer anderen Karte mit größerem Maßstab wird als Generalisierung bezeichnet [Li 2008]. Die Generalisierung spielt eine übergeordnete Rolle bei der Kartenherstellung [Dorgu & Ulugtekin 2006]. Aus diesem Grund widmen sich viele Forscher der Bewertung von Ergebnissen der Generalisierung, um die Qualität der raumbezogenen Daten zu bewerten. Breunig et al. [2007] stellen ein Modell für eine Analyse der Topologie in den Detailierungsstufen auf Basis der orientierten hierarchischen d-Generalisierten Karten dar. Haunert und Sester [2008] definieren ein semantisches Distanzmaß, um die Ergebnisse der Generalisierung auszuwerten und die logische Konsistenz sowie die semantische Genauigkeit der Generalisierung zu gewährleisten.

3.1.2 Externe Qualitätsprüfung

Die externen Verfahren ermitteln die Datenqualität mit Hilfe von externen Referenzinformationen. Im Prinzip können alle Arten von Informationen, welche die gleichen Objekte wie der zu prüfende Datenbestand beinhalten und redundante Informationen in verschiedenartigen Formen bilden, als Referenz zur Qualitätsprüfung genutzt werden. In Abbildung 3.4 wird eine Übersicht über Datenquellen gegeben, die häufig bei der externen Qualitätsbewertung eingesetzt werden. Abhängig von den Arten der Referenzdaten werden unterschiedliche Qualitätsmerkmale berechnet.

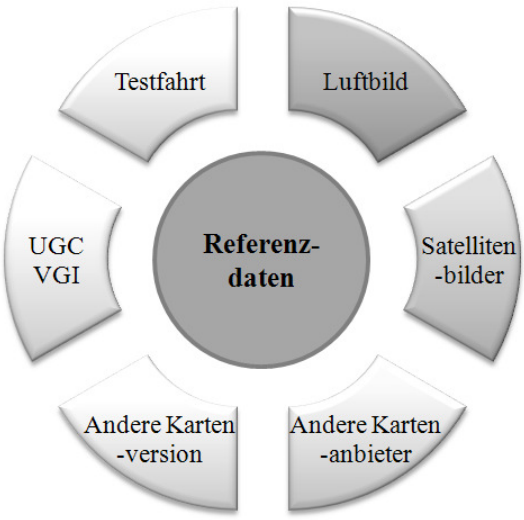


Abbildung 3.4: Mögliche Referenzdaten bei der externen Qualitätsbewertung

Referenzstrecken

Referenzstrecken (Testfahrten) können von Kartenlieferanten, Systemanbietern und Automobilherstellern selbständig oder gemeinsam erfasst werden. Dabei handelt es sich um einen Vergleich zwischen der realen Welt und der zu prüfenden digitalen Karte. Aus diesem Grund können i.d.R. alle Qualitätsmerkmale untersucht werden. Allerdings sind Testfahrten sehr kosten- und zeitaufwendig und können aus diesem Grund nur stichprobenweise durchgeführt werden. Trotz des hohen Aufwandes spielen Referenzstrecken bei der Freigabe einer digitalen Karte eine übergeordnete Rolle.

Basierend auf die in der Testfahrt entdeckten systematischen Fehler sind Qualitätsregeln abzuleiten und Werkzeuge zur automatischen Prüfung anschließend zu entwickeln. Die während der Testfahrten aufgezeichneten Sensordaten lassen sich im Labor einspielen, um den Aufwand der Qualitätsprüfung zu reduzieren. So lassen sich automatisierbare und systematische Prüfumfänge von der Straße ins Labor verlegen [Chen 2006]. In der Laborumgebung werden am Prüfling, d.h. dem Navigationssystem mit einer neuen Softwareversion und/oder einer neuen Kartenversion, rechnergesteuert reale Fahrten nachgefahren. Die dabei aufgezeichneten Navigations- und CAN-Daten (Informationen, die zwischen Navigation und Kombiinstrument fließen) werden anschließend im Auswerteschritt mit den entsprechenden Daten des Referenzsystems verglichen. Darüber hinaus kann die Verknüpfung zwischen Referenzstrecken und einer digitalen Karte mit Wegesuche berechnet und anschließend verglichen werden.

Photogrammmetrie und Fernerkundung

Unter Fernerkundung versteht man im weiteren Sinne die Gesamtheit der Verfahren zur Gewinnung von Information über die Objekte der realen Welt ohne körperliche bzw. physikalische Kontakte [Lo & Yeung 2002]. Photogrammmetrie und Satellitengeodäsie sind der Fernerkundung zugeordnet. Bereits seit Jahrzehnten werden Fernerkundungsdaten zur Verifikation und Fortführung von bestehenden Vektordaten (z.B. GDF oder ATKIS) eingesetzt. Nach [Heipke et al. 2008] sind Luft- und Satellitenbilder die primäre Informationsquelle für die Verifikation. In erster Linie sind Objekte aus den verschiedenartigen Fernerkundungsdaten zu extrahieren. Für diesen Zweck werden unterschiedliche automatische bzw. halbautomatische Ansätze entwickelt: z.B. Extraktion von Linien- und Flächenobjekten aus Luftbildern [Hinz 2003], aus Satellitenbildern [Klang 1998] und aus SAR (Synthetic Aperture Radar) Daten [Dell'Acqua et al. 2002]. Eine Übersicht über automatische Objektextraktionsverfahren aus unterschiedlichen Arten von Fernerkundungsdaten ist in [Mayer et al. 2008] zu ermitteln.

In [Walter 1999; Walter 2004] wird ein Ansatz zur Objektextraktion aus multispektralen Bildern mittels objektbasierter Klassifikation entwickelt, um Änderungen in ATKIS Daten herauszufinden. Aus bereits existierenden ATKIS Vektordaten werden Trainingsgebiete automatisch abgeleitet, um die zeitaufwendige manuelle Erfassung der Trainingsgebiete zu vermeiden. Mit einer überwachten Klassifikation werden Linien- und Flächenobjekte aus Satellitenbildern extrahiert. Weiterhin werden die klassifizierten Objekte mit den Kriterien Prozentzahl, Homogenität und Form der Pixel beschrieben, um Änderungen zu detektieren. Eine Voraussetzung für den Ansatz ist, dass keine massiven Änderungen in der realen Wert im Vergleich zum bestehenden Datensatz stattfinden.

In Rahmen des Projekts WIPKA-QS (Wissensbasierter Photogrammetrisch-Kartographischer Arbeitsplatz zur QualitätsSicherung) wird ein flexibles und einfach konfigurierbares System entwickelt, um ATKIS Daten mit Luft- und Satellitenbildern zu vergleichen. Ein graphbasierter Ansatz zur Unterstützung der Operatoren bei der Verifikation wird in [Gerke et al. 2004]

vorgestellt. Die Verifikation basiert auf einer zweistufigen Extraktion von Straßenobjekten innerhalb einer Region, die durch zu verifizierende ATKIS-Objekte festgelegt wird. Kriterien zur Berechnung der Vollständigkeit, geometrischen Genauigkeit, Genauigkeit der Attribute sowie temporäre Genauigkeit werden definiert. Eine Erweiterung zur Verifikation der Flächenobjekte wird in [Busch et al. 2006] vorgestellt. Die Ergebnisse der Verifikation werden mit einem Ampelsystem in unterschiedlichen Farben visualisiert. Becker et al. [2008] beschreiben einen Ansatz zur multihierarchischen Qualitätsanalyse auf Basis des Bildinterpretationssystems GEOAIDA (*Geo Automatic Image Data Analyzer*). Die Strategie der Bildanalyse und die Datenmodelle werden in GEOAIDA als baumähnliche semantische Netzwerke dargestellt.

Verfahren zur Fortführung und Erweiterung der Fahrzeugnavigationsdaten (GDF) werden in [May 2002] präsentiert und verglichen. Durch Einsatz von hochaufgelösten Luftbildern als Referenzdaten, die eine Genauigkeit im Dezimeter-Bereich liefern, lassen sich sowohl die geometrische Genauigkeit als auch die Vollständigkeit von Straßenobjekten überprüfen und verbessern. Darüber hinaus können zusätzliche Attribute von Straßen (z.B. Anzahl der Fahrspuren) überprüft werden.

User Generated Content (UGC)

Der Begriff *User Generated Content* (nutzergenerierter Inhalt) ist im Zusammenhang mit dem Begriff Web 2.0 entstanden. Das Web 2.0 motiviert Web-Benutzer für einen Beitrag im Internet. In den vergangenen Jahren hat UGC eine rasante Verbreitung in unterschiedlichen Webseiten erfahren. Ein führendes Beispiel dafür ist die Webseite der *Wikimedia* mit der freien Enzyklopädie *Wikipedia* [Goodchild 2008]. Digitale Globen (z.B. *Google Earth* und *Virtual Earth*) und preiswerte Positionierungstechnologien (GPS) haben das Interesse von Benutzern an einem Beitrag für die Erfassung von raumbezogenen Daten geweckt [McDougall 2009]. Nutzergenerierte raumbezogene Inhalte werden in [Goodchild 2007] als *Volunteered Geographic Information (VGI)* bezeichnet. Eine Charakterisierung von unterschiedlichen positiven und negativen Motivationen von Benutzern wird in [Coleman et al. 2009] präsentiert.

"This network of human sensors has over 6 billion components, each an intelligent synthesizer and interpreter of local information" [Goodchild 2007].

Die nutzergenerierten raumbezogenen Inhalte werden für verschiedene Communities mit vielfältigen Plattformen bzw. Werkzeugen bereitgestellt. Dazu zählen z.B. *OpenStreetMap* und *WikiMedia* [Longueville et al. 2009]. Im Folgenden werden unterschiedliche Untersuchungen über die Nutzung von nutzergenerierten Inhalten zur Qualitätsprüfung und -verbesserung von digitalen Karten vorgestellt. Dabei sind manuelle und automatische Verfahren zu unterscheiden. Die Nutzung einer Karte zur Qualitätsprüfung, die auf nutzergenerierten Inhalten basiert, wird in der vorliegenden Arbeit als Kartenvergleich klassifiziert und im nächsten Abschnitt diskutiert.

Um die manuelle Erfassung von nutzergenerierten Inhalten zu ermöglichen, wurde eine Reihe von Plattformen bzw. Werkzeugen von unterschiedlichen Unternehmen entwickelt. Beispielsweise erlaubt *Google Maps* den Benutzern, die Position von Landmarken zu editieren. Die Firma *NavTeq* stellt das Tool *NAVTEQ Map Reporter* zur Verfügung, um die Meldung von Kartendatenfehlern zu ermöglichen. Zum gleichen Zweck bietet *TeleAtlas* ein Werkzeug mit der Bezeichnung *TeleAtlas Map Insight* an [TeleAtlas 2009b]. Die nutzergenerierten Inhalte, die aus der rasch wachsenden Community resultieren, gewinnen immer mehr an Bedeutung bei der Qualitätsprüfung und Fortführung von digitalen Karten [TeleAtlas 2009c]. Zum Beispiel wurden bis Dezember 2008 mehr als fünf Millionen Einträge von den TeleAtlas-Benutzern erzeugt, während insgesamt nur

1,500 Einträge im Juli 2007 generiert wurden. Abbildung 3.5 links visualisiert ein Beispiel für die Bestimmung einer geometrischen Änderung einer Straße in Polen. Die gelben (dicken) Linien repräsentieren die Autobahnen in der alten Kartenversion und die schwarzen Punkte sind GPS-Positionen von den Navigationsgeräten der Benutzer. Die roten (dünnen) Linien stellen die neue Geometrie der Autobahnen dar. In Abbildung 3.5 rechts wird eine neue Straße durch nutzergenerierte Inhalte identifiziert. Die roten Punkte und die blauen Linien repräsentieren die Änderung in der realen Welt.

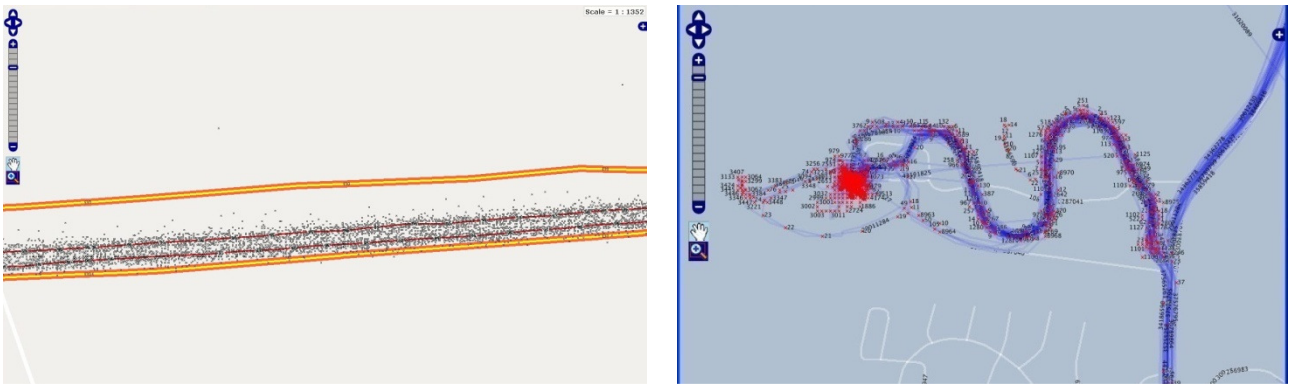


Abbildung 3.5: Nutzung der UGCs bei TeleAtlas (aus [TeleAtlas 2009b])

Das Projekt *MapGenerator* widmet sich der Erstellung einer Straßenkarte unter Nutzung von weltweit aufgezeichneten GPS-Positionen von Benutzern [Brüntrup et al. 2005]. Die GPS-Positionen werden gefiltert und in einer Zentralbank gespeichert. Durch Kombination von GPS-Positionen ergibt sich eine freie verfügbare Straßenkarte, die insbesondere für unbekannte Gebiete interessant ist und zur Qualitätsprüfung von bestehenden digitalen Karten eingesetzt werden kann. Ein ähnliches Konzept zur Prüfung der Plausibilität von digitalen Karten durch Nutzung von aufgezeichneten GPS-Positionen wird in [Franz 2008] vorgestellt. Aus den aufgezeichneten Positionen wird ein Streckenmodell abgeleitet, welches zur Überprüfung der Datenqualität genutzt wird.

Im Rahmen des *FeedMap/ActMap* Projekts wird eine sogenannte *FeedMap*-Schleife von Automobilherstellern, Automobillieferanten, Kartenlieferanten, Lieferanten für Location-Based-Content (LBC) und öffentlichen Behörden zur Verstärkung der Zusammenarbeit zwischen Kartenlieferanten und Service-Benutzern entwickelt [Visintainer et al. 2008]. Das Hauptziel besteht darin, nachhaltige Quellen für die Kartenaktualisierung bereitzustellen und den Prozess der Kartenaktualisierung zu beschleunigen (siehe Abbildung 3.6). Durch Vergleich des Fahrverhaltens von Benutzern (aufgezeichnete Sensordaten im Fahrzeug) mit der Kartendatenbank im Fahrzeug können Abweichungen detektiert werden, die als *Map Devivation Report (MDR)* bezeichnet werden und an das *FeedMap Service Centre (FMSC)* gesendet werden. Die MDRs werden vom *FeedMap Service Centre (FMSC)* zusammengefasst und statistisch ausgewertet. Die daraus resultierenden *Map Deviation Alerts (MDA)* werden den Kartenlieferanten zur Verifizierung zur Verfügung gestellt. Die qualifizierten Fortführungen werden durch das *ActMap Service Centre (AMSC)* an die Service-Benutzer übertragen. Das Konzept wird von unterschiedlichen Firmen in fünf Testgebieten ausprobiert und die Ergebnisse werden in [Landwehr et al. 2008] zusammengefasst. Nicht nur geometrische Abweichungen, sondern auch Abweichungen der Attribute können entdeckt werden. Zu den detektierbaren Attributen zählen z.B. Einbahnstraße, Geschwindigkeitseinschränkung und Abbiegerestriktion, die sich auf das Fahrverhalten beziehen.

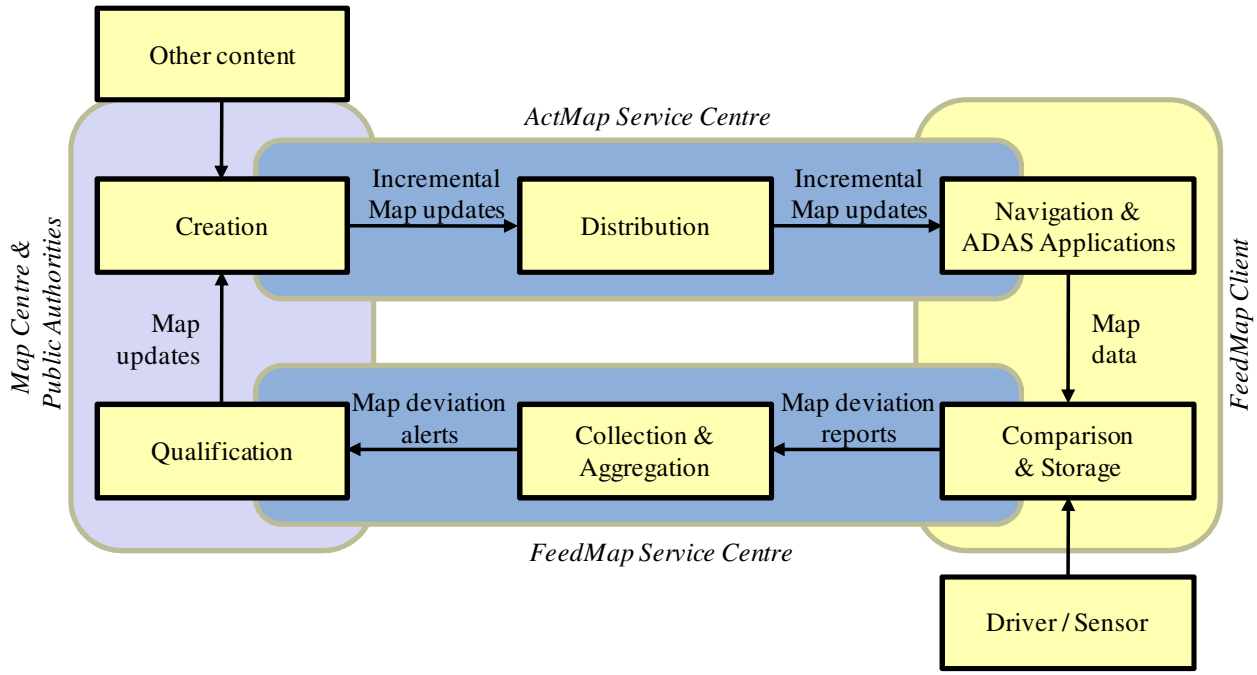


Abbildung 3.6: FeedMap-Schleife (aus [Visintainer et al. 2008])

Das Projekt *ActMap* beschäftigt sich mit der Spezifikation des inkrementellen Updates von digitalen Karten im Fahrzeug, da die Fortführungen aufgrund der Zeit- und Übertragungseinschränkung nicht durch ein vollständiges Kartenupdate übertragen werden können [Otto et al. 2004]. Für die Aktualisierungen werden unterschiedliche Aktualisierungsmethoden vorgestellt: inkrementelles Update und partielles Update. Das inkrementelle Update besteht aus drei grundlegenden Operationen: *Insert, Modify* und *Delete* [Anders et al. 2007] und wird in vielen Fällen im Zusammenhang mit der Versionskontrolle durchgeführt [Spéry 1998; Peerbocus et al. 2002; Ying et al. 2007]. Da oft nur ein bestimmtes Gebiet für einen Benutzer von Interesse ist, lässt sich die gesamte Karte in Kacheln aufteilen. Durch ein partielles Update kann eine unabhängige Kartenaktualisierung für einzelne Kacheln durchgeführt werden. Um Inkonsistenzen auf der Grenze der Kacheln nach dem partiellen Update zu vermeiden, werden Aktualisierungen auch von benachbarten Kacheln übertragen, falls eine Fortführung über mehrere Kacheln verteilt ist.

Die Nutzung von nutzergenerierten Inhalten bietet nachhaltige Quellen zur Qualitätsprüfung und -verbesserung von raumbezogenen Daten an. Die Prüfmöglichkeit der Qualitätsmerkmale hängt von Art und Umfang der Benutzerbeiträge ab. Aus diesem Grund ist eine systematische Qualitätskontrolle schwierig. Darüber hinaus können Attribute durch Nutzung von dynamischen Einträgen oder aufgezeichneten GPS-Positionen nur teilweise überprüft werden.

Kartenvergleich

Unter dem Begriff *Kartenvergleich* wird in der vorliegenden Arbeit die Qualitätsprüfung einer digitalen Karte (Datensatz) durch Vergleich mit einer anderen digitalen Karte verstanden. Der Kartenvergleich ist ein wichtiger Ansatz zur Integration von heterogenen Datenquellen. Dabei lassen sich Verfahren unterscheiden, welche einen Vergleich ohne Zuordnung und mit Zuordnung durchführen.

Der räumliche Operator „Puffer“ ist beim Kartenvergleich für verschiedene Zwecke nützlich und kann z.B. zur Identifikation von Abweichungen in heterogenen Datensätzen eingesetzt werden. Um

zwei Kartenversionen ohne eine Zuordnung zu vergleichen und Abweichungen (z.B. neu digitalisierte Objekte) herauszufinden, ist ein Puffer (z.B. ein Meter breit) um die ältere Kartenversion zu legen. Die außerhalb des Puffers liegenden Objekte in der neueren Kartenversion sind die neu erfassten Objekte. Darüber hinaus wird ein Ansatz mit iterativem Puffer von Goodchild und Hunter [1997] vorgestellt, um die geometrische Genauigkeit von linienförmigen Objekten zu berechnen. Der Prozentwert der Länge eines linienförmigen Objekts in einem Datensatz, der innerhalb eines Puffers um das gleiche Objekt von einem anderen Datensatz liegt, variiert mit der Änderung der Pufferbreite. Die geometrische Genauigkeit wird durch die Pufferbreite reflektiert, wobei der vorgegebene Prozentwert erreicht wird. Der Ansatz eignet sich sowohl für Auswertung einzelner linienförmigen Objekte als auch für Auswertung einer gesamten Karte.

Ein Vergleich ohne Zuordnung zwischen einer VGI-basierten Karte (*OpenStreetMap*) und *Ordnance Survey* (OS) Vektordaten wird von Haklay [2008] durchgeführt. In beiden Datensätzen steht das Straßennetz in Großbritannien als Vektorformat zur Verfügung. Die geometrische Genauigkeit wird am Beispiel von Autobahnen mit dem oben beschriebenen Verfahren von Goodchild und Hunter [1997] berechnet. Mit einer Pufferbreite von sechs Metern wird ein Prozentwert von 90% erzielt. Zur Ermittlung der Vollständigkeit wird ein gitterbasierter Ansatz entwickelt. Zunächst wird der gesamte Kartenbereich in kleine Gitterzellen (z.B. 1×1 km) aufgeteilt. Im nächsten Schritt wird die Gesamtlänge in den Gitterzellen für beide Datensätze berechnet. Die Vollständigkeit wird auf Basis der Längendifferenz in den einzelnen Gitterzellen ausgewertet.

Ein Ansatz zur Auswertung der geometrischen Genauigkeit von linienförmigen Objekten im Rasterformat wird in [Seo & O'hara 2009] präsentiert. Die linienförmigen Objekte werden vom Vektorformat zum Rasterformat konvertiert und ein punkt- sowie ein linienbasierter Ansatz zur Zuordnung und Auswertung entwickelt. Beim punktbasierten Ansatz werden die Pixel mit einem Puffer zugeordnet. Sind mehrere Zuordnungskandidaten vorhanden, dann wird der Zuordnungskandidat mit minimalem Abstand bestimmt. Die Ergebnisse der Zuordnung werden interaktiv verbessert. Aus dem Abstand der zugeordneten Pixel werden Verschiebungsvektoren berechnet. Beim linienbasierten Ansatz werden aus den Vektordaten vier Layer (*Klasse*, *Länge*, *Distanz* und *Orientierung*) erzeugt. Der Layer *Klasse* bestimmt, ob ein Pixel ein Liniensegment enthält, während der Layer *Länge* die Länge des Liniensegments innerhalb eines Pixels repräsentiert. Die Layer *Distanz* und *Orientierung* werden mit Hilfe von Puffern berechnet. Pixel aus dem Layer *Klasse* werden ausgewählt und mit Kriterien der *Länge*, *Distanz* und *Orientierung* zugeordnet. Als Ergebnis werden ebenfalls Verschiebungsvektoren aus den Differenzen der *Länge*, *Distanz* und *Orientierung* ermittelt. Ein Vorteil des Verfahrens ist, dass die topologischen Relationen zwischen den Objekten bei der Auswertung nicht zu berücksichtigen sind.

Beim Kartenvergleich mit einer Zuordnung können weitere Qualitätsmerkmale (z.B. thematische Genauigkeit) untersucht werden. In [Nitz 2004] wird die Qualität der punktförmigen Objekte (POIs) von unterschiedlichen Kartenanbietern ausgewertet. POIs werden einerseits zu POIs einer Referenzkarte mit geometrischen (z.B. Puffer) und thematischen Kriterien (z.B. POI Name, Straßenname, Hausnummer und POI Typ) zugeordnet und verglichen. Anderseits werden die POIs zu Kanten einer Referenzkarte zugeordnet, um die Attribute der POIs (z.B. Straßenname und Postleitzahl) zu überprüfen. Zum Vergleich von flächenförmigen Objekten werden Ansätze zur Auswertung von Polygonkanten in [Gombosi et al. 2003] oder unter Nutzung von Fuzzylogik in [Fritz & See 2005] vorgestellt.

Zuordnungsverfahren für linienförmige Objekte werden in Kapital 3.2.1 ausführlich beschrieben. Zur Bestimmung der korrespondierenden Objekte werden unterschiedliche Ähnlichkeitsmaße bezüglich der Geometrie, Topologie und Thematik eingesetzt, die gleichzeitig die Qualität der

Daten angeben. Eine Zusammenfassung von Ähnlichkeitsmaßen findet sich in [Samal et al. 2004]. Dabei werden kontextabhängige Ähnlichkeiten bezüglich der Zeichenkette, Skalar, Position und Form sowie kontextunabhängige Ähnlichkeiten auf Basis der Nachbarschaftgraphen unterschieden. Zusammenfassend lassen sich die Ähnlichkeitsmaße wie folgt aufteilen:

- *Geometrische Ähnlichkeit* lässt sich durch Auswertung der geometrischen Eigenschaften von linienförmigen Objekten bestimmen: z.B. Anfangs- und Endposition, Winkelunterschied und Längenunterschied. Darüber hinaus stehen verschiedene Distanzfunktionen für diesen Zweck zur Verfügung: z.B. Hausdorff-Distanz [Hangouet 1995; Deng et al. 2007] und Fréchet-Distanz [Devogele 2002].

- *Topologische Ähnlichkeit* ist anhand der topologischen Informationen von korrelierenden Objekten zu berechnen. In [Volz 2006a] werden Adjazenz- und Inzidenzangaben (z.B. die Anzahl der verbundenen Kanten) sowie graphbasierte Merkmale (z.B. Richtung der Kanten) zur Bestimmung der topographischen Ähnlichkeit vorgeschlagen.

- *Thematische (Semantische) Ähnlichkeit* ergibt sich durch Vergleich der Attribute von zugeordneten Objekten. Dabei sind Verfahren zur Auswertung der verschiedenen Skalen und Datentypen zu unterscheiden.

- *Aggregation von Ähnlichkeiten*: Für spezifische Zielsetzung werden die verschiedenen Ähnlichkeitsmaße zusammengesetzt. Volz [2006a] berechnet die Ähnlichkeit eines Zuordnungspaars durch Kombination von geometrischen, topologischen und thematischen Ähnlichkeiten, um die Relation von Mehrfachrepräsentationen zu modellieren.

Zur Auswertung der Inkonsistenz in Mehrfachrepräsentationen wird ein wissensbasierter Ansatz von Sheeren et al. [2009] vorgestellt. Das Verfahren MECO (Method for Evaluating COnsistency) analysiert die Konsistenz in heterogenen Repräsentationen mit Hilfe von direkten und indirekten Regeln, die über die Komponente MACO (Method for Acquiring knowledge to evaluate COnsistency) durch direkte Klassifikation oder Trainieren von der Spezifikation bzw. den Daten ermittelt werden. Im Vergleich zu den direkten Regeln beinhalten die indirekten Regeln zusätzliche Bedingungen. So lassen sich die geometrischen und thematischen Inkonsistenzen bestimmen.

Die Möglichkeit der Qualitätsprüfung ändert sich beim Kartenvergleich mit Zuordnung je nach Ähnlichkeit der Datensätze. Eine flächendenke Qualitätskontrolle für einen gesamten Datensatz ist in diesem Fall möglich. Trotz der Vielzahl an diskutierten Arbeiten hat bislang kein nennenswerter Einzug der Qualitätsprüfung durch Kartenvergleich stattgefunden. Nach [Sheeren et al. 2009] beziehen sich nur einige Untersuchungen auf die Auswertung von Inkonsistenzen in MRDB und es werden häufig nur die topologischen Relationen untersucht [Rodríguez 2005]. Weiterhin wird bisher ein Kartenvergleich mit einer Zuordnung zwischen zwei Datensätzen mit hoher Ähnlichkeit und hohen Redundanzen nicht untersucht. Ein Vergleich mit einer Zuordnung zwischen einem kommerziellen Datensatz und einem VGI-basierten Datensatz mit Zuordnung existiert ebenfalls nicht. Aus diesen Gründen besteht noch Forschungsbedarf auf dem Gebiet der Qualitätsprüfung durch Kartenvergleich.

3.1.3 Indirekte Qualitätsprüfung

Die indirekten Verfahren ermitteln die Datenqualität mit Hilfe von externen Informationen, wozu z.B. Qualitätsberichte über den zu prüfenden Datenbestand zählen [ISO19114 2003]. Dabei spielen Ansätze mittels der Ontologie eine bedeutende Rolle [Devillers & Jeansoulin 2006]. Die externen Kenntnisse von Benutzern (*Problemontologie*) und die Kenntnisse von Produkten

(*Produktontologie*) sind zu gewinnen (Abbildung 3.7). Die Ermittlung der Qualität erfolgt durch Vergleich der Problem- und Produktontologie.

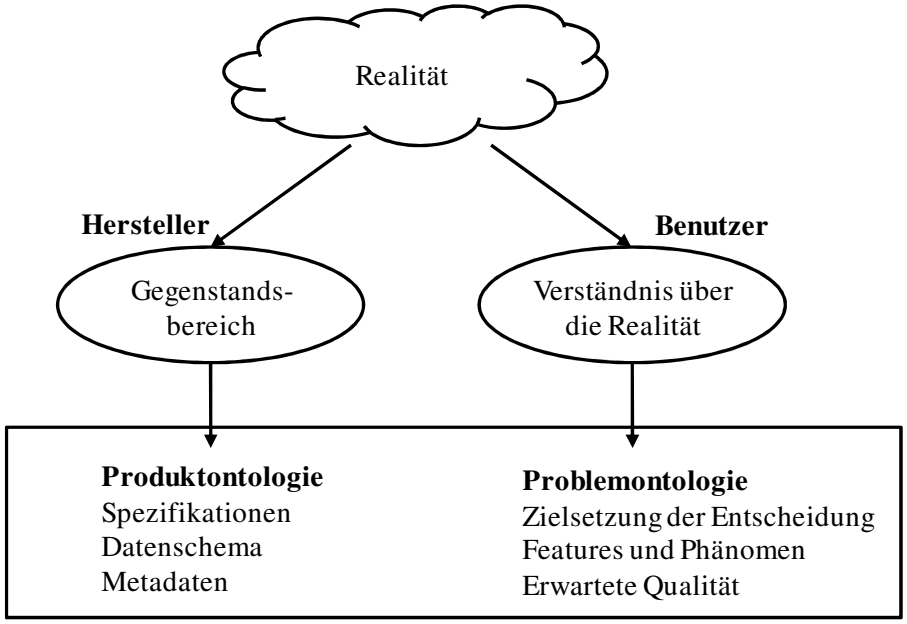


Abbildung 3.7: Ontologische Ansätze für indirekte Qualitätsprüfung (aus [Devillers & Jeansoulin 2006])

3.1.4 Qualitätsbericht und Visualisierung der Qualität

Metadaten werden häufig zum Austausch von Qualitätsinformationen in einer strukturierten Weise eingesetzt [Devillers & Jeansoulin 2006]. Unterschiedliche Modelle werden hinsichtlich der Granularität entwickelt. Beispielsweise schlägt ISO 19115 [2003] eine hierarchische Struktur zur Speicherung der Metadaten in folgenden Detailierungsstufen vor: *Dataset Series*, *Dataset*, *Feature Type*, *Feature Instance*, *Attribute Type* und *Attribute Instance*. Im Rahmen des Projekts *EuroRoads* wird eine Plattform zur Bereitstellung von Straßendaten in Europa eingerichtet [EuroRoadS 2006] und ein Austauschformat mit Berücksichtigung von Qualitätsinformationen auf Basis der Extensible Markup Language (XML) bzw. Geography Markup Language (GML) definiert [Wikström 2006].

Die Qualität der Daten lässt sich mit unterschiedlichen Verfahren (z.B. Ampelsystem, Smiley-Symbol) visualisieren, um den Benutzern die Qualitätsinformationen effizient zu vermitteln und eine interaktive Analyse zu ermöglichen. Devillers et al. [2007] präsentieren das Tool „OLAP“ zur hierarchischen Visualisierung von Qualitätsinformationen in unterschiedlichen Granularitäten. Objekte mit unterschiedlichen Qualitätsklassen werden graphisch in verschiedenen Farben dargestellt. Eine Navigation in den unterschiedlichen Granularitäten wird erlaubt. Sulo et al. [2005] entwickeln ein Werkzeug mit der Bezeichnung „Davis“ zur Visualisierung und Markierung von Daten mit schlechter Qualität in Tabellenform. Darüber hinaus bietet es sich an, Qualitätsinformationen der Vektordaten im Rasterformt zu visualisieren [Haklay 2008; Seo & O'hara 2009].

3.2 Integration

„*Integration beschreibt den Prozess der Abbildung von Daten unterschiedlicher Herkunft und verschiedenartiger Modellierung oder Struktur in einem gemeinsamen Datenmodell zum Zwecke des gleichartigen und gleichzeitigen Zugriffs durch die Anwender*“ [Bill & Zehner 2001]. Im Grunde unterscheiden sich zwei Arten der Integration: visuelle und physikalische Integration. Die visuelle Integration führt heterogene Datensätze durch Abfragen zusammen (Föderierte Datenbank), während die physikalische Integration die Datensätze in einem Zentraldatensatz verschmilzt (Conflation). Im Vergleich zu der physikalischen Integration besitzt die visuelle Integration eine bessere Aktualität und Flexibilität. Allerdings ist die Antwortzeit der Analyse durch visuelle Integration langsamer als durch physikalische Integration [Su 2005]. In der Folge werden zwei Forschungsprojekte auf dem Gebiet der Datenintegration beschrieben: NEXUS und INSPIRE.

Im Rahmen des NEXUS-Projekts wird eine Plattform entwickelt, um offene bzw. verteilte Umgebungen für ortsbezogene Dienste und Anwendungen in einer gemeinsamen Umgebung darzustellen und zu verarbeiten [Volz et al. 2000]. Damit diese untereinander kommunizieren können, müssen sie eine einheitliche Sicht auf das System haben. Die NEXUS-Plattform besteht aus drei Schichten: der Anwendungs-, der Föderations- und der Dienstebene. Die oberste Schicht wird durch die Anwendungen gebildet. Darunter liegt die sogenannte Föderationsebene, innerhalb der sich NEXUS-Knoten und ein Adressverzeichnis der Dienstgeber befinden. Die Dienstgeber befinden sich dann in Form von Spatial Model Server innerhalb der Dienstebene. Volz [2006a] fokussiert in seiner Dissertation darauf, raumbezogene Daten von offenen Geodateninfrastrukturen an die NEXUS-Plattforme bereitzustellen. Mehrfachrepräsentationen der heterogenen Datensätze sollen an dieser Stelle zusammengeführt und gemeinsam verarbeitet werden können. Die konzeptionelle Datenmodellierung der heterogenen Ausgangsdatensätze von GDF, ATKIS und ALK wird untersucht und gegenübergestellt. Ein globales Schema zur Abbildung der heterogenen Datensätze wird anschließend entwickelt. Die Mehrfachrepräsentationen werden mit Schema-Matching Verfahren ausgewertet. Die Ähnlichkeit der heterogenen Repräsentationen ergibt sich aus geometrischen, topologischen und thematischen Ähnlichkeitsmaßen. Darüber hinaus wird eine Netzwerkanalyse in den heterogenen Datensätzen mittels Einführung von Übergangsknoten entwickelt.

Das Projekt INSPIRE (INfrastructure for SPatial InfoRmation in Europe) wurde von der Europäischen Kommission spezifiziert und ist im Jahr 2007 in Kraft aufgetreten [Inspire 2009]. INSPIRE beschäftigt sich mit der Realisierung einer europäischen Dateninfrastruktur mit integrierten raumbezogenen Informationsdiensten. Am Anfang des Projekts konzentriert sich INSPIRE auf raumbezogene Daten (Geodaten) aus dem Umweltbereich und wird zukünftig auf weitere Bereiche (z.B. Landwirtschaft, Verkehr) ausgedehnt. Die Zielsetzung von INSPIRE findet sich in [Seifert 2006]. Um die grenzüberschreitende Nutzbarkeit und Kompatibilität der Dateninfrastrukturen der Mitgliedsstaaten zu sichern, werden gemeinsame Implementing Rules (Durchführungsbestimmungen) in folgenden fünf Bereichen zum Jahr 2012 entwickelt:

- *Metadata* (Metadaten) dienen zur Beschreibung der Datensätze und berücksichtigen die unterschiedlichen Ländersprachen. Zu diesem Zweck sind die ISO Normen (z.B. ISO 19115 und ISO 19119) zu verfolgen.

- *Data Specifications* definieren ein einheitliches Datenmodell, um raumbezogene Daten aus den unterschiedlichen Anwendungsbereichen und Staaten zu harmonisieren und die Interoperabilität der Daten zu verbessern.

- *Network Services* spezifizieren die Schnittstellen zur Bereitstellung der Daten im Internet.

- *Data and Service Sharing* stellen die Daten und Services bereit.
- *Monitoring and Reporting* überwachen die Umsetzung des Projekts und erzeugen Reports.

3.2.1 Zuordnung

Die Zuordnung von raumbezogenen Daten ist schon seit Mitte der 1980er Jahre ein Thema im Bereich der Geoinformatik und wurde bereits von vielen Forschern aus unterschiedlichen Gesichtspunkten untersucht (siehe Abbildung 3.8). Grundsätzlich wird zwischen Zuordnung auf Schemaebene und Zuordnung auf Objektebene bzw. Instanzebene unterschieden [Dunkars 2003; Sheeren et al. 2009]. Für Schemazuordnung werden Ansätze aus dem Datenbankbereich mit Mediation [Rahm 2001] oder mit Hilfe von Ontologien [Uitermark et al. 1999] entwickelt. Ferner lassen sich Korrespondenzen von Schemata mittels der Zuordnungsergebnisse auf der Objektebene bestimmen [Volz 2006a]. Zuordnungsansätze auf der Objektebene werden jeweils für Datensätze in ähnlichen oder unterschiedlichen Maßstäben vorgestellt. Weiterhin unterscheiden sich Zuordnungsverfahren von verschiedenen Objektarten: *Punkt*, *Linie* und *Fläche*. Die punktförmigen Objekte lassen sich einfach mit Puffer oder Distanzfunktionen zuordnen. Ein Verfahren für die Flächenzuordnung mittels Schwerpunkte wird von Kraft [1995] vorgeschlagen. Chen [2006] implementiert eine Methode für die Zuordnung einer einzelnen Linie (z.B. GPS-Tracking) zu Kanten einer Referenzkarte mit kürzester Wegesuche. In der Folge werden Zuordnungsansätze für linienförmige Objekte in heterogenen Datensätzen ausführlich beschrieben.

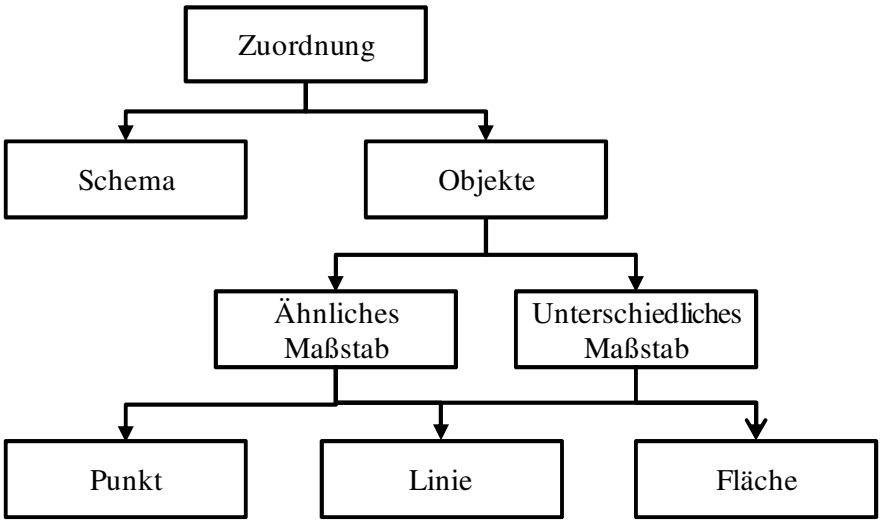


Abbildung 3.8: Ansätze der Zuordnung

Für die Zuordnung von Objekten im ähnlichen Maßstab wurde das Verfahren „Buffer growing“ von Walter [1997] entwickelt und mit Datensätzen ATKIS und GDF getestet. Die geometrischen Abweichungen in den zwei Datensätzen werden durch eine Vorverarbeitung mit einem Rubber-Sheeting Verfahren [Gillmann 1985] reduziert. Die Zuordnungskandidaten sind mit wachsendem Puffer aufzustellen. Mit Hilfe von geometrischen (z.B. Winkelunterschied, Längenunterschied), topologischen und thematischen Einschränkungen wird der beste Zuordnungskandidat bestimmt. Die Relationen der Zuordnung lassen sich in *1:1*, *1:n*, *n:1* und *n:m* untergliedern.

Das Verfahren „Buffer Growing“ wird von unterschiedlichen Forschern erweitert. Um die Genauigkeit der Zuordnung zu erhöhen, werden in [Zhang & Meng 2006] die topologischen

Unterschiede in heterogenen Datensätzen vor der Zuordnung reduziert. Gleichzeitig schlagen Zhang und Meng [2006] ein unsymmetrisches „Buffer Growing“ vor, um die lokalen geometrischen Abweichungen in den Datensätzen zu berücksichtigen und die Aufstellung von falschen Zuordnungskandidaten zu vermeiden. Eine andere Erweiterung erfolgt in [Zhang et al. 2007] durch Erkennung der Strukturen von Straßenobjekten (z.B. Kreisverkehr, Parallelstraßen, Rampe) mit räumlichen und semantischen Kriterien, um die Unsicherheit der Datenmodellierung in den heterogenen Datensätzen zu beachten. Volz [2006b] implementiert eine iterative Zuordnung auf Basis des „Buffer Growing“ Verfahrens. Im ersten Schritt werden die Zuordnungspaare mit der Relation *1:1* ermittelt. Im Anschluss daran ist die Zuordnung mit der Relation *1:2* zu berechnen. So lässt sich die Zuordnung iterativ durchführen, bis alle Zuordnungskandidaten behandelt werden.

Lüscher und Burghardt [2006] präsentieren ein Verfahren für die Zuordnung von linienförmigen Objekten in unterschiedlichen Maßstäben, womit Verknüpfungen zwischen zwei Repräsentationen VECTOR25 (1:25 000) und VECTOR200 (1:200 000) der Schweizer Landkarten berechnet werden. Zuordnungskandidaten für Knoten und Kanten (Straßen) werden zuerst mit einem Puffer selektiert. Anschließend werden Kantenkandidaten durch Auswertung der Objektklasse und Knotenkandidaten mit topologischen und geometrischen Kriterien (Knotengrad und Zwischenwinkelsumme) ausgefiltert. Daraus ergeben sich *1:1* Knotenverknüpfungen. Die Kantenverknüpfungen werden durch Berechnung des sogenannten nächstbenachbarten Wegs zwischen den verknüpften Knoten nach der Gewichtung der Hausdorff-Distanz ermittelt. In der Regel handelt es sich um *n:1* Zuordnungen zwischen Kanten im hochaufgelösten und im weniger detaillierten Datensatz.

Ein ähnlicher Zuordnungsansatz zum gleichen Zweck wird in [Mustière & Devogele 2008] vorgestellt und als *NetMatcher* bezeichnet. Knoten im weniger detaillierten Datensatz können zu mehreren Knoten und Kanten im hochaufgelösten Datensatz zugeordnet werden. Dasselbe gilt ebenso für Kanten im weniger detaillierten Datensatz. Zunächst wird eine grobe Zuordnung zwischen Knoten mit geometrischen Kriterien durchgeführt. Daran schließt sich eine grobe Zuordnung zwischen Kanten in beiden Datensätzen anhand der Hausdorff-Distanzen an. Auf Basis der groben Knoten- und Kantenzuordnung werden die Knotenpaare mittels der topologischen Ähnlichkeit bestimmt. Die Ergebnisse der Knotenzuordnung werden in drei Stufen klassifiziert: vollständig, unvollständig und unmöglich. Zum Schluss sind die Kanten mittels kürzester Wegesuche unter der Bedingung zuzuordnen, dass der Inhalt der Fläche, welche ein Zuordnungspaar bildet, am kleinsten sein muss. Der Ansatz wurde mit den Datensätzen BD CARTO (1:100 000 oder 1:250 000) und BD-TOPO (1:25 000) in Frankreich getestet.

Aus anderen Aspekten werden weitere Ansätze für die Zuordnung von linienförmigen Objekten entwickelt. Beispielsweise wird in [Ripperda 2004] eine graphbasierte Zuordnung durch Suche nach Teilgraph-Isomorphismen vorgestellt. Daraus ergibt sich die Zuordnung der korrespondierenden Objekte. Weiterhin werden Zuordnungen von linienförmigen Objekten mit einem knotenbasiertem Ansatz [Bofinger 2001; Safra et al. 2006] und Zuordnungen von Vektorobjekten im Rasterformat [Seo & O’hara 2009] präsentiert. Darüber hinaus schlägt Dunkars [2003] vor, den besten Zuordnungskandidaten durch die n-dimensionale Euklidische Distanz auf Basis der Semantik, Geometrie, Topologie und Beziehung der internen Objekte zu bestimmen. Ferner wird der beste Zuordnungskandidat in [Olteanu 2007] mit Hilfe der Wahrscheinlichkeitstheorie bestimmt.

3.2.2 Conflation

Unter Conflation (Verschmelzung) versteht man im engeren Sinne einen Satz von Funktionen und Prozeduren, der die Linien eines Datensatzes nach denen einer anderen ausrichtet und dann die Linienattribute des einen Datensatzes auf den anderen überträgt [Kappas 2001]. Das erste

interaktive und iterative Conflation-System wurde von Lynch und Saalfeld [1985] entwickelt, um einen dritten Datensatz mit besserer Qualität durch Kombination von zwei Datensätzen zu erzeugen. Conflation wird in [Yuan & Tao 1999] in *horizontale* und *vertikale* Conflation untergliedert. Darüber hinaus wird *interne* Conflation in [Blasby et al. 2003] als eine weitere Kategorie der Conflation definiert.

- *Horizontale Conflation* dient zur Eliminierung von Diskrepanzen in überlappenden Bereichen für zwei benachbarte Datensätze.

- *Vertikale Conflation* widmet sich der Behandlung von Diskrepanzen in zwei Datensätzen, welche Objekte in einem gleichen Gebiet abbilden, und lässt sich weiterhin in Vektor-Vektor, Vektor-Raster und Raster-Raster Conflation aufteilen.

- *Interne Conflation* beschäftigt sich mit Beseitigung von Diskrepanzen in einem Datensatz (z.B. Überlappungen von Flächen).

Conflation unterteilt sich grundsätzlich in zwei Aufgaben [Lupien & Moreland 1987]: Zuordnung der Features und Ausrichtung der Features. Die Ausrichtung der Features wird generell mittels Triangulation sowie Rubber-Sheeting Verfahren [Gillmann 1985] durchgeführt. Allerdings ist eine lokale Rubber-Sheeting Transformation mit Knoten als Ankerobjekte nur für einfache Fälle geeignet. Aus diesem Grund schlagen Doytsher et al. [2001] vor, lineare Features statt Punkt-Features als Ankerobjekte (Referenzobjekte) bei der lokalen Rubber-Sheeting Transformation zu benutzen. Auf diese Weise wird die Form der transformierten Objekte beibehalten. Haunert [2005] interpoliert zusätzliche Punkte für die Rubber-Sheeting Transformation, um die Verteilung der Kontrollpunkte zu verbessern. Dadurch werden die existierenden geometrischen Differenzen in den heterogenen Datensätzen bis zu einem gewissen Grade berücksichtigt.

In [Deretsky & Rdony 1993] wird ein automatischer Vektor-Vektor Conflationansatz unter Nutzung der Verkettungen von Kanten vorgestellt. Schnittpunkte der Verkettungen werden zugeordnet und als Relationen betrachtet. So wird der gesamte Bereich des Datensatzes in zugeordnete Zellen aufgeteilt. Die Geometrie der zugeordneten Verkettungen wird mit einer nicht linearen Transformation in einen gemeinsamen Datensatz transformiert. Zur Verschmelzung der nicht zugeordneten Objekte in einzelnen Zellen werden ebenfalls spezifische Filter auf Basis der Geometrie und Attribute entwickelt. Cobb et al. [1998] entwickeln ein regelbasiertes System für Conflation unter Berücksichtigung der Datenqualität und der Maßstäbe der Eingangsdatensätze. Yuan und Tao [1999] präsentieren eine komponentenbasierte Conflationstrategie, um die Komplexibilität der Entwicklung zu reduzieren und die einzelnen Komponenten der Conflation unabhängig für spezielle Anwendungen nutzen zu können.

Ein Conflationansatz mit dem Name „Best Map“ wird von Edwards und Simpson [2002] für Datensätze mit unterschiedlichen Auflösungen entwickelt. Zu diesem Zweck werden räumliche und thematische Regeln entwickelt. In hochaufgelösten Gebieten werden hochaufgelöste Daten verwendet. Außerhalb dieser Gebiete werden weniger detaillierte Daten genutzt. In Übergangsbereichen werden Konnektivitätsvektoren erzeugt, um Objekte von hochaufgelösten Daten und von weniger detaillierten Daten zu verknüpfen. Die weniger detaillierten Daten, die im hochaufgelösten Gebiet verfügbar und nicht zugeordnet sind, werden ebenfalls in den resultierenden Datensatz aufgenommen. Heutzutage stehen sowohl kommerzielle Produkte wie z.B. *MapMerger* und *Conflex* [Chen et al. 2006] als auch Open-Source-Software für die automatische Conflation von Vektordaten zur Verfügung. Zu Open Source zählt z.B. die im Rahmen des JUMP-Projekts entwickelte Software *RoadMatcher* [Blasby et al. 2003; Horn 2007].

Chen et al. [2006] unterscheiden zwei Arten von Vektor-Raster Conflation. Zum einen werden die Objekte zunächst aus Rasterdaten extrahiert. Anschließend sind die Korrespondenzen mit verschiedenen Zuordnungsverfahren zu detektieren. Zum Schluss sind Objekte mit traditionellen Conflationansätzen zu verschmelzen. Zum anderen werden einzelne Vektorobjekte zu korrespondierenden Objekten in Rasterdaten mit Snakes-basierten Verfahren ausgerichtet.

4 Globale Datenmodellierung

In diesem Kapitel wird zunächst die konzeptionelle Modellierung von *Geographic Data File* (GDF) und *OpenStreetMap* (OSM) miteinander verglichen. Daran schließt sich die Vorstellung des entwickelten globalen Datenmodells und die Abbildung der heterogenen Datenmodelle in das globale Datenmodell an.

4.1 Datenmodellierung

Im Allgemeinen wird der Vorgang der Modellierung in fünf Schritte aufgeteilt (siehe Abbildung 4.1). Bei der Objektauswahl sind Objekte der realen Welt festzulegen, die in einem Datenmodell abgebildet werden. Weiterhin sind geometrische und topologische Strukturen von ausgewählten Objekten und Attribute der Objekte zu definieren. Um eine bessere Übersichtlichkeit zu gewinnen, wird häufig eine große Anzahl von Objekten einer spezifischen Thematik (z.B. Verkehr) in verschiedenen Objektklassen gruppiert. Schließlich sind Beziehungen von Objekten bzw. Objektklassen aufzustellen [Möser et al. 2004].

Abbildung 4.1: Vorgang der Modellierung

4.1.1 GDF - Geographic Data File

Im Bereich der Fahrzeugnavigation werden momentan hauptsächlich zwei Datenmodelle weltweit verwendet. Zu einem kommt das physikalische Speicherformat mit der Bezeichnung KIWI zum Einsatz, welches vom Japanischen KIWI-Konsortium entwickelt wurde [KIWI 2000]. Navigationsdaten werden in KIWI hierarchisch strukturiert, um einen schnellen Datenzugriff zu ermöglichen. Zum anderen wird das *Geographic Data File* verwendet, welches sowohl ein konzeptionelles sowie logisches Datenmodell als auch ein Standardaustauschformat für Navigationsdaten zur Verfügung stellt. Die erste Entwurfsversion (GDF 1.0) wurde im Jahr 1988 als Produkt vom EUREKA Projekt DEMETER herausgegeben. Weiterhin wurden einige Zwischenversionen von GDF freigegeben. Im Jahr 1995 wurde GDF 3.0 vom CEN/TC278 offiziell als europäischer Standard akzeptiert [ERTICO 1995].

Phänomene der realen Welt werden im GDF-Datenmodell durch *Features*, *Attribute* und *Relationen* modelliert (siehe Abbildung 4.2). Im Mittelpunkt des konzeptionellen Datenmodells von GDF stehen *Features*, welche geographische Objekte der realen Welt repräsentieren und exakt zu einer *Feature Class* sowie einem *Feature Theme* zugeordnet werden. Features werden in drei geometrische Arten unterteilt: *Punkt*, *Linie* und *Fläche*. Eigenschaften von Features werden als *Attribute* modelliert. Beziehungen zwischen Features werden als *Relationen* dargestellt. Regelungen zur Erfassung und Speicherung der Features, Attribute und Relationen werden in GDF 3.0 in verschiedenen Katalogen definiert. Folgende Featurethemen werden im Featurekatalog abgebildet

[ERTICO 1995]: 1) *Roads and Ferries*, 2) *Administrative Areas*, 3) *Settlements and Named Areas*, 4) *Land Cover and Use*, 5) *Brunnels*, 6) *Railways*, 7) *Waterways*, 8) *Road Furniture*, 9) *Services*, 10) *Public Transport* und 11) *General Features.*

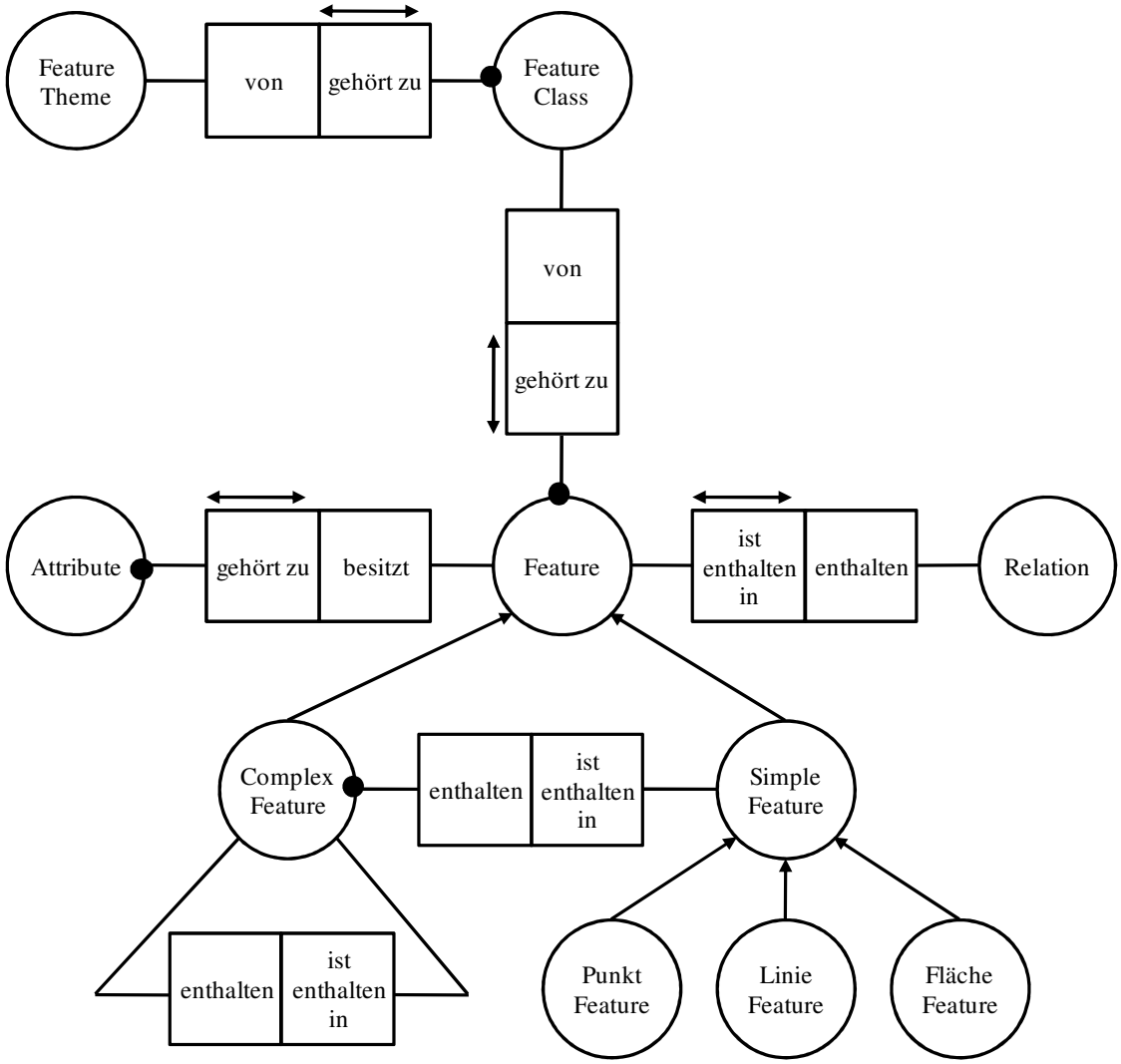


Abbildung 4.2: Konzeptionelles Datenmodell von GDF (nach [ERTICO 1995] und [Walter 1997])

Das konzeptionelle Datenmodell von GDF repräsentiert die Objekte in drei hierarchischen Stufen (Levels) [ISO14825 2004]:

- *Level 0* modelliert die grundlegenden graphischen Primitive (Punkt, Linie und Fläche).
- *Level 1* bildet die Objekte als Simple Feature (z.B. *Road Element*, *Junction*, *Ferry Connection*) ab und ist z.B. für die Routenberechnung geeignet.
- *Level 2* modelliert die Objekte als Complex Features (z.B. *Road*, *Intersection*) und eignet sich z.B. für die Kartendarstellung. Ein Complex Feature kann aus mehreren Simple Features und/oder Complex Features zusammengesetzt sein.

Abbildung 4.3 zeigt ein Beispiel für die Abbildung von Objekten in Level 1 und Level 2. Der Kreisverkehr wird in Level 1 mit vier linienförmigen Features modelliert (Abbildung 4.3a) und in Level 2 als ein Complex Feature (Intersection) abgebildet (Abbildung 4.3b).

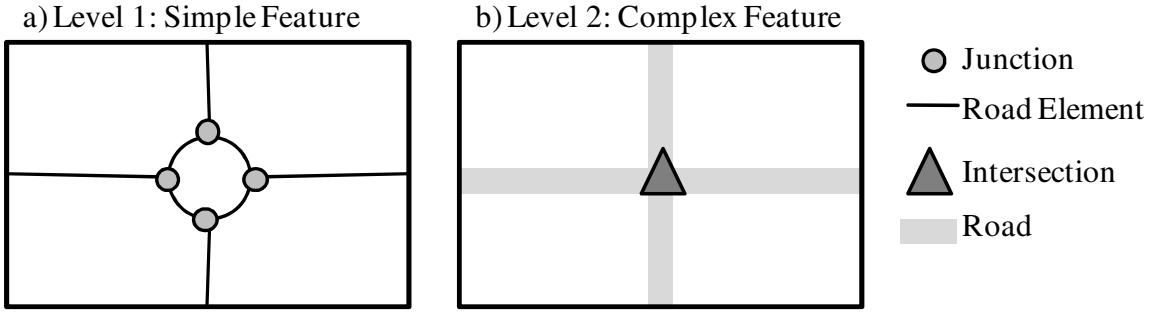


Abbildung 4.3: Darstellung eines Kreisverkehrs in GDF-Levels

Im GDF-Attributkatalog sind zwei Typen von Attributen zu unterscheiden: einfache Attribute und zusammengesetzte (composite) Attribute. Zu zusammengesetzten Attributen gehören z.B. zeitabhängige Attribute und segmentierte Attribute zur Beschreibung eines Teils (Segments) von *Road Elements*. In einer Relation können Features von unterschiedlichen Feature Classes (z.B. *Service* entlang eines *Road Elements*) oder von gleichen Feature Classes (z.B. Restriktion für Manöver mit mindestens zwei *Road Elements*) abgebildet werden. Darüber hinaus kann eine Relation durch Attribute beschrieben werden [ERTICO 1995].

Im Jahr 2004 wurde eine übergearbeitete GDF-Version (GDF 4.0) von ISO/TC 204 herausgegeben [ISO14825 2004]. Im Vergleich zu GDF 3.0 finden sich folgende Änderungen in GDF 4.0 [Essen & Hiestermann 2005]:

- *Konzeptionelle Änderungen*: Das topologische Modell wird in GDF 4.0 erweitert. In GDF 3.0 wird nur die sogenannte Volltopologie spezifiziert, die topologische Relationen zwischen Punkten, Linien und Flächen explizit definiert (Planarer Graph). Allerdings ist dies nicht für alle Applikationen notwendig. Aus diesem Grund werden in GDF 4.0 die Topologie der Konnektivität (Relationen zwischen Punkten und Linien ohne Berücksichtigung von Flächen) für eine effiziente Netzwerkanalyse und nicht-explizite Topologie (keine Definition von räumlichen Relationen) für eine effiziente Kartendarstellung spezifiziert. Darüber hinaus werden zwei-Byte Charakter zur Unterstützung von Sprachen wie z.B. Chinesisch, Japanisch und Arabisch definiert und das Sub-Attributmodell für zusammengesetzte Attribute verbessert.

- *Änderungen der Inhaltsdefinition*: Dazu zählen die Spezifikation der nicht hierarchischen administrativen Flächenstruktur und die Erweiterung des Adressenmodells.

Für Fahrassistenzsysteme, Personennavigation sowie neue Dateninhalte (insbesondere 3D-Daten) stellt ISO die XGDF (eXtended GDF) als eine zukünftige Weiterentwicklung von GDF 4.0 vor [Essen & Hiestermann 2005]. Weiterhin sind z.B. Techniken zur Verbesserung der Interoperabilität in der zukünftigen GDF-Version zu spezifizieren.

NavTeq

Das US-amerikanische Unternehmen *NavTeq* (Navigation Technologies) erfasst seit 1985 weltweit Navigationsdaten im GDF-Datenmodell. Ein internes Datenformat (NavTeq Core Map Database) wird im Herstellungsprozess bei NavTeq verwendet. Weiterhin werden verschiedene standardisierte

Datenformate als Produkte bereitgestellt, um diverse Anforderungen von Kunden zu erfüllen (siehe Tabelle 4.1).

Produkte	Datenformat	Inhalt des Produkts
GDF 3.0	ACSII Datei Struktur, sequenziell sortiert nach Rekordtypen	Vollständig
RDF (Relational Data Format)	Relationale Repräsentation der NAVTEQ Datenbank, Auslesen in Oracle oder SQL Server möglich	Vollständig
SIF+(Standard Interchange Format+)	ACSII Datei Struktur, sequenziell sortiert nach LinkID	Vollständig
NAVSTREETS	Layerbasierte Repräsentation, verfügbar in ESRI® Shapefile Format und MapInfo® Table Format	Nicht vollständig (z.B. Complex Feature, City Modell und 3D Landmark fehlen)

Tabelle 4.1: Gegenüberstellung der Produkte von NavTeq (nach [NavTeq 2007])

TeleAtlas

TeleAtlas ist ein niederländisch-belgisches Unternehmen und wurde im Jahr 1984 gegründet. Aus unterschiedlichen Anforderungen an Funktionalitäten von Navigationsdaten werden die Produkte *MultiNet*, *ConnectPlus* und *Connect* zur Verfügung gestellt (siehe Tabelle 4.2). Alle Produkte stehen in GDF-AS (ASCII-Sequentiell), GDF-AR (ASCII-Relational), Shapefile und Oracle Spatial Format zur Verfügung [Davie & McCullar 2009].

Funktionalität	*MultiNet*	*ConnectPlus*	*Connect*
Routenplanung und Kartendarstellung	√	√	√
Grundfunktion der Navigation	√	√	
Turn-by-Turn Navigation	√		

Tabelle 4.2: Gegenüberstellung der Produkte von TeleAtlas (nach [Davie & McCullar 2009])

4.1.2 OpenStreetMap

Das Projekt *OpenStreetMap* startete im Jahr 2004 und hat das Ziel, freie geographische (raumbezogene) Daten (wie zum Beispiel Straßenkarten) zu erstellen. Zur Erfassung und Verwaltung der Daten wurden spezielle Werkzeuge und Datenmodelle bzw. Datenformate entwickelt. Menschen (OSM-Community), die die Daten freiwillig erfassen und fortführen, sind allerdings für das Projekt am bedeutendsten [Ramm & Topf 2009]. Aus aufgezeichneten GPS-Positionen oder Referenzdaten (z.B. Luftbilder) werden OSM-Daten erfasst. Darüber hinaus können frei verfügbare raumbezogene Daten (z.B. US-Tiger-Straßendaten) importiert werden.

OSM-Daten werden in einem zentralen Datenbankserver (PostgreSQL) in London gespeichert. Auf die Daten kann mit verschiedenen Editoren (z.B. JOSM - Java OpenStreetMap Editor) oder über

APIs (Application Programming Interface) zugegriffen werden. Darüber hinaus ist es möglich, OSM-Daten in verschiedenen Formaten zu exportieren (siehe Abbildung 4.4). Eine Region (minimale bzw. maximale Länge und Breite) ist durch Eingabe der Koordination oder Anpassung der Größe eines Rechtecks auf der Karte festzulegen. In der vorliegenden Arbeit werden OSM-Daten über die Regionsdefinition im XML-Format (*.*OSM* Datei) bereitgestellt.

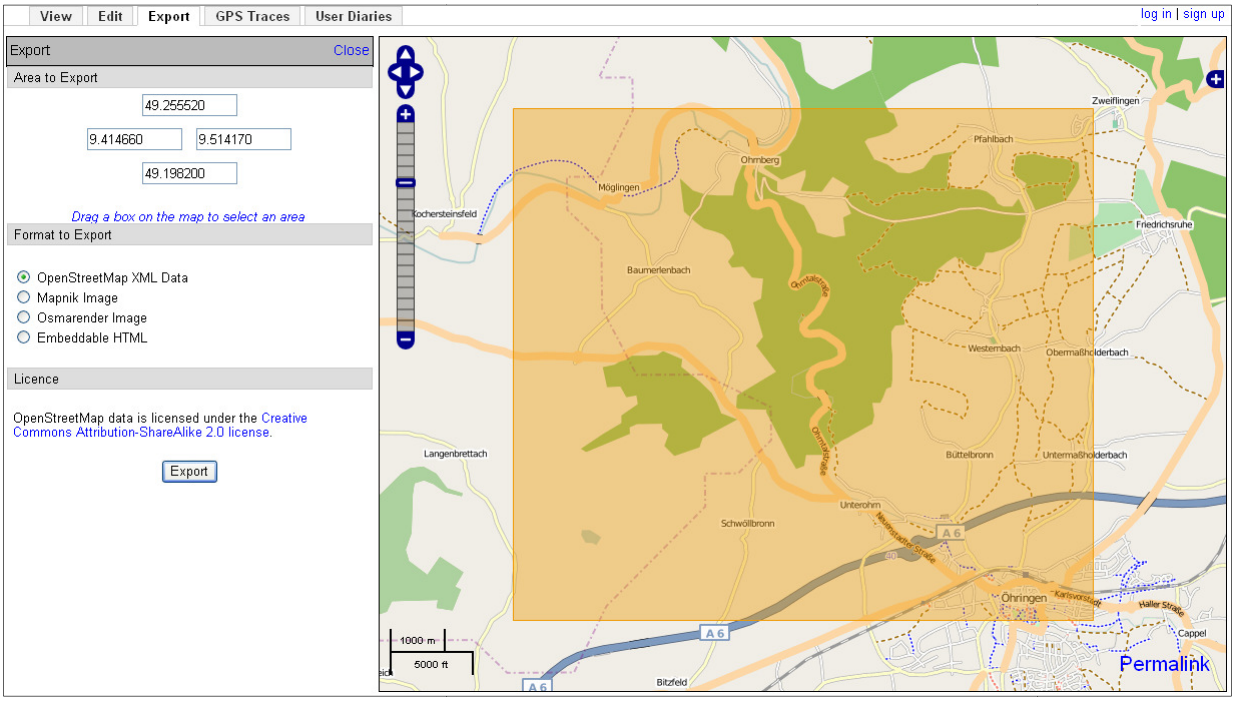


Abbildung 4.4: Datenexport von OpenStreetMap (aus [OpenStreetMap 2008])

Das OSM-Datenmodell besteht aus zwei Objekttypen *Node* (Knoten) und *Way* (Weg) sowie einem Datentyp *Relation* (Abbildung 4.5) [Ramm & Topf 2009]. Weiterhin werden *Tags* zur Beschreibung der Objekttypen und des Datentyps definiert. Jeder Objekt- bzw. Datentyp enthält eine eindeutige numerische Identifikation (ID), eine beliebige Menge von *Tags* und Informationen über die Datenerfassung (z.B. Herkunft, historische Änderungen).

- Ein *Node* verfügt über geographische Länge und Breite und dient vornehmlich zur Gestaltung der Zwischenpunkte von *Ways*. Weiterhin werden POIs (Point of Interest) als *Node* mit spezifischen *Tags* abgebildet.

- Ein *Way* besteht aus mindestens zwei *Nodes* und stellt hauptsächlich linienförmige Objekte wie z.B. Straßen dar. Der Objekttyp *Area* (Fläche) liegt im OSM-Datenmodell nicht vor und wird durch geschlossene *Ways* mit spezifischen *Tags* (z.B. „water“) modelliert.

- Eine *Relation* bildet die Beziehung zwischen Objekttypen ab und ist eine Kombination von einer beliebigen Menge von *Nodes*, *Ways* und anderen *Relationen*.

- *Tags* sind Attribute zur Beschreibung eines *Nodes*, eines *Ways* oder einer *Relation* und bestehen aus einem Schüssel (*key*) und einem Wert (*value*).

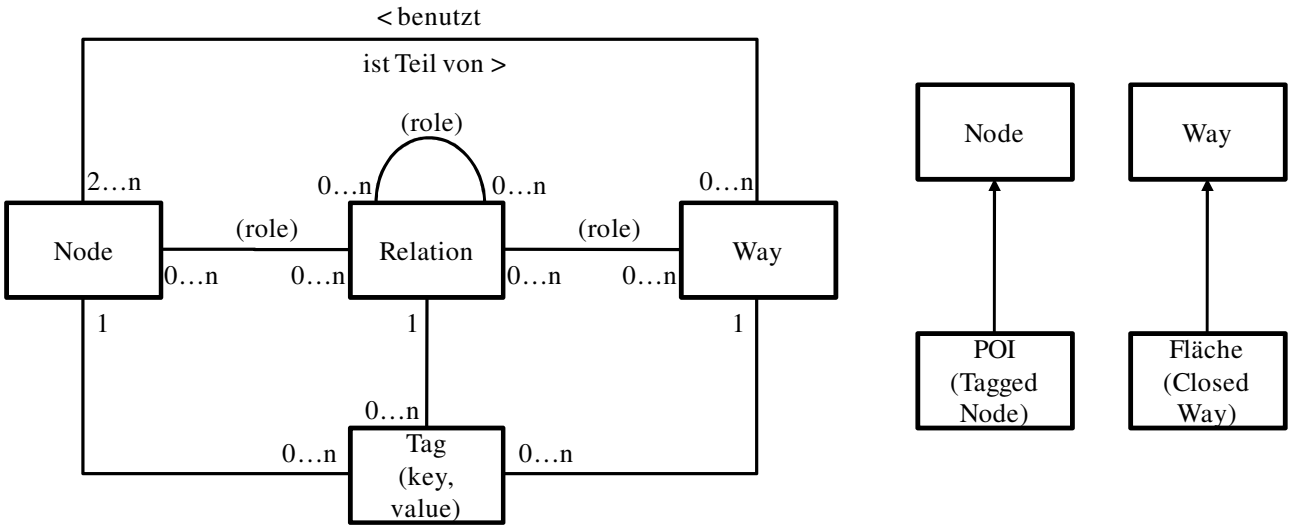


Abbildung 4.5: Vereinfachtes Datenmodell von OpenStreetMap (aus [Ramm & Topf 2009])

Im OSM-Datenmodell besteht keine strenge Spezifikation für *Tags*. Aus diesem Grund können alle möglichen Objekte der realen Welt als *Node* oder *Way* mit entsprechenden *Tags* abgebildet werden. In Abbildung 4.6 wird die hierarchische Gruppierung von Objektklassen im OSM-Datenmodell dargestellt. Hierbei handelt es sich grundsätzlich um eine dreistufige Hierarchie (z.B. Way, Highway, Road). Objekte und Relationen lassen sich von einer .osm Datei mittels Erkennung von spezifischen *Tags* abstrahieren und in verschiedene Objektklassen (Layers) bzw. Tabellen aufteilen. Aus dem Objekttyp *Node* sind POIs und Koordinaten der Zwischenpunkte von *Ways* zu ermitteln. Aus *Ways* werden linienförmige Objektklassen (z.B. *Street*, *Railway*) und flächenförmige Objektklassen (z.B. *Building*, *Landuse*) berechnet. Aus *Relationen* sind z.B. Tabellen für topologische Relationen und Restriktionen zu erzeugen.

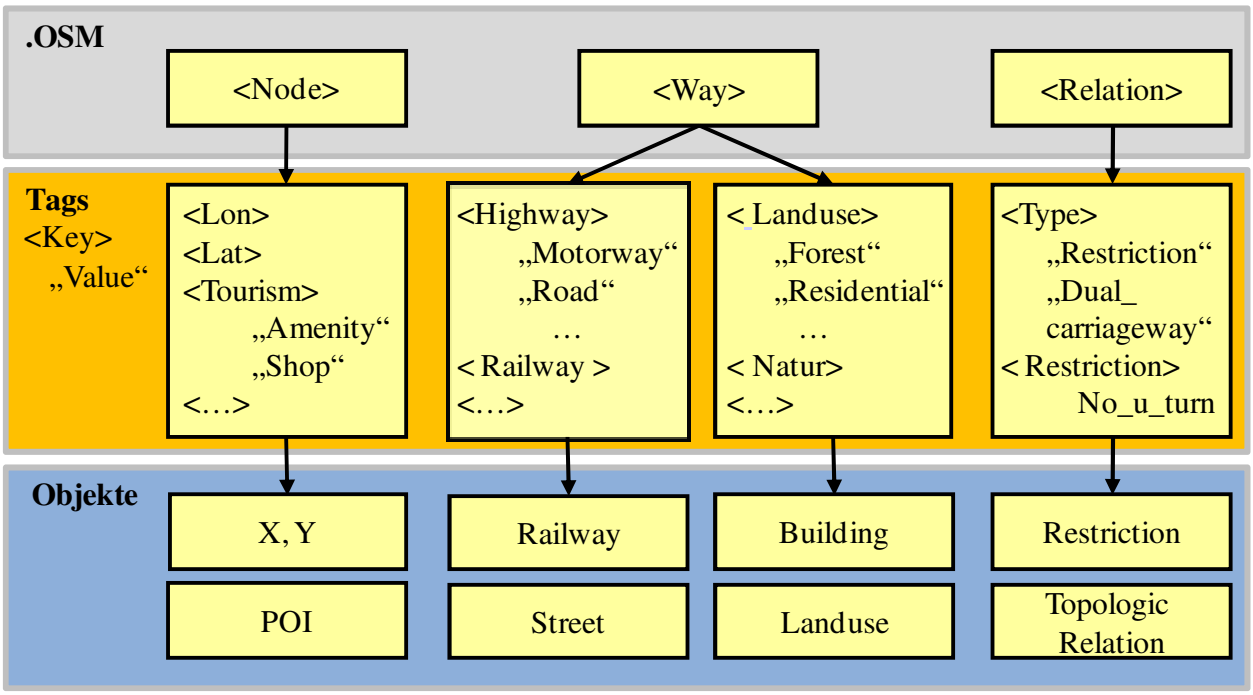


Abbildung 4.6: Objektgruppierung in OSM und Ermittlung von Objekten in einer .osm Datei

4.1.3 Gegenüberstellung der Konzepte zur Modellierung

Im Folgenden werden die Konzepte der Modellierung von Straßenobjekten im GDF- und OSM-Datenmodell verglichen. Tabelle 4.3 gibt eine Übersicht der konzeptionellen Modellierung in beiden Datenmodellen.

Modellierung	**GDF**	**OSM**
Objektauswahl	Relevante Objekte für Navigation	Beliebige Objekte
Benutzerdefinierte Objekte, Attribute und Relationen	Möglich durch Verwendung von vorbehaltenden Codes	Möglich durch erweiterte Tags
Objektstruktur	Objekte in drei unterschiedlichen Komplexitätsstufen (Levels)	Keine spezifische Stufenstruktur vorhanden, Codierung von komplexen Objekten mit Hilfe von Relationen möglich
Objekttyp	Punkt, Linie, Fläche	Node, Way
Objektmerkmale	Attribut	Tags
Zusammengesetzte Attribute	Möglich	Möglich
Zeitabhängige Attribute	Möglich	Möglich
Segmentierte Attribute	Möglich	Nicht möglich
Objektgruppierung	Zwei hierarchische Stufen	Drei hierarchische Stufen
Objektbeziehung	Möglich	Möglich
Attribute für Relationen	Möglich	Möglich

Tabelle 4.3: Gegenüberstellung der Modellierung im GDF- und OSM-Datenmodell

Im Allgemeinen werden Straßen im GDF- und OSM-Datenmodell als linienförmige Objekte und Kreuzungen als punktförmige Objekte modelliert. Aus diesem Gesichtspunkt ist die Modellierung in beiden Datenmodellen ähnlich. Allerdings ist die konzeptionelle Modellierung in GDF komplexer [Ramm & Topf 2009]:

- Bei der Objektauswahl werden relevante Objekte für die Navigationsapplikation im GDF-Datenmodell abgebildet, während das OSM-Datenmodell beliebige Objekte durch erweiterbare *Tags* darstellen kann. Allerdings sind benutzerdefinierte Objekte in GDF durch Verwendung von vorbehaltenden Codes möglich (Bereich: 90-99).

- Objekte im GDF-Datenmodell werden in drei unterschiedlichen Komplexitätsstufen modelliert, während im OSM-Datenmodell keine spezifische Stufenstruktur besteht. Allerdings ist die Modellierung von komplexen Objekten im OSM-Datenmodell mit Hilfe von Relationen möglich. Ferner werden flächenförmige Objekte im OSM-Datenmodell mit geschlossenen *Ways* und spezifischen *Tags* modelliert.

- Das Attributkonzept ist in beiden Datenmodellen ähnlich [OpenStreetMap 2008]. Dennoch werden segmentierte Attribute im OSM-Datenmodell nicht unterstützt. *Ways* sind in diesem Fall in mehrere *Ways* zu zerlegen [Ramm & Topf 2009].

- Im GDF-Datenmodell werden die Objektklassen in zwei hierarchischen Stufen dargestellt [Kieler et al. 2007], während es sich um eine dreistufige Hierarchie im OSM-Datenmodell handelt.

- Die Modellierung der Objektbeziehung ist in beiden Datenmodellen ähnlich.

Trotz der Verwendung von gleichen bzw. ähnlichen Datenmodellen existieren Unterschiede in den drei Datensätzen, die im nächsten Abschnitt durch eine Abbildung der unterschiedlichen Datenmodellierungen in ein globales Datenmodell minimiert werden.

4.2 Entwicklung des globalen Datenmodells

Im Rahmen der vorliegenden Arbeit stehen NAVSTREETS von NavTeq in *Shapefile*, MultiNet von TeleAtlas in *Shapefile* und OSM-Daten durch Export in XML-Format zur Verfügung (siehe Abbildung 4.7). Durch eine Transformation werden die heterogenen Ausgangsdatenmodelle in das übergeordnete Datenmodell abgebildet, welches auf dem GDF-Datenmodell basiert. In dem übergeordneten Datenmodell finden die Zuordnung, Qualitätsanalyse und Datenverschmelzung statt.

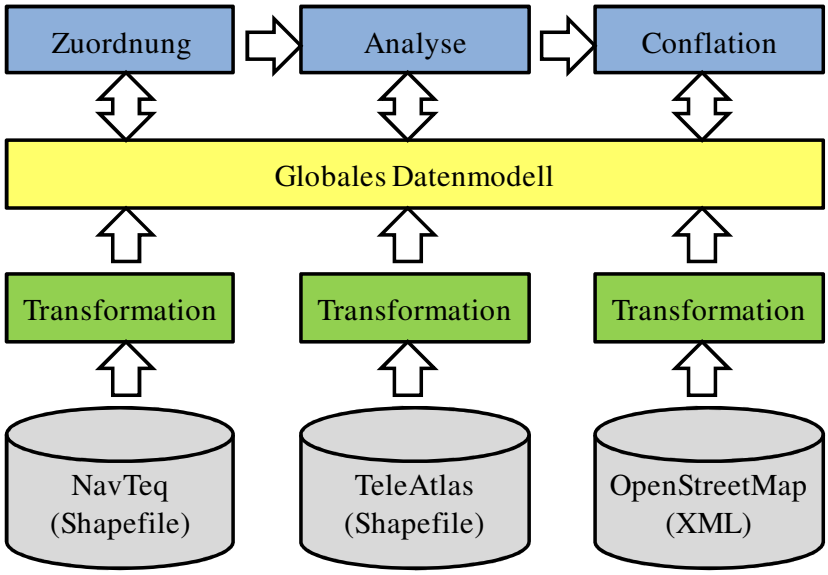


Abbildung 4.7: Entwicklung eines globalen Schema

Aufgrund des großen Umfangs können nicht alle Objekte, die im GDF- und OSM-Datenmodell abgebildet sind, in der vorliegenden Arbeit untersucht werden. So werden Objekte des Straßenverkehrs für die Untersuchung ausgewählt, weil sie für Navigationsanwendungen am relevantesten sind. Außerdem können nicht alle Attribute untersucht werden. In Anhang A werden die ausgewählten Attribute und ihre Ausprägungen in den verschiedenen Datenquellen ausführlich beschrieben.

Das globale Datenmodell repräsentiert Straßenobjekte in zwei hierarchischen Stufen: einfache Objekte (Knoten und Kanten) und komplexe Objekte (Komplexknoten und Komplexkanten). Da komplexe Objekte (Complex Features) in NavTeq (NAVSTREETS) und in den exportierten OSM-Daten nicht vorhanden sind, werden in diesem Abschnitt nur einfache Objekte berechnet. Im Folgenden wird zunächst auf die Vorteile und Notwendigkeit der Entwicklung des globalen Datenmodells eingegangen. Der Rest des Kapitels widmet sich der semantischen Homogenisierung und der Abbildung von Objekten in das globale Datenmodell.

4.2.1 Motivation

Die unterschiedliche Modellierung in den Datensätzen führt zu Problemen bei der Datenintegration. Mittels des übergeordneten Datenmodells wird der Aufwand der Datenintegration reduziert. Zusammenfassend ist die Entwicklung des übergeordneten Datenmodells aus folgenden Aspekten empfehlenswert:

- *Vereinfachung des Datenmodells*: Das GDF-Datenmodell ist sehr komplex und nicht leicht nachvollziehbar. Darüber hinaus ist das OSM-Datenmodell flexibel und erweiterbar. Das globale Datenmodel ermöglicht eine bessere Übersichtlichkeit und eine vereinfachte Modellierung [Volz 2006a].

- *Minimierung der semantischen Heterogenität*: Mit dem globalen Datenmodell werden die Unterschiede der semantischen Darstellungen in den heterogenen Datensätzen minimiert.

- *Minimierung der Heterogenität der Modellierung*: Straßenobjekte mit mehreren *Straßennamen* oder einer zusätzlichen *Road Number* werden in NAVSTREETS als mehrere Objekte modelliert. Dies führt zu Problemen bei der Integration (z.B. bei der Berechnung der Zuordnungsrelation).

- *Beseitigung von Fehlern der Erfassung*: In manchen Situationen wird die topologische Modellierung in OSM nicht streng eingehalten. Abbildung 4.8 zeigt eine Gegenüberstellung der topologischen Modellierung in TeleAtlas (links) und OpenStreetMap (rechts). An vielen Stellen sind die Straßen (*Ways*) in OpenStreetMap nicht getrennt.

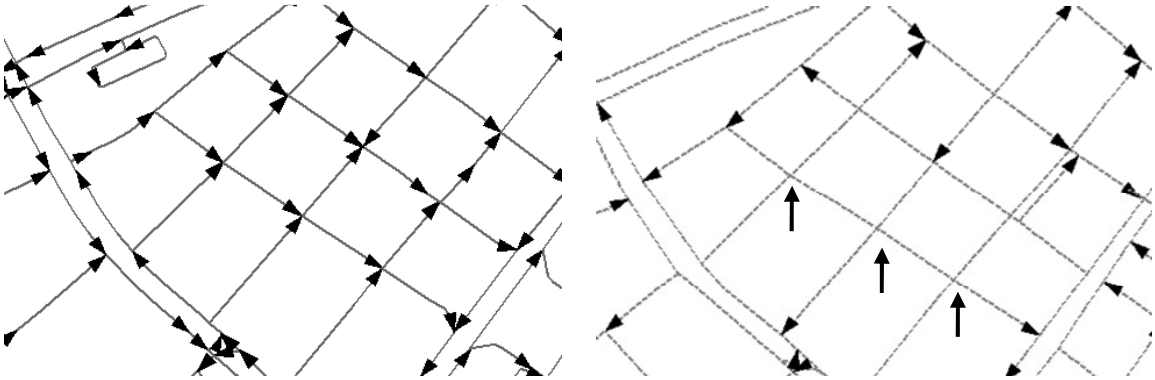


Abbildung 4.8: Topologische Modellierung in TeleAtlas (links) und in OpenStreetMap (rechts)

- *Speicherung von Ergebnissen der Zuordnung und Qualitätsanalyse*: Ergebnisse der Zuordnung und Qualitätsanalyse sind im globalen Datenmodell aufzunehmen.

- *Speicherung von neuen Informationen*: Neue Informationen, die durch die Datenintegration entstanden sind, müssen ebenfalls im globalen Datenmodell abspeichert werden.

4.2.2 Semantische Homogenisierung

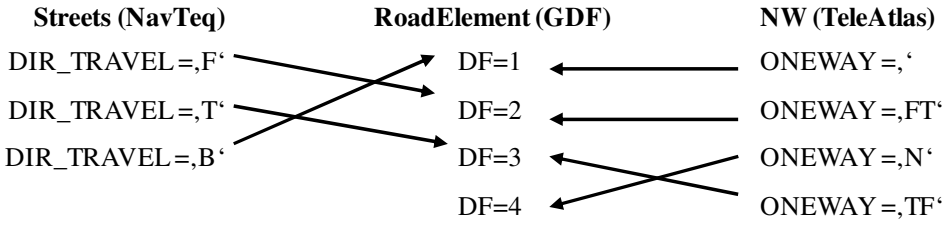
Ansätze zur automatischen Bestimmung von semantischen Korrespondenzen in heterogenen Datensätzen wurden bereits in vielen Arbeiten vorgestellt. Infolgedessen werden die semantischen Korrespondenzen in der vorliegenden Arbeit manuell bestimmt. Danach werden die heterogenen semantischen Darstellungen in den unterschiedlichen Datenquellen mit folgenden Operationen

homogenisiert, um einen einfachen Vergleich von Attributen zu ermöglichen [Bauer & Günzel 2000]:

- *Anpassung von Datentypen*: Unterschiedliche Datentypen von Attributen in den verschiedenen Datenquellen sind zu einem gleichen Datentyp zu konvertieren (z.B. Text nach Numerisch konvertieren).

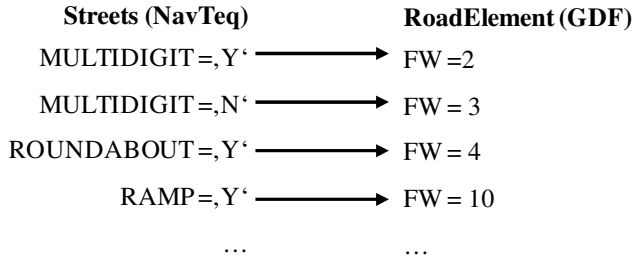
- *Vereinheitlichung von Zeichenketten*: Unterschiedliche Schreibweisen für Zeichenketten (z.B. Sonderzeichen, Groß- und Kleinbuchstaben) in den verschiedenen Datenquellen sind zu vereinheitlichen.

- *Konvertierung von Kodierung*: Die unterschiedlichen Kodierungen für Attributwerte sind zu vereinheitlichen. Im folgenden Beispiel werden die Attributwerte von *Direction of Traffic Flow (DF)* vereinheitlicht:



- *Berechnung von abgeleiteten Werten:* Von vorhandenen Attributwerten (Intervall) sind neue Werte (Ordinal) abzuleiten. Beispielsweise wird das Attribut *Speed Category* in TeleAtlas numerisch erfasst, während es in NavTeq in unterschiedlichen Kategorien erfasst wird.

- *Kombination/Separierung von Attributen*: Mehrere Attribute werden in einem Attribut zusammengefasst, welches mehrere Attributwerte besitzt (z.B. *FW- Form of Way*):



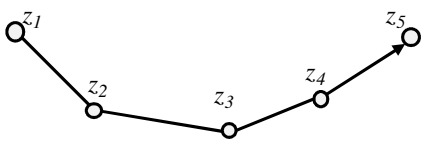
Allerdings lässt sich nicht jede semantische Heterogenität (wie z.B. Attribute mit unterschiedlichen Klassifikationen in den heterogenen Datenquellen) beseitigen. Dazu zählt z.B. das Attribut *Functional Road Class* (fünf Klassen in NavTeq und zehn Klassen in TeleAtlas).

4.2.3 Transformationsregeln

Um eine einheitliche Modellierung von Objekten in den verschiedenen Datenquellen zu erzielen, werden Objekte bei der Transformation zusammengefasst oder aufgeteilt. So sind folgende Regeln bei der Transformation von Objekten zu verwenden: *1:1*, *1:n* und *n:1*. Attribute der Objekte sind dann mit den vorgestellten Verfahren der semantischen Homogenisierung in das übergeordnete Datenmodell abzubilden.

Transformationsregeln für NavTeq

Straßenobjekte (z.B. Road Element, Ferry Connection) werden in NAVSTREETS in der FeatureClass *Streets* gruppiert, welche alle relevanten Attribute inklusive Hausnummern enthält. Straßen, die mehr als einen *Straßennamen* oder eine zusätzliche *Road Number* enthalten, werden als mehrere Objekte mit einer gleichen *Link_ID* modelliert. Abbildung 4.9 stellt ein Beispiel für die Abbildung von NavTeq-Objekten in das übergeordnete Datenmodell dar. Das Straßenobjekt verfügt über einen *Straßenname* („Rotebühlplatz“) sowie eine *Road Number* („B27A“) und wird in NAVSTREETS als zwei linienförmige Objekte mit einer gleichen *Link_ID* erfasst. Darüber hinaus werden fünf punktförmige Objekte (1 Anfangsknoten, 3 Zwischenpunkte und 1 EndKnoten) in FeatureClass *Z_Level* erzeugt. Nach der Transformation werden die zwei linienförmigen Objekte in einem Objekt gruppiert (Relation *n:1*). Aus *Z_Level* sind Knoten zu ermitteln. So wird der Anfangs- und Endknoten (z_1 und z_5) als Knoten in das übergeordnete Datenmodell abgebildet.

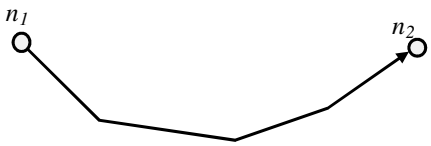


Streets

Objekt	LINK_ID	ST_NAME	…
1	52854178	Rotebühlplatz	…
2	52854178	B27A	…

Z_Level

Objekt	LINK_ID	POINT_NUM	NODE_ID	…
1	52854178	1	528522055	…
2	52854178	2	0	…
3	52854178	3	0	…
4	52854178	4	0	…
5	52854178	5	528522090	…



Kanten

Objekt	LINK_ID	NAME	RN	…
1	52854178	Rotebühlplatz	B27A	…

Knoten

Objekt	NODE_ID	…
1	528522055	…
2	528522090	…

Abbildung 4.9: Abbildung von NavTeq-Objekten in das übergeordnete Datenmodell

In Tabelle 4.4 sind die Transformationsregeln für NavTeq-Objekte zu ermitteln. Aus einer oder mehreren *Streets* wird eine Kante generiert. Weiterhin werden nicht alle Objekte von *Z_Level* als Knoten abgebildet.

NavTeq (NAVSTREETS)	Relation	Globales Datenmodell
Streets	1:1, n:1	*Kanten*
Z_Level	1:1, 1:*	*Knoten*

Tabelle 4.4: Transformationsregeln für NavTeq (in Anlehnung an [Volz 2006a])

Transformationsregeln für TeleAtlas

FeatureClasses *JC* (Junction), *NW* (Network) und *GC* (GeoCode) in MultiNet werden bei der Transformation ins globale Datenmodell verwendet. Ähnlich wie *Streets* in NAVSTREETS werden alle relevanten Objekte (Road Element, Ferry usw.) des Straßenverkehrs in *NW* (Network)

gespeichert. Weiterhin sind Attribute (wie z.B. Hausnummern) von der FeatureClass *GC* zu ermitteln (siehe Abbildung 4.10).

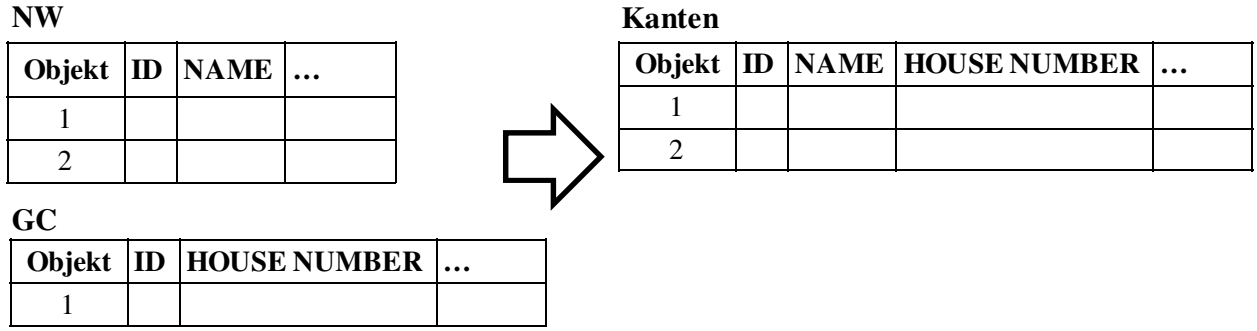


Abbildung 4.10: Abbildung von TeleAtlas-Objekten in das übergeordnete Datenmodell

Aus Tabelle 4.5 sind die Transformationsregeln für TeleAtlas-Objekte zu entnehmen. Durch Zusammenführung von FeatureClass *NW* und *GC* werden Kanten generiert. Von *JC* werden Knoten ermittelt. Die Anzahl der Objekte bleibt nach der Transformation unverändert.

TeleAtlas (MultiNet)	Relation	Globales Datenmodell
NW (Network)	1:1	*Kanten*
GC (GeoCode)		
JC (Junction)	1:1	*Knoten*

Tabelle 4.5: Transformationsregeln für TeleAtlas

Transformationsregeln für OpenStreetMap

Wie bereits in Abbildung 4.6 dargestellt, sind Straßenobjekte anhand ihrer *Tags* von einer exportierten .OSM Datei zu extrahieren und in unterschiedlichen Objektklassen bzw. Tabellen abzuspeichern. Für Navigationszwecke wie z.B. Routenberechnung mit OSM-Daten ist das in Abbildung 4.8 dargestellte Problem zu beseitigen [Schmitz et al. 2008].

Hierzu wird ein vierstufiger Vorverarbeitungsprozess eingeführt (siehe Abbildung 4.11). Im ersten Schritt werden die Anfangs- und Endknoten von *Ways* ermittelt. Im Anschluss daran sind die Schnittpunkte der *Ways* nach der Geometrie und Topologie zu berechnen. Zur Berechnung des Schnittpunkts werden Determinanten eingesetzt (siehe [Berg et al. 2008]). Weiterhin sind die *Ways* anhand der ermittelten Schnittpunkte sowie Anfangs- und Endknoten zu zerlegen. Zum Schluss werden Attribute der alten *Ways* in die neu generierten Kanten transformiert. Attribute wie z.B. Länge der Kanten sind allerdings neu zu berechnen. Abbildung 4.11a stellt die originalen *Ways* ohne Schnittpunkte dar. Die ermittelten Anfangs- und Endknoten werden in Abbildung 4.11b in Grün (hell) dargestellt, während die berechneten Schnittpunkte in Abbildung 4.11c in Rot (dunkel) gezeichnet werden. Die Kanten nach der Zerlegung sind in Abbildung 4.11d zu sehen.

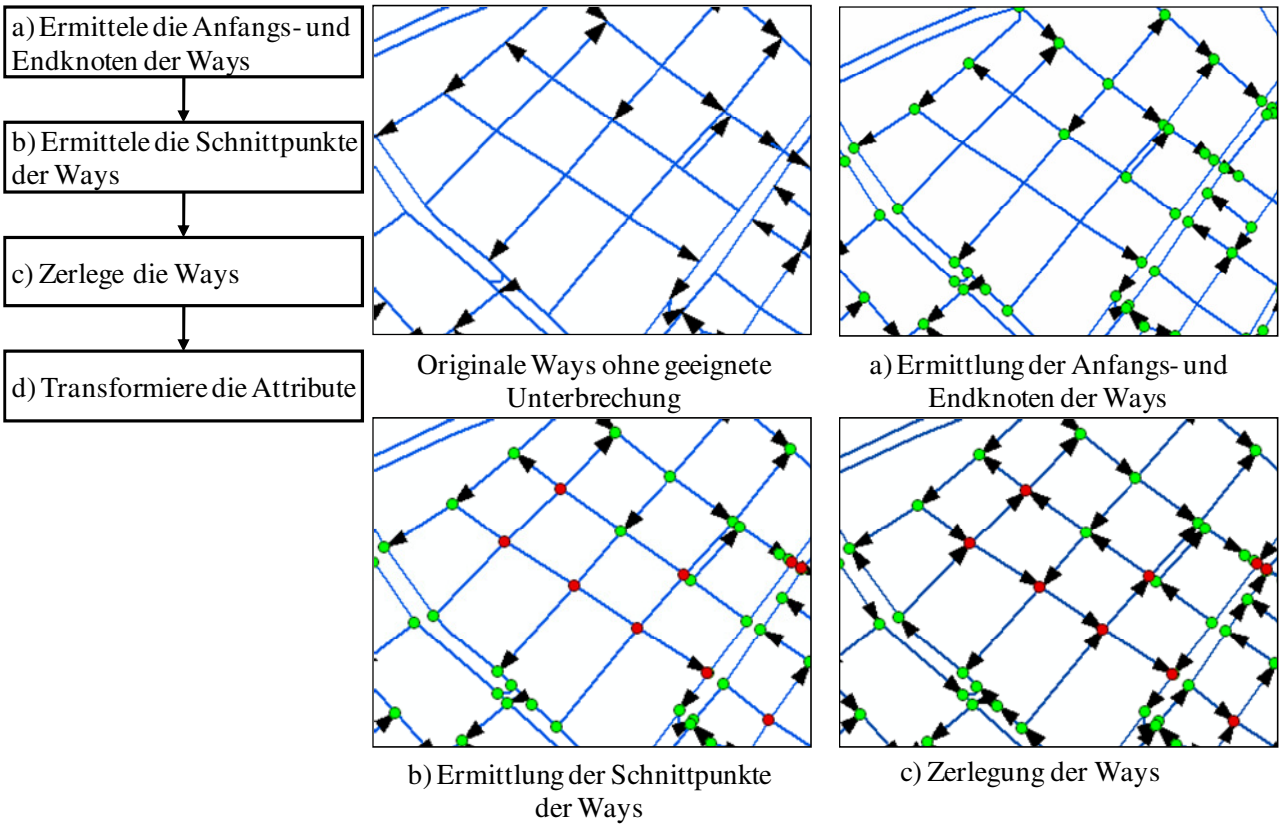


Abbildung 4.11: Beseitigung der topologischen Probleme in OSM

Die Regeln zur Abbildung der OSM-*Ways* und -*Nodes* ins globale Datenmodell werden in Tabelle 4.6 zusammenfasst. Die *Ways* können in eine bzw. mehrere Kanten zerlegt werden. Vorhandene Anfangs- und Endknoten von *Ways* werden als Knoten aufgenommen. Darüber hinaus werden die berechneten Schnittpunkte ebenfalls als Knoten hinzugefügt.

OpenStreetMap	Relation	Globales Datenmodell
Way	1:1, 1:n	*Kanten*
Node	1:1, *:1	*Knoten*

Tabelle 4.6: Transformationsregeln für OSM

5 Zuordnung

Dieses Kapitel widmet sich der Bestimmung von Korrespondenzen in den verschiedenen Datensätzen, um Ausgangsdaten für die nachfolgende Qualitätsanalyse und Datenverschmelzung bereitzustellen. Zunächst wird auf die manuelle Bestimmung von Kantenkorrespondenzen eingegangen. Anschließend wird die Form der Zuordnungspaare nach der Topologie klassifiziert und automatisch berechnet. Zum Schluss werden Knoten in den verschiedenen Datensätzen automatisch zugeordnet.

5.1 Ansatz der Zuordnung

In der vorliegenden Arbeit wird das bereits im Kapitel 3.2.1 beschriebene Zuordnungsmodell „Buffer Growing“ verwendet. Mit diesem Zuordnungsmodell sind lediglich Zuordnungen zwischen Kanten möglich. Unterschiedliche geometrische Modellierungen in den verschiedenen Datensätzen führen zu problematischen Situationen bei der Zuordnung. Beispielsweise lässt sich die Kante B_3 in GDF mit diesem Zuordnungsmodell nicht zuordnen (siehe Abbildung 5.1).

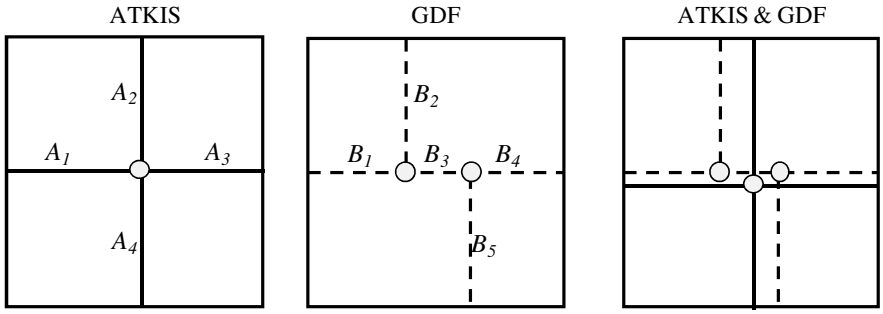


Abbildung 5.1: Problematische Situation der Zuordnung (aus [Walter 1997])

Um eine Zuordnung für die oben genannte Situation zu ermöglichen und die Vollständigkeit sowie die Genauigkeit der Zuordnung zu verbessern, wird das Zuordnungsmodell „Buffer Growing“ so erweitert, dass nicht nur Zuordnungen zwischen Kanten, sondern auch Zuordnungen zwischen Knoten und Kanten erlaubt sind. Auf diese Weise können Unterschiede der geometrischen Modellierungen bei der Zuordnung berücksichtigt werden. Weiterhin können alle möglichen Kantenkandidaten vollständig erfasst werden. In Abbildung 5.2 links wird der Knoten a_1 im Datensatz A zu der Kante B_1 im Datensatz B zugeordnet. Der Knoten a_1 im Datensatz A in Abbildung 5.2 rechts wird zu Kanten B_1, B_2, B_3 und B_4 im Datensatz B zugeordnet.

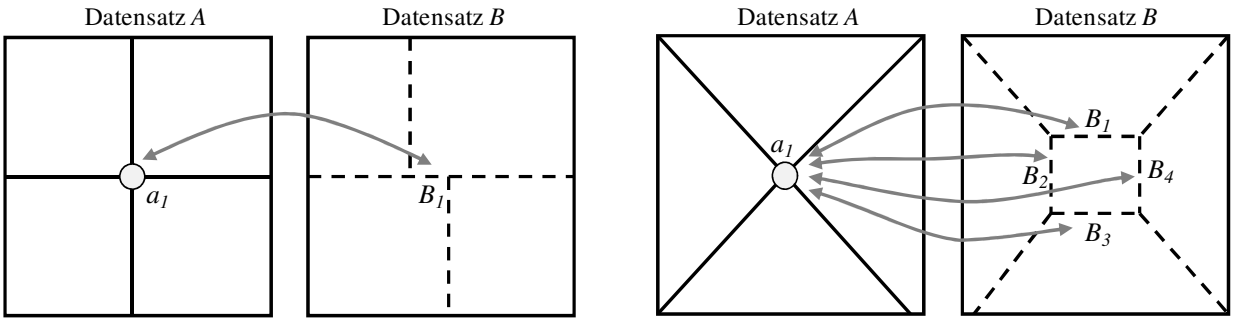


Abbildung 5.2: Zuordnung zwischen einem Knoten und einer Kante (links) und zwischen einem Knoten und mehreren Kanten (rechts)

Zusammenfassend werden die Zuordnungsrelationen des erweiterten Zuordnungsmodells in acht Gruppen untergliedert. Den Zuordnungspaaren zwischen Kanten wird eine der folgenden Relationen *1:1*, *1:n*, *n:1* und *n:m* zugewiesen. Die Zuordnungen zwischen Knoten und Kanten lassen sich mit den Relationen *p:1*, *p:n*, *1:p* und *n:p* modellieren.

5.2 Manuelle Zuordnung

Im Folgenden wird zunächst auf die ausgewählten Testgebiete eingegangen. Anschließend wird ein Werkzeug für die manuelle Zuordnung zur Bestimmung der Korrespondenzen von Kanten vorgestellt. Zum Schluss werden die Ergebnisse der manuellen Zuordnung zusammengefasst und anhand von Beispielen diskutiert.

5.2.1 Beschreibung der Testgebiete

In der vorliegenden Arbeit werden zwei Testgebiete in Süddeutschland für die Untersuchung verwendet. Das Testgebiet I befindet sich in der Innenstadt von Stuttgart und das Testgebiet II liegt im ländlichen Raum in Öhringen. In beiden Testgebieten stehen Straßendaten jeweils von NavTeq (NT), TeleAtlas (TA) und OpenStreetMap (OSM) zur Verfügung.

In Tabelle 5.1 sind Informationen über die zwei Testgebiete und Versionen der drei Datensätze dargestellt. Außerdem wird die Anzahl der Objekte vor und nach der Transformation aufgelistet. Graphische Übersichten der Testgebiete finden sich in Anhang B.

- NavTeq: Nach der Transformation werden 1510 *Streets* in 1291 Kanten im Testgebiet I und 819 *Streets* in 690 Kanten im Testgebiet II gruppiert. Darüber hinaus werden 946 Knoten von 3705 *Z_Levels* im Testgebiet I und 582 Knoten von 3711 *Z_Levels* im Testgebiet II berechnet.

- TeleAtlas: Die Anzahl der Objekte in TeleAtlas bleibt nach der Abbildung ins globale Datenmodell unverändert. So stehen 1708 Kanten und 1243 Knoten im Testgebiet I sowie 1991 Kanten und 1537 Knoten im Testgebiet II nach der Transformation zur Verfügung.

- OpenStreetMap: Insgesamt werden 2264 Kanten aus 987 *Ways* im Testgebiet I und 340 Kanten aus 173 *Ways* im Testgebiet II erzeugt. Außerdem werden 273 Schnittpunkte im Testgebiet I und 92 Schnittpunkte im Testgebiet II als Knoten aufgenommen.

	NavTeq (NT) (Q1/2005)		**TeleAtlas (TA) (Q1/2006)**		**OpenStreetMap (OSM) (10/2008)**	
	Kanten	Knoten	Kanten	Knoten	Kanten	Knoten
Testgebiet I (2*2 km)	1291 (1510)	946 (3705)	1708	1243	2264 (987)	1526 (1253)
Testgebiet II (5*6 km)	690 (819)	582 (3711)	1991	1537	340 (173)	325 (233)

Tabelle 5.1: Informationen über die zwei Testgebiete

5.2.2 Werkzeug für manuelle Zuordnung

Wie bereits angedeutet, kann die Zuordnung von Kanten auch mit einem automatischen Ansatz durchgeführt werden. Allerdings liefern automatische Zuordnungsansätze an manchen Stellen

ungenaue Ergebnisse, die wiederum zu einer Ungenauigkeit bei der Qualitätsanalyse und der Datenverschmelzung führen können. Um solche Probleme zu vermeiden und den Fokus der Arbeit auf die Aufgaben nach der Zuordnung zu legen, wird die Zuordnung manuell durchgeführt.

Für die manuelle Zuordnung wurde ein Werkzeug unter *ArcGIS* und *Visual Basic Application* (VBA) entwickelt. Die Bedienoberfläche des Werkzeugs ist in Abbildung 5.3 veranschaulicht. Eine Zuordnung erfolgt durch Auswahl von korrespondierenden Kanten bzw. Knoten in der Karte. Die IDs der selektierten Kanten oder Knoten werden dann in zwei Listen (*Source ID* und *Target ID*) eingetragen. Anschließend wird eine manuelle Bewertung für das Zuordnungspaar vom Operator vergeben. Eine Anmerkung zu einem Zuordnungspaar für zukünftige Auswertungen ist möglich. Ergebnisse der Zuordnung werden in einer Zuordnungstabelle gespeichert.

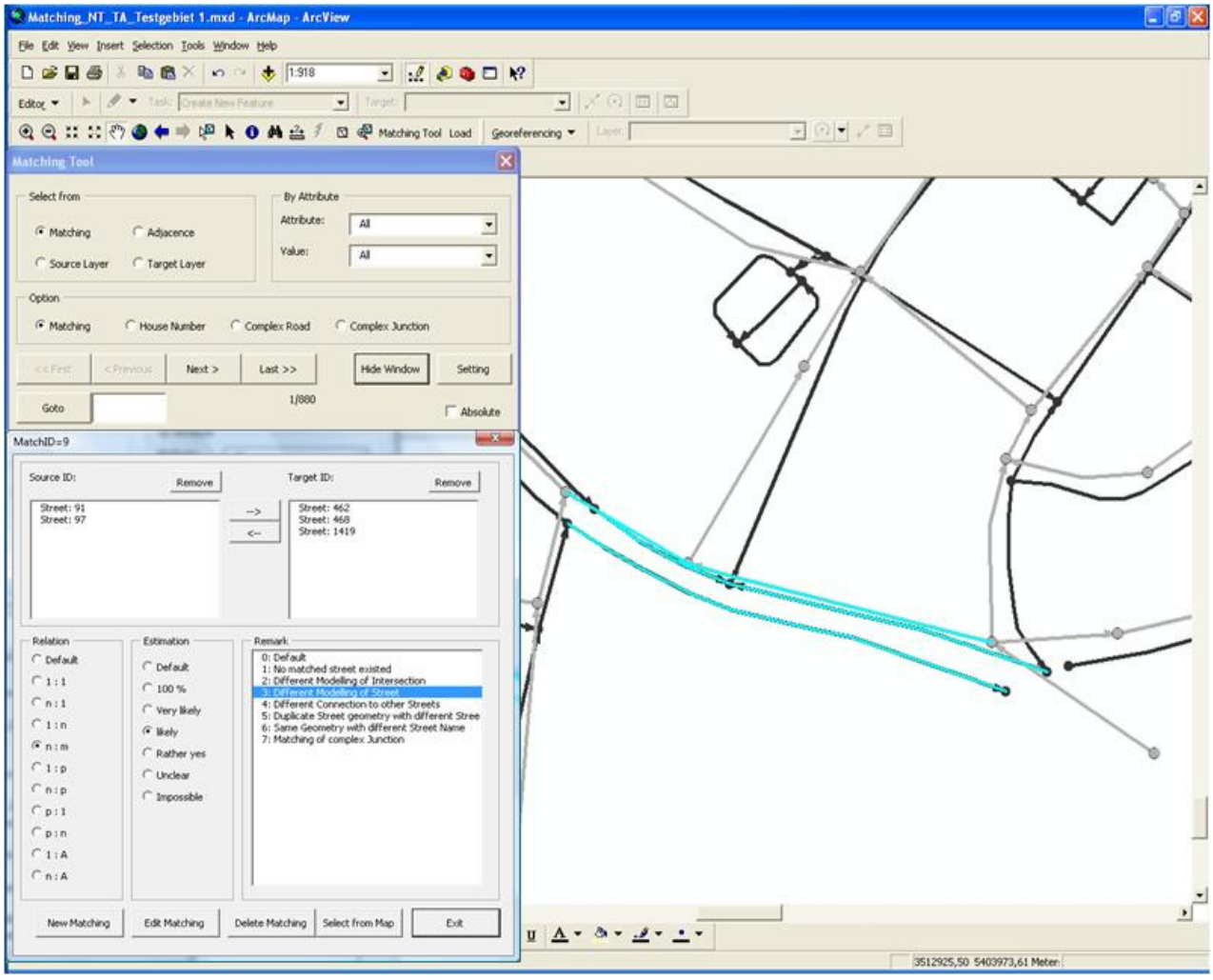


Abbildung 5.3: Werkzeug für die manuelle Zuordnung

5.2.3 Ergebnisse der Zuordnung

Tabelle 5.2 stellt die Zuordnungsergebnisse zwischen NavTeq und TeleAtlas in den zwei Testgebieten dar. Es wird deutlich, dass Unterschiede bezüglich der geometrischen Modellierung in den zwei Datensätzen existieren. In Tabelle 5.2 wird nicht nur die Anzahl der Zuordnungspaare sondern auch die Anzahl der betroffenen Kanten in den zwei Datensätzen aufgelistet. Dies ermöglicht einen ersten Einblick in die unterschiedliche geometrische Modellierung. Insgesamt werden 880 Zuordnungspaare im Testgebiet I und 588 Zuordnungspaare im Testgebiet II manuell

gebildet. Davon werden mehr als 50% der Zuordnungspaare in beiden Testgebieten mit der Relation *1:1* zugeordnet. Außerdem finden 5,2% der Zuordnungen im Testgebiet I und 3,7% der Zuordnungen im Testgebiet II zwischen Knoten und Kanten statt.

Relation	**Testgebiet I**			**Testgebiet II**		
(NT&TA)	Zuordnungspaare	NavTeq	TeleAtlas	Zuordnungspaare	NavTeq	TeleAtlas
1:1	477 (54,2%)	477	477	352 (59,9%)	352	352
n:1	70 (8,0%)	142	70	20 (3,4%)	44	20
1:n	160 (18,2%)	160	377	130 (22,1%)	130	374
n:m	127 (14,4%)	309	370	64 (10,9%)	142	205
1:p	12 (1,4%)	12	-	6 (1,0%)	6	-
n:p	3 (0,3%)	10	-	-	-	-
p:1	28 (3,2%)	-	28	16 (2,7%)	-	16
p:n	3 (0,3%)	-	6	-	-	-
*1:**	-	181	-	-	16	-
**:1*	-	-	380	-	-	1024
Gesamt	880	1291	1708	588	690	1991

Tabelle 5.2: Zuordnungsergebnisse zwischen NavTeq und TeleAtlas

Die Zuordnungsergebnisse zwischen TeleAtlas und OpenStreetMap sind aus Tabelle 5.3 zu entnehmen. Insgesamt werden 874 Zuordnungspaare im Testgebiet I und 174 Zuordnungspaare im Testgebiet II erzeugt. Hiervon werden 4,8% der Zuordnungen im Testgebiet I und 5,7% der Zuordnungen im Testgebiet II zwischen Knoten und Kanten durchgeführt. Im Vergleich zu Tabelle 5.2 werden nur 30,5% der Zuordnungspaare im Testgebiet II mit der Relation *1:1* gebildet. Die Kanten im Testgebiet II in OpenStreetMap sind häufig länger als in TeleAtlas. Aus diesem Grund werden 36,2% der Zuordnungspaare im Testgebiet II mit der Relation *n:1* erfasst.

Relation	**Testgebiet I**			**Testgebiet II**		
(TA&OSM)	Zuordnungspaare	TeleAtlas	OSM	Zuordnungspaare	TeleAtlas	OSM
1:1	410 (47,0%)	410	410	53 (30,5%)	53	53
n:1	135 (15,4%)	333	135	63 (36,2%)	277	63
1:n	149 (17,0%)	149	352	9 (5,2%)	12	21
n:m	138 (15,8%)	394	418	39 (22,4%)	154	116
1:p	21 (2,4%)	21	-	4 (2,3%)	4	0
n:p	7 (0,8%)	15	-	-	-	0
p:1	13(1,5%)	-	13	3 (1,7%)	-	3
p:n	1 (0,1%)	-	4	3 (1,7%)	-	12
*1:**	-	386	-	-	1491	-
**:1*	-	-	932	-	-	72
Gesamt	874	1708	2264	174	1991	340

Tabelle 5.3: Zuordnungsergebnisse zwischen TeleAtlas und OpenStreetMap

Bei der Zuordnung werden die Zuordnungspaare nach Gesichtspunkten der topologischen Ähnlichkeit, geometrischen Ähnlichkeit und thematischen Ähnlichkeit manuell bewertet. In Anhang C werden die Kriterien und die Ergebnisse der manuellen Bewertung der Zuordnungspaare dargestellt. Die Ähnlichkeit von NavTeq und TeleAtlas ist höher als die Ähnlichkeit von TeleAtlas und OpenStreetMap. Die Ähnlichkeit im Stadtgebiet (Testgebiet I) ist offenbar höher als im ländlichen Raum (Testgebiet II).

Im Folgenden werden die Ergebnisse der Zuordnung anhand von Beispielen diskutiert und Ursachen für die unterschiedliche geometrische Modellierung untersucht.

Abbildung 5.4 zeigt drei Zuordnungen zwischen Knoten und Kanten (mit der Relation *p:1* bzw. *1:p*). Nur die betroffenen Kanten und Knoten im Zuordnungspaar sind beschriftet. Abhängig von den Operatoren und der Genauigkeit der Datenerfassung, können kurze Kanten (Straßenabschnitte mit einer Länge kleiner als zehn Meter) als Knoten oder als Kanten erfasst werden (siehe Abbildung 5.4a). Komplexe Kreuzungen werden von verschiedenen Operatoren unterschiedlich erfasst. Unterschiedliche geometrische Modellierungen in der Einmündung sind in Abbildung 5.4b zu sehen. In manchen Bereichen könnten physikalische Trennungen der Straße dynamisch hinzugefügt bzw. entfernt werden. Dies führt ebenfalls zu unterschiedlichen Modellierungen für diese Straße zu unterschiedlichen Zeitpunkten (siehe Abbildung 5.4c).

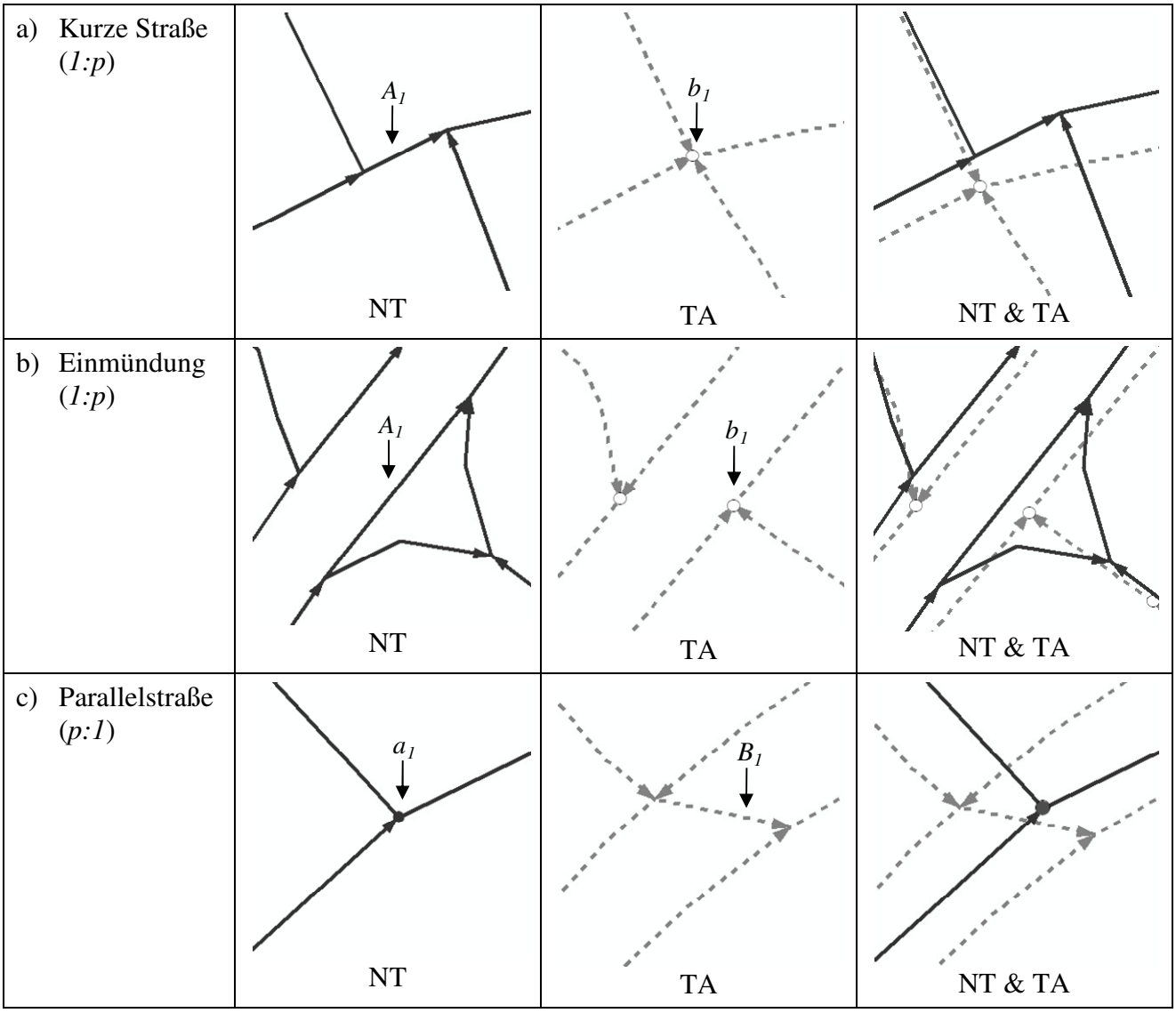


Abbildung 5.4: Zuordnung zwischen Knoten und Kanten (Relation *p:1* oder *1:p*)

Beispiele für die Zuordnung mit der Relation *p:n* oder *n:p* sind in Abbildung 5.5 dargestellt. Zuordnungen dieser Art und Weise erlauben einen großen geometrischen Unterschied in den Datensätzen. Ein undefinierter Platz kann ebenfalls von Operatoren unterschiedlich digitalisiert werden. Der Platz in Abbildung 5.5a ist mit vier Kanten (A_1, A_2, A_3, A_4) in NavTeq modelliert, während er in TeleAtlas mit nur einem Knoten b_1 erfasst wird. Abbildung 5.5b stellt die

Modellierungen einer Kreuzung da. In NavTeq wird diese als drei Kanten (A_1, A_2, A_3) und in TeleAtlas als ein Knoten b_1 modelliert. Abbildung 5.5c veranschaulicht die unterschiedlichen Modellierungen eines Kreuzungsinnenbereichs. Der Knoten a_1 in NavTeq wird zwei Kanten (B_1, B_2) in TeleAtlas zugeordnet. Die Kreuzung in Abbildung 5.5d wird in OpenStreetMap sehr komplex digitalisiert (B_1, B_2, B_3, B_4, B_5), während sie in TeleAtlas mit nur einem Knoten a_1 modelliert ist. Es handelt sich evtl. um einen Digitalisierungsfehler in OpenStreetMap, weil die Kanten B_1, B_2, B_3, B_4 und B_5 nur zwei bis fünf Meter lang sind.

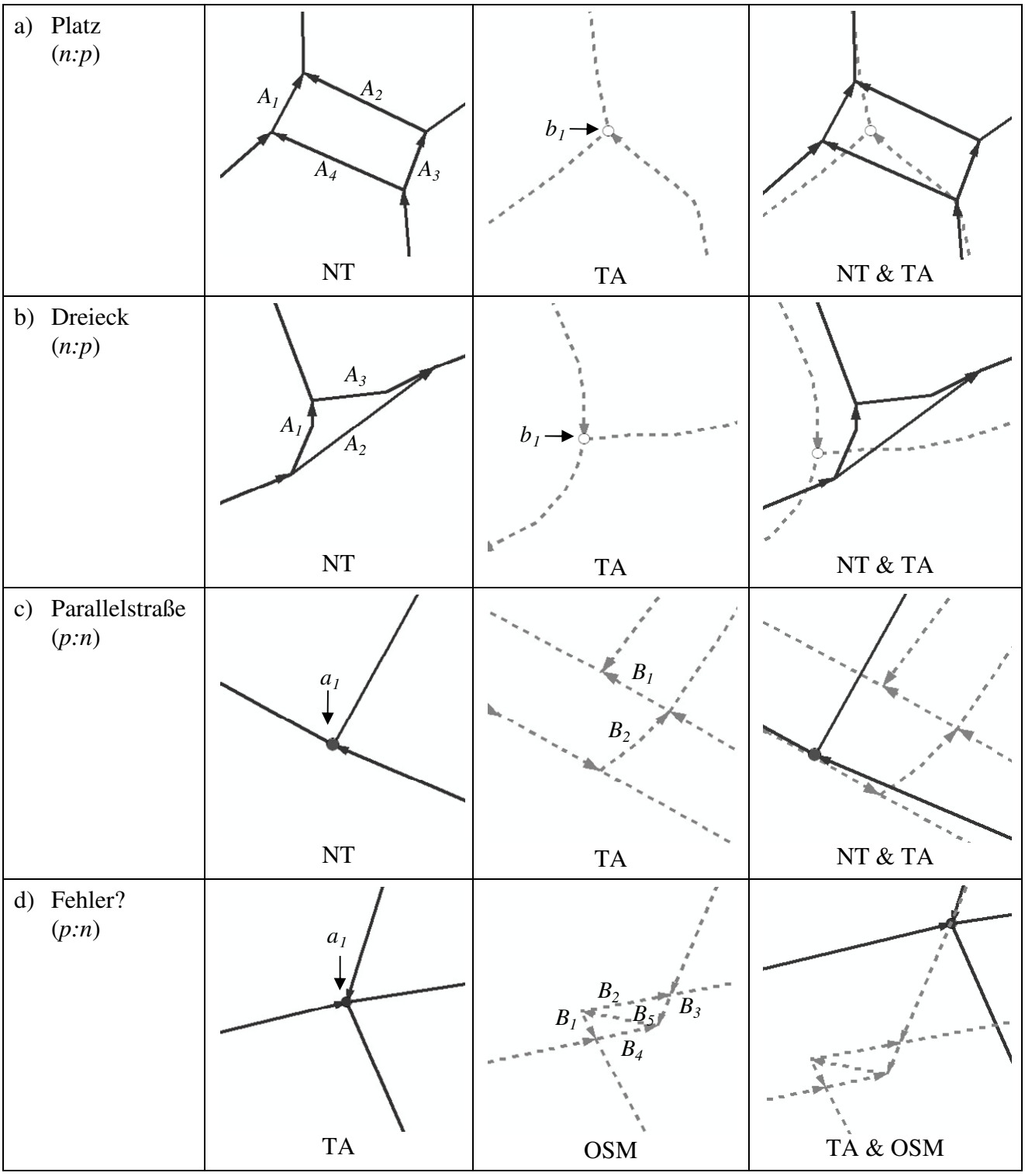


Abbildung 5.5: Zuordnung zwischen Knoten und Kanten (Relation *p:n* oder *n:p*)

Trotz der Verwendung des komplexen Zuordnungsmodells ist es an manchen Stellen schwierig, die Unterschiede der Modellierung in den verschiedenen Datensätzen zu behandeln. Abbildung 5.6 stellt eine solche kritische Stelle dar. Bei der manuellen Zuordnung wird die Kante A_2 in TeleAtlas zu der Kante B_2 in OpenStreetMap zugeordnet (siehe Abbildung 5.6a). Weiterhin wird bei der manuellen Zuordnung eine Anmerkung zu diesem Zuordnungspaar gemacht, damit es bei der weiteren Verarbeitung als eine Zuordnung zwischen einem Komplexknoten und einem anderen Komplexknoten (*p:p*) behandelt wird (siehe Abbildung 5.6b). So erhält man eine ähnliche Modellierung der zwei Datensätze.

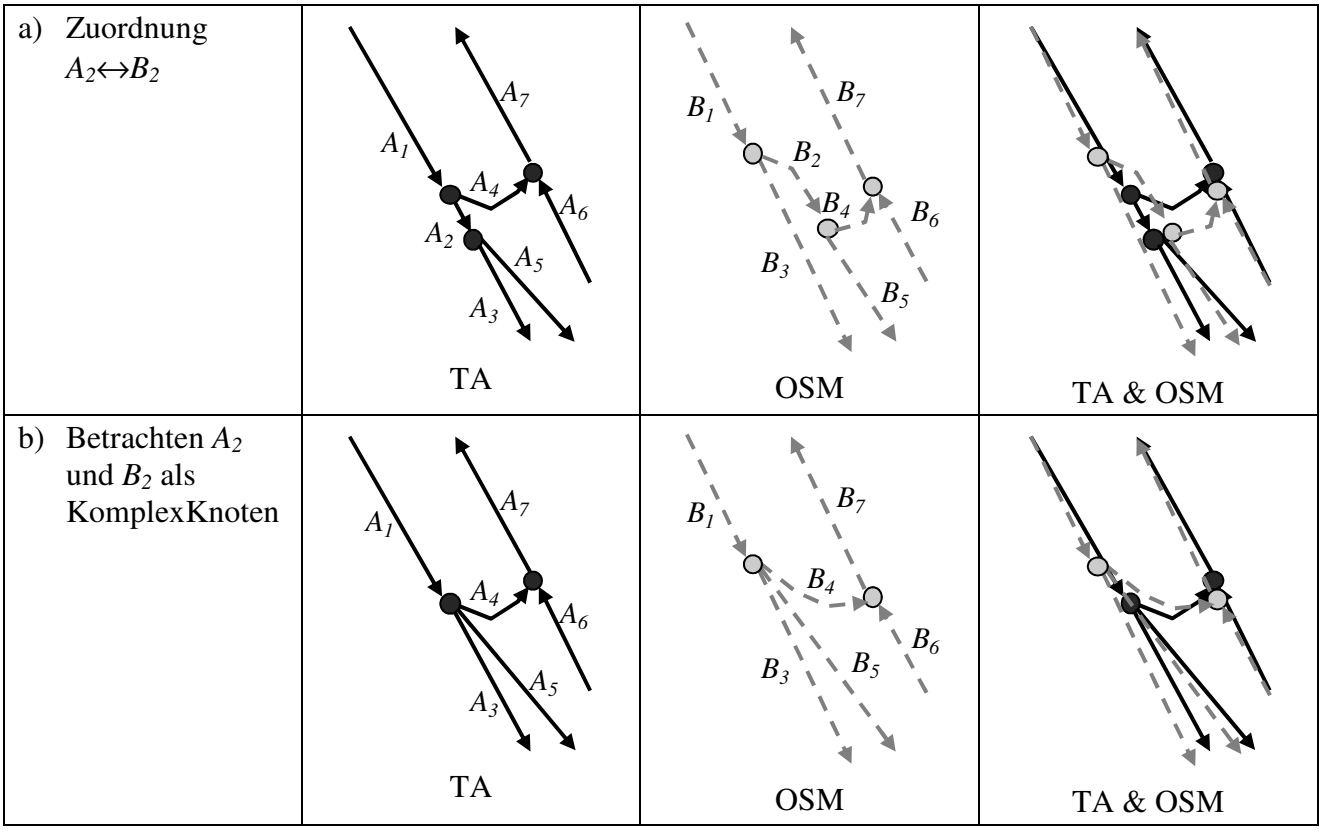


Abbildung 5.6: Kritische Stelle bei der manuellen Zuordnung

5.3 Formzuordnung

Unterschiede der geometrischen Modellierung in den verschiedenen Datensätzen werden mit dem Zuordnungsmodell erfasst und im Folgenden durch Berechnung und Zuordnung der Form der Zuordnungspaare interpretiert. Die Formerkennung ist ebenfalls ein notwendiger Prozess für die Zuordnung von korrespondierenden Knoten und die Verschmelzung der Daten.

5.3.1 Ansatz der Formerkennung

Für die Formerkennung werden die Formen nach der Topologie klassifiziert. Es werden acht grundlegende Formklassen definiert, die in Abbildung 5.7 dargestellt sind. Darüber hinaus werden die Formklassen „Mix“ und „Point“ als zusätzliche Kategorien eingeführt, um alle möglichen Formen behandeln zu können. Die Klasse „Mix“ besteht aus mindestens zwei grundlegenden Formklassen. Da beliebige Kombinationen von Formklassen vorkommen können, wird die Klasse

„Mix“ nicht mehr weiter unterteilt. Aufgrund der ähnlichen Datenmodellierung in den Datensätzen ist zu erwarten, dass die Klasse „Mix“ nicht häufig vorkommt.

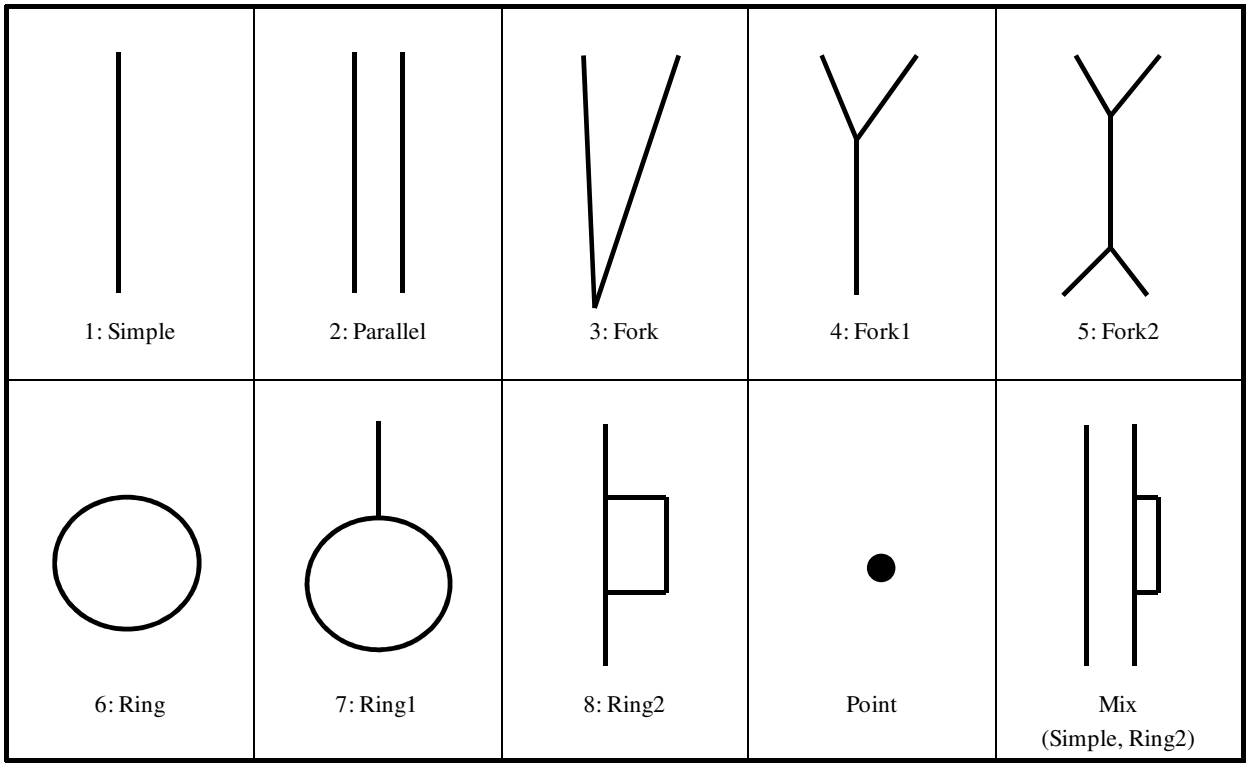


Abbildung 5.7: Klassen der Form

Zur Bestimmung der Formklassen werden die Kanten eines Zuordnungspaares von jedem Datensatz zunächst in ein Netzwerk konvertiert, welches in der vorliegenden Arbeit als Mini-Netzwerk bezeichnet wird. Im Anschluss daran ist der Grad der Knoten zu ermitteln, die im Mini-Netzwerk enthalten sind. Anhand des Knotengrads werden die Knoten wie folgt klassifiziert:

- *Anfangs-* oder *Endknoten*: Knotengrad ist gleich eins.
- *Zwischenpunkt*: Knotengrad ist gleich zwei.
- *Zwischenknoten*: Knotengrad ist größer als zwei.

Im nächsten Schnitt wird das Mini-Netzwerk in unterschiedliche Abschnitte aufgeteilt. Aufgrund der Existenz von Zwischenpunkten sind Verbindungen zwischen den Anfangs-, End- und Zwischenknoten nicht immer direkt ermittelbar. Diese Verbindungen werden mit dem Floyd-Algorithmus berechnet, welcher kürzeste Wege zwischen allen Knoten ermittelt [Sedgewick 1995]. Verbindungen zwischen den Anfangs-, End- und Zwischenknoten im Mini-Netzwerk werden in vier Typen von Abschnitten untergliedert:

- *Begin*: Abschnitt von einem *Anfangsknoten* zu einem *Zwischenknoten*
- *Middle*: Abschnitt von einem *Zwischenknoten* zu anderem *Zwischenknoten*
- *End*: Abschnitt von einem *Zwischenknoten* zu einem *Endknoten*
- *Whole*: Abschnitt von einem *Anfangsknoten* zu einem *Endknoten*

Anhand der unterschiedlichen Knoten und Abschnittstypen lassen sich die Formklassen in den Zuordnungspaaren ermitteln:

- „Simple“: 1 *Whole*
- „Parallel“: 2 *Whole*
- „Fork“: 2 *Begin* bzw. 2 *End*
- „Fork1“: 2 *Begin* und 1 *End* bzw. 2 *End* und 1 *Begin*
- „Fork2“: 2 *Begin*, 1 *Middle* und 2 *End*
- „Ring“: n *Zwischenpunkte*
- „Ring1“: 1 *Anfangsknoten* und 1 *Zwischenknoten*
- „Ring2“: 1 *Begin*, 2 *Middle* und 1 *End*
- „Point“: 1 *Knoten*
- „Mix“: sonstig

Abbildung 5.8 zeigt ein Beispiel für die Formerkennung. Aus sieben Kanten eines Zuordnungspaars im Datensatz *A* wird ein Mini-Netzwerk gebildet. Nach der Berechnung des Knotengrads werden *Anfangsknoten* (a_1, a_6), *Endknoten* (a_5, a_8), *Zwischenknoten* (a_3, a_4) und *Zwischenpunkte* (a_2, a_7) ermittelt. Nach der Berechnung von möglichen Verbindungen zwischen den Knoten (a_1, a_3, a_4, a_5, a_6, a_8) lässt sich das Mini-Netzwerk in fünf Abschnitte (1 *Begin*, 2 *Middle*, 1 *End* und 1 *Whole*) einteilen. Daraus ergeben sich eine Formklasse „Ring2“ (1 *Begin*, 2 *Middle*, 1 *End*) und eine Formklasse „Simple“ (1 *Whole*). Die Formklasse des Mini-Netzwerks im Datensatz *A* entspricht dann der Klasse „Mix“.

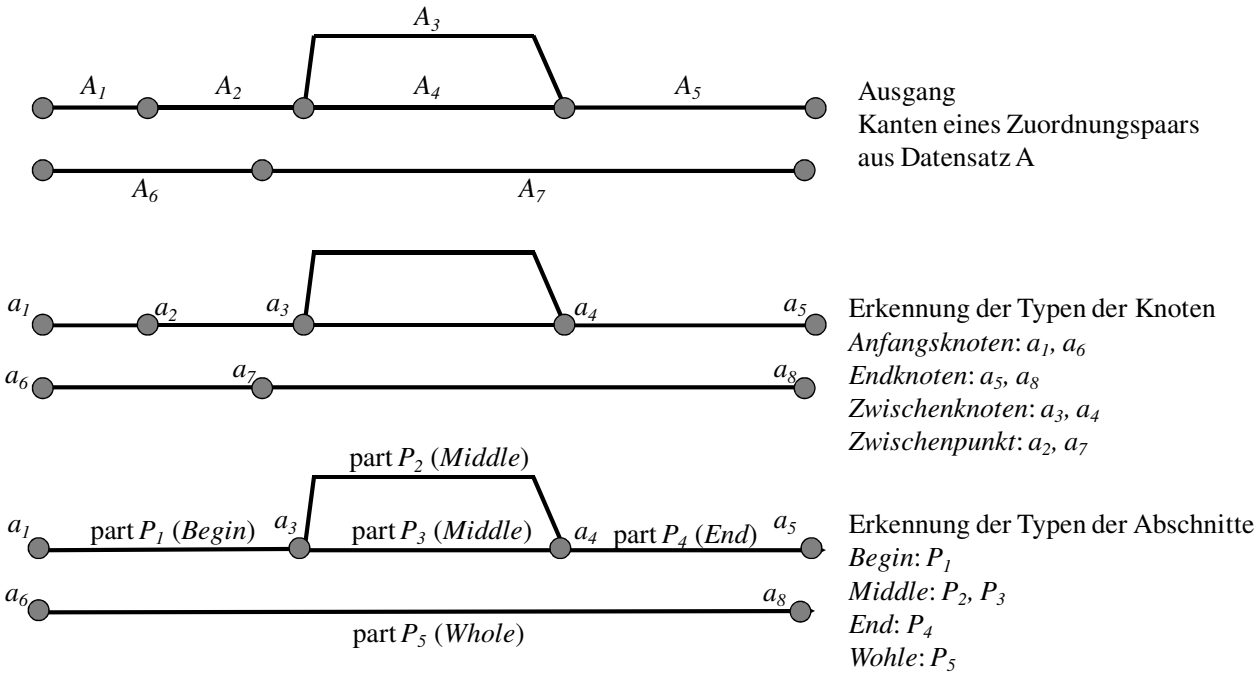


Abbildung 5.8: Beispiel für die Formerkennung

5.3.2 Ergebnisse der Formzuordnung

In diesem Abschnitt werden die Ergebnisse der Formerkennung und Formzuordnung zusammengefasst. Die Formzuordnung erfolgt durch die Berechnung und Gegenüberstellung von Formklassen der Zuordnungspaare. Abbildung 5.9 zeigt ein Beispiel für die Formzuordnung zwischen „Fork1“ mit drei Kanten (A_1, A_2, A_3) in TeleAtlas und „Simple“ mit einer Kante (B_1) in OpenStreetMap.

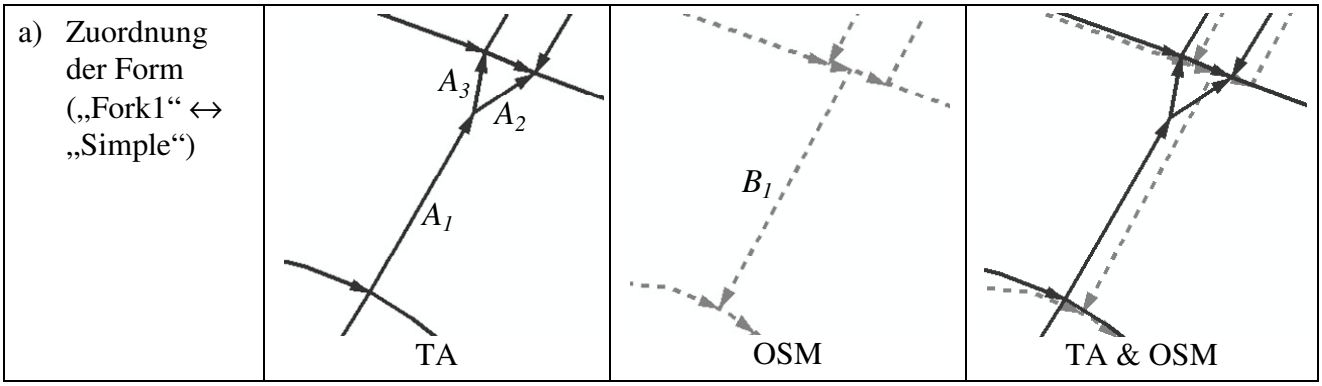


Abbildung 5.9: Beispiel für Formzuordnung zwischen TeleAtlas und OpenStreetMap

In Tabelle 5.4 werden die Ergebnisse der Formzuordnung zwischen NavTeq und TeleAtlas im Testgebiet I zusammengefasst. Ergebnisse im Testgebiet II sind aus Tabelle 5.5 zu entnehmen. Im Testgebiet I sind alle Formklassen zu finden. 90,8% der Formzuordnungen im Testgebiet I und 95,1% der Formzuordnung im Testgebiet II finden zwischen „Simple“ und „Simple“ statt. Die Formklasse „Mix“ kommt im Testgebiet I in TeleAtlas nur mit einer geringen Häufigkeit (0,1%) vor. Im Testgebiet II wird eine Formzuordnung zwischen „Parallel“ und „Fork1“ gefunden. Insgesamt gibt es nur 0,4% Formzuordnungen im Testgebiet I und 0,2% im Testgebiet II, welche die Formklasse „Simple“ weder in NavTeq noch in TeleAtlas aufweisen.

	Point (TA)	*Simple (TA)*	*Parallel (TA)*	*Fork0 (TA)*	*Fork1 (TA)*	*Fork2 (TA)*	*Ring0 (TA)*	*Ring1 (TA)*	*Mix (TA)*
Point (NT)	-	30 (3,5%)	-	1 (0,1%)	-	-	3 (0,3%)	-	-
Simple (NT)	12 (1,4%)	799 (90,8%)	12 (1,4%)	-	12 (1,4%)	1 (0,1%)	1 (0,1%)	1 (0,1%)	1 (0,1%)
Parallel (NT)	-	2 (0,2%)	-	-	-	-	-	-	-
Fork0(NT)	-	3 (0,3%)	-	-	-	-	-	-	-
Fork1(NT)	-	1 (0,1%)	-	-	-	-	-	-	-
Ring0 (NT)	-	3 (0,3%)	-	-	-	-	-	-	-
Ring2 (NT)	-	1 (0,1%)	-	-	-	-	-	-	-

Tabelle 5.4: Formzuordnung zwischen NavTeq und TeleAtlas im Testgebiet I

	Point (TA)	*Simple (TA)*	*Parallel (TA)*	*Fork1 (TA)*
Point (NT)	-	16 (2,7%)	-	-
Simple (NT)	6 (1,0%)	559 (95,1%)	2 (0,3%)	2 (0,3%)
Parallel (NT)	-	-	-	1 (0,2%)
Fork0 (NT)	-	1 (0,2%)	-	-
Fork1 (NT)	-	1 (0,2%)	-	-

Tabelle 5.5: Formzuordnung zwischen NavTeq und TeleAtlas im Testgebiet II

In Tabelle 5.6 und Tabelle 5.7 sind die Ergebnisse der Formzuordnung zwischen TeleAtlas und OpenStreetMap im Testgebiet I und Testgebiet II dargestellt. Die Klasse „Mix“ tritt im Testgebiet I mit einer Häufigkeit 0,1% auf. 92,0% der Formklassen im Testgebiet I und 90,2% im Testgebiet II befinden sich in beiden Datensätzen vom Typ „Simple“. 99,8% der Formzuordnungen im Testgebiet I und 96,6% der Formzuordnungen im Testgebiet II enthalten mindestens die Formklasse „Simple“ in TeleAtlas oder in OpenStreetMap.

	Point (OSM)	*Simple (OSM)*	*Parallel (OSM)*	*Fork0 (OSM)*	*Fork1 (OSM)*	*Ring1 (OSM)*
Point (TA)	-	13 (1,5%)	-	-	-	1 (0,1%)
Simple (TA)	28 (3,2%)	803 (92,0%)	-	1 (0,1%)	3 (0,3%)	-
Parallel (TA)	-	8 (0,9%)	-	-	-	-
Fork0 (TA)	-	1 (0,1%)	-	-	-	-
Fork1 (TA)	-	12 (1,4%)	1 (0,1%)	-	-	-
Fork2 (TA)	-	1 (0,1%)	-	-	-	-
Ring1 (TA)	-	1 (0,1%)	-	-	-	-
Mix (TA)	-	1 (0,1%)	-	-	-	-

Tabelle 5.6: Formzuordnung zwischen TeleAtlas und OpenStreetMap im Testgebiet I

	Point (OSM)	*Simple (OSM)*	*Parallel (OSM)*	*Fork1 (OSM)*	*Ring0 (OSM)*	*Mix (OSM)*
Point (TA)	-	3 (1,7%)	-	-	2 (1,1%)	1 (0,6%)
Simple (TA)	4 (2,4%)	157 (90,2%)	-	1 (0,6%)	-	-
Parallel (TA)	-	1 (0,6%)	-	-	-	-
Fork1 (TA)	-	2 (1,1%)	3 (1,7%)	-	-	-

Tabelle 5.7: Formzuordnung zwischen TeleAtlas und OpenStreetMap im Testgebiet II

Zusammenfassend ist festzuhalten, dass die geometrische Modellierung bei der Datenerfassung trotz der Verwendung des gleichen bzw. ähnlichen Datenmodells zu einem gewissen Grad unsicher ist. Die Ergebnisse zeigen, dass es mehr Formklassen im Stadtgebiet (Testgebiet I) im Vergleich zum ländlichen Raum (Testgebiet II) gibt. Es wird festgestellt, dass die Datenerfassung von komplexen Kreuzungen im Stadtgebiet häufig unterschiedlich ist. Aus diesem Grund werden mehr Formklassen verwendet, um die unterschiedlichen Datenerfassungen zu interpretieren.

5.4 Automatische Knotenzuordnung

Letztendlich sind Korrespondenzen zwischen Knoten mit Hilfe der Ergebnisse der Form- und Kantenzuordnung zu bestimmen. Analog zur Kantenzuordnung lässt sich die Knotenzuordnung mit den Relationen (*1:1*, *1:n*, *n:1* und *n:m*) erfassen. In der Folge werden die Korrespondenzen von Knoten mit den Relationen *1:1* und *1:n* in zwei Schritten bestimmt (siehe Abbildung 5.10). Die Ermittlung von *n:m* Knotenzuordnungen wird in Kapitel 6.1.1 diskutiert. Im ersten Schritt werden korrespondierende Knoten im Datensatz *B* für Knoten im Datensatz *A* gesucht. Im nächsten Schritt sind Korrespondenzen im Datensatz *A* für Knoten im Datensatz *B* mit dem gleichen Verfahren zu ermitteln.

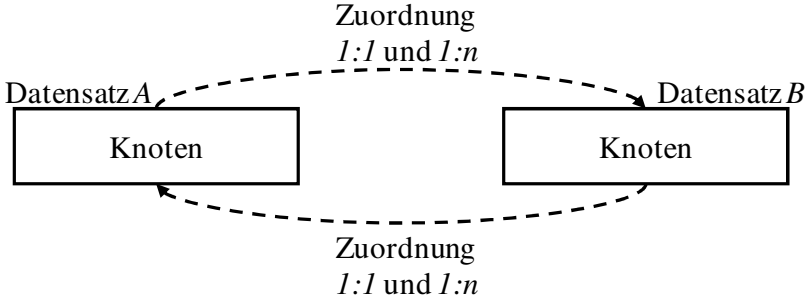


Abbildung 5.10: Zwei Schritte der Knotenzuordnung

Für jeden Knoten im Datensatz *A* und im Datensatz *B* werden die Korrespondenzen berechnet. In Abbildung 5.11 wird ein Beispiel für die Zuordnung von Knoten dargestellt. Als Ergebnis der Knotenzuordnung werden zwei Listen erzeugt. Dabei sind drei Fälle zu unterscheiden:

1. Ein Knoten wird bei der manuellen Zuordnung einer bzw. mehreren Kanten zugeordnet (Relation *p:1* oder *p:n*). Dieser Knoten wird dann zu allen Knoten des Zuordnungspaars im anderen Datensatz zugeordnet. Zum Beispiel wird der Knoten a_2 im Datensatz *A* zu der Kante B_3 im Datensatz *B* zugeordnet (Relation *p:1*). So wird a_2 bei der Knotenzuordnung zu allen Knoten des Zuordnungspaars im Datensatz *B* (b_3, b_4) zugeordnet.

2. Ein Knoten ist ein Anfangs- bzw. Endknoten von Zuordnungspaaren. In diesem Fall werden alle Zuordnungspaare selektiert, welche diesen Knoten enthalten. Weiterhin werden die Anfangs- bzw. Endknoten dieser Zuordnungspaare einander zugeordnet. Beispielsweise werden die Zuordnungspaare ($A_3A_4 \leftrightarrow B_5B_6$, $A_5 \leftrightarrow B_{10}$, $A_6A_7 \leftrightarrow B_7B_8B_9$) zur Berechnung der Korrespondenzen für den Knoten b_5 selektiert. So wird b_5 zu den Knoten a_4 und a_9 zugeordnet. Der Knoten b_6 wird dem Knoten a_4 zugeordnet, da nur die Zuordnungspaare ($A_3A_4 \leftrightarrow B_5B_6$, $A_6A_7 \leftrightarrow B_7B_8B_9$) selektiert werden.

3. Ein Knoten ist ein Zwischenknoten in einem Zuordnungspaar. Dabei werden die Korrespondenzen dieses Knotens mittels der Entfernung berechnet. Beispielsweise wird der Knoten a_5 im Datensatz *A* nach der Entfernung mit Schwellwerten dem Knoten b_7 im Datensatz *B* zugeordnet ($A_6A_7 \leftrightarrow B_7B_8B_9$).

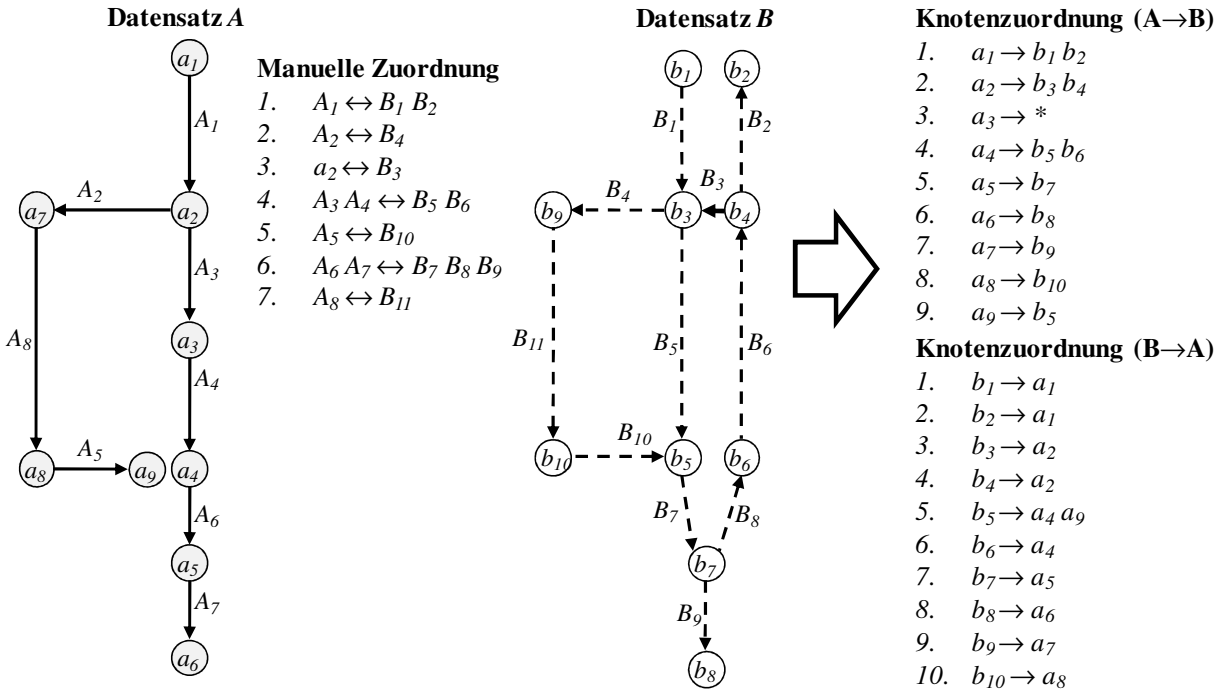


Abbildung 5.11: Beispiel für die Knotenzuordnung

Die Ergebnisse der Knotenzuordnung zwischen NavTeq und TeleAtlas werden in Tabelle 5.8 zusammengefasst. 34,5% der NavTeq-Knoten und 48,2% der TeleAtlas-Knoten im Testgebiet I sowie 18,9% der NavTeq-Knoten und 68,1% der TeleAtlas-Knoten im Testgebiet II können nicht zugeordnet werden. Weiterhin sind 60,0% der NavTeq-Knoten und 49,1% der TeleAtlas-Knoten im Testgebiet I mit der Relation *1:1* zugeordnet. Darüber hinaus werden 5,5% der NavTeq-Knoten und 2,7% der TeleAtlas-Knoten im Testgebiet I mit der Relation *1:n* zugeordnet. Die Anzahl der *1:1*

Zuordnungen und der *1:n* Zuordnungen sind in den zwei Datensätzen nicht identisch. Um eine gleiche Anzahl von Knotenzuordnungen zu erzielen, ist die *n:m* Knotenzuordnung notwendig.

	Testgebiet I		**Testgebiet II**	
	NT → TA	TA → NT	NT → TA	TA → NT
*1:**	326 (34,5%)	599 (48,2%)	110 (18,9%)	1047 (68,1%)
1:1	568 (60,0%)	610 (49,1%)	446 (76,6%)	480 (31,2%)
1:n	53 (5,5%)	34 (2,7%)	26 (4,5%)	10 (0,7%)

Tabelle 5.8: Ergebnisse der Knotenzuordnung zwischen NavTeq und TeleAtlas

Aus Tabelle 5.9 sind die Ergebnisse der Knotenzuordnung zwischen TeleAtlas und OpenStreetMap zu entnehmen. Analog zur Tabelle 5.8 können viele Knoten in beiden Datensätzen nicht zugeordnet werden (im Extremfall werden 90,8%).

	Testgebiet I		**Testgebiet II**	
	TA → OSM	OSM → TA	TA → OSM	OSM → TA
*1:**	601 (48,4%)	896 (58,7%)	1395 (90,8%)	179 (55,1%)
1:1	590 (47,4%)	545 (35,7%)	135 (8,8%)	139 (42,7%)
1:n	52 (4,2%)	85 (5,6%)	7 (0,4%)	7 (2,2%)

Tabelle 5.9: Ergebnisse der Knotenzuordnung zwischen TeleAtlas und OpenStreetMap

Aufgrund der unterschiedlichen Modellierung kann eine Unsicherheit der Knotenzuordnung auftreten. In Abbildung 5.12 sind zwei Beispiele für die Knotenzuordnung dargestellt (Formzuordnung zwischen „Simple“ und „Parallel“). In Abbildung 5.12a handelt es sich um eine *1:n* Zuordnung ($b_1 \leftrightarrow a_1, a_2$). Eine Verbindung zwischen Knoten a_1 und a_2 existiert in NavTeq. So ist die topologische Modellierung in den zwei Datensätzen ähnlich. In Abbildung 5.12b werden Knoten a_1 und a_2 in NavTeq zu Knoten b_1 in TeleAtlas zugeordnet. Hierbei existiert jedoch keine Verbindung zwischen Knoten a_1 und a_2.

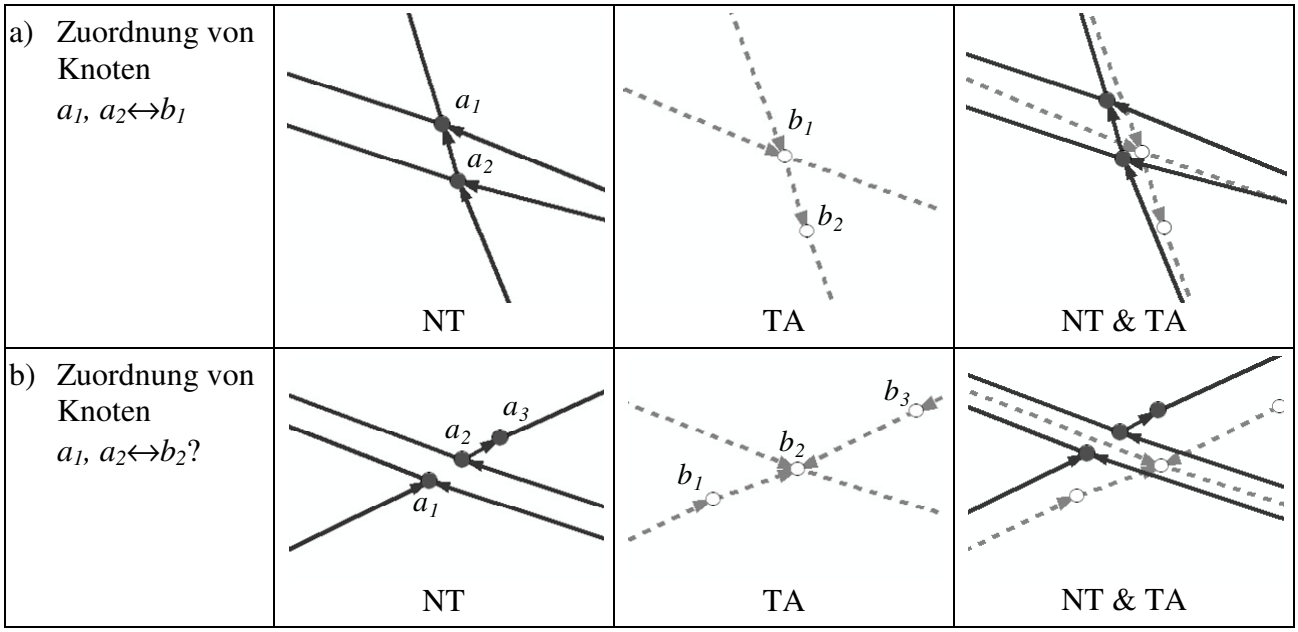


Abbildung 5.12: Kritische Stellen bei der Knotenzuordnung

6 Qualitätsanalyse

In diesem Kapitel wird die Qualität der Daten unter Verwendung von unterschiedlichen Ähnlichkeitsmaßen ausgewertet. Eine hohe Ähnlichkeit spiegelt eine hohe relative Qualität wider. Zunächst wird eine globale Qualitätsauswertung auf der Ebene des Datensatzes vorgestellt. Danach erfolgt eine lokale Qualitätsauswertung auf der Ebene der Zuordnungspaare. Eine Diskussion der erzielten Ergebnisse schließt dieses Kapitel ab.

6.1 Globale Qualitätsauswertung

Die globale Qualitätsauswertung beschreibt die Qualität des Gesamtdatensatzes. Hierbei werden folgende Merkmale untersucht:

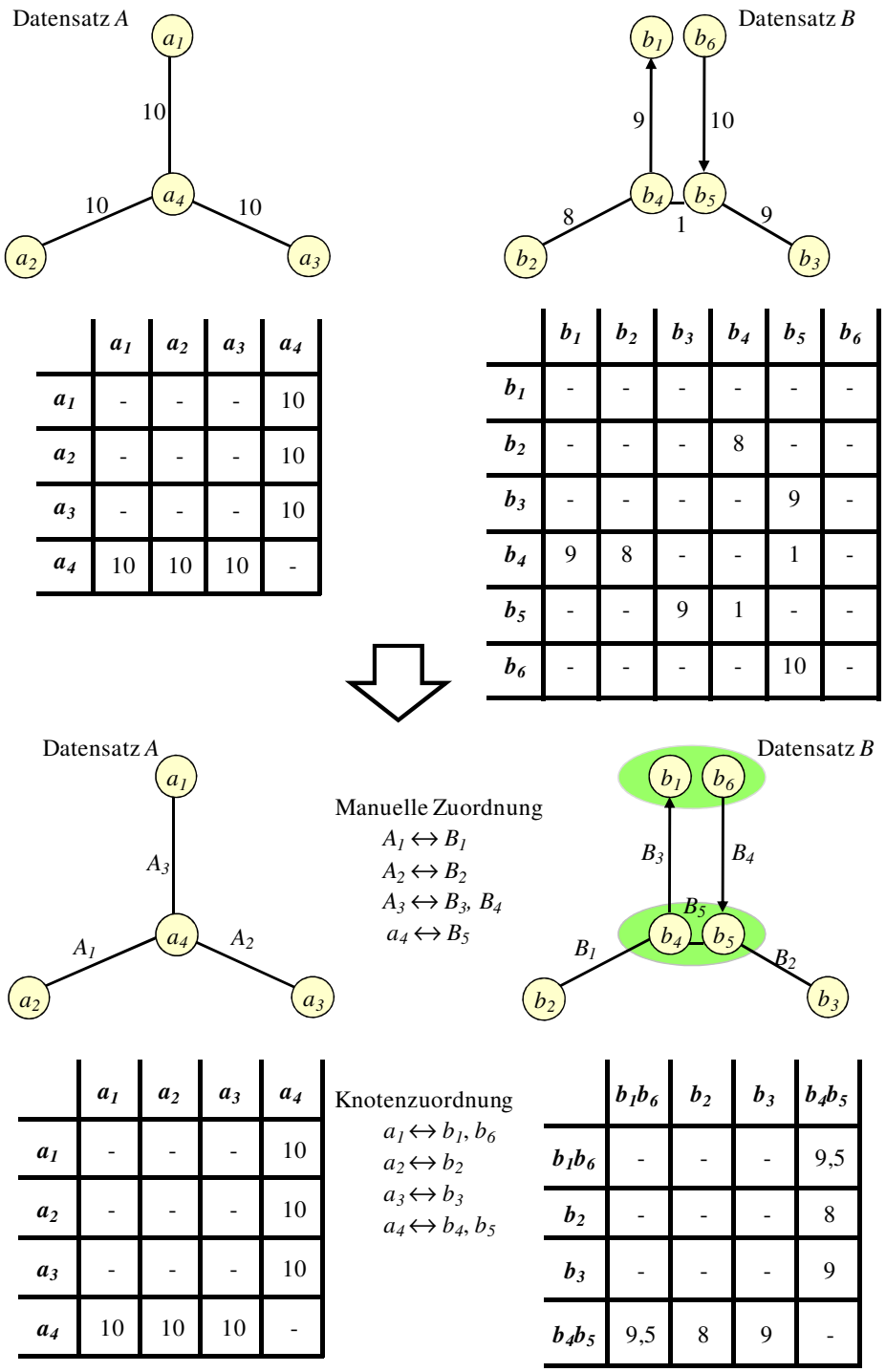
- Geometrische Ähnlichkeit
- Vollständigkeit
- Topologische Ähnlichkeit

Die Auswertung der geometrischen und topologischen Ähnlichkeit erfolgt durch einen Vergleich der Adjazenzmatrizen. Eine Voraussetzung dafür ist eine einheitliche Darstellung in den Adjazenzmatrizen. Das heißt, dass die Dimensionen der Adjazenzmatrizen identisch sein und ihre Zellen gleiche Objekte repräsentieren müssen. Differenzen bei der geometrischen Modellierung in den Datensätzen führen jedoch zu unterschiedlichen Dimensionen der Adjazenzmatrizen. Aus diesem Grund werden hierzu komplexe Objekte eingeführt, die mit Hilfe der Zuordnungsergebnisse zu berechnen sind.

Abbildung 6.1 zeigt ein Beispiel für die Berechnung der Adjazenzmatrizen mit gleicher Dimension. Die unterschiedliche Anzahl von Knoten (4 Knoten im Datensatz *A* und 6 Knoten im Datensatz *B*) führt zu unterschiedlichen Dimensionen der Adjazenzmatrizen. Nach der Berechnung der komplexen Objekte werden die Knoten (b_4, b_5) und (b_1, b_6) im Datensatz *B* jeweils in einen Komplexknoten und die Kanten B_3 und B_4 in eine Komplexkante überführt. Die Knoten (a_2, a_3) und (b_2, b_3) repräsentieren jeweils einen Komplexknoten. Die mittlere Länge der Komplexkanten wird in die entsprechenden Zellen der Adjazenzmatrix eingetragen. So entstehen zwei vergleichbare Adjazenzmatrizen.

6.1.1 Ermittlung von komplexen Objekten

Eine gleiche Dimension von Adjazenzmatrizen erfordert eine gleiche Anzahl von Komplexknoten in den Datensätzen. Die Berechnung basiert auf den Ergebnissen der Knoten- und Kantenzuordnung. Bei der Ermittlung der Komplexknoten werden die *n:m* Knotenzuordnungen betrachtet. Dabei wird z.B. ein Knoten vom Datensatz *A* selektiert und in eine Liste (Liste *A*) eingetragen. Weiterhin werden die Korrespondenzen im Datensatz *B* für diesen Knoten in eine zweite Liste (Liste *B*) eingetragen. Im nächsten Schritt sind die Korrespondenzen im Datensatz *A* für alle Knoten in der Liste *B* zu ermitteln und in die Liste *A* einzutragen. Es wird so lange iteriert, bis keine Änderung in der Liste *A* und Liste *B* stattfindet. Der Komplexknoten im Datensatz *A* besteht aus den Knoten in der Liste *A* und der Komplexknoten im Datensatz *B* aus den Knoten in der Liste *B*. Die gemittelte Geometrie der Knoten, die in einem Komplexknoten enthalten sind, wird als die Geometrie dieses Komplexknotens eingesetzt.



	a_1	a_2	a_3	a_4
a_1	-	-	-	10
a_2	-	-	-	10
a_3	-	-	-	10
a_4	10	10	10	-

	b_1	b_2	b_3	b_4	b_5	b_6
b_1	-	-	-	-	-	-
b_2	-	-	-	8	-	-
b_3	-	-	-	-	9	-
b_4	9	8	-	-	1	-
b_5	-	-	9	1	-	-
b_6	-	-	-	-	10	-

	a_1	a_2	a_3	a_4
a_1	-	-	-	10
a_2	-	-	-	10
a_3	-	-	-	10
a_4	10	10	10	-

	b_1b_6	b_2	b_3	b_4b_5
b_1b_6	-	-	-	9,5
b_2	-	-	-	8
b_3	-	-	-	9
b_4b_5	9,5	8	9	-

Abbildung 6.1: Umformung der Adjazenzmatrizen

In Abbildung 6.2 werden zwei Beispiele für die Berechnung der Komplexknoten aus den in Abbildung 5.11 ermittelten Knotenzuordnungen dargestellt. Für den Knoten a_1 im Datensatz A werden korrespondierende Knoten b_1 und b_2 im Datensatz B gefunden (siehe Beispiel 1). Da keine weiteren Korrespondenzen im Datensatz A für die Knoten b_1 und b_2 entdeckt werden, wird jeweils ein Komplexknoten aus dem Knoten a_1 im Datensatz A und aus den Knoten b_1 und b_2 im Datensatz B berechnet. Für den Knoten a_4 im Datensatz A werden korrespondierende Knoten b_5 und b_6 im Datensatz B ermittelt (siehe Beispiel 2). Weiterhin wird ein anderer korrespondierender Knoten a_9 im Datensatz A für die Knoten b_5 und b_6 entdeckt. Eine *n:m* Knotenzuordnung ($a_4a_9 \leftrightarrow b_5b_6$) wird erzeugt und Komplexknoten werden jeweils aus den Knoten in der Liste A und Liste B berechnet.

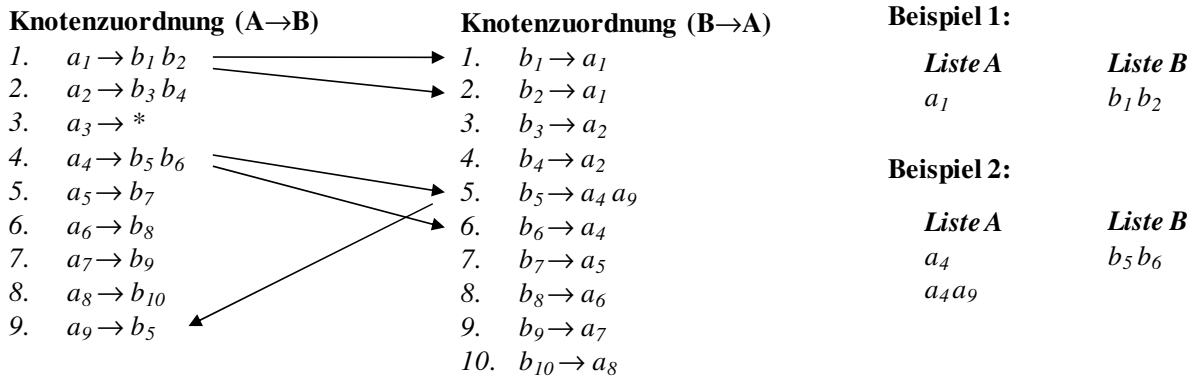


Abbildung 6.2: Berechnung der Komplexknoten (vgl. Abbildung 5.11)

Die Komplexkanten sind aus den Ergebnissen der Kantenzuordnung zu ermitteln. Für Zuordnungspaare mit der Relation *p:1*, *p:n*, *1:p* oder *n:p* werden nur Komplexknoten berechnet. Weiterhin werden ebenfalls nur Komplexknoten für die Zuordnungspaare mit der Anmerkung bei der manuellen Zuordnung als Zuordnung zwischen einem Komplexknoten und einem anderen Komplexknoten (*p:p*) berechnet (siehe Abbildung 5.6). Im Allgemeinen wird die Geometrie von Komplexkanten wie folgt berechnet:

1. Zunächst sind Mittellinien jeweils für die Abschnitte *Begin*, *Middle* und *End* zu berechnen.

2. Aus den ermittelten Mittellinien ergibt sich ein Abschnitt *Whole*.

3. Die Komplexkante ist durch Mittelung von allen Abschnitten *Whole* zu ermitteln.

Tabelle 6.1 fasst die Ergebnisse der Berechnung der komplexen Objekte von NavTeq und TeleAtlas zusammen. Die Anzahl der Knoten und Kanten, die in den komplexen Objekten enthalten sind, werden ebenfalls aufgelistet. Insgesamt werden 587 Komplexknoten im Testgebiet I und 462 Komplexknoten im Testgebiet II erzeugt.

	Testgebiet I		**Testgebiet II**	
	NavTeq	TeleAtlas	NavTeq	TeleAtlas
Knoten	620	644	472	490
Kanten	1110	1328	674	967
Komplexknoten	587	587	462	462
Komplexkanten	831	831	566	566

Tabelle 6.1: Ergebnisse der Berechnung der komplexen Objekte (NavTeq & TeleAtlas)

In Tabelle 6.2 sind die Ergebnisse der Berechnung der komplexen Objekte von TeleAtlas und OpenStreetMap dargestellt.

	Testgebiet I		**Testgebiet II**	
	TeleAtlas	OpenStreetMap	TeleAtlas	OpenStreetMap
Knoten	642	630	142	146
Kanten	1322	1332	500	268
Komplexknoten	575	575	131	131
Komplexkanten	819	819	164	164

Tabelle 6.2: Ergebnisse der Berechnung der komplexen Objekte (TeleAtlas & OpenStreetMap)

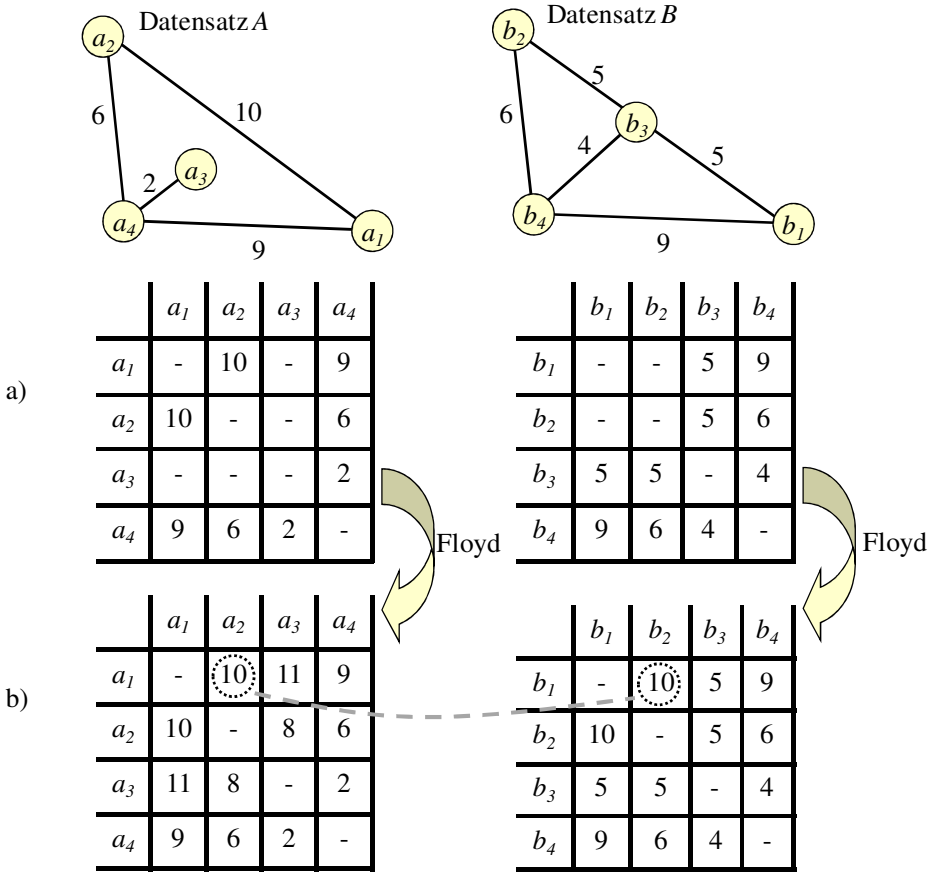
6.1.2 Geometrische Ähnlichkeit

Die Auswertung der geometrischen Ähnlichkeit erfolgt durch einen Vergleich der Zellen der Adjazenzmatrizen. Im Folgenden wird zunächst eine Vorverarbeitung für die Adjazenzmatrizen und anschließend die Auswertung der Adjazenzmatrizen vorgestellt.

Vorverarbeitung

Aus den ermittelten Komplexknoten und Komplexkanten sind die Adjazenzmatrizen zu berechnen. Topologische Unterschiede in den Datensätzen können zu großen Differenzen in den Adjazenzmatrizen führen. Zum Beispiel besitzen manche Zellen in einer Adjazenzmatrix aufgrund der topologischen Unterschiede einen Wert, in der anderen Adjazenzmatrix aber nicht. Um dieses Problem zu vermeiden, wird eine Vorverarbeitung mit dem Floyd-Algorithmus eingeführt [Sedgewick 1995].

Abbildung 6.3 zeigt ein Beispiel für die Minimierung der topologischen Unterschiede. Die Verbindungen (a_1, a_3) und (a_2, a_3) im Datensatz *A* sind vor der Vorverarbeitung nicht verfügbar und die Verbindung (b_1, b_2) im Datensatz *B* existiert ebenfalls nicht. Mit dem Floyd-Algorithmus werden alle kürzesten Wege zwischen zwei beliebigen Knoten ermittelt. Die Kosten der kürzesten Wege werden in die Zellen der Adjazenzmatrizen eingetragen. Die Adjazenzmatrizen vor und nach der Vorverarbeitung sind in Abbildung 6.3 dargestellt. Nach der Vorverarbeitung sind die topologischen Differenzen entfernt.



	a_1	a_2	a_3	a_4
a_1	-	10	-	9
a_2	10	-	-	6
a_3	-	-	-	2
a_4	9	6	2	-

	b_1	b_2	b_3	b_4
b_1	-	-	5	9
b_2	-	-	5	6
b_3	5	5	-	4
b_4	9	6	4	-

	a_1	a_2	a_3	a_4
a_1	-	10	11	9
a_2	10	-	8	6
a_3	11	8	-	2
a_4	9	6	2	-

	b_1	b_2	b_3	b_4
b_1	-	10	5	9
b_2	10	-	5	6
b_3	5	5	-	4
b_4	9	6	4	-

Abbildung 6.3: Minimierung der topologischen Unterschiede mit dem Floyd-Algorithmus

Auswertung der Adjazenzmatrix

Die maximale und mittlere Abweichung der Zellen in den Adjazenzmatrizen werden als geometrische Ähnlichkeitsmaße berechnet. Um die Länge der Kanten zu berücksichtigen, werden sowohl die absolute Abweichung (Meter) als auch die relative Abweichung (Prozent) ermittelt. Die absolute Abweichung für eine Zelle *(i, j)* in den Adjazenzmatrizen wird wie folgt berechnet:

$$Diff_{Abs}(i,j) = |Adj_A(i,j) - Adj_B(i,j)| \quad \textbf{(6.1)}$$

Die relative Abweichung für eine Zelle *(i, j)* gibt das Verhältnis zwischen der absoluten Abweichung und der maximalen Kantenlänge an:

$$Diff_{Rel}(i,j)(\%) = \frac{Diff_{Abs}(i,j)}{Max\big(Adj_A(i,j), Adj_B(i,j)\big)} * 100 \quad \textbf{(6.2)}$$

Tabelle 6.3 fasst die Ergebnisse der Auswertung der geometrischen Ähnlichkeit zusammen. Insgesamt werden 1706 Elemente in den Adjazenzmatrizen von NavTeq und TeleAtlas im Testgebiet I und 1134 Elemente im Testgebiet II ausgewertet. Mit dem Floyd-Algorithmus werden 46 Einträge im Testgebiet I und 30 Einträge im Testgebiet II in die Adjazenzmatrizen von NavTeq und TeleAtlas hinzugefügt. Die gemittelte absolute Abweichung von NavTeq und TeleAtlas im Testgebiet I beträgt 11,5 Meter und im Testgebiet II 27,6 Meter. Allerdings ist die relative Abweichung von NavTeq und TeleAtlas im Testgebiet II (12,2%) und im Testgebiet I (13,4%) ähnlich. Durch die Vorverarbeitung werden 126 Einträge im Testgebiet I und 24 Einträge im Testgebiet II in die Adjazenzmatrizen von TeleAtlas und OpenStreetMap hinzugefügt. Die gemittelte absolute Abweichung von TeleAtlas und OpenStreetMap beträgt 14,1 Meter im Testgebiet I und 108,7 Meter im Testgebiet II. Die gemittelte relative Abweichung von TeleAtlas und OpenStreetMap in den zwei Testgebieten ist ebenfalls ähnlich. Die großen absoluten Abweichungen sind auf große Abweichungen am Rand des Testgebiets zurückzuführen.

	NT & TA		**TA & OSM**	
	Testgebiet I	Testgebiet II	Testgebiet I	Testgebiet II
Anzahl der Elemente	1706	1134	1736	338
	(46)	(30)	(126)	(24)
Gemittelte Abweichung	11,5 m	27,6 m	14,1 m	108,7 m
	(13,4%)	(12,2%)	(17,2%)	(18,1%)
Maximale Abweichung	237 m	1397 m	639 m	2105 m
	(89,0%)	(89,1%)	(87,6%)	(91,9%)

Tabelle 6.3: Ergebnisse der Auswertung der geometrischen Ähnlichkeit

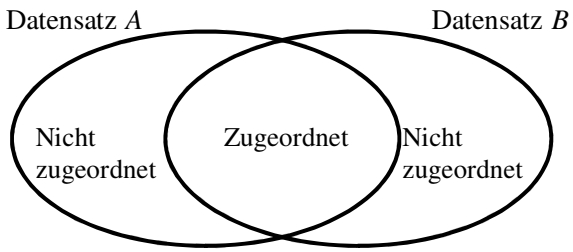
Für die zukünftige Entwicklung sind weitere Merkmale wie z.B. die Korrelation der Adjazenzmatrizen zu berechnen. Weiterhin sind die großen Längenunterschiede am Rand des Testgebiets zu minimieren.

6.1.3 Vollständigkeit

Wie bereits in Kapitel 2.3 vorgestellt, lässt sich die Vollständigkeit hierarchisch aus unterschiedlichen Aspekten berechnen. Eine einfache Auswertung der Vollständigkeit ist der

Vergleich der Gesamtlänge von Kanten. Zu diesem Zweck wird der Kartenbereich in kleinere Gitterzellen aufgeteilt [Haklay 2008].

Die Vollständigkeit in einer Gitterzelle von Datensatz *A* ist das Verhältnis zwischen der Gesamtlänge von Datensatz *A* und der Gesamtlänge der Vereinigungsmenge von Datensatz *A* und Datensatz *B* in der Gitterzelle. Hierbei muss berücksichtigt werden, dass die zugeordneten Elemente nicht zweimal betrachtet werden. Daher berechnet sich die Vereinigungsmenge aus den nicht zugeordneten Elementen von *A* und *B* plus die zugeordneten Elemente der Schnittmenge $A \cap B$. Deren Länge ergibt sich angenähert aus $\frac{1}{2}(A_1 + B_1)$.



$$A_1 = \sum Länge^{A}_{Zugeordnet}$$
$$A_2 = \sum Länge^{A}_{Nicht\ zugeordnet}$$
$$B_1 = \sum Länge^{B}_{Zugeordnet}$$
$$B_2 = \sum Länge^{B}_{Nicht\ zugeordnet}$$

Die Vollständigkeit wird wie folgt berechnet:

$$Vollständigkeit_A(\%) = \frac{A}{A \cup B} * 100 \approx \frac{A_1 + A_2}{\frac{1}{2}(A_1 + B_1) + A_2 + B_2} * 100 \qquad \textbf{(6.3)}$$

Da die zwei Testgebiete relativ klein sind, werden sie nicht mehr in kleinere Gitterzellen aufgeteilt. Die Ergebnisse der Auswertung der Vollständigkeit zwischen NavTeq und TeleAtlas werden in Tabelle 6.4 dargestellt. NavTeq und TeleAtlas haben eine ähnliche Vollständigkeit im Testgebiet I. Im Testgebiet II verfügt TeleAtlas eine höhere Vollständigkeit als NavTeq. Die graphischen Darstellungen der Datensätze finden sich in Anhang B.

	Testgebiet I		**Testgebiet II**	
	NavTeq	TeleAtlas	NavTeq	TeleAtlas
Vollständigkeit	87,1%	88,5%	42,7%	99,4%

Tabelle 6.4: Ergebnisse der Auswertung der Vollständigkeit (NavTeq & TeleAtlas)

Tabelle 6.5 fasst die Ergebnisse der Auswertung der Vollständigkeit zwischen TeleAtlas und OpenStreetMap zusammen. Im Testgebiet II sind die Datensätze TeleAtlas und OpenStreetMap ebenfalls unterschiedlich (siehe Anhang B).

	Testgebiet I		**Testgebiet II**	
	TeleAtlas	OpenStreetMap	TeleAtlas	OpenStreetMap
Vollständigkeit	67,5%	89,3%	94,8%	32,7%

Tabelle 6.5: Ergebnisse der Auswertung der Vollständigkeit (TeleAtlas & OpenStreetMap)

6.1.4 Topologische Ähnlichkeit

Die topologische Ähnlichkeit wird anhand der Exzentrizität von Knoten berechnet. Die Exzentrizität *e(x)* eines Knotens *x* ist die maximale Distanz von kürzesten Wegen in einem

Netzwerk, die vom Knoten *x* zu allen anderen Knoten führen. Die Exzentrizität von allen Knoten in einem Netzwerk lässt sich durch einen Vektor der Exzentrizität (N×1) repräsentieren, wobei *N* die Anzahl der Knoten ist. Für die Netzwerke in Abbildung 6.3 werden folgende zwei Vektoren der Exzentrizität ermittelt:

$$\vec{E}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 11 \\ 10 \\ 11 \\ 9 \end{pmatrix} \text{ und } \vec{E}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 5 \\ 9 \end{pmatrix}$$

Unter topologischer Ähnlichkeit ist die Korrelation von Vektoren der Exzentrizität zu verstehen. Die Korrelation *r* für einen Vektor *A* und Vektor *B* lässt sich im Allgemeinen wie folgt berechnen, wobei *N* die Anzahl der Elemente von Vektoren, x_1 ein Element vom Vektor *A* und x_2 ein Element vom Vektor *B* ist [Wong & Lee 2005]:

$$r = \frac{N\sum(x_1x_2) - (\sum x_1)(\sum x_2)}{\sqrt{[N\sum x_1^2 - (\sum x_1)^2][N\sum x_2^2 - (\sum x_2)^2]}} \quad \textbf{(6.4)}$$

Da die Anzahl der Knoten auf der Ebene der einfachen Objekte in den heterogenen Datensätzen nicht identisch ist, lässt sich die Korrelation nicht bestimmen. Aus diesem Grund wird die Exzentrizität auf der Ebene der komplexen Objekte berechnet. Tabelle 6.6 fasst die Ergebnisse der Korrelation der Exzentrizität von allen Komplexknoten in den zwei Testgebieten zusammen. Die topologische Ähnlichkeit zwischen NavTeq und TeleAtlas ist höher als die zwischen TeleAtlas und OpenStreetMap. Weiterhin ist die topologische Ähnlichkeit im Stadtgebiet (Testgebiet I) höher als im ländlichen Raum (Testgebiet II), weil die drei Datensätze im Testgebiet II unterschiedlich sind (siehe Anhang B).

	NT & TA		**TA & OSM**	
	Testgebiet I	Testgebiet II	Testgebiet I	Testgebiet II
Korrelation der Exzentrizität	0,929	0,840	0,854	0,663

Tabelle 6.6: Korrelation der Exzentrizität

6.2 Lokale Qualitätsauswertung

Die lokale Qualitätsauswertung beurteilt die Zuordnungspaare nach folgenden Aspekten: Ähnlichkeit der Form, geometrische Ähnlichkeit, topologische Ähnlichkeit und Ähnlichkeit von Attributen.

6.2.1 Ähnlichkeit der Form

Die Ermittlung der Ähnlichkeit der Form basiert auf den Ergebnissen der Formzuordnung (siehe Kapitel 5.3). Die verschiedenartigen Typen von Abschnitten in einem Zuordnungspaar erhalten unterschiedliche Gewichte:

- Abschnitt *Whole*: Gewicht=1
- Abschnitt *Begin*, *Middle* und *End*: Gewicht= $\frac{1}{Anzahl\ der\ Abschnitttypen}$

Zum Beispiel enthält die Formklasse „Fork1“ (siehe Abbildung 5.7) drei Abschnitte (1 *Begin*, 2 *End*). Jeder Abschnitt erhält dann ein Gewicht $\frac{1}{2}$ und das gesamte Gewicht beträgt $\frac{3}{2}$. Die Ähnlichkeit der Form eines Zuordnungspaars entspricht dann dem Verhältnis der Gewichte:

$$Ähnlichkeit_{Form} = \frac{\sum Gewicht_A}{\sum Gewicht_B} \quad \textbf{(6.5)}$$

Die Ähnlichkeit der Form gibt die Komplexität der geometrischen Modellierung an und lässt sich in folgende drei Klassen unterteilen:

- $Ähnlichkeit_{Form} = 1$: Die Komplexität der Modellierung eines Zuordnungspaares ist in den zwei Datensätzen gleich.
- $Ähnlichkeit_{Form} < 1$: Die Modellierung eines Zuordnungspaares ist im Datensatz *A* einfacher als im Datensatz *B*.
- $Ähnlichkeit_{Form} > 1$: Die Modellierung eines Zuordnungspaares ist im Datensatz *A* komplexer als im Datensatz *B*.

Die Ergebnisse der Auswertung der Formähnlichkeit sind in Tabelle 6.7 dargestellt. Mehr als 90% der Zuordnungspaare zwischen NavTeq und TeleAtlas in den zwei Testgebieten sind aus dem Aspekt der Komplexität der Modellierung identisch. Ebenfalls haben mehr als 90% der Zuordnungspaare zwischen TeleAtlas und OpenStreetMap eine gleiche Form. Bei der manuellen Zuordnung zwischen NavTeq und TeleAtlas wurden allerdings nur 54% der Zuordnungspaare im Testgebiet I und 60% im Testgebiet II mit der Relation *1:1* erfasst. Weiterhin wurden ebenfalls nur 47% der Zuordnungspaare zwischen TeleAtlas und OpenStreetMap im Testgebiet I und 30% im Testgebiet II *1:1* zugeordnet. Die drei Datensätze repräsentieren die gleichen Objekte mit ähnlichen Datenmodellen. Aus diesem Grund bietet die Ähnlichkeit der Form eine plausiblere Aussage für die geometrische Modellierung als die Zuordnungsrelationen.

	NT & TA		**TA & OSM**	
	Testgebiet I	Testgebiet II	Testgebiet I	Testgebiet II
$Ähnlichkeit_{Form} < 1$	60 (6,8%)	25 (4,3%)	45 (5,1%)	11 (6,3%)
$Ähnlichkeit_{Form} = 1$	818 (93,0%)	562 (95,5%)	814 (93,2%)	158 (90,8%)
$Ähnlichkeit_{Form} > 1$	2 (0,2%)	1 (0,2%)	15 (1,7%)	5 (2,9%)

Tabelle 6.7: Ergebnisse der Auswertung der Formähnlichkeit

6.2.2 Geometrische Ähnlichkeit

Die geometrische Ähnlichkeit wird in der vorliegenden Arbeit anhand der Hausdorff-Distanz berechnet. Besitzt ein Zuordnungspaar eine unterschiedliche Anzahl von Abschnitten, dann werden die Komplexkanten des Zuordnungspaars bei der Auswertung betrachtet. Die Hausdorff-Distanz ist nach (6.6) zu berechnen [Hangouet 1995]:

$$\delta H(A, B) = \mathrm{Max}\{\mathrm{Min}[|a - b|]\} \quad \textbf{(6.6)}$$

In Anhang D sind die Häufigkeitsverteilungen der Hausdorff-Distanz in den zwei Testgebieten dargestellt. Die durchschnittliche Hausdorff-Distanz zwischen NavTeq und TeleAtlas in beiden Testgebieten liegt im Bereich von zehn Metern. Die Verteilungen in den zwei Testgebieten sind ähnlich. Die durchschnittliche Hausdorff-Distanz zwischen TeleAtlas und OpenStreetMap in den zwei Testgebieten liegt im Bereich zwischen zehn und fünfzehn Metern. Allerdings sind die Verteilungen in den zwei Testgebieten unterschiedlich. Die Unterschiede der Länge am Rand führen ebenfalls zu großen Hausdorff-Distanzen.

6.2.3 Topologische Ähnlichkeit

Bei der globalen Qualitätsauswertung wurde die topologische Ähnlichkeit für den gesamten Datensatz ermittelt. Im Folgenden wird die topologische Ähnlichkeit für jedes Zuordnungspaar mit der *Erreichbarkeit* berechnet. Bei einem zusammenhängenden Netzwerk existiert i.d.R. immer ein Weg zwischen zwei Knoten. Aus diesem Grund wird die Erreichbarkeit durch die Distanz (Kosten des kürzesten Wegs) zwischen zwei Knoten repräsentiert. Die Erreichbarkeit eines Knotens x in einem Netzwerk wird durch einen Vektor von Kosten repräsentiert, welche Distanzen von kürzesten Wegen von allen anderen Knoten zu diesem Knoten x darstellen. Da die Anzahl der Knoten in den Datensätzen nicht identisch ist, wird die Erreichbarkeit mit Hilfe der komplexen Objekte berechnet. Dabei werden kürzeste durchschnittliche Kosten von Komplexknoten zu einem Zuordnungspaar ermittelt. Für jeden Datensatz werden die Kosten von den im Komplexknoten enthaltenden Knoten zu den im Zuordnungspaar enthaltenden Knoten berechnet und anschließend gemittelt.

In Abbildung 6.4 wird ein Beispiel für die Berechnung der kürzesten durchschnittlichen Kosten von einem Komplexknoten zu einem Zuordnungspaar dargestellt. Die Kanten A_1 und A_2 im Datensatz A werden zu der Kante B_1 im Datensatz B zugeordnet. Für den Datensatz A werden Kosten der kürzesten Wege vom Komplexknoten (a_4) zu den Knoten a_1, a_2 und a_3 berechnet und die kleinsten Kosten ausgewählt. Der Komplexknoten im Datensatz B besteht aus zwei Knoten. Für den Datensatz B werden kürzeste Wege jeweils von Knoten b_3 und b_4 zu Knoten b_1 und b_2 berechnet und die kleinsten Kosten anschließend gemittelt. Die kürzesten durchschnittlichen Kosten vom Komplexknoten zum Zuordnungspaar betragen 5 im Datensatz A und 5,5 im Datensatz B.

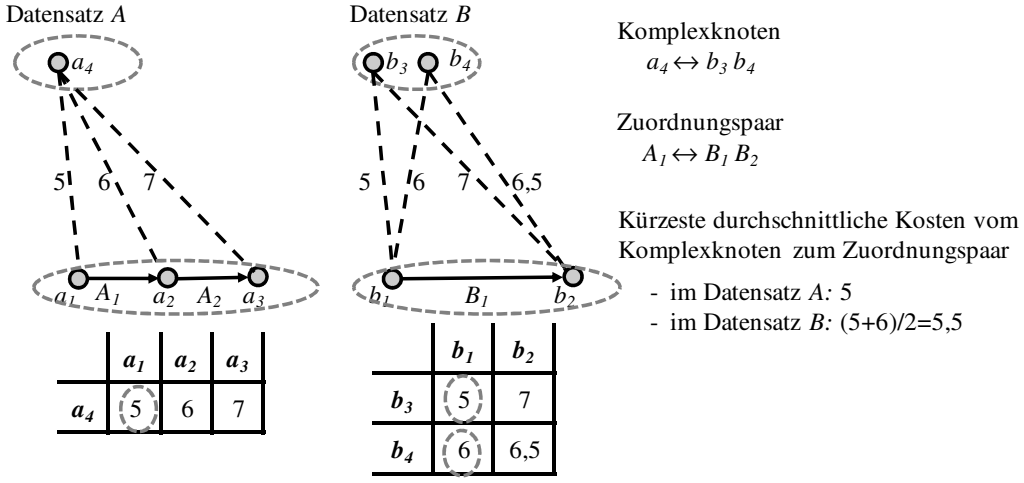


Abbildung 6.4: Berechnung der kürzesten durchschnittlichen Kosten von einem Komplexknoten zu einem Zuordnungspaar

Von allen Komplexknoten zu einem Zuordnungspaar werden die kürzesten durchschnittlichen Kosten berechnet. Daraus ergibt sich ein Vektor (N×1) in jedem Datensatz, wobei N die Anzahl der

Komplexknoten ist. Die topologische Ähnlichkeit dieses Zuordnungspaars wird durch die Korrelation der Vektoren repräsentiert.

In Anhang D sind die Ergebnisse der Auswertung der topologischen Ähnlichkeiten in den zwei Testgebieten zu entnehmen. Die topologischen Ähnlichkeiten in den zwei Testgebieten sind meistens sehr hoch (>0,9), aber die topologischen Ähnlichkeiten von einigen Zuordnungspaaren im Testgebiet II sind niedrig (z.B. <0,5), weil die drei Datensätze im Testgebiet II unterschiedlich sind. Die topologische Ähnlichkeit im Stadtgebiet (Testgebiet I) ist höher als im ländlichen Raum (Testgebiet II). Darüber hinaus ist die topologische Ähnlichkeit zwischen NavTeq und TeleAtlas höher als die zwischen TeleAtlas und OpenStreetMap.

6.2.4 Ähnlichkeit von Attributen

Je nach Typen von Attributen sind unterschiedliche Verfahren zur Auswertung der Ähnlichkeit von Attributen einzusetzen (siehe Tabelle 6.8). Nominal- und Ordinalskalen werden an dieser Stelle nicht unterschieden. Da die semantischen Korrespondenzen bereits bestimmt wurden, lassen sich Skalen (Attribute) mit gleichen begrenzten bzw. unbegrenzten Wertebereichen direkt vergleichen. Die Auswertung von Skalen mit unterschiedlichen Wertebereichen wird mit Hilfe einer Konfusionsmatrix durchgeführt. Zusammengesetzte Attribute (wie z.B. Hausnummer) sind in einzelne Attribute zu zerlegen und anschließend auszuwerten.

Typen von Attributen	**Auswertungsverfahren**	**Beispiele von Attributen**
Skalen mit gleichen begrenzten Wertebereichen	Direktvergleich	Tollroad Ferry Ownership
Skalen mit gleichen unbegrenzten Wertebereichen	Direktvergleich	Straßenname Road Number
Skalen mit unterschiedlichen Wertebereichen	Konfusionsmatrix	Functional Road Class
Zusammengesetzte Attribute	Auswertung nach der Zerlegung von Attributen	Hausnummer

Tabelle 6.8: Auswertungsverfahren für unterschiedliche Attributtypen

Direkt vergleichbare Attribute erhalten eine Ähnlichkeit von *1* oder *0*. Für nicht direkt vergleichbare Attribute wird die Ähnlichkeit anhand einer Konfusionsmatrix berechnet. Beispielsweise verfügt ein Attribut über *k* Kategorien im Datensatz *A* und *m* Kategorien im Datensatz *B*. So entsteht eine Konfusionsmatrix (siehe Tabelle 6.9), wobei O_{ij} die Anzahl von Attributzuordnungen zwischen Kategorie *i* im Datensatz *A* und Kategorie *j* im Datensatz *B*, O_{i+} die Summe der Zeile *i*, O_{+j} die Summe der Spalte *j* und *N* die Summe von allen Elementen der Konfusionsmatrix ist.

	Datensatz *B*				
Datensatz *A*	***1***	***2***	**…**	***m***	***Summe ∑***
1	O_{11}	O_{12}	…	O_{1m}	O_{1+}
2	O_{21}	O_{22}	…	O_{2m}	O_{2+}
…	…	…	…	…	…
k	O_{k1}	O_{k2}	…	O_{km}	O_{k+}
Summe ∑	O_{+1}	O_{+2}	…	O_{+m}	*N*

Tabelle 6.9: Konfusionsmatrix für Skalen mit unterschiedlichen Klassifikationen in den Datensätzen

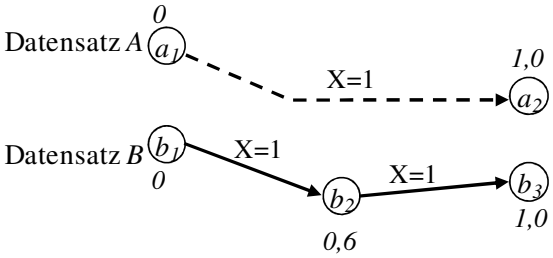
Die Ähnlichkeit $s_{i,j}$ für nicht direkt vergleichbare Attribute lässt sich auf Basis der Konfusionsmatrix wie folgt berechnen:

$$s_{i,j} = \frac{O_{i,j}}{O_{+j}} \tag{6.7}$$

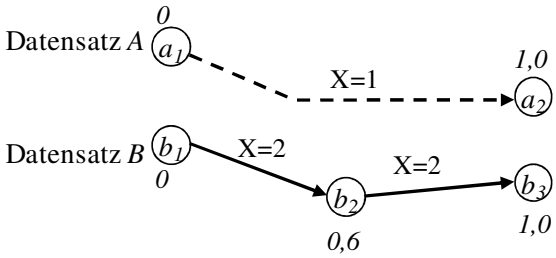
Die Ähnlichkeit einer Skala in einem Zuordnungspaar (unterschiedliche Attributwerte in den verschiedenen Abschnitten) wird nach Gleichung (6.8) berechnet, wobei $d_{i,j}$ die Distanz entlang des Zuordnungspaars von einem Segment mit der Kategorie *i* im Datensatz *A* und der Kategorie *j* im Datensatz *B* und $s_{i,j}$ die Ähnlichkeit zwischen der Kategorie *i* im Datensatz *A* und der Kategorie *j* im Datensatz *B* ist:

$$Ähnlichkeit_{Att} = \frac{\sum(d_{i,j} * s_{i,j})}{\sum(d_{i,j})} \tag{6.8}$$

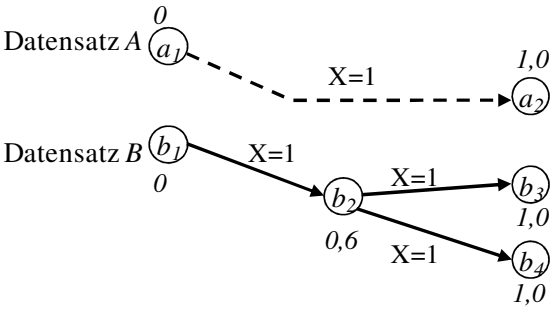
Abbildung 6.5 veranschaulicht vier Beispiele für die Ermittlung der Ähnlichkeit von einem Attribut „X“, welches gleiche Kategorien (zwei Kategorien: „*1*“ und „*2*“) im Datensatz *A* und Datensatz *B* enthält. Distanzen entlang des Zuordnungspaars werden dargestellt. Besitzt das Attribut in einem Zuordnungspaar jeweils im Datensatz *A* und Datensatz *B* genau eine Kategorie, dann ist die Ähnlichkeit gleich *0* oder *1* (siehe Abbildung 6.5a, b und c). Ansonsten ergibt sich eine Ähnlichkeit zwischen *0* und *1* (siehe Abbildung 6.5d).



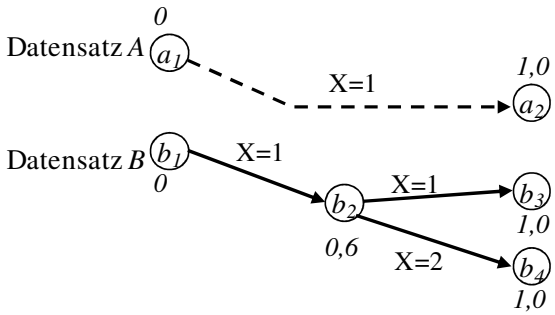
a) Das Attribut „X“ besitzt genau eine Kategorie im Datensatz *A* und Datensatz *B*.



b) Das Attribut „X“ hat unterschiedliche Kategorien im Datensatz *A* und Datensatz *B*.



c) Das Attribut „X“ besitzt genau eine Kategorie im Datensatz *A* und Datensatz *B*.



d) Das Attribut „X“ hat unterschiedliche Kategorien im Datensatz *A* und Datensatz *B*.

Abbildung 6.5: Beispiele für die Ermittlung der Ähnlichkeit von Attributen

a) $$Ähnlichkeit_X = \frac{(0{,}6 * s_{1,1} + 0{,}4 * s_{1,1})}{(0{,}6 + 0{,}4)} = \frac{(0{,}6 * 1 + 0{,}4 * 1)}{(0{,}6 + 0{,}4)} = 1{,}0$$

b) $$Ähnlichkeit_X = \frac{(0{,}6 * s_{1,2} + 0{,}4 * s_{1,2})}{(0{,}6 + 0{,}4)} = \frac{(0{,}6 * 0 + 0{,}4 * 0)}{(0{,}6 + 0{,}4)} = 0{,}0$$

c) $$Ähnlichkeit_X = \frac{(0{,}6 * s_{1,1} + 0{,}4 * s_{1,1} + 0{,}4 * s_{1,1})}{(0{,}6 + 0{,}4 + 0{,}4)} = \frac{(0{,}6 * 1 + 0{,}4 * 1 + 0{,}4 * 1)}{(0{,}6 + 0{,}4 + 0{,}4)} = 1{,}0$$

d) $$Ähnlichkeit_X = \frac{(0{,}6 * s_{1,1}) + (0{,}4 * s_{1,1}) + (0{,}4 * s_{1,2})}{(0{,}6 + 0{,}4 + 0{,}4)} = \frac{(0{,}6 * 1) + (0{,}4 * 1) + (0{,}4 * 0)}{(0{,}6 + 0{,}4 + 0{,}4)}$$
$$\approx 0{,}71$$

Auswertung einer Skala mit gleichen begrenzten Wertebereichen

Eine Skala mit gleichen begrenzten Wertebereichen kann zwei Kategorien (binäre Klassifikation) oder mehr als zwei Kategorien enthalten. In der Folge werden Attribute mit binärer Klassifikation als Beispiele ausgewertet. Die gemittelte Ähnlichkeit eines Attributs von allen Zuordnungspaaren wird als ein Gesamtmaß der Ähnlichkeit dieses Attributs berechnet.

Tabelle 6.10 fasst die Ergebnisse der Auswertung der Ähnlichkeit von Attributen mit binärer Klassifikation zusammen. Wie zu erkennen ist, sind Attribute *Toll Road* und *Ferry* in NavTeq und TeleAtlas in den zwei Testgebieten hundertprozentig identisch. Darüber hinaus sind die gemittelten Ähnlichkeiten von *Ownership* in NavTeq und TeleAtlas ebenfalls sehr hoch (>99%). Da die Attribute *Toll Road* und *Ferry* in OpenStreetMap in den zwei Testgebieten nicht vorhanden sind, wird nur die Ähnlichkeit von *Ownership* zwischen TeleAtlas und OpenStreetMap ausgewertet. Im Testgebiet II ist das Attribut *Ownership* hundertprozentig identisch und im Testgebiet I wird ebenfalls eine sehr hohe Ähnlichkeit (99,8%) erzielt.

	NT & TA						**TA & OSM**	
	Testgebiet I			Testgebiet II			Testgebiet I	Testgebiet II
	Toll Road	*Ferry*	*Ownership*	*Toll Road*	*Ferry*	*Ownership*	*Ownership*	*Ownership*
Ähnlichkeit = 1	833 (100%)	833 (100%)	830 (99,7%)	566 (100%)	566 (100%)	561 (99,1%)	830 (99,8%)	164 (100%)
Ähnlichkeit ∈ (0,1)	0 (0%)	0 (0%)	2 (0,2%)	0 (0%)	0 (0%)	1 (0,2%)	0 (0%)	0 (0%)
Ähnlichkeit = 0	0 (0%)	0 (0%)	1 (0,1%)	0 (0%)	0 (0%)	4 (0,7%)	2 (0,2%)	0 (0%)
Gemittelte Ähnlichkeit	1,0	1,0	0,998	1,0	1,0	0,993	0,998	1,0

Tabelle 6.10: Ergebnisse der Auswertung der Ähnlichkeit von Skalen mit gleichen begrenzten Kategorien (binärer Klassifikation)

Auswertung einer Skala mit gleichen unbegrenzten Wertebereichen

Skalen mit gleichen unbegrenzten Klassifikationen wie z.B. *Straßenname* und *Road Number* lassen sich nach der semantischen Homogenisierung (siehe Kapitel 4.2.2) direkt vergleichen.

In Tabelle 6.11 sind die Ähnlichkeiten von Skalen mit gleichen unbegrenzten Wertebereichen dargestellt. Die Ähnlichkeit des Attributs *Road Number* zwischen TeleAtlas und OpenStreetMap wird nicht berechnet, weil es in OpenStreetMap nicht verfügbar ist. Die Ähnlichkeit von *Straßennamen* im Stadtgebiet (Testgebiet I) ist höher als im ländlichen Raum (Testgebiet II). Darüber hinaus ist die Ähnlichkeit von *Straßenname* zwischen NavTeq und TeleAtlas höher als die zwischen TeleAtlas und OpenStreetMap. Insgesamt haben 78 Zuordnungspaare (48,1%) zwischen TeleAtlas und OpenStreetMap im Testgebiet II unterschiedliche *Straßennamen*. Weiterhin enthalten ca. 26% der Zuordnungspaare zwischen NavTeq und TeleAtlas in den zwei Testgebieten unterschiedliche *Road Number*. Die Ähnlichkeit der Attribute, die über keinen Wert in beiden Datensätzen verfügen, wird nicht berechnet. Beispielsweise wird die Ähnlichkeit des Attributs *Road Number* für 53,8% der Zuordnungspaare zwischen NavTeq und TeleAtlas im Testgebiet I nicht berechnet. Bei der Berechnung der gemittelten Ähnlichkeit werden sie nicht berücksichtigt.

	NT & TA				**TA & OSM**	
	Testgebiet I		Testgebiet II		Testgebiet I	Testgebiet II
	Straßenname	*Road Number*	*Straßenname*	*Road Number*	*Straßenname*	*Straßenname*
Ähnlichkeit= 1	686 (82,4%)	163 (19,6%)	358 (63,3%)	144 (25,5%)	611 (73,5%)	54 (33,3%)
Ähnlichkeit∈ (0,1)	26 (3,1%)	3 (0,3%)	29 (5,1%)	4 (0,7%)	49 (5,9%)	14 (7,4%)
Ähnlichkeit= 0	112 (13,4%)	219 (26,3%)	148 (26,1%)	149 (26,3%)	165 (19,8%)	78 (48,1%)
Nicht ermittelbare Ähnlichkeit	9 (1,1%)	448 (53,8%)	31 (5,5%)	269 (47,5%)	7 (0,8%)	18 (11,2%)
Gemittelte Ähnlichkeit	0,853	0,430	0,698	0,495	0,778	0,424

Tabelle 6.11: Ergebnisse der Auswertung der Ähnlichkeit von Skalen mit gleichen unbegrenzten Wertebereichen

Auswertung einer Skala mit unterschiedlichen Wertebereichen

Für Skalen mit unterschiedlichen Wertebereichen wird das Attribut *Functional Road Class* als Beispiel ausgewertet. Da das Attribut *Functional Road Class* in OpenStreetMap nicht verfügbar ist, findet die Auswertung nur zwischen NavTeq und TeleAtlas statt. Mit Hilfe der Zuordnungsergebnisse wird die Konfusionsmatrix berechnet. Für ein Zuordnungspaar mit k unterschiedlichen Kategorien in NavTeq und m unterschiedlichen Kategorien in TeleAtlas wird der Wert $\frac{1}{k*m}$ jeweils in die entsprechenden Zellen der Konfusionsmatrix eingetragen. Tabelle 6.12 stellt die Konfusionsmatrix für *Functional Road Class* aus den Zuordnungsergebnissen vom Testgebiet I dar. Die Beschreibungen der unterschiedlichen Kategorien für *Functional Road Class* in NavTeq und TeleAtlas sind in Anhang A zu ermitteln.

		TeleAtlas										
		-1	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**∑**
NavTeq	**1**	0	0	0	0	0	0	0	0	0	0	0
	2	0,5	0	90,25	23,75	0	0,25	0	0	1,25	0	116
	3	0	0	3,25	48	0	70,25	15	1	3	0	140,5
	4	0	0	0	1	0	42	8	4	19	0	74
	5	1	0	7	7,75	0	9,5	9	33,5	433,75	32	533,5
	∑	1,5	0	100,5	80,5	0	122	32	38,5	457	32	864

Tabelle 6.12: Konfusionsmatrix für Functional Road Class (NavTeq & TeleAtlas)

Für Skalen mit unterschiedlichen Wertebereichen ist der Chi-Quadrat Koeffizient χ^2 wie folgt zu berechnen, um ein zusätzliches Ähnlichkeitsmaß zu ermitteln [Wong & Lee 2005]:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \qquad \textbf{(6.9)}$$

Dabei ist die erwartete Häufigkeit für das Element *(i, j)*:

$$E_{i,j} = \frac{O_{i+} \cdot O_{+j}}{N} \qquad \textbf{(6.10)}$$

Ist der ermittelte Chi-Quadrat Koeffizient größer als $\chi^2_{(k-1)(m-1);(1-\alpha)}$ mit einem Freiheitsgrad *f=(k-1)(m-1)* und einer Irrtumswahrscheinlichkeit α, dann wird die Annahme akzeptiert, dass das Attribut in beiden Datensätzen ähnlich ist. Die Ermittlung des Chi-Quadrat Koeffizients χ^2 setzt voraus, dass jedes $E_{i,j}$ ungleich 0 ist ($\forall E_{i,j} \neq 0$). Aus diesem Grund werden die Zeilen ($O_{i+} = 0$) und Spalten ($O_{+j} = 0$) von der Konfusionsmatrix entfernt. Tabelle 6.13 stellt die erwarteten Häufigkeiten für *Functional Road Class* dar. Die Kategorie *1* in NavTeq und die Kategorien *0* und *3* in TeleAtlas werden von der Konfusionsmatrix entfernt (vgl. Tabelle 6.12). Danach wird die Dimension der Konfusionsmatrix reduziert.

		TeleAtlas							
		-1	**1**	**2**	**4**	**5**	**6**	**7**	**8**
NavTeq	**2**	0,20 (0,5)	13,49 (90,25)	10,81 (23,75)	16,38 (0,25)	4,30 (0)	5,17 (0)	61,36 (1,25)	4,30 (0)
	3	0,24 (0)	16,34 (3,25)	13,9 (48)	19,84 (70,25)	5,20 (15)	6,26 (1)	74,32 (3)	5,20 (0)
	4	0,13 (0)	8,61 (0)	6,89 (1)	10,45 (42)	2,74 (8)	3,30 (4)	39,14 (19)	2,74 (0)
	5	0,93 (1)	62,06 (7)	49,71 (7,75)	75,33 (9,5)	19,76 (9)	23,77 (33,5)	282,19 (433,75)	19,76 (32)

Tabelle 6.13: Reduzierte Matrix von erwarteten Häufigkeiten zur Ermittlung des Chi-Quadrat Koeffizients

Aus Tabelle 6.14 sind die Ergebnisse der Auswertung der Ähnlichkeit von *Functional Road Class* zu entnehmen. Neben dem Chi-Quadrat Koeffizient wird die Anzahl der Kategorien (*k* Kategorien in NavTeq und *m* Kategorien in TeleAtlas) nach der Reduktion aufgelistet. Da die ermittelten Koeffizienten χ^2 in den zwei Testgebieten größer als $\chi^2_{21;(1-0.01)} = 39{,}932$ (α=0,01) sind, wird die

Annahme akzeptiert, dass das Attribut in den zwei Datensätzen ähnlich ist. Die Ähnlichkeit von *Functional Road Class* im Testgebiet II ist höher als im Testgebiet I, da das Attribut über weniger Kategorien im Testgebiet II (3 Kategorien in NavTeq und 6 Kategorien in TeleAtlas) als im Testgebiet I (4 Kategorien in NavTeq und 8 Kategorien in TeleAtlas) verfügt.

	Testgebiet I	**Testgebiet II**
χ^2 (k×m)	1242,5 (4×8)	1302,6 (3×6)
Gemittelte Ähnlichkeit (s)	0,776	0,858

Tabelle 6.14: Ergebnisse der Auswertung der Ähnlichkeit von Functional Road Class

Auswertung von zusammengesetzten Attributen

Am Beispiel von Hausnummern werden zusammengesetzte Attribute ausgewertet. Zur Speicherung von Hausnummern werden i.d.R. zwei Datenstrukturen verwendet: Hausnummernbereiche [ISO14825 2004] und punktbasierte Hausnummern. Hausnummernbereiche sind für regelmäßige Hausnummern geeignet und werden als Attribute den Kanten zugeordnet. Im Gegensatz dazu eignen sich punktbasierte Hausnummern für unregelmäßige Hausnummern (z.B. in China).

Da in OpenStreetMap keine Hausnummern in den zwei Testgebieten verfügbar sind, findet die Auswertung von Hausnummern nur zwischen NavTeq und TeleAtlas statt. In den zwei Testgebieten stehen allerdings nur Hausnummernbereiche in NavTeq und TeleAtlas zur Verfügung. Um die Geometrie von Hausnummern vergleichen zu können, werden die Hausnummernbereiche anhand der Distanz entlang der Kante in einzelne Hausnummernpunkte umgewandelt (siehe Abbildung 6.6).

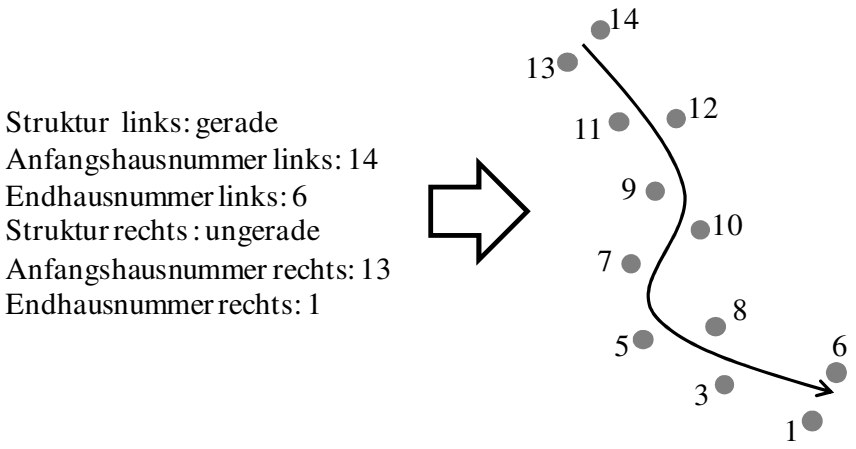


Abbildung 6.6: Umwandlung eines Hausnummernbereichs in einzelnen Hausnummernpunkten

In Abbildung 6.7 werden Ausschnitte von umgewandelten Hausnummernpunkten dargestellt. Es ist zu erkennen, dass Hausnummern in den zwei Datensätzen unterschiedlich sind.

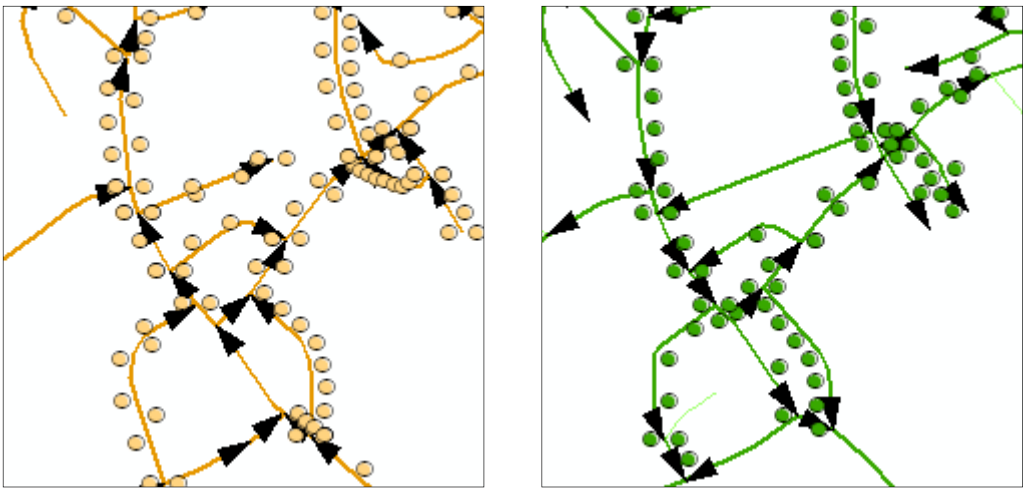


Abbildung 6.7: Ergebnisse der Umwandlung von Hausnummern in NavTeq (links) und TeleAtlas (rechts)

Im ersten Schritt der Auswertung von Hausnummern sind die Korrespondenzen der Hausnummernpunkte in den zwei Datensätzen zu finden. Zu diesem Zweck können zwei Verfahren eingesetzt werden. Einerseits können die Hausnummernpunkte mit geometrischen (Puffer) und thematischen Kriterien (Straßenname und Hausnummer) zugeordnet werden. Andererseits erfolgt die Zuordnung von Hausnummernpunkten auf Basis der Kantenzuordnung. Im diesem Fall sind die Korrespondenzen von Hausnummern innerhalb eines Zuordnungspaares zu finden. Aufgrund der verfügbaren Kantenzuordnung wird das zweite Verfahren in der vorliegenden Arbeit verwendet. Nach der Zuordnung werden die Hausnummern aus folgenden Aspekten untersucht:

1. *Hausnummernstruktur der Zuordnungspaare*: Die linke und rechte Hausnummernstruktur von einem Zuordnungspaar muss in den zwei Datensätzen identisch sein.

2. *Topologie der Hausnummernpunkte*: Die Hausnummernpunkte müssen in den zwei Datensätzen auf der gleichen Seite von Kanten liegen.

3. *Geometrie der Hausnummernpunkte*: Die Hausnummernpunkte müssen sich auf der gleichen Position in den zwei Datensätzen befinden.

Tabelle 6.15 fasst die Ergebnisse der Auswertung von Hausnummernstrukturen der Zuordnungspaare zusammen. Hausnummernstrukturen stehen in 31,5% der Zuordnungspaare im Testgebiet I und 44,7% im Testgebiet II in beiden Datensätzen nicht zur Verfügung. Weiterhin sind 25,8% der linken Hausnummernstrukturen und 13,3% der rechten Hausnummernstrukturen der Zuordnungspaare im Testgebiet I nur in einem Datensatz verfügbar. Dasselbe gilt für ca. 10% der linken und rechten Hausnummernstrukturen der Zuordnungspaare im Testgebiet II. Insgesamt sind 10,3% der linken und 13,0% der rechten Hausnummernstrukturen der Zuordnungspaare im Testgebiet I unterschiedlich. Das gilt auch für ca. 17% der Zuordnungspaare im Testgebiet II. Nur 32,4% der linken und 42,2% der rechten Hausnummernstrukturen im Testgebiet I sowie 28,8% der linken und 27,4% der rechten Hausnummernstrukturen im Testgebiet II sind identisch.

	Testgebiet I		Testgebiet II	
	Struktur links	Struktur rechts	Struktur links	Struktur rechts
Nicht verfügbar in beiden Datensätzen	277 (31,5%)	277 (31,5%)	263 (44,7%)	263 (44.7%)
Nur verfügbar in einem Datensatz	227 (25,8%)	117 (13,3%)	56 (9,5%)	61 (10,4%)
Unterschiedliche Struktur	91 (10,3%)	114 (13,0%)	100 (17,0%)	103 (17,5%)
Identische Struktur	285 (32,4%)	372 (42,2%)	169 (28,8%)	161 (27,4%)

Tabelle 6.15: Ergebnisse der Auswertung von Hausnummernstrukturen der Zuordnungspaare

In Tabelle 6.16 sind die Ergebnisse der topologischen Auswertung der Hausnummernpunkte zu ermitteln. Die Hausnummernpunkte können zugeordnet oder nicht zugeordnet sein. Insgesamt werden 2224 Hausnummernpunkte (82,5% in NavTeq und 77,1% in TeleAtlas) im Testgebiet I und 1206 Hausnummernpunkte (67,0% in NavTeq und 78,5% in TeleAtlas) im Testgebiet II mit gleicher Topologie zugeordnet. Weiterhin werden 70 Hausnummernpunkte im Testgebiet I und 61 im Testgebiet II mit falscher Topologie zugeordnet.

	Testgebiet I		Testgebiet II	
	NavTeq	TeleAtlas	NavTeq	TeleAtlas
Nicht Zugeordnet	350 (13,0%)	561 (19,5%)	533 (29,6%)	268 (17,5%)
Zugeordnet mit identischer Topologie	2224 (82,5%)	2224 (77,1%)	1206 (67,0%)	1206 (78,5%)
Zugeordnet mit falscher Topologie	70 (2,5%)	70 (2,4%)	61 (3,4%)	61 (4,0%)

Tabelle 6.16: Ergebnisse der Auswertung der Topologie der Hausnummernpunkte

Abbildung 6.8 stellt die Häufigkeitsverteilung der geometrischen Abweichungen der zugeordneten Hausnummernpunkte dar. Die meisten Abweichungen liegen im Bereich von 20 Metern. Allerdings beträgt die maximale geometrische Abweichung über 200 Meter.

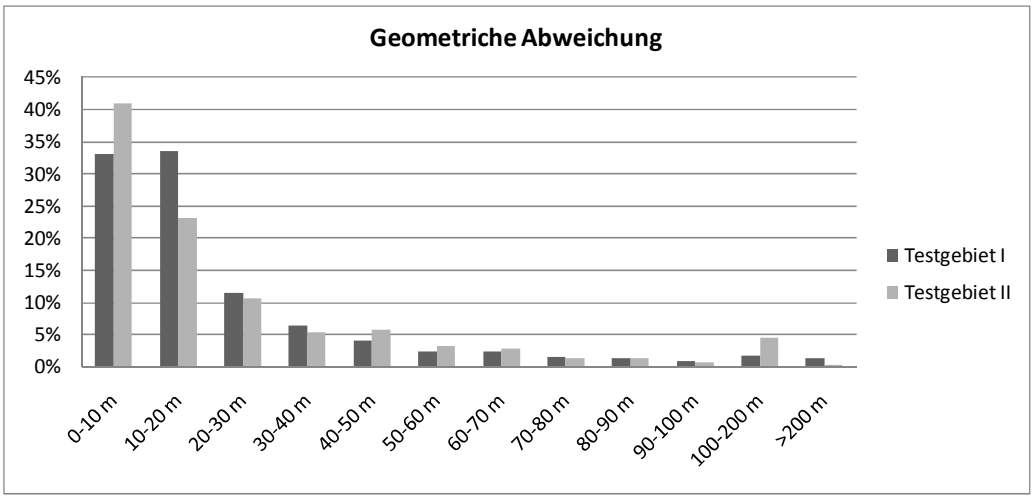


Abbildung 6.8: Ergebnisse der Auswertung der geometrischen Abweichungen der Hausnummernpunkte

6.3 Diskussion der Ergebnisse

Aus den Ergebnissen der globalen Qualitätsauswertung ist festzustellen, dass TeleAtlas im Vergleich zu NavTeq und OpenStreetMap eine höhere Vollständigkeit im Testgebiet II besitzt. Die

Abweichung der Vollständigkeit reduziert sich auf der Ebene der komplexen Objekte. Weiterhin ist die geometrische und topologische Ähnlichkeit im Stadtgebiet höher als im ländlichen Raum. Darüber hinaus ist die geometrische und topologische Ähnlichkeit zwischen NavTeq und TeleAtlas höher als die zwischen TeleAtlas und OpenStreetMap.

Unterschiedliche Verfahren zur Auswertung der Ähnlichkeit eines Zuordnungspaars hinsichtlich der Form, Geometrie, Topologie und Attribute wurden vorgestellt. Die Ähnlichkeit der Form betrachtete die Komplexität der Modellierung in den heterogenen Datensätzen und kann z.B. als Gewichte bei einer Conflation (Verschmelzung) eingesetzt werden. Darüber hinaus wurden die geometrische Ähnlichkeit anhand der Hausdorff-Distanz und die topologische Ähnlichkeit hinsichtlich der Erreichbarkeit für ein Zuordnungspaar berechnet, um eine detaillierte Untersuchung von Zuordnungspaaren zu ermöglichen. Attribute wurden in verschiedene Typen aufgeteilt. Die Ergebnisse zeigen, dass Attribute mit gleichen begrenzten Wertebereichen (insbesondere mit binärer Klassifikation) eine höhere Ähnlichkeit im Vergleich zu anderen Attributtypen haben, da weniger Unsicherheiten bei der Datenerfassung dieser Attribute existieren. Die Ähnlichkeit von Attributen mit gleichen unbegrenzten Wertebereichen ist relativ niedriger. Nach der Untersuchung werden folgende Kenntnisse gewonnen:

- Ein Attribut kann mehrere Werte besitzen, z.B. unterschiedliche Straßennamen „Königstraße“ und „Rotebühlstraße“ für eine gleiche Straße. Bei der Datenerfassung wird der Straßenname von verschiedenen Erfassern unterschiedlich erfasst.

- Ein Attribut kann mit unterschiedlicher Semantik von verschiedenen Erfassern erfasst werden, z.B. „Leonhardstraße“ und „Leonhardplatz“ für dasselbe Objekt.

- Typische Unterschiede von *Road Number* sind z.B. „B27“ in NavTeq und „B27A“ in TeleAtlas.

- Im ländlichen Raum wird manchmal *Road Number* statt *Straßenname* für das Attribut *Straßenname* erfasst.

- In manchen Fällen ist ein Attribut nur in einem Datensatz verfügbar. In diesem Fall lässt sich das Attribut bei der Datenverschmelzung von einem Datensatz zu anderem Datensatz übertragen, um die Qualität der Daten zu verbessern.

Für direkt vergleichbare Attribute lässt sich eine Qualitätsaussage einfacher treffen, weil die Ähnlichkeit von Attributen meistens gleich Null oder Eins ist. Bei Attributen mit unterschiedlichen Wertebereichen wird in den meisten Fällen eine Ähnlichkeit zwischen Null und Eins ermittelt. Die Auswertung von Hausnummern zeigte, dass für einzelne Attribute auch eine geometrische Genauigkeit berechnet werden kann.

In der Zukunft sind weitere Attributtypen zu untersuchen, die eine Kombination der vorgestellten Typen sind. Zum Beispiel ist das Attribut *Direction of Flow* von der Richtung der Kanten abhängig (siehe Anhang A). Weiterhin ist die Kategorie „4“ von *Direction of Flow* in TeleAtlas ein zusammengesetztes Attribut. Daher müssen die Richtungen der Kanten und die semantischen Korrespondenzen für die Kategorie „4“ in NavTeq und OpenStreetMap bestimmt werden. Danach kann dieses Attribut mit dem vorgestellten Verfahren ausgewertet werden.

7 Datenverschmelzung

Dieses Kapitel befasst sich mit der Verschmelzung von verschiedenen Datensätzen, um die Qualität der Daten zu verbessern. Zunächst wird ein übergeordneter Ansatz der Datenverschmelzung präsentiert. Danach erfolgt eine detaillierte Vorstellung von unterschiedlichen Verfahren zur Verschmelzung von zugeordneten und nicht zugeordneten Kanten und Knoten.

7.1 Ansatz der Verschmelzung

Mit Hilfe der Zuordnungsergebnisse lassen sich die Datensätze mit unterschiedlichen Kriterien in Cluster aufteilen, die in Abbildung 7.1 dargestellt werden (Kanten in Großbuchstaben und Knoten in Kleinbuchstaben). Insgesamt werden m Cluster im Datensatz A und n Cluster im Datensatz B generiert. Die Datenverschmelzung findet auf der Ebene der Cluster statt. Die Anzahl der Cluster für zugeordnete Kanten und Knoten muss in den zwei Datensätzen identisch sein. So werden $i\epsilon(1,k)$ Cluster für zugeordnete Kanten und Knoten in den zwei Datensätzen erzeugt. Für nicht zugeordnete Kanten und Knoten werden $j\epsilon(k+1,m)$ Cluster im Datensatz A und $j\epsilon(k+1,n)$ Cluster im Datensatz B berechnet.

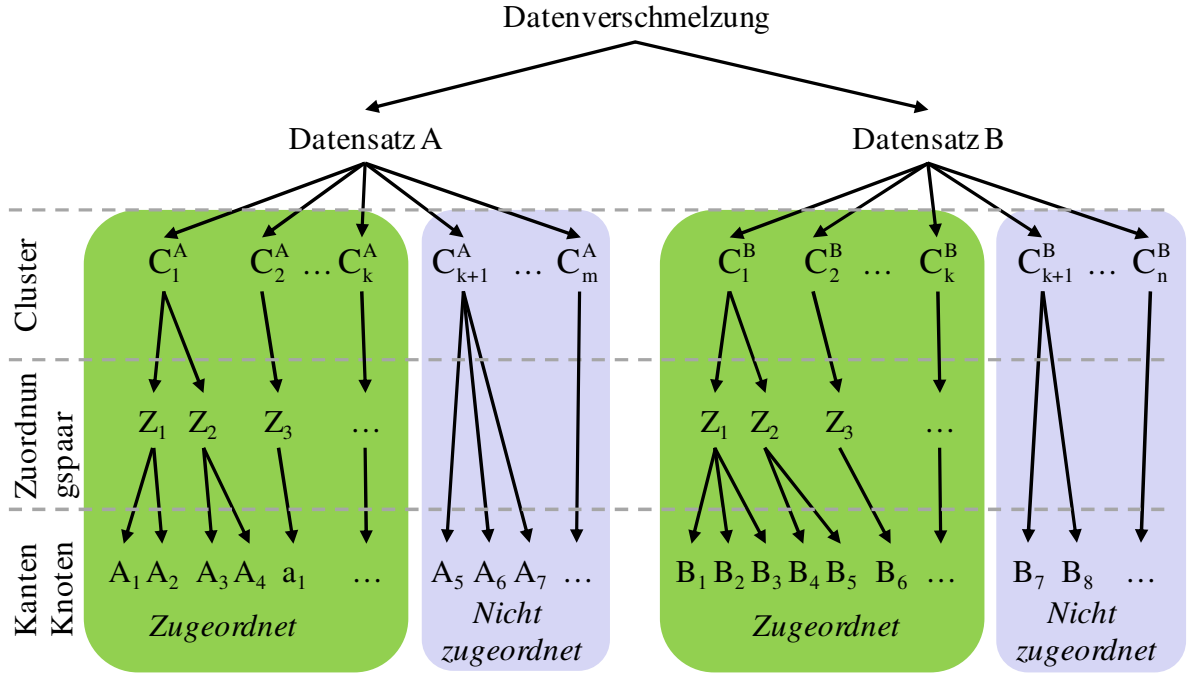


Abbildung 7.1: Cluster zur Datenverschmelzung

Cluster werden an dieser Stelle mit C gekennzeichnet. So gilt für die Datensätze A und B:

$$Datensatz\ A = \sum_{i=1}^{k} C_i^A + \sum_{j=k+1}^{m} C_j^A \text{ und } Datensatz\ B = \sum_{i=1}^{k} C_i^B + \sum_{j=k+1}^{n} C_j^B \qquad \textbf{(7.1)}$$

Die einzelnen Cluster sind mit verschiedenartigen Verfahren zu verschmelzen. Ergebnisse der Verschmelzung sind in einem dritten Datensatz (Enddatensatz) zu speichern, welcher eine Summe

von Clustern vom Datensatz *A* und Datensatz *B* darstellt. So gilt (7.2) für den Enddatensatz, wobei der Operator ⊕ an dieser Stelle als Verschmelzungsoperator definiert wird:

$$Enddatensatz = \sum_{i=1}^{k}\left(C_i^A \oplus C_i^B\right) \oplus \sum_{j=k+1}^{m} C_j^A \oplus \sum_{j=k+1}^{n} C_j^B \qquad (7.2)$$

7.2 Datenverschmelzung von zugeordneten Kanten und Knoten

In diesem Abschnitt werden Verfahren zur Verschmelzung von zugeordneten Kanten und Knoten diskutiert. Zum Zwecke der Datenverschmelzung werden häufig Mittellinien berechnet. Die Unterschiede der geometrischen Modellierung führen jedoch zu Problemen bei der Berechnung der Mittellinien. Aus diesem Grund werden in der vorliegenden Arbeit Zuordnungspaare, die unterschiedliche Formen in den Datensätzen haben, in verschiedene Cluster eingeteilt und mit einer Transformation verschmolzen. Die Zuordnungspaare, die sich durch Berechnung der Mittellinie verschmelzen lassen, werden in der Arbeit auch als Cluster betrachtet. Infolgedessen kann ein Cluster in der Arbeit aus einem Zuordnungspaar oder mehreren Zuordnungspaaren bestehen. Zusammenfassend werden folgende zwei Methoden bei der Datenverschmelzung eingesetzt:

1. *Bildung der Mittellinie* für Cluster mit einem Zuordnungspaar, welches gleiche Form in den Datensätzen hat.

2. *Transformation von Cluster* für Cluster mit einem bzw. mehreren Zuordnungspaaren, welche unterschiedliche Form in den Datensätzen haben.

Um die Problematik aufgrund der unterschiedlichen geometrischen Modellierung zu lösen, wird die Komplexität der geometrischen Modellierung als Parameter bei der Datenverschmelzung eingesetzt (siehe Abbildung 7.2). So ist ein Cluster mit einer *einfachen* oder *komplexen* geometrischen Modellierung (Form) in den Enddatensatz zu transformieren. Ein anderer denkbarer Parameter ist die geometrische Ähnlichkeit (die Hausdorff-Distanz). Überschreitet die Hausdorff-Distanz eines Zuordnungspaars, welches in einem Cluster enthalten ist, einen vorgegebenen Schwellwert, dann wird die Datenverschmelzung für diesen Cluster durchgeführt. Darüber hinaus können unterschiedliche Gewichte für die verschiedenen Datensätze bei der Datenverschmelzung definiert werden, um die Qualitätsunterschiede der Datensätze zu berücksichtigen.

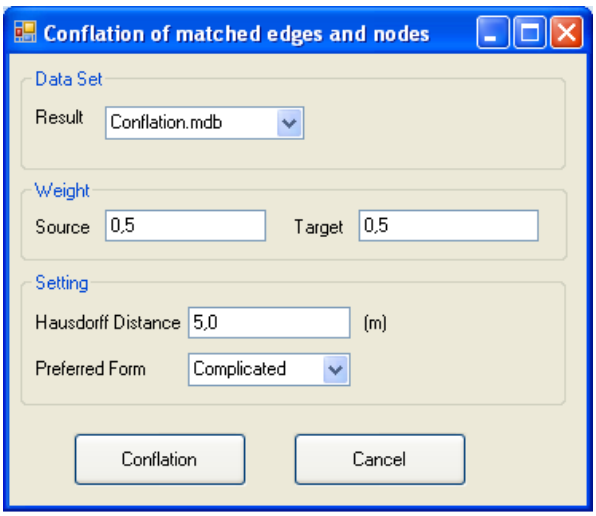


Abbildung 7.2: Parameter bei der Datenverschmelzung von zugeordneten Kanten und Knoten

Im Folgenden werden zunächst *Verbindungsknoten* ermittelt, welche die Cluster miteinander verbinden. Anschließend wird auf die Datenverschmelzung mittels *Bildung der Mittellinie* und *Transformation von Cluster* eingegangen. Zum Schluss werden die Ergebnisse der Datenverschmelzung anhand von Beispielen diskutiert.

7.2.1 Ermittlung von Verbindungsknoten

Verbindungsknoten werden mit Hilfe der Ergebnisse der Knoten- und Kantenzuordnung berechnet. Um die Komplexität der Datenverschmelzung zu minimieren, ist die Größe der Cluster so klein wie möglich zu halten. Es werden folgende Knoten als Verbindungsknoten definiert:

1. Knoten, die bei der Knotenzuordnung mit der Relation *1:1* zugeordnet sind.
2. Knoten, die bei der manuellen Zuordnung zu Kanten zugeordnet sind.

Zur Ermittlung der Verbindungsknoten sind im ersten Schritt die Knotenpaare mit der Relation *1:1* aufzufinden. Für diese Knoten wird ihr Mittelpunkt berechnet und in die Liste der Verbindungsknoten eingetragen. Im Weiteren werden die Zuordnungspaare mit der Relation *p:1*, *p:n*, *1:p* oder *n:p* behandelt. In Abbildung 7.3 werden die Verbindungsknoten mit gestrichelten Linien dargestellt. Der Knoten a_1 im Datensatz *A* ist bei der manuellen Zuordnung zu der Kante B_1 im Datensatz *B* zugeordnet. Nach der automatischen Knotenzuordnung wird der Knoten a_1 dann den Knoten b_1 und b_2 zugeordnet. Ist eine einfache Form bevorzugt, dann wird der Mittelpunkt von a_1, b_1 und b_2 in die Liste der Verbindungsknoten hinzugefügt. Ist die komplexe Form als Parameter eingestellt, werden die Knoten b_1 und b_2 mit einem Translationsvektor transformiert und in die Liste der Verbindungsknoten eingetragen. Gleichzeitig lässt sich die Kante B_1 mit dem gleichen Translationsvektor in den Enddatensatz transformieren.

Im Folgenden wird die Berechnung der Geometrie von Verbindungsknoten für Zuordnungspaare mit der Relation *p:1*, *p:n 1:p* bzw. *n:p* vorgestellt. Es wird angenommen, dass die Form im Datensatz *A* einfacher ist. Das heißt, dass ein Knoten a im Datensatz *A* zu mehreren Knoten b_i im Datensatz *B* zugeordnet wird, wobei $i\epsilon(1,n)$. Wird die einfache Form bevorzugt, dann ist die Geometrie des Verbindungsknotens c nach Gleichung (7.3) zu berechnen:

$$c = a + \frac{1}{n}\sum_{i=1}^{n} b_i \tag{7.3}$$

Ist allerdings die komplexe Form eingestellt, dann wird zunächst ein Translationsvektor $\vec{v}$ nach (7.4) berechnet:

$$\vec{v} = \left(\frac{1}{n}\sum_{i=1}^{n} b_i\,, a + \frac{1}{n}\sum_{i=1}^{n} b_i\right) \tag{7.4}$$

Weiterhin werden die Verbindungsknoten c_i im Enddatensatz durch Transformation der Knoten b_i mit dem Translationsvektor $\vec{v}$ ermittelt:

$$c_i = \vec{v} \cdot b_i, i\epsilon(1,n). \tag{7.5}$$

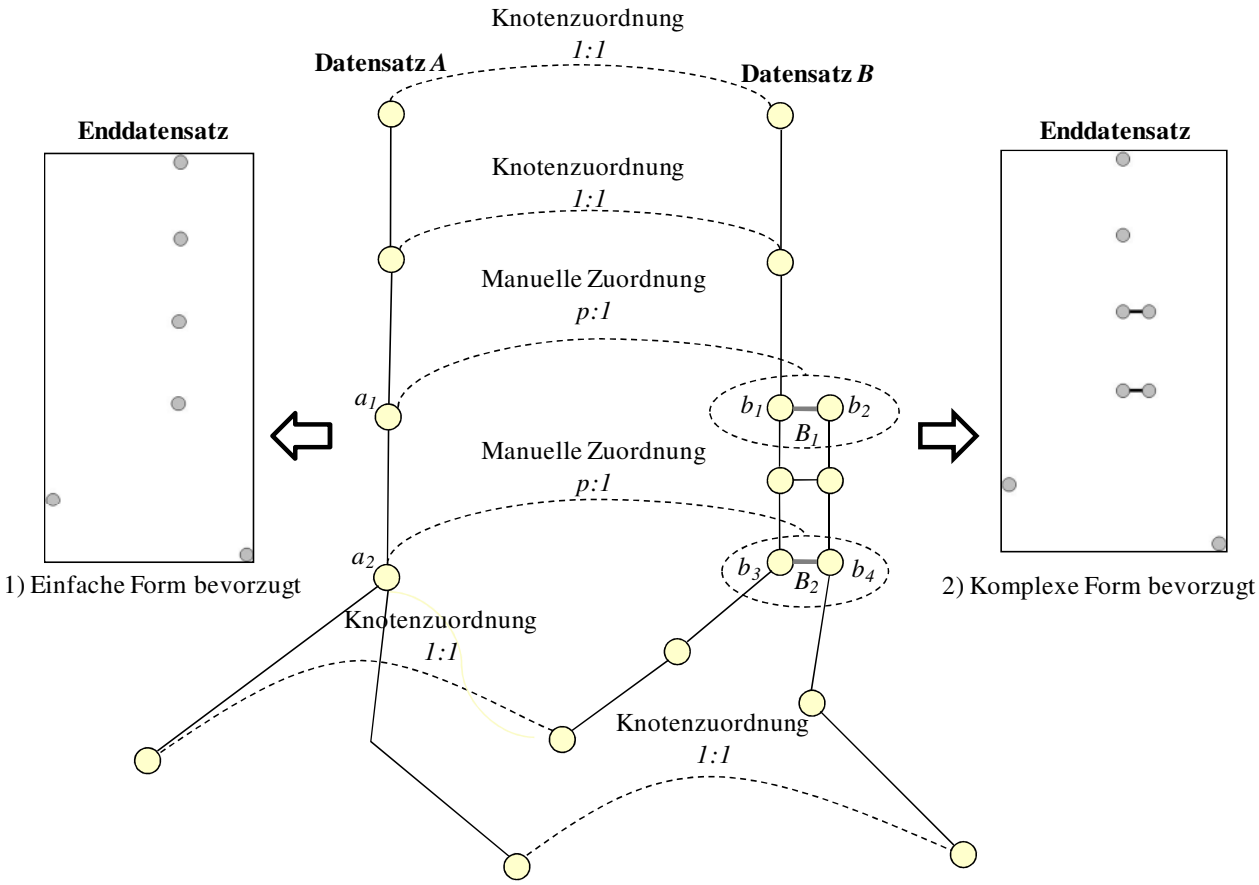


Abbildung 7.3: Ermittlung von Verbindungsknoten

Werden unterschiedliche Gewichte für die Datensätze eingesetzt und die einfache Form eingestellt, dann ist die Geometrie des Verbindungsknotens c nach Gleichung (7.6) zu berechnen, wobei ω_A und ω_B die Gewichte für die Datensätze *A* und *B* sind:

$$c = \omega_A a + \frac{\omega_B}{n} \sum_{i=1}^{n} b_i \qquad \textbf{(7.6)}$$

Wird die komplexe Form bevorzugt, dann ist der Translationsvektor $\vec{v}$ nach (7.7) zu ermitteln:

$$\vec{v} = \left(\frac{1}{n}\sum_{i=1}^{n} b_i\,,\ \omega_A a + \frac{\omega_B}{n}\sum_{i=1}^{n} b_i\right) \qquad \textbf{(7.7)}$$

7.2.2 Bildung der Mittellinie

Das Verfahren *Bildung der Mittellinie* setzt voraus, dass die Cluster gleiche Form in den Datensätzen haben. Aus diesem Grund werden alle Zuordnungspaare mit der Formzuordnung „Simple“ zu „Simple“ selektiert. Zunächst wird geprüft, ob der Anfangs- und Endknoten eines Zuordnungspaars ein Verbindungsknoten ist. Wenn ja, dann ist dieses Zuordnungspaar ein Cluster, welcher über Bildung der Mittellinie zu verschmelzen ist. In Abbildung 7.4 werden die Ergebnisse der Datenverschmelzung durch Bildung der Mittellinie veranschaulicht (vgl. Abbildung 7.3). Die Zuordnungspaare ($A_1 \leftrightarrow B_3$; $A_2 \leftrightarrow B_4$; $A_3 \leftrightarrow B_5 B_6$; $A_4 \leftrightarrow B_7 B_8$) sind die Cluster, die durch Bildung der

Mittellinie verschmolzen werden. Je nachdem, welcher Parameter für die geometrische Modellierung eingestellt ist, ergeben sich zwei unterschiedliche Ergebnisse (siehe *Enddatensatz*). Der Enddatensatz mit der Einstellung in einfacher Form verfügt über eine geringe Anzahl der Kanten und Knoten als der Enddatensatz mit der Einstellung in komplexer Form.

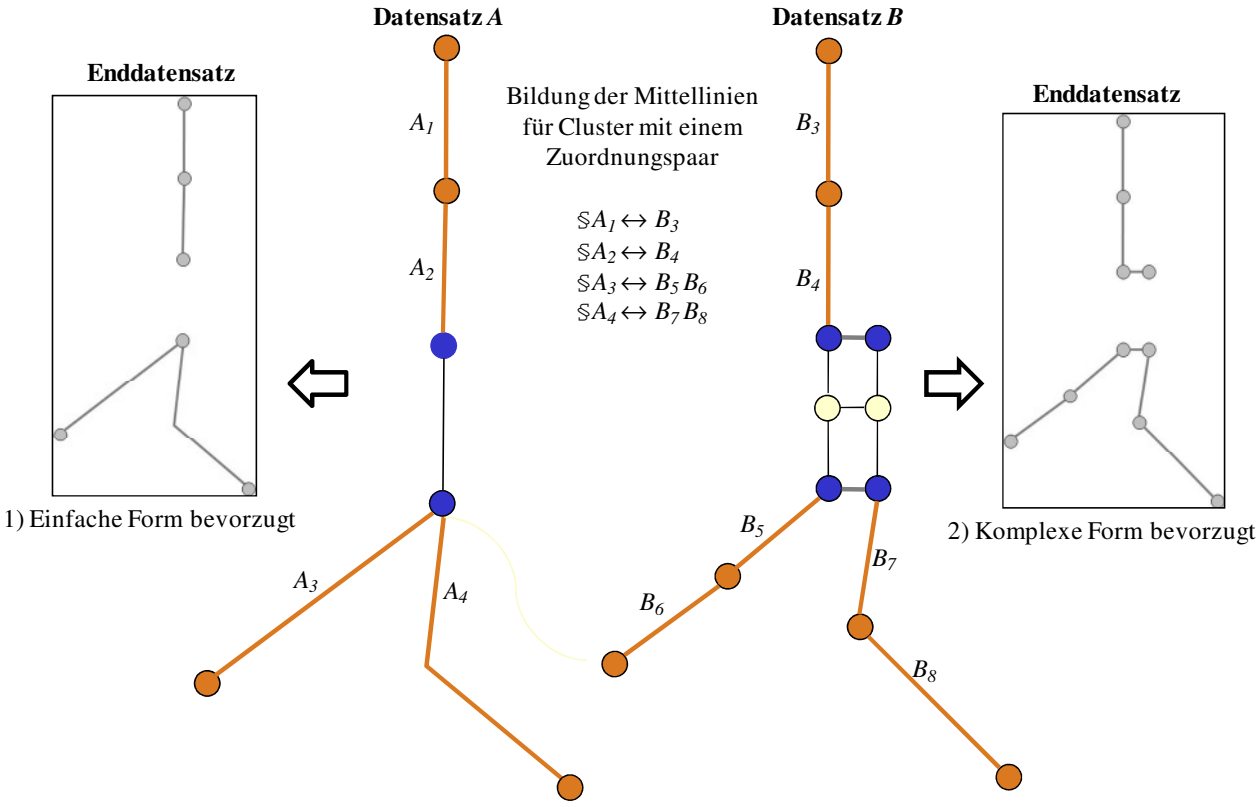


Abbildung 7.4: Datenverschmelzung über Bildung der Mittellinie (vgl. Abbildung 7.3)

Zur Berechnung der Mittellinie von zwei Polylinien wird in den meisten Fällen versucht, eine gleiche Anzahl von Zwischenpunkten für die zwei Polylinien zu erzielen. Zu diesem Zweck können die Zwischenpunkte z.B. nach der Distanz entlang der Polylinie mit vorgegebenen Schritten (z.B. 100) berechnet werden [Keul 1998]. So entstehen jeweils 100 Zwischenpunkte für jede Polylinie, die gleichmäßig entlang der Polylinie verteilt sind. Die Mittellinie ergibt sich aus den 100 Mittelpunkten der Zwischenpunkte der zwei Polylinien. Das Verfahren ist zwar sehr einfach zu implementieren, kann allerdings zu einer großen Anzahl von Zwischenpunkten führen. Außerdem wird die Form bzw. der Verlauf der Polylinie nicht berücksichtigt.

In der vorliegenden Arbeit wird die Lotdistanz zur Interpolation der Zwischenpunkte verwendet (siehe Abbildung 7.5). Zunächst ist die Lotdistanz von allen Zwischenpunkten von der Linie *A* zu der Linie *B* zu berechnen und dann umgekehrt (siehe Schritt *1*). Für die originalen und interpolierten Zwischenpunkte wird die Distanz entlang der Polylinie berechnet. Anschließend werden diese Punkte nach den ermittelten Distanzen sortiert und zugeordnet (siehe Schritt *2*). Weiterhin sind die Mittelpunkte mit den entsprechenden Gewichten für die Datensätze zu berechnen. Die Bildung der Mittellinie erfolgt durch Verbindung der Mittelpunkte (siehe Schritt *3*). Die Anzahl der Zwischenpunkte der Mittellinie ist dann die Summe der Anzahl der Zwischenpunkte von Linie *A* und Linie *B*. Um die Konnektivität nach der Datenverschmelzung beizubehalten, werden die Anfangs- und Endknoten der Mittellinie an die Verbindungsknoten angepasst. Das Verfahren lässt sich sehr einfach mit den vorhandenen ArcGIS-Funktionen implementieren und liefert sehr gute Ergebnisse.

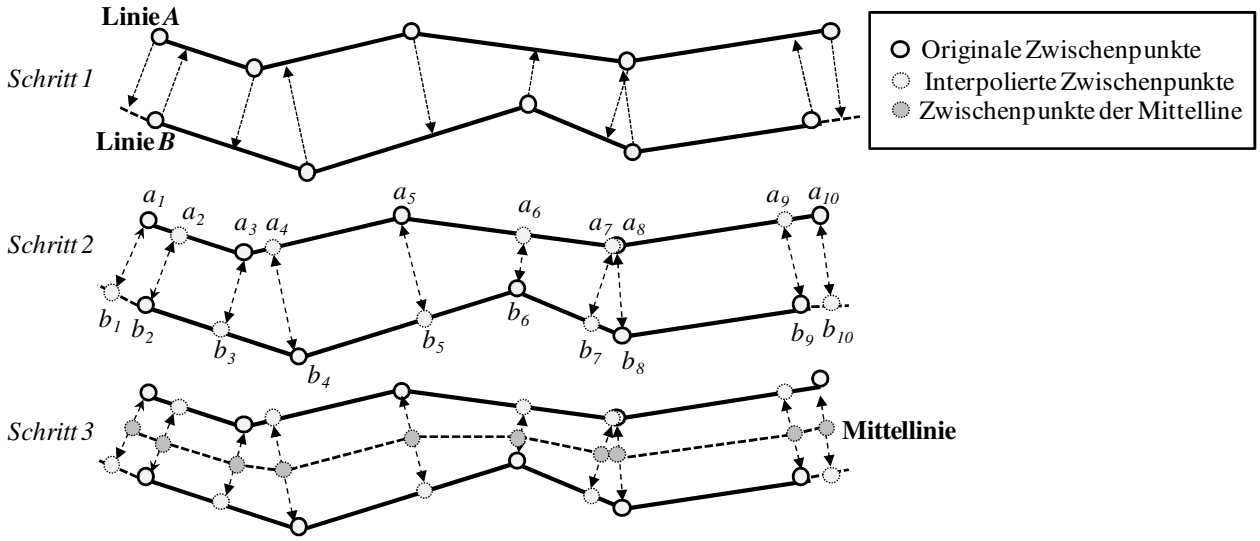


Abbildung 7.5: Ermittlung der Mittellinie aus zwei Polylinien

Für Cluster mit unterschiedlicher Konnektivität kann die Mittellinie verlängert werden, um die Konnektivität im Enddatensatz zu verbessern. In Abbildung 7.6 werden zwei Zuordnungspaare ($A_1 \leftrightarrow B_1$ und $A_2 \leftrightarrow B_2$) erfasst. Die Kante A_1 im Datensatz A schließt keine weiteren Kanten am Knoten a_2 an. Jedoch schließt die Kante B_1 im Datensatz B die Kante B_2 am Knoten b_2 an. Nach der Berechnung von Mittellinien wird die Konnektivität im Enddatensatz gleich an den Datensatz A angepasst (siehe Abbildung 7.6a). Um die Konnektivität wie im Datensatz B zu erhalten, wird die Kante B_1 in zwei Abschnitte (b_1b_4) und (b_4b_2) aufgeteilt (siehe Abbildung 7.6b). Eine Mittelinie (c_1c_2) von Abschnitten (a_1a_2) und (b_1b_4) wird berechnet. Anschließend wird der Abschnitt (b_4b_2) mit der Knotenzuordnung ($b_4 \rightarrow c_2$ und $b_2 \rightarrow c_3$) in den Enddatensatz transformiert.

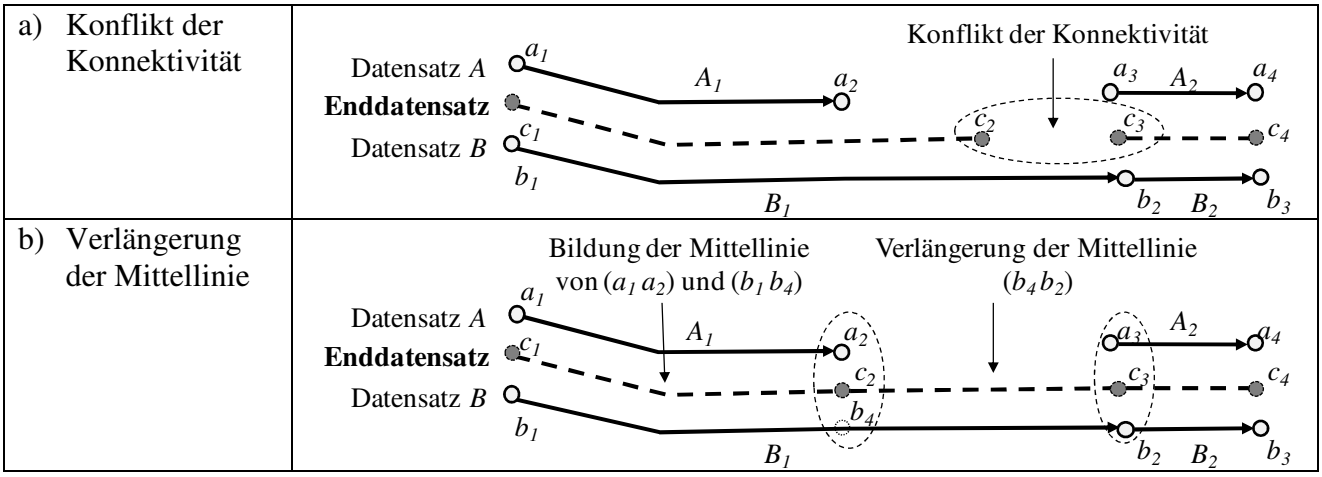


Abbildung 7.6: Verlängerung der Mittellinie zum Beibehalten der Konnektivität

7.2.3 Transformation von Cluster

Die verbliebenen Zuordnungspaare sind mit der Methode *Transformation von Cluster* zu verschmelzen (siehe Abbildung 7.7). Kanten der verbliebenen Zuordnungspaare werden selektiert und in eine Liste (*Kantenliste*) eingetragen (siehe Schritt *a*). Zunächst wird eine einelementige Liste (*Liste für Cluster*) erzeugt, die aus einer beliebigen Kante der *Kantenliste* besteht (siehe Schritt *c*).

Anhand der Konnektivität werden die Kanten in verschiedene Cluster aufgeteilt. Um die Größe der Cluster zu reduzieren, werden die Verbindungsknoten als Abbruchbedingungen eingeführt. Es wird so lange iteriert, bis keine Kanten in der *Kantenliste* gefunden werden, die mit den Kanten in der *Liste für Cluster* verbunden sind (siehe Schritt *d*). Eine Kante wird aus der *Liste für Cluster* selektiert und dann geprüft, ob der Anfangsknoten dieser Kante ein Verbindungsknoten ist (siehe Schritt *e* und *f*). Wenn nein, dann werden alle Kanten in der *Kantenliste*, die diese Kante am Anfangsknoten verbinden, selektiert und in die *Liste für Cluster* eingetragen (siehe Schritt *g*). Gleichzeitig werden diese Kanten aus der *Kantenliste* entfernt. Genauso wird der Endknoten dieser Kante behandelt (siehe Schritt *h* und *i*).

Anschließend sind die Parameter der Transformation mit Hilfe der gefundenen Verbindungsknoten zu berechnen, die in diesem Cluster enthalten sind (siehe Schritt *j*). Weiterhin werden die Gewichte der geometrischen Modellierung des Clusters im Datensatz *A* und Datensatz *B* berechnet (siehe Schritt *k*). Anhand der Transformationsparameter sind die Kanten im Cluster vom Datensatz *A* bzw. Datensatz *B* in den Enddatensatz zu transformieren (siehe Schritt *l*). Die Gewichte für die Datensätze wurden bereits bei der Berechnung der Verbindungsknoten verwendet und werden daher bei der Berechnung der Parameter der Transformation berücksichtigt.

a) Bilde eine Liste aus Kanten der verbliebenen Zuordnungspaare (*Kantenliste*)			
b) Bis die *Kantenliste* leer ist			
	c) Bilde eine einelementige Liste (*Liste für Cluster*)		
	d) Bis keine Kanten gefunden werden, die mit den Kanten in der *Liste für Cluster* verbunden sind		
		e) Selektiere eine Kante aus der *Liste für Cluster*	
		f) Ist der Anfangsknoten dieser Kante ein *Verbindungsknoten* ?	
		Nein	*Ja*
		g) Selektiere alle Kanten in der *Kantenliste*, die diese Kante am Anfangsknoten verbinden. Trage die gefunden Kanten in die *Liste für Cluster* ein und entfernt sie aus der *Kantenliste*.	Nichts tun.
		h) Ist der Endknoten dieser Kante ein *Verbindungsknoten* ?	
		Nein	*Ja*
		i) Selektiere alle Kanten in der *Kantenliste*, die diese Kante am Endknoten verbinden. Trage die gefunden Kanten in die *Liste für Cluster* ein und entfernt sie aus der *Kantenliste*.	Nichts tun.
	j) Berechnung der Parameter der Helmert-Transformation auf Basis der *Verbindungsknoten* in der *Liste für Cluster*		
	k) Berechnung der Gewichte der geometrischen Modellierung im Datensatz *A* und Datensatz *B*		
	l) Transformation des Clusters vom Datensatz *A* oder Datensatz *B* je nach der Komplexität der geometrischen Modellierung		

Abbildung 7.7: Ablaufdiagramm für Transformation von Cluster

In Abbildung 7.8 wird ein Beispiel für die Datenverschmelzung mit dem Verfahren Transformation von Cluster dargestellt (vgl. Abbildung 7.4). Abhängig vom Parameter für die geometrische Modellierung werden die Kante (A_5) im Datensatz *A* oder die Kanten ($B_9B_{10}B_{11}B_{12}B_{13}$) im Datensatz *B*, die in Blau (Dunkel) dargestellt sind, in den Enddatensatz transformiert. Aus diesem Grund sind zwei unterschiedliche Ergebnisse nach der Datenverschmelzung möglich. Die Geometrie der

Verbindungsknoten soll nach der Transformation unverändert bleiben. So sind die Anfangs- und Endknoten der transformierten Kanten an die Verbindungsknoten anzupassen.

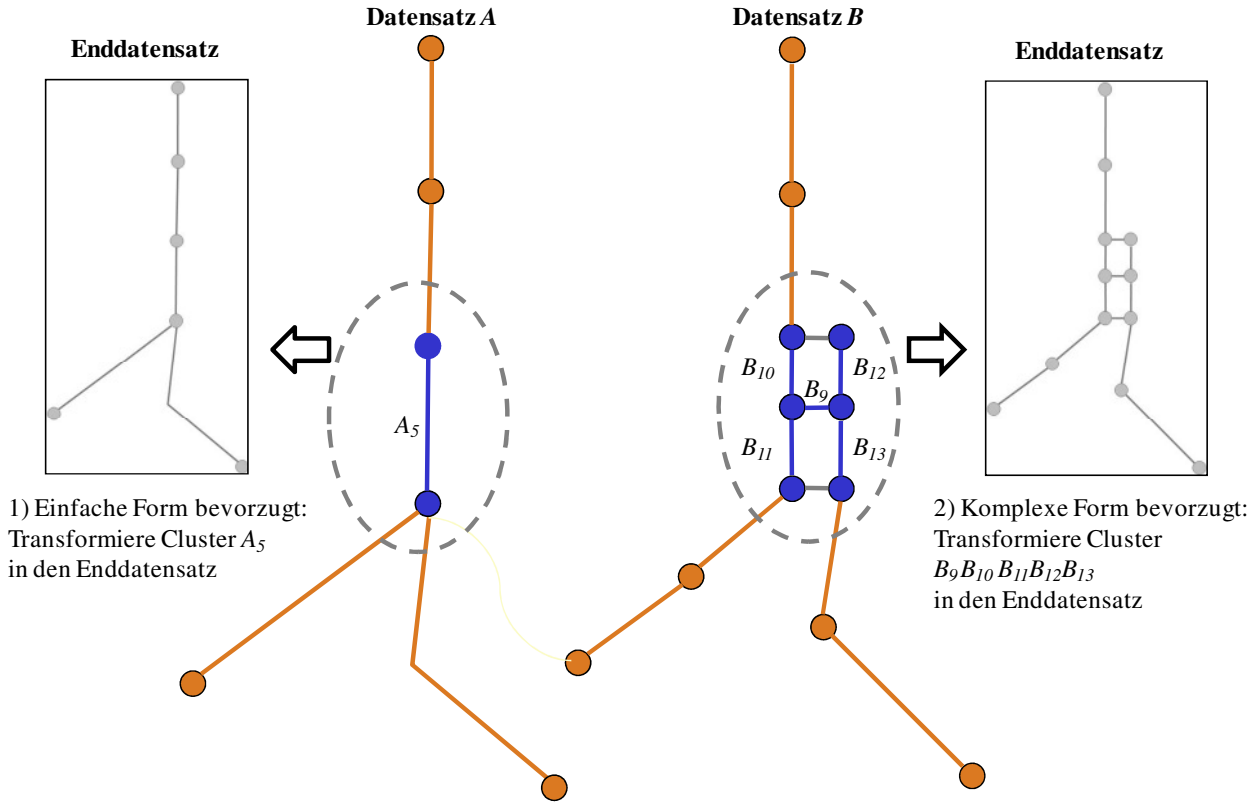


Abbildung 7.8: Datenverschmelzung über Transformation von Cluster

7.2.4 Diskussion der Ergebnisse

In diesem Abschnitt werden die Ergebnisse der Verschmelzung von den zugeordneten Kanten und Knoten diskutiert. In Abbildung 7.9 werden Beispiele der Datenverschmelzung mit dem Verfahren Bildung der Mittellinie dargestellt. In Abbildung 7.9a handelt es sich um eine einfache Datenverschmelzung von NavTeq und TeleAtlas in einem Wohngebiet. Der Kreisverkehr in Abbildung 7.9b wird in NavTeq und TeleAtlas jeweils mit fünf Kanten modelliert. Bei der Zuordnung wurden dann fünf Zuordnungspaare mit der Relation *1:1* erfasst. Die einzelnen Zuordnungspaare werden mit dem Verfahren Bildung der Mittellinie verschmolzen. In Abbildung 7.9c wird das Ergebnis der Datenverschmelzung für den gleichen Kreisverkehr in TeleAtlas und OpenStreetMap dargestellt. In Abbildung 7.9d wird die Mittellinie im Enddatensatz verlängert, um die Konnektivität (wie in OpenStreetMap) beizubehalten.

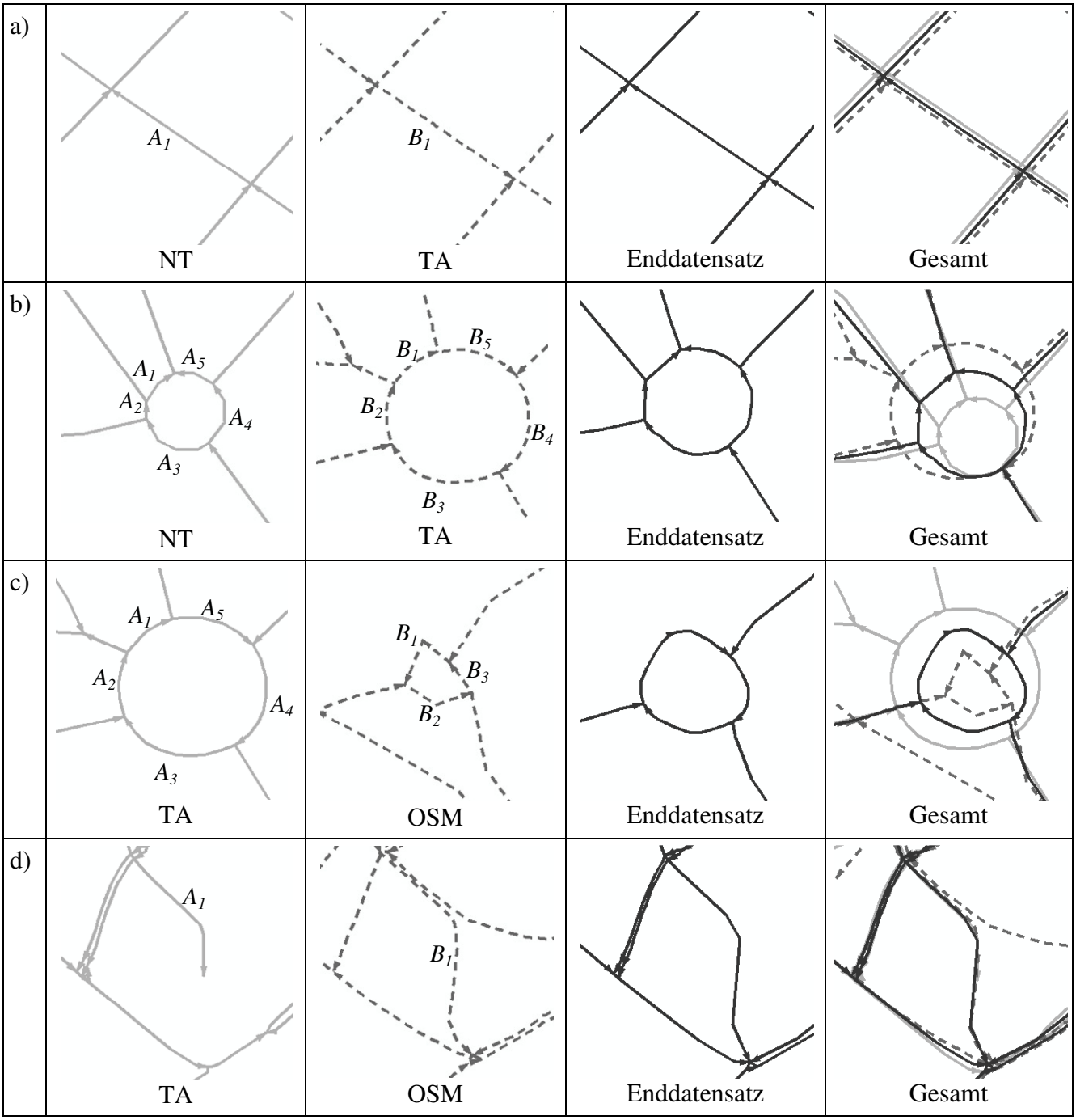


Abbildung 7.9: Beispiele der Datenverschmelzung über Bildung der Mittellinie

In Abbildung 7.10 sind Beispiele der Datenverschmelzung über Transformation von Cluster dargestellt. Dabei ist die komplizierte Form bevorzugt. In Abbildung 7.10a werden die Kanten von NavTeq in den Enddatensatz transformiert, da es sich um eine Formzuordnung zwischen „Fork1“ in NavTeq und „Simple“ in TeleAtlas handelt. In Abbildung 7.10b ist der Kreisverkehr in TeleAtlas mit einem Knoten und in OpenStreetMap mit vier Kanten modelliert. Da die komplexe Form bevorzugt ist, ergeben sich vier Kanten im Enddatensatz. In Abbildung 7.10c wird ebenfalls die komplexe Form „Parallel“ in den Enddatensatz übertragen.

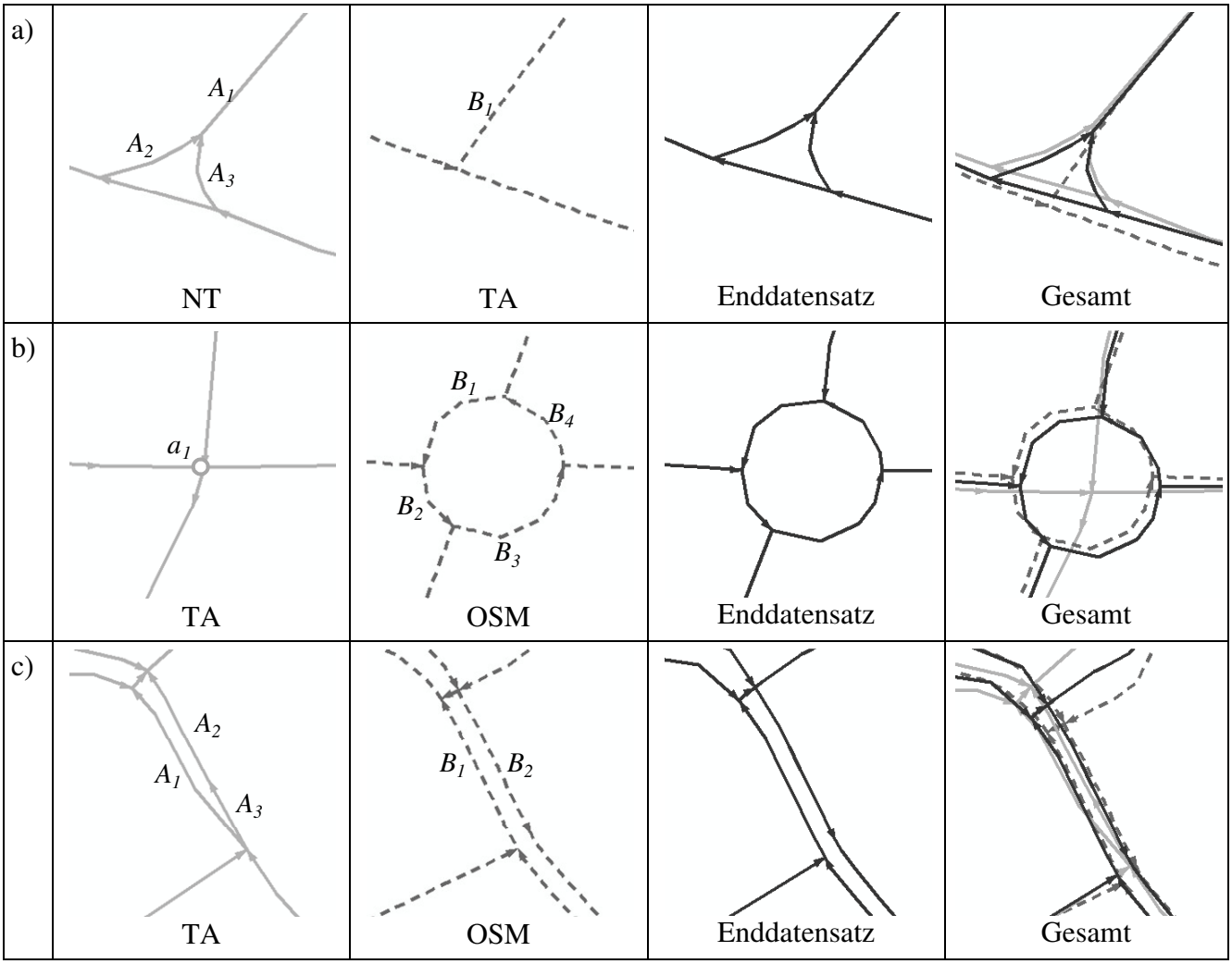


Abbildung 7.10: Beispiele der Datenverschmelzung über Transformation von Cluster

In Tabelle 7.1 werden die Zuordnungspaare nach den unterschiedlichen Verfahren der Datenverschmelzung klassifiziert. Die Zuordnungspaare zwischen Knoten und Kanten wurden bei der Ermittlung von Verbindungsknoten behandelt. Die meisten Cluster (>87%) in den zwei Testgebieten wurden über Bildung der Mittellinie verschmolzen. Die Größe der Cluster (Anzahl der Zuordnungspaare), die mit der Transformation von Cluster verschmolzen wurden, wird ebenfalls dargestellt. Aufgrund der unterschiedlichen Erfassungen von komplexen Kreuzungen sind die Cluster im Stadtgebiet (Testgebiet I) größer als im ländlichen Raum (Testgebiet II). Alle Cluster zur Verschmelzung von TeleAtlas und OpenStreetMap im Testgebiet II enthalten nur ein Zuordnungspaar. Bei der Verschmelzung von TeleAtlas und OpenStreetMap im Testgebiet I wurde allerdings ein Cluster mit 17 Zuordnungspaaren erzeugt.

		NavTeq & TeleAtlas		TeleAtlas & OpenStreetMap	
		Testgebiet I	Testgebiet II	Testgebiet I	Testgebiet II
Zuordnung von Knoten und Kanten		46 (5,62%)	22 (3,79%)	41 (4,89%)	10 (5,75%)
Bildung der Mittellinie		729 (89,01%)	545 (93,80%)	757 (90,34%)	152 (87,35%)
Transformation von Cluster		44 (5,37%)	14 (2,41%)	40 (4,77%)	12 (6,90%)
	1 Zuordnungspaar	28 (3,42%)	11 (1,89%)	30 (3,61%)	12 (6,90%)
	2 Zuordnungspaare	-	1 (0,17%)	1 (0,12%)	-
	3 Zuordnungspaare	4 (0,49%)	-	2 (0,24%)	-
	4 Zuordnungspaare	5 (0,61%)	2 (0,34%)	-	-
	5 Zuordnungspaare	2 (0,24%)	-	2 (0,24%)	-
	6 Zuordnungspaare	3 (0,37%)	-	2 (0,24%)	-
	7 Zuordnungspaare	-	-	1 (0,12%)	-
	8 Zuordnungspaare	1 (0,12%)	-	-	-
	9 Zuordnungspaare	1 (0,12%)	-	-	-
	17 Zuordnungspaare	-	-	1 (0,12%)	-

Tabelle 7.1: Cluster bei der Datenverschmelzung von zugeordneten Kanten und Knoten

Die Ergebnisse zeigen, dass der Ansatz sowohl für Stadtgebiete als auch für den ländlichen Raum geeignet ist. Abbildung 7.11 stellt eine kritische Stelle bei der Datenverschmelzung von zugeordneten Kanten und Knoten dar. In diesem Fall liegt eine Straße unter einer anderen Straße. Bei der Datenerfassung wurden in der gleichen Lageposition zwei Knoten (b_1, b_2) mit unterschiedlicher Höhe in TeleAtlas erfasst. Die Höhe der Knoten wird bei der Datenverschmelzung nicht berücksichtigt. Die zwei Knoten b_1 und b_2 werden als unterschiedliche Verbindungsknoten behandelt und bei der Verschmelzung mit unterschiedlichen Parametern transformiert, weil sie in unterschlichen Clustern enthalten sind. Nach der Verschmelzung ist die Lage der zwei Knoten unterschiedlich. Um das Problem zu lösen, sind die zwei Knoten in der Zukunft als ein Verbindungsknoten bei der Datenverschmelzung zu betrachten.

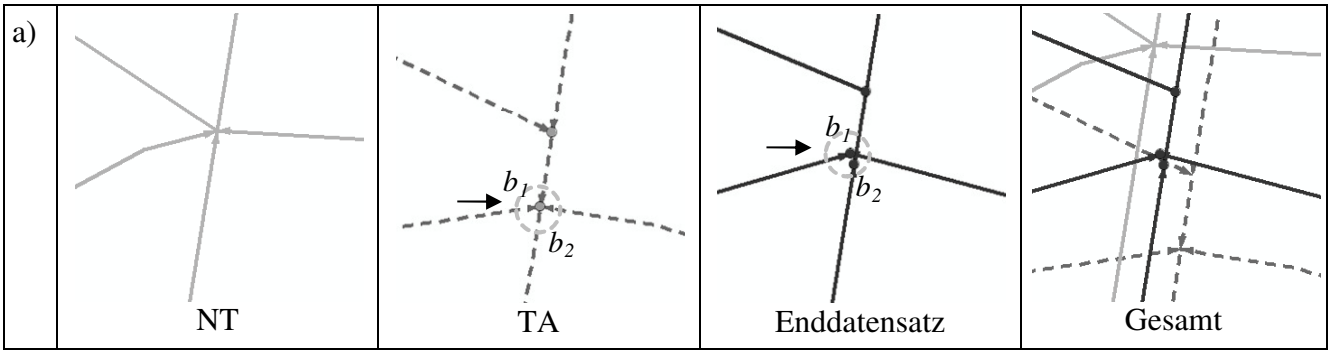


Abbildung 7.11: Kritische Stelle bei der Datenverschmelzung

Nach der Verschmelzung der Geometrie sind die Attribute in den Enddatensatz zu übertragen. Direkt vergleichbare Attribute lassen sich ohne weitere Verarbeitung direkt übertragen. Entsteht ein Konflikt, dann können z.B. die Attribute mit höherer Aktualität in den Enddatensatz übertragen werden. Attribute mit unterschiedlichen Klassifikationen sind mit der Konfusionsmatrix zu homogenisieren. Weiterhin können Attribute, die nur in OpenStreetMap verfügbar sind (wie z.B. minimale und maximale zugelassene Höhe), in den Enddatensatz transformiert werden.

7.3 Datenverschmelzung von nicht zugeordneten Kanten und Knoten

In diesem Abschnitt wird die Datenverschmelzung von nicht zugeordneten Kanten und Knoten präsentiert. Zunächst wird der Ansatz der Datenverschmelzung erklärt. Danach erfolgt eine Diskussion über die Ergebnisse der Datenverschmelzung.

7.3.1 Ansatz zur Verschmelzung von nicht zugeordneten Kanten und Knoten

Der Ansatz zur Verschmelzung von nicht zugeordneten Kanten und Knoten wird in Abbildung 7.12 dargestellt. Aus den nicht zugeordneten Kanten in jedem Datensatz (Datensatz *A* oder *B*) wird jeweils eine *Kantenliste* erzeugt. Im ersten Schritt sind die Kanten in der *Kantenliste* in verschiedene Cluster aufzuteilen. Anschließend sind *Verlinkungsknoten* für einzelne Cluster zu ermitteln und Transformationsparameter daraus zu berechnen. Zum Schluss werden die einzelnen Cluster mit den ermittelten Transformationsparametern in den Enddatensatz übertragen. Der Prozess wird so lange iteriert, bis alle Kanten in der *Kantenliste* behandelt werden.

Bilde eine Liste aus den nicht zugeordneten Kanten (*Kantenliste*)			
Bis die *Kantenliste* leer ist			
	Bilde eine einelementige Liste (*Liste für Cluster*)		
	Bis keine Kanten in der *Kantenliste* gefunden werden, die mit den Kanten in der *Liste für Cluster* verbunden sind		
		Selektiere eine Kante aus der *Kantenliste*	
		Ist diese Kante mit den Kanten in der *Liste für Cluster* verbunden *?*	
		Ja	Nein
		Trage diese Kante in die *Liste für Cluster* ein und entferne sie aus der *Kantenliste*	Nichts tun
	Suche den Verlinkungsknoten in diesem Cluster		
	Berechnung der Transformationsparameter		
	Übertragung des Clusters		

Abbildung 7.12: Ansatz zur Verschmelzung von nicht zugeordneten Kanten und Knoten

Bildung von Clustern

Analog zu Kapitel 7.2.3 werden die Cluster nach der Konnektivität der Kanten berechnet (siehe Abbildung 7.12). Zunächst wird eine einelementige Liste (*Liste für Cluster*) erzeugt, die aus einer beliebigen Kante der *Kantenliste* besteht. Die anderen Kanten in der *Kantenliste*, die mit dieser Kante verbunden sind, werden selektiert und in die *Liste für Cluster* eingetragen. Gleichzeitig werden diese Kanten aus der *Kantenliste* entfernt. Es wird so lange iteriert, bis keine Kanten in der *Kantenliste* gefunden werden, die mit den Kanten in der *Liste für Cluster* verbunden sind.

Ermittlung von Verlinkungsknoten

Verlinkungsknoten sind Knoten, die in einem Cluster enthaltenen sind und sich auf einer zugeordneten Kante befinden. Aus diesem Grund weisen diese Knoten eine Verlinkung zwischen den Datensätzen auf. Zunächst sind diese Knoten im Cluster (z.B. im Datensatz *A*) zu finden. Danach sind die Korrespondenzen für diese Knoten im Enddatensatz mit Hilfe der Ergebnisse der Knotenzuordnung zu ermitteln. Werden keine Korrespondenzen für einen Knoten im Enddatensatz gefunden, dann ist dieser Knoten im Enddatensatz zu interpolieren. Dabei werden zwei Methoden (Distanz entlang der Kante und kürzeste Distanz) eingesetzt. Zunächst wird ein Knoten nach der Distanz entlang der Kante berechnet. Dann wird die Entfernung zwischen den Knoten in den zwei Datensätzen berechnet. Überschreitet die Entfernung einen vorgegebenen Schwellwert, dann wird der Knoten nach der kürzesten Distanz interpoliert.

Berechnung von Transformationsparametern

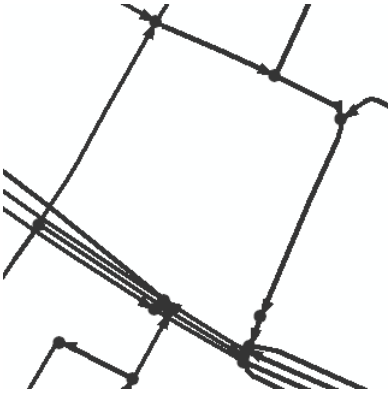
Abhängig von der Anzahl der gefundenen Verlinkungsknoten in einem Cluster wird eine unterschiedliche Anzahl von Transformationsparametern berechnet:

- Anzahl der Verlinkungsknoten= 1: 2 Parameter (2 Translationen)
- Anzahl der Verlinkungsknoten= 2: 4 Parameter (1 Maßstab, 1 Rotation, 2 Translationen)
- Anzahl der Verlinkungsknoten= 3: 6 Parameter (2 Maßstäbe, 2 Rotationen, 2 Translationen)

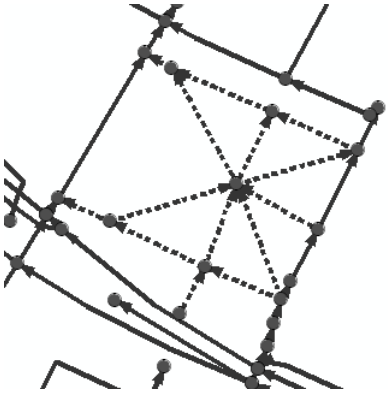
Übertragung der Cluster

Im letzten Schritt sind die Kanten im Cluster mit den ermittelten Transformationsparametern in den Enddatensatz zu übertragen. Die Geometrie der Verlinkungsknoten im Enddatensatz soll nach der Transformation unverändert bleiben, um die Konnektivität mit anderen Kanten beibehalten zu können. Die interpolierten Verlinkungsknoten sind als neue Knoten im Enddatensatz zu speichern. Gleichzeitig sind die Kanten zu zerlegen, worauf sich die interpolierten Verlinkungsknoten befinden. Weiterhin sind Attribute (z.B. Länge der Kante sowie ID des Anfangs- und Endknotens) für die neuen Kanten zu berechnen.

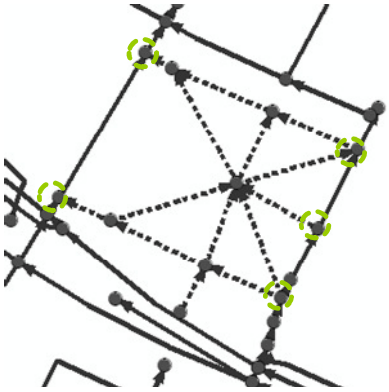
In Abbildung 7.13 werden die einzelnen Schritte der Datenverschmelzung von den nicht zugeordneten Kanten und Knoten anhand eines Beispiels veranschaulicht. Dabei werden die nicht zugeordneten Kanten und Knoten von NavTeq in den Enddatensatz übertragen, welcher nach der Datenverschmelzung der zugeordneten Kanten und Knoten von NavTeq und TeleAtlas entstanden ist (siehe Abbildung 7.13a). Abbildung 7.13b zeigt die Ergebnisse der Clusterbildung von den nicht zugeordneten Kanten in NavTeq, die mit gestrichelten Linien dargestellt werden. Die Knoten, die mit einem Kreis in Abbildung 7.13c dargestellt werden, sind die Verlinkungsknoten in NavTeq. Da keine korrespondierenden Knoten im Enddatensatz verfügbar sind, werden diese Knoten nach der Distanz entlang der Kante interpoliert. Die übertragenen Kanten im Cluster werden mit gestrichelten Linien in Abbildung 7.13d dargestellt.



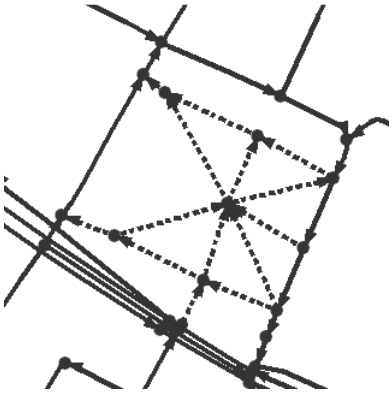
a) Enddatensatz nach der Datenverschmelzung von zugeordneten Kanten und Knoten



b) Bildung der nicht zugordneten Kanten in NavTeq in einem Cluster



c) Berechnung der Verlinkungsknoten des Clusters in NavTeq



d) Ergebnisse der Datenverschmelzung nach der Übertragung und der Zerlegung von Kanten

Abbildung 7.13: Beispiel für die Datenverschmelzung von nicht zugeordneten Kanten und Knoten

Der Ansatz zur Verschmelzung von nicht zugeordneten Kanten und Knoten kann ebenfalls eingesetzt werden, um z.B. die nicht zugeordneten Kanten und Knoten vom Datensatz *A* in den Datensatz *B* zu übertragen (siehe [Chen et al. 2006]). Dabei sind die Kanten und Knoten im Datensatz *B* als Basisdaten zu verwenden. Die nicht zugeordneten Kanten und Knoten im Datensatz *A* sind in Cluster aufzuteilen und anschließend in den Basisdatensatz (Datensatz *B*) zu übertragen.

7.3.2 Diskussion der Ergebnisse

Im Folgenden werden zunächst die Ergebnisse der Bildung von Clustern diskutiert. Anschließend werden Konflikte im Enddatensatz vorgestellt, die nach der Datenverschmelzung entstehen.

Ergebnisse der Bildung von Clustern

Tabelle 7.2 fasst die Ergebnisse der Bildung von Clustern für die nicht zugeordneten Kanten und Knoten von NavTeq und TeleAtlas zusammen. Es werden 52 Cluster aus 213 nicht zugeordneten NavTeq-Kanten im Testgebiet I berechnet. Durchschnittlich besteht ein Cluster in NavTeq aus 4,1 Kanten und maximal aus 39 Kanten. Weiterhin werden 90 Cluster aus 233 TeleAtlas-Kanten im Testgebiet I erzeugt. Im Testgebiet II werden 11 Cluster aus 14 NavTeq-Kanten und 134 Cluster aus 631 TeleAtlas-Kanten gebildet. Darüber hinaus wird ein Cluster mit 134 TeleAtlas-Kanten im Testgebiet II generiert.

	Testgebiet I		Testgebiet II	
	NavTeq	TeleAtlas	NavTeq	TeleAtlas
Anzahl der Cluster	52	90	11	134
Anzahl der Kanten	213	233	14	631
Gemittelte Größe	4,1	2,6	1,2	4,7
Maximale Größe	39	25	2	134

Tabelle 7.2: Ergebnisse der Bildung von Clustern (NavTeq & TeleAtlas)

In Tabelle 7.3 sind die Ergebnisse der Bildung von Clustern für die nicht zugeordneten Kanten und Knoten von TeleAtlas und OpenStreetMap dargestellt. Insgesamt werden 89 Cluster in TeleAtlas und 95 Cluster in OpenStreetMap im Testgebiet I erzeugt. Im Testgebiet II werden 80 Cluster in TeleAtlas und 11 Cluster in OpenStreetMap berechnet. Weiterhin wird ein Cluster mit 510 TeleAtlas-Kanten im Testgebiet II erzeugt.

	Testgebiet I		Testgebiet II	
	TeleAtlas	OpenStreetMap	TeleAtlas	OpenStreetMap
Anzahl der Cluster	89	95	80	11
Anzahl der Kanten	202	603	1483	68
Gemittelte Größe	2,3	6,3	18,5	6,78
Maximale Größe	24	204	510	18

Tabelle 7.3: Ergebnisse der Bildung von Clustern (TeleAtlas & OpenStreetMap)

Konflikte bei der Datenverschmelzung

Die Cluster von den verschiedenen Datensätzen werden mit unterschiedlichen Parametern in den Enddatensatz transformiert, welcher nach der Datenverschmelzung von den zugeordneten Kanten und Knoten entstanden ist. Da die Datensätze das gleiche Gebiet repräsentieren, können Konflikte nach der Datenverschmelzung entstehen:

1. *Einfache Überschneidung*: Kanten von den verschiedenen Datensätzen überschneiden sich nach der Datenverschmelzung. Zum Beispiel werden die nicht zugeordneten Kanten (A_1 in TeleAtlas und B_1, B_2, B_3 in OpenStreetMap) in Abbildung 7.14a jeweils zu einem Cluster zusammengefasst und in den Enddatensatz transformiert. Nach der Datenverschmelzung überschneiden sich die Kanten A_1 und B_3 im Enddatensatz.

2. *Multiple Überschneidungen*: Kanten von den verschiedenen Datensätzen überschneiden sich an mehreren Stellen im Enddatensatz. Beispielsweise überschneidet die Kante B_1 in TeleAtlas in Abbildung 7.14b die Kanten A_3 und A_4 in NavTeq.

3. *Unterschiedliche Repräsentation*en: Nach der Datenverschmelzung können unterschiedliche Repräsentationen für dieselben Objekte der realen Welt entstehen. In Abbildung 7.14c wurden die Kanten A_1 und A_2 in NavTeq sowie B_1 in TeleAtlas aufgrund der großen Unterschiede bei der manuellen Zuordnung nicht zugeordnet. Nach der Datenverschmelzung wird festgestellt, dass diese Kanten über gleiche Anfangs- und Endknoten verfügen. Aus diesem Grund können sie dieselben Objekte repräsentieren.

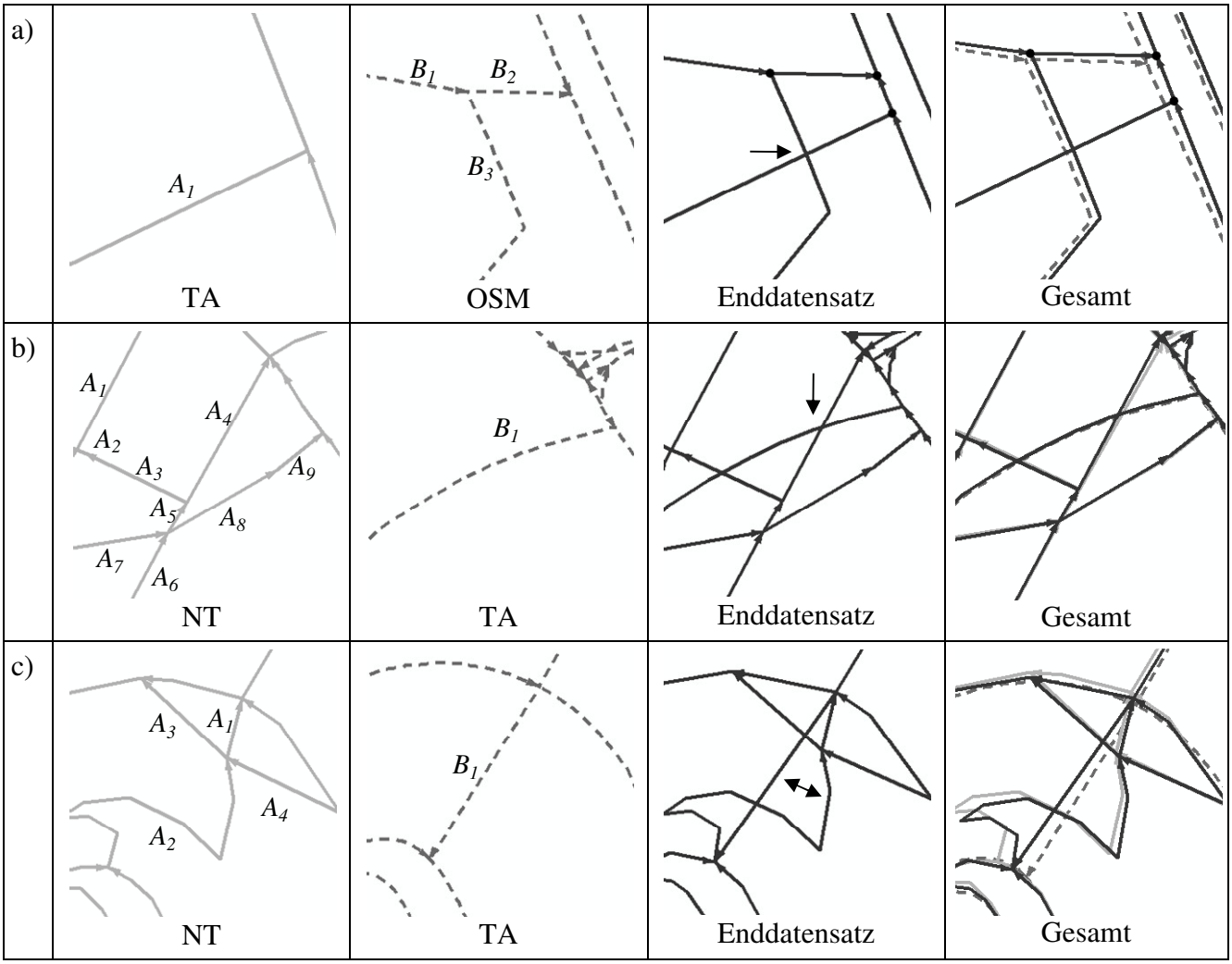


Abbildung 7.14: Konflikte bei der Datenverschmelzung

Um die Konflikte zu lösen und eine aussagekräftige Entscheidung zu treffen, sind zusätzliche Informationen (wie z.B. Luftbilder) einzusetzen. Weiterhin können unterschiedliche Regelungen eingesetzt werden. Zum Beispiel bestehen Konflikte zwischen zwei Clustern (Cluster C^A vom Datensatz A und Cluster C^B vom Datensatz B). Als Lösung können einerseits die beiden Cluster in den Enddatensatz übertragen werden. Danach sind Schnittpunkte zu berechnen und die betroffenen Kanten aufzuteilen. Andererseits sind die Cluster aus unterschiedlichen Gesichtspunkten zu bewerten:

- *Aktualität des Datensatzes*: Konflikte werden nach der Aktualität des Datensatzes gelöst. Ist der Datensatz A aktueller, dann ist der Cluster C^A in den Enddatensatz zu übertragen. Sonst wird der Cluster C^B in den Enddatensatz übertragen.

- *Eigenschaft des Clusters*: Cluster sind nach ihren Eigenschaften (z.B. Kantenlänge, Attribute) zu bewerten. Daraus ergeben sich unterschiedliche Gewichte für den Cluster C^A und C^B. Der Cluster mit einem größeren Gewicht ist in den Enddatensatz zu übertragen.

8 Zusammenfassung und Ausblick

8.1 Zusammenfassung

Die vorliegende Arbeit zielt darauf, die Qualität von Navigationsdaten durch Integration von verschiedenen Datenquellen zu überprüfen und zu verbessern. Untersucht wurde dabei die Integration zweier kommerzieller Datensätze (NavTeq und TeleAtlas) mit hoher Ähnlichkeit und hohen Redundanzen und einer kommerziellen (TeleAtlas) und eines kostenfreien Datensatzes (OpenStreetMap). Durch die Transformation in ein übergeordnetes Datenmodell wurde die Heterogenität der Datenmodellierung minimiert. Gleichzeitig konnte das Problem der topologischen Modellierung in OpenStreetMap gelöst werden, welches an fehlenden Unterbrechungen von Kanten liegt. Allerdings konnte nicht jede Art der semantischen Heterogenität mit der semantischen Homogenisierung beseitigt werden: z.B. Heterogenität aufgrund unterschiedlicher Anzahl von Klassifikationen eines Attributs in den Datensätzen.

Im weiteren Verlauf der Arbeit wurden die Korrespondenzen in den Datensätzen am Beispiel von Testgebieten im Stadtgebiet (Testgebiet I) und im ländlichen Raum (Testgebiet II) bestimmt. Um genaue Zuordnungsergebnisse zu erzielen, wurden die Kantenkorrespondenzen manuell bestimmt. Durch die Erweiterung des „Buffer Growing“ konnten alle Zuordnungskandidaten der Kanten vollständig erfasst und Unterschiede der geometrischen Modellierung zugelassen werden. Aus den Ergebnissen der manuellen Zuordnung ist festzustellen, dass ca. 4-5% der Zuordnungen zwischen Knoten und Kanten stattfinden. Diese Zuordnungen sind insbesondere im Kreuzungsbereich zu finden.

Trotz der Verwendung des komplexen Zuordnungsmodells konnten einige Stellen mit sehr stark unterschiedlichen Erfassungen, die zu unterschiedlichen topologischen Modellierungen führen, nur problematisch zugeordnet werden. Solche Zuordnungen wurden als Zuordnungen zwischen Komplexknoten und Komplexknoten für die weitere Auswertung markiert, um eine gleiche topologische Modellierung ableiten zu können. Dazu zählen 0,5% der Zuordnungen zwischen NavTeq und TeleAtlas und 1,5% der Zuordnungen zwischen TeleAtlas und OpenStreetMap im Stadtgebiet. Aus den Ergebnissen der manuellen Zuordnung ist festzustellen, dass das vorgestellte Zuordnungsmodell Unterschiede der geometrischen Modellierung besser als Unterschiede der topologischen Modellierung tolerieren kann.

In der vorliegenden Arbeit wurde ein Betrag zur Untersuchung der unterschiedlichen geometrischen Modellierungen in den Datensätzen geleistet. Durch die Erkennung von vorgegebenen Formklassen konnten die geometrischen Modellierungen der Datensätze gegenübergestellt werden. Aus den Ergebnissen ist zu entnehmen, dass mehr als 90% der Zuordnungspaare eine gleiche Form in den Datensätzen haben, da die geometrische Datenmodellierung in den drei Datensätzen ähnlich ist. Weiterhin wurden mehr Formklassen im Stadtgebiet als im ländlichen Raum gefunden. Es wird festgestellt, dass komplexere Verkehrssituationen im Stadtgebiet oft unterschiedlich erfasst wurden.

Die Qualitätsanalyse fand sowohl auf der Ebene des Datensatzes als auch auf der Ebene der Zuordnungspaare statt. Auf der Ebene des Datensatzes wurden die globale geometrische und topologische Ähnlichkeit sowie die Vollständigkeit berechnet. Dabei wurden komplexe Objekte anhand der Zuordnungsergebnisse berechnet und eine gleiche Darstellung in den Adjazenzmatrizen erzielt. Um eine gleiche Dimension der Adjazenzmatrizen zu gewährleisten, sind genaue und vollständige Zuordnungsergebnisse erforderlich. Für jedes Zuordnungspaar wurden die Ähnlichkeit

der Form sowie die lokale geometrische und topologische Ähnlichkeit berechnet. Dadurch wurde eine detaillierte Untersuchung hinsichtlich des Zuordnungspaars für die globale geometrische und topologische Ähnlichkeit ermöglicht. Ein weiterer Aspekt der Arbeit liegt in der Entwicklung der Verfahren zur Auswertung der Ähnlichkeit von Attributen. Die einzelnen Attribute wurden entweder direkt verglichen oder über eine Konfusionsmatrix ausgewertet. Dabei wurde die Distanz der unterschiedlichen Abschnitte des Zuordnungspaars berücksichtigt. Die Ergebnisse zeigen, dass die Attribute mit wenigen Kategorien eine höhere Ähnlichkeit besitzen, weil die Unsicherheit bei der Datenerfassung für diese Attribute niedriger ist. Die Ähnlichkeit von Attributen mit gleichen unbegrenzten Wertebereichen wie z.B. Straßennamen ist relativ niedriger. Darüber hinaus ist die Ähnlichkeit dieses Attributs im Stadtgebiet höher als im ländlichen Raum.

Basierend auf den Ergebnissen der Qualitätsanalyse wird festgestellt, dass die globale geometrische und topologische Ähnlichkeit zwischen den kommerziellen Datenbeständen (NavTeq und TeleAtlas) höher als die zwischen dem kommerziellen (TeleAtlas) und dem kostenfreien Datenbestand (OpenStreetMap) ist. Ein möglicher Grund ist, dass das gleiche Datenmodell in den zwei kommerziellen Datenbeständen verwendet wird. Weiterhin sind die Ähnlichkeiten im Stadtgebiet höher als im ländlichen Raum.

Es wurde ebenfalls ein Beitrag geleistet, Verfahren zur Datenverschmelzung unter Berücksichtigung der unterschiedlichen geometrischen Modellierungen zu entwickeln. Dabei wurde ein clusterbasierter Ansatz vorgestellt. Cluster für zugeordnete und für nicht zugeordnete Kanten und Knoten sind zu unterscheiden. Für die zugeordneten Kanten und Knoten wurde eine gleiche Anzahl von Clustern in den Datensätzen ermittelt, die aus einem bzw. mehreren Zuordnungspaaren bestehen. Allerdings kann die Anzahl von Clustern für nicht zugeordnete Kanten und Knoten in den Datensätzen, die anhand der Konnektivität von Kanten berechnet werden, unterschiedlich sein. Die Ergebnisse der Datenverschmelzung können in der Zukunft mit weiteren Referenzdaten kontrolliert werden.

Die Datenverschmelzung für Zuordnungspaare mit unterschiedlichen geometrischen Modellierungen (Form) in den Datensätzen ist mit bisherigen Ansätzen schwierig, da Mittellinien in diesem Fall nicht berechnet werden können. In der vorliegenden Arbeit wurden zwei Arten von Clustern für die zugeordneten Kanten und Knoten erzeugt. Für die Cluster aus Zuordnungspaaren mit gleicher Form in den Datensätzen wurden Mittellinien berechnet. Für die Cluster aus Zuordnungspaaren, die unterschiedliche Formen in den Datensätzen haben, wurde die Komplexität der geometrischen Modellierung der Cluster anhand der ermittelten Ähnlichkeit der Form bewertet. Eine einfache bzw. komplexe geometrische Modellierung dieser Cluster wurde ausgewählt und durch eine Transformation in den Enddatensatz übertragen. Verbindungsknoten, welche die Cluster miteinander verbinden, wurden in der Arbeit definiert und anhand der Zuordnungsergebnisse berechnet, um die Parameter der Transformation zu berechnen. So lässt sich die Geometrie der Cluster sowohl mit gleicher als auch mit unterschiedlicher Form verbessern. Die Ergebnisse der Datenverschmelzung sind von der Genauigkeit der Verbindungsknoten abhängig. In der Zukunft können mehr Kriterien zur Qualifikation der Verbindungsknoten eingesetzt werden.

Die Cluster für die nicht zugeordneten Kanten und Knoten wurden ebenfalls durch eine Transformation in den Enddatensatz übertragen. Dabei wurden Knoten in einem Cluster, woran zugeordnete Kanten diesen Cluster anschließen, als Verlinkungsknoten definiert, um die Parameter der Transformation abzuleiten. Mit dem vorgestellten Ansatz blieben die Konnektivität und die Form der Cluster nach der Datenverschmelzung unverändert. Weiterhin ist der Ansatz für Cluster mit einer unterschiedlichen Anzahl von Kanten geeignet. Die Genauigkeit der Datenverschmelzung hängt genauso von der Genauigkeit der Verlinkungsknoten ab. Um die Ergebnisse zu verbessern, können Distanzen der Verlinkungsknoten, die aus der unterschiedlichen Geometrie eines

Verlinkungsknotens in den Datensätzen zu berechnen sind, in weiterführenden Arbeiten als Gewicht bei der Transformation verwendet werden.

8.2 Ausblick

In der zukünftigen Entwicklung ist die Komponente der Zuordnung mit dem vorgestellten Zuordnungsmodell zu automatisieren. Die in der Arbeit erzielten Ergebnisse können als Referenzdaten verwendet werden, um zu untersuchen, inwieweit fehlerhafte Ergebnisse aus einer automatischen Zuordnung die Komponenten der Qualitätsanalyse und der Datenverschmelzung auswirken. Am Beispiel von linienförmigen Straßenobjekten wurde die Qualität der Daten untersucht und verbessert. Ansätze zur Qualität von anderen Objekten wie z.B. flächenförmigen Objekten oder Relationen sind zu entwickeln.

Ein weiterer Forschungsaspekt liegt in der Qualitätsuntersuchung durch Integration von Datensätzen mit unterschiedlichen Maßstäben. Dabei kann der hoch detaillierte Datensatz zunächst mit einer Generalisierung abstrahiert werden. Die Zuordnung mit dem vorgestellten Modell findet zwischen dem wenig detaillierten und dem generalisierten Datensatz statt. Daraus ist die Zuordnung zwischen dem wenig und hoch detaillierten Datensatz abzuleiten. Ein weiterer wichtiger Aspekt ist die Qualitätsanalyse der Daten. Ferner wird eine Integration von mehreren Datensätzen bislang noch nicht erreicht, was eine Erweiterung des Zuordnungsmodells auf mehreren Datensätzen erfordern würde. Danach können die vorgestellten Ansätze zur Qualitätsanalyse und Datenverschmelzung angewendet werden.

Weitere Komponenten zur Visualisierung der Qualität und Lösung der Konflikte sind zu entwickeln. Dabei können Werkzeuge zur hierarchischen Visualisierung der Ergebnisse der Qualitätsanalyse entwickelt werden. Ein einfaches Beispiel ist die Darstellung auf Basis der ermittelten Ähnlichkeiten mit einem Ampelsystem, um einen schnellen Einblick auf die Qualität zu vermitteln und eine schnelle Identifikation von Fehlerstellen zu ermöglichen. Konflikte, die durch eine Datenverschmelzung entstehen können, wurden in der Arbeit anhand von Beispielen diskutiert. Regelungen zur Lösung der Konflikte wurden vorgeschlagen. Um eine aussagekräftige Lösung anzubieten, sind diese Konflikte durch Einsatz von anderen Datenquellen (z.B. aktuelle Luftbilder oder Satellitenbilder) zu beheben.

Weiterhin können andere Arten von externen Informationen zur Qualitätsprüfung und -verbesserung eingesetzt werden. In der Arbeit wurde eine kostenfreie digitale Karte verwendet, die sich aus nutzergenerierten Inhalten ergibt. Die nutzergenerierten Inhalte spielen allerdings eine immer wichtigere Rolle bei der Erfassung von raumbezogenen Daten. Aus diesem Grund besteht noch Forschungsbedarf, weitere Arten der nutzergenerierten Inhalte (z.B. einzelne GPS-Tracks) als Informationsquellen einzusetzen. Darüber hinaus können externe Informationen wie z.B. Google-Streetview integriert werden, um die Qualität von Verkehrszeichen oder Fahrspuren zu überprüfen.

Die Qualitätsuntersuchung durch Datenintegration ist nicht nur ein Thema für zweidimensionale Daten, sondern auch für dreidimensionale Daten. Dadurch kann die Qualität von dreidimensionalen Daten untersucht werden. Die Integration von zwei- und dreidimensionalen Daten bietet auch eine Möglichkeit zur Qualitätsuntersuchung an. Dabei können z.B. Regeln zur Konsistenzprüfung definiert werden.

Literaturverzeichnis

Anders, K.-H., M. Sester & J. Bobrich [2007]: Incremental Update in an MRDB. *In*: Proceedings of the 23rd International Cartographic Conference, Moscow, Russia, on CDROM.

Balley, S., C. Parent & S. Spaccapietra [2004]: Modeling Geographic Data with Multiple Representations. International Journal of Geographical Information Science, 18/4, 327-352.

Bauer, A. & H. Günzel [2000]: Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung (1. Auflage). dpunkt Verlag, Heidelberg, 579 S.

Becker, C., M. Ziems, T. Büschenfeld, C. Heipke, S. Müller, J. Ostermann & M. Pahl [2008]: Multi-Hierarchical Quality Assessment of Geo-Spatial Data. *In*: Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, XXXVII/B2, 779-786.

Berg, M., O. Cheong, M. Kreveld & M. Overmars [2008]: Computational Geometry: Algorithms and Applications. Springer, Berlin, 386 S.

Bill, R. & M. Zehner [2001]: Lexikon der Geoinformatik. Herbert Wichmann Verlag, Heidelberg, 319 S.

Blasby, D., M. Davis, D. Kim & P. Ramsey [2003]: GIS Conflation using Open Source Tools. Jump-Project Whitepaper.
http://www.jump-project.org/assets/JUMP_Conflation_Whitepaper.pdf
Zugriff: 20.02.2007

Bofinger, J.-M. [2001]: Analyse und Implementierung eines Verfahrens zur Referenzierung geographischer Objekte. Diplomarbeit, Institut für Photogrammetrie, Universität Stuttgart, unveröffentlicht.

Brassel, K., F. Bucher, E. M. Stephan & A. Vckovshi [1995]: Completeness. *In*: S. C. Guptill & J. Morrison (eds.), Elements of Spatial Data Quality, Elsevier, Oxford, 81-108.

Breunig, M., A. Thomsen, B. Broscheit, E. Butwilowski & U. Sander [2007]: Representation and Analysis of Topology in Multi-Representation Databases. *In*: PIA 2007, Photogrammetric Image Analysis, München, 36, 167-172.

Brüntrup, R., S. Edelkamp, S. Jabbar & B. Scholz [2005]: Incremental Map Generation with GPS Traces. *In*: Proceedings of the IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, 574-579.

Busch, A., M. Gerke, D. Grünreich, C. Heipke, P. Helmholz, C.-E. Liedtke & S. Müller [2006]: Automated Verification of a Topographic Dataset using Ikonos Imagery. *In*: Proceedings of the IntArchPhRS, Goa, XXXVI/4, 134-139.

Butenuth, M., G. v. Gösseln, M. Tiedge, C. Heipke, U. Lipeck & M. Sester [2007]: Integration of Heterogeneous Geospatial Data in a Federated Database. ISPRS Journal of Photogrammetry and Remote Sensing, 62/5, 328-346.

Chen, C.-C., C. A. Knoblock & C. Shahabi [2006]: Automatically Conflating Road Vector Data with Orthoimagery. GeoInformatica, 2006/10, 495-530.

Chen, H. [2006]: Entwicklung von Auswerteprogrammen für das Testen von Navigationsfunktionen. Diplomarbeit, Institut für Photogrammetrie, Universität Stuttgart, unveröffentlicht.

Chen, H., V. Walter & D. Fritsch [2006]: Quality Inspection and Quality Improvement by Map Fusion. *In*: Proceedings of The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, XXXVII, 467-472.

Chrisman, N. R. [1983]: The Role of Quality Information in the Longterm Functioning of a Geographic Information System. Cartographica, 21/2-3, 79-87.

Claussen, H. [1996]: Qualitätsbeurteilung digitaler Karten für Fahrzeug-Navigationssysteme. GIS, 9/5, 23-29.

Cobb, M. A., M. J. Chung, H. Foley, F. E. Petry, K. B. Shaw & H. V. Miller [1998]: A Rule-Based Approach for the Conflation of Attributed Vector Data. Geoinformatica, 2/1, 7-35.

Cockcroft, S. [1997]: A Taxonomy of Spatial Data Integrity Constraints. GeoInformatica, 1/4, 327-343.

Cockcroft, S. [2001]: Modelling Spatial Data Integrity Rules at the Metadata Level. *In*: Proceedings of the 6th International Conference on GeoComputation, Brisbane, Australia, on CDROM.

Coleman, D. J., Y. Georgiadou & J. Labonte [2009]: Volunteered Geographic Information: the Nature and Motivation of Produsers. *In*: Proceedings of the GSDI 11 World Conference, Rotterdam, Netherlands, on CDROM.

Czommer, R. [2000]: Leistungsfähigkeit fahrzeugautonomer Ortungsverfahren auf der Basis von Map-Matching-Techniken. Dissertation, Deutsche Geodätische Kommission (DGK), Reihe C, Nr. 535.

Davie, P. & C. McCullar [2009]: Tele Atlas Has It Covered: MultiNet, ConnectPlus and Connect. Coverage White Paper.
http://www.teleatlas.com/stellent/groups/public/documents/content/ta_ct021036.pdf
Zugriff: 15.09.2009

Dell'Acqua, F., P. Gamba & G. Lisini [2002]: Extraction and Fusion of Street Networks from Fine Resolution SAR Data. *In*: Proceedings of the IGARSS, Toronto, Canada, 89-91.

Deng, M., Z. L. Li & X. Y. Chen [2007]: Extended Hausdorff Distance for Spatial Objects in GIS. International Journal of Geographical Information Science, 21/4, 459-475.

Deretsky, Z. & U. Rdony [1993]: Automatic Conflation of Digital Maps. *In*: Proceedings of the IEEE - IEE Vehicle Navigation & Information Systems Conference, Ottawa, A27-A29.

Devillers, R., Y. Bédard, R. Jeansoulin & B. Moulin [2007]: Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use of Spatial Data. International Journal of Geographical Information Science, 21/3, 261-282.

Devillers, R. & R. Jeansoulin [2006]: Fundamentals of Spatial Data Quality. Wiley-ISTE, 312 S.

Devogele, T. [2002]: A New Merging Process for Data Integration Based on the Discrete Fréchet Distance. *In*: Proceedings of the ISPRS Commission IV Symposium: Geospatial Theory, Processing and Applications, Ottawa, Canada, on CDROM.

Dorgu, A. O. & N. Ulugtekin [2006]: Car Navigation Map Design in terms of Multiple Representations. *In*: Proceedings of the First International Conference on Cartography & GIS, Borovets, Bulgaria, on CDROM.

Doytsher, Y., S. Filin & E. Ezra [2001]: Transformation of Datasets in a Linear-Based Map Conflation Framework. Surveying and Land Information Systems, 61/3, 159-169.

Dunkars, M. [2003]: Matching of Datasets. *In*: Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science (ScanGIS 2003), Espoo, Finland, 67-78.

Edwards, D. & J. Simpson [2002]: Integration and Access of Multi-Source Vector Data. *In*: Proceedings of the Joint International Symposium of Geospatial Theory, Processing and Application, Ottawa, Canada, on CD-ROM.

Egenhofer, M. & R. D. Franzosa [1991]: Point-Set Topological Spatial Relations. International Journal of Geographical Information Science, 5/2, 161-174.

Egenhofer, M. J. [1994b]: Evaluating Inconsistencies among Multiple Representations. *In*: Proceedings of the 6th International Synposium on Spatial Data Handling, Edinburgh, Scotland, 901-920.

Egenhofer, M. J., D. M. Mark & J. Herring [1994a]: The 9-Intersection: Formalism and its Use for Natural-Language Spatial Predicates. National Center for Geographic Information and Analysis, Report 94-1.

ERTICO [1995]: GDF - Geographic Data Files 3.0.
http://www.ertico.com/en/page_archive/gdf_-_geographic_data_files.htm
Zugriff: 01.09.2009

Essen, R. & V. Hiestermann [2005]: "X-GDF" - The ISO Model of Geographic Information for ITS. *In*: Proceedings of the ISPRS Workshop on Service and Application of Spatial Data Infrastructure, Hangzhou, China, XXXVI, 59-64.

EuroRoadS [2006]: EU-Projekt EuroRoadS (Pan-European Road Data Solution) (2004-2006).
www.euroroads.org
Zugriff: 01.09.2009

Fang, X. [2008]: Konzeption und Realisierung einer automatisierten Validierung der geometrischen Verteilung von Points of Interest (POIs) in digitalen Straßenkarten. Diplomarbeit, Institut für Photogrammetrie, Universität Stuttgart, unveröffentlicht.

Franz, B. [2008]: Plausibilitätsprüfung von Karten für Navigationssysteme auf Basis aufgezeichneter Strecken. Diplomarbeit, Fachhochschul-Masterstudiengang SOFTWARE ENGINEERING, Hagenberg, Austria.

Fritsch, D., M. Glemser, U. Klein, M. Sester & G. Strunz [1998]: Zur Integration von Unsicherheit bei Vektor- und Rasterdaten. GIS - GeoInformationssysteme, 11/4, 26-35.

Fritz, S. & L. See [2005]: Comparison of Land Cover Maps using Fuzzy Agreement. International Journal of Geographical Information Science, 19/7, 787-807.

Gerke, M., M. Butenuth, C. Heipke & F. Willrich [2004]: Graph-Supported Verification of Road Databases. ISPRS Journal of Photogrammetry and Remote Sensing, 58/3-4, 152-165.

Gillmann, D. [1985]: Triangulations for Rubber-Sheeting. *In*: Proceedings of the 7th International Symposium on Computer Assisted Cartography (AutoCarto 7), Washington, D.C., 191-199.

Glemser, M. [2001]: Zur Berücksichtigung der geometrischen Objektunsicherheit in der Geoinformatik. Dissertation, Deutsche Geodätische Kommission (DGK), Reihe C, Nr. 539.

Gombosi, M., B. Zalik & S. Krivograd [2003]: Comparing Two Sets of Polygons. International Journal of Geographical Information Science, 17/5, 431-443.

Gong, P. & L. Mu [2000]: Error Detection through Consistency Checking. Geographic Information Sciences, 6/2, 188-193.

Goodchild, M. [1995]: Attribute Accuracy. *In*: S. C. Guptill & J. Morrison (eds.), Elements of Spatial Data Quality, Elsevier, Oxford, 59-79.

Goodchild, M. F. [2007]: Citizens as Sensors: the World of Volunteered Geography. GeoJournal, 69/4, 211-221.

Goodchild, M. F. [2008]: Spatial Accuracy 2.0. *In*: Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, 1-7.

Goodchild, M. F. & G. J. Hunter [1997]: A Simple Positional Accuracy Measure for Linear Features. International Journal of Geographical Information Science, 11/3, 299-306.

Guptill, S. C. & J. L. Morrison, Eds. [1995]: Elements of Spatial Data Quality (The International Cartographic Association). Elsevier, Oxford, 250 S.

Hadzilacos, T. & N. Tryfona [1992]: A Model for Expressing Topological Integrity Constraints in Geographic Databases. Lecture Notes in Computer Science, 639, 252-268.

Hake, G., D. Grünreich & L. Meng [2002]: Kartographie. Walter de Gmbh Gruyter, Berlin, New York, 604 S.

Haklay, M. [2008]: How Good is OpenStreetMap information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the rest of England. http://www.ucl.ac.uk/~ucfamha/OSM%20data%20analysis%20070808_web.pdf Zugriff: 07.07.2009

Hampe, M. [2007]: Integration einer multiskaligen Datenbank in eine Webservice-Architektur. Dissertation, Deutsche Geodätische Kommission (DGK), Reihe C, Nr. 605.

Hangouet, J. F. [1995]: Computation of the Hausdorff Distance between Plane Vector Polylines. *In*: Proceedings of the 10th International Symposium on Computer-Assisted Cartography, Bethesda, 4, 1-10.

Haunert, J.-H. [2005]: Link based Conflation of Geographic Datasets. *In*: Proceedings of the 8th ICA Workshop on Generalisation and Multiple Representation, La Coruna, Spanien, on CDROM.

Haunert, J.-H. & M. Sester [2008]: Assuring Logical Consistency and Semantic Accuracy in Map Generalization. Photogrammetrie - Fernerkundung - Geoinformation (PFG), 2008/3, 165-173.

Heipke, C., P. A. Woodford & M. Gerke [2008]: Updating Geospatial Databases from Images. *In*: Z. Li, J. Chen & E. Baltsavias (eds.), Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences: 2008 ISPRS Congress Book, 335-362.

Hinz, S. [2003]: Automatische Extraktion urbaner Straßennetze aus Luftbildern. Dissertation, Deutsche Geodaetische Kommission, Reihe C, Nr. 580.

Horn, M. [2007]: Conflation zur Erzeugung von Fußgängernavigationsdaten. Diplomarbeit, Institut für Kartographie und Geoinformatik, Universität Hannover.

Inspire [2009]: Inspire Home Page.
http://inspire.jrc.ec.europa.eu
Zugriff: 28.07.2009

ISO9000 [2005]: Qualitätsmanagement - Grundlagen und Begriffe. Beuth Verlag, Berlin.

ISO14825 [2004]: GDF-Geographic Data Files-Version 4. Beuth Verlag, Berlin.

ISO19113 [2002]: Geographic Information - Quality Principles. Beuth Verlag, Berlin.

ISO19114 [2003]: Geographic Information - Quality Evaluation Procedures. Beuth Verlag, Berlin.

ISO19115 [2003]: Geoinformation - Metadaten. Beuth Verlag, Berlin.

ISO19118 [2005]: Geographic Information - Encoding. Beuth Verlag, Berlin.

ISO/TC211 [2009]: Standards Guide ISO/TC 211 Geographic Information/Geomatics.
http://www.isotc211.org/Outreach/ISO_TC%20_211_Standards_Guide.pdf
Zugriff: 02.11.2009

Joos, G. [2000]: Zur Qualität von objektstrukturierten Geodaten. Dissertation, Schriftenreihe des Studienganges Geodäsie und Geoinformation der Universität der Bundeswehr München, Heft 66/2000.

Kappas, M. [2001]: Geographische Informationssysteme. Westermann Schulbuch Verlag, Braunschweig, 315 S.

Kaufmann, T. & T. Wiltschko [2005]: Evaluation Schema for Information Quality. EuroRoads, D 2.5.

Keul, E. [1998]: Anwendung der parametrischen Distanzfunktion zur Bestimmung der Mittelachse eines Polygons. Studienarbeit, Institut für Photogrammetrie, Universität Stuttgart, unveröffentlicht.

Kieler, B., M. Sester, H. Wang & J. Jiang [2007]: Semantic Data Integration: Data of Similar and Different Scales. Photogrammetrie-Fernerkundung-Geoinformation (PFG), 6, 447-457.

KIWI [2000]: Input for ISO Physical Storage Format V1.22.
http://www.kiwi-w.org/format_english/format_kihon.html
Zugriff: 20.05.2009

Klang, D. [1998]: Automatic Detection of Changes in Road Databases using Satellite Imagery. *In*: Proceedings of the International Archives of Photogrammetry and Remote Sensing, Stuttgart, 32/4, 293-298.

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

Krauß, S. [1998]: Qualitative Beschreibung von unsicheren topologischen Relationen innerhalb des Minimum-/Maximum-Modells. Diplomarbeit, Institut für Photogrammetrie, Unversität Stuttgart, unveröffentlicht.

Landwehr, M., M. Flament, S. Durekovic, J. Loewenau, V. Naumann, H. Deragarden, V. Meliga, B. Thomas, M. Landwehr, G. Thomaidis, P. Issacson, U. Haspel & F. Visintainer [2008]: FeedMAP Test Report. FeedMAP, WP5.

Leser, U. & F. Naumann [2007]: Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Dpunkt Verlag, 464 S.

Li, Z. [2008]: Multi-Scale Modelling and Representation of Geospatial Data. *In*: Z. Li, J. Chen & E. Baltsavias (eds.), Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences: 2008 ISPRS Congress Book, 265-278.

Lo, C. P. & A. K. W. Yeung [2002]: Concepts and Techniques in Geographic Information Systems. Prentice Hall, Upper Saddle River, NJ, 492 S.

Longueville, B., N. Ostländer & C. Keskitalo [2009]: Addressing Vagueness in Volunteered Geographic Information (VGI) - A Case Study. *In*: Proceedings of the GSDI 11 World Conference, Rotterdam, the Netherlands, on CDROM.

Lüscher, P. & D. Burghardt [2006]: Matching von Straßendaten stark unterschiedlicher Maßstäbe und Aufbau einer MRDB. Mitteilungen des Bundesamtes für Kartographie und Geodäsie, Band 36: Arbeitsgruppe Automation in der Kartographie - Tagung 2005, 79-88.

Lupien, A. E. & W. H. Moreland [1987]: A General Approach to Map Conflation. *In*: Proceedings of the 8th International Symposium on Computer Assisted Cartography (AutoCarto 8), Maryland, 630-639.

Lynch, M. & A. Saalfeld [1985]: Conflation: Automated Map Compilation, a Video Game Approach. *In*: Proceedings of the 7th International Symposium on Computer Assisted Cartography (AutoCarto 7), Washington, D.C., 343-352.

Mäs, S. [2008]: Checking the Integrity of Spatial Semantic Integrity Constraints. *In*: Dagstuhl Seminar: Constraint Databases, Geometric Elimination and Geographic Information Systems, Schloss Dagstuhl, Germany, Nr. 07212.

May, I. [2002]: Fortführung und Erweiterung von GDF (Geographic Data File) als Datengrundlage für Autonavigationssysteme. Dissertation, Universitätsverlag der TU Berlin.

Mayer, H., S. Hinz & W. Stilla [2008]: Automated Extraction of Roads, Buildings and Vegetation from Multi-Source Data. *In*: Z. Li, J. Chen & E. Baltsavias (eds.), Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences: 2008 ISPRS Congress Book, 213-226.

McDougall, K. [2009]: The Potential of Citizen Volunteered Spatial Information for Building SD. *In*: Proceedings of the GSD 11 World Conference, Rotterdam, the Netherlands, on CDROM.

Möhlenbrink, W., T. Wiltschko & T. Kaufmann [2006]: Zur Nutzung der Geodaten in zukünftigen Thelematiksystemen. Vermessung & Geoinformation, 1+2/2006, 112-119.

Möser, M., G. Müller, H. Schlemmer & W. Reinhardt [2004]: Handbuch Ingenieurgeodäsie: Raumbezogene Informationssysteme. Wichmann, Heidelberg, 226 S.

Mostafavi, M.-A., G. Edwards & R. Jeansoulin [2004]: An Ontology-Based Method for Quality Assessment of Spatial Data Bases. *In*: Proceedings of the Third International Symposium on Spatial Data Quality, Bruck an der Leitha, Austria, 28a, 49-66.

Mustière, S. & T. Devogele [2008]: Matching Networks with Different Levels of Detail. Geoinformatica, 12/4, 435-453.

NavTeq [2005]: NAVSTREETS Product Guide for Arcview (Version 3.5.0). NAVTEQ, 131 S.

NavTeq [2007]: Data Extraction Format Guide.
http://www.nn4d.com/site/global/learn/nt_map_data_formats/p_map_data_formats.jsp
Zugriff: 15.09.2009

Nitz, I. [2004]: Entwicklung eines Tools zur Qualitätsprüfung von Points of Interest mit ArcGIS. Diplomarbeit, Universität Dresden.

Olteanu, A.-M. [2007]: Matching Geographical Data using the Theory of Evidence. *In*: Proceedings of the XXIII International Cartographic Conference (ICC), Moskau, on CDROM.

Oort, P. [2005]: Spatial Data Quality: from Description to Application. Dissertation, NCG, Nederlandse Commissie voor Geodesie. Publications on Geodesy 60.

OpenStreetMap [2008]: OpenStreetMap Homepage.
http://www.openstreetmap.org/
Zugriff: 21.10.2008

Otto, H.-U., L. Beuk, M. Aleksić, J. Meier, J. Löwenau, M. Flament, A. Guarise, A. Bracht, L. Capra, K. Bruns & H. Sabel [2004]: ActMAP Specification. ActMap, D 3.2.

Paiva, J. A. C. [1998]: Topological Equivalence and Similarity in Multi-Representation Geographic Databases. Dissertation, Spatial Information Science and Engineering, University of Maine.

Parent, C. & S. Spaccapietra [2000]: Database Integration: the Key to Data Interoperability. *In*: M. Papazoglou, S. Spaccapietra & Z. Tari (eds.), Advances in Object-oriented Data Modelling, 221-253.

PAS1071 [2007]: Qualitätsmodell für die Beschreibung von Geodaten. Deutschen Dachverbandes für Geoinformation, Beuth Verlag, Berlin.

Peerbocus, A., G. Jomier & T. Badard [2002]: A Methodology for Updating Geographic Databases using Map Versions. *In*: Proceedings of the 10th International Symposium on Spatial Data Handling, Ottawa, Canada, 305-335.

Rahm, E., Bernstein, P. A [2001]: A Survey of Approaches to Automatic Schema Matching. The VLDB Journal, 10/4, 334-350.

Ramm, F. & J. Topf [2009]: OpenStreetMap - Die freie Weltkarte nutzen und mitgestalten (2. überarbeitete und erweiterte Auflage). Lehmanns Media, Berlin, 352 S.

Ripperda, N. [2004]: Graphbasiertes Matching in räumlichen Datenbanken. Diplomarbeit, Institut für Informationssysteme, Universität Hannover.

Rodríguez, A. [2005]: Inconsistency Issues in Spatial Databases. Inconsistency Tolerance, Lecture Notes in Computer Science, 3300, 237-269.

Safra, E., Y. Kanza, Y. Sagiv & Y. Doytsher [2006]: Efficient Integration of Road Maps. *In*: Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic information Systems, Arlington, Virginia, USA, 59-66.

Samal, A., S. Seth & K. Cueto [2004]: A Feature-Based Approach to Conflation of Geospatial Sources. International Journal of Geographical Information Science, 18/5, 459-489.

Schlott, S. [1997]: Fahrzeugnavigation: Routenplanung, Positionsbestimmung, Zielführung. Die Bibliothek der Technik 144, Moderne Industrie, 69 S.

Schmitz, S., A. Zipf & P. Neis [2008]: New Applications Based on Collaborative Geodata - the Case of Routing. *In*: Proceedings of the XXVIII INCA International Congress on Collaborative Mapping and SpaceTechnology, Gandhinagar, Gujarat, India, on CDROM.

Sedgewick, R. [1995]: Algorithmen in C++. Addison-Wesley Deutschland GmbH, 742 S.

Seifert, M. [2006]: INSPIRE - Geodaten für Europa. *In*: 11. Münchner Fortbildungsseminar Geoinformationssysteme, München.

Seo, S. & C. G. O'hara [2009]: Quality Assessment of Linear Data. International Journal of Geographical Information Science, 23/12, 1503-1525.

Servigne, S., T. Ubeda, A. Puricelli & R. Laurini [2000]: A Methodology for Spatial Consistency Improvement of Geographic Databases. Geoinformatica, 4/1, 7-34.

Sester, M. [2000]: Maßstabsabhängige Darstellungen in digitalen räumlichen Datenbeständen. Habilitation, Fakultät Bauingenieur- und Vermessungswesen, Universität Stuttgart. Deutsche Geodätische Kommission, Reihe C, Nr. 544, München 2001.

Sester, M. [2008]: Multiple Representation Databases. *In*: Z. Li, J. Chen & E. Baltsavias (eds.), Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences: 2008 ISPRS Congress Book, 279-288.

Sheeren, D., S. Mustiére & J.-D. Zucker [2009]: A Data-Mining Approach for Assessing Consistency between Multiple Representations in Spatial Databases. International Journal of Geographical Information Science, 23/8, 961-992.

Sonnen, D. [2007]: Emerging Issue: Spatial Data Quality. Directions Magazine, January 2007.
http://www.directionsmag.com/article.php?article_id=2372&trv=1
Zugriff: 01.08.2009

Spaccapietra, S., C. Parent & C. Vangenot [2000]: GIS Databases: From Multiscale to MultiRepresentation. Lecture Notes in Computer Science, 1864/2000, 57-70.

Spéry, L. [1998]: Spatial Data Transfer in the Case of Update. International Archives of Photogrammetry and Remote Sensing, 4/32, 586-593.

Steyer, R., B. Feser & F.-J. Knelangen [2004]: Qualität von Daten im Straßen- und Verkehrswesen. Bundesministerium für Verkehr, Bau- und Wohnungswesen (BMVBW), Abteilung Straßenbau, Straßenverkehr, Bonn, 133 S.

Stoter, J. & S. Zlatanova [2004]: Multiple Representations in DBMS: Two Algorithms. *In*: Proceedings of the International Society for Photogrammetry and Remote Sensing, Istanbul, 222-227.

Su, G. [2005]: Überblick zum Thema Informationsintegration.
http://www.suguangya.com/profile/Information%20Integration.pdf
Zugriff: 13.05.2009

Sulo, R., S. Eick & R. Grossman [2005]: DaVis: A Tool for Visualizing Data Quality.
http://www.rgrossman.com/dl/proc-095.pdf
Zugriff: 20.05.2009

Taylor, F. W. [1911]: The Principles of Scientific Management. Harper Bros, New York.

TeleAtlas [2005]: Tele Atlas MultiNet™ Shapefile 4.3.1 Format Specifications v1.0. Tele Atlas NV and Tele Atlas North America, Inc., 172 S.

TeleAtlas [2009a]: Tele Atlas MultiNet™ Map Database Production.

http://www.geopost.ch/Texte_GP/PDF_GP/MN%20MapDatabase.pdf
Zugriff: 23.07.2009

TeleAtlas [2009b]: TeleAtlas Homepage.
http://www.teleatlas.com/
Zugriff: 15.04.2008

TeleAtlas [2009c]: The Power of Community. White Paper.
http://www.teleatlas.com/stellent/groups/public/documents/content/ta_ct023328.pdf
Zugriff: 20.04.2009

Uitermark, H., P. Oosterom, N. J. I. Mars & M. Molenaar [1999]: Ontology-Based Geographic Data Set Integration. *In*: Proceedings of the International Workshop on Spatio-Temporal Database Management, Edinburgh, 60-78.

Ulugtekin, N. N., A. O. Dogru & R. C. Thomson [2004]: Modelling Urban Road Networks Integrating Multiple Representations of Complex Road and Junction Structures. *In*: Proceedings of the 12th International Conferences on Geoinformatics, Gavle, Sweden, 757-764.

Visintainer, F., M. Darin, M. Flament, S. Durekovic, H.-U. Otto, J. Loewenau, V. Naumann, H. Andersson, V. Meliga, B. Thomas, M. Landwehr, M. Bimpas, P. Isaksson & U. Haspel [2008]: Final Requirements and Strategies for Map Feedback. FeedMAP, D 2.2.

Volz, S. [2006a]: Modellierung und Nutzung von Relationen zwischen Mehrfachrepräsentationen in Geo-Informationssystemen. Dissertation, Institut für Photogrammetrie, Universität Stuttgart.

Volz, S. [2006b]: An Iterative Approach for Matching Multiple Representations of Street Data. *In*: Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data, Hannover, Germany, 101-110.

Volz, S., M. Sester, D. Fritsch & D. Klinec [2000]: Nexus - eine Plattform für ortsabhängige, verteilte Geodatennutzung. Publikationen der Deutschen Gesellschaft für Photogrammetrie und Fernerkundung, Bd. 8., 137-150.

Walter, V. [1997]: Zuordnung von raumbezogenen Daten - am Beispiel ATKIS und GDF, Deutsche Geodätische Kommission (DGK), München.

Walter, V. [1999]: Automatic Classification of Remote Sensing Data for GIS Database Revision. *In*: Proceedings of the ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, 32/4, 641-648.

Walter, V. [2004]: Object-Based Classification of Remote Sensing Data for Change Detection. ISPRS Journal of Photogrammetry and Remote Sensing, 58/3-4, 225-238.

Wikström, L. [2006]: Final Specification of Road Network Exchange Format. EuroRoads, D 6.1.

Wilhelmy, H., A. Hüttermann & P. Schröder [2002]: Kartographie in Stichworten (Taschenbuch). Borntraeger, Berlin, Stuttgart, 391 S.

Wiltschko, T. [2004]: Sichere Information durch infrastrukturgestützte Fahrerassistenzsysteme zur Steigerung der Verkehrssicherheit an Straßenknotenpunkten. Dissertation, Fortschritt-Bericht VDI, Reihe 12, Nr. 570.

Winter, S. [1994]: Uncertainty of Topological Relations in GIS. *In*: Proceedings of the ISPRS Commission III Symposium, Munich, 30, 924-930.

Wong, D. W. S. & J. Lee [2005]: Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS (Gebundene Ausgabe). John Wiley & Sons, Hoboken, New Jersey, 464 S.

Ying, S., L. Li, X. Liu, H. Zhao & D. Li [2007]: Change-Only Modeling in Navigation Geo-Databases. *In*: Proceedings of the ISPRS Workshop on Updating Geo-spatial Databases with Imagery & the 5th ISPRS Workshop on DMGISs, Urumchi, Xingjiang, China, 207-213.

Yuan, S. & C. Tao [1999]: Development of Conflation Components. *In*: Proceedings of the Geoinformatics'99 Conference, Ann Arbor, Michigan, USA, 39/3, 363-372.

Zhang, M. & L. Meng [2006]: Implementation of a Generic Road-Matching Approach for the Integration of Postal Data. *In*: Proceedings of the 1st ICA Workshop on Geospatial Analysis and Modeling, Vienna, Austria, 141-154.

Zhang, M., L. Meng & H. Qian [2007]: A Structure-Oriented Matching Approach for the Integration of Different Road Networks. *In*: Proceedings of the XXIII International Cartographic Conference (ICC), Incremental Updating and Versioning of Spatial Data, Moscow, Russia, CDROM.

Anhang A Heterogene Beschreibungen der Attribute in den Datensätzen

Im Folgenden werden die Beschreibungen der Attribute für die linienförmigen Straßenobjekte in den Datensätzen aufgelistet. Das globale Datenmodell basiert auf dem GDF 3.0 Datenmodell.

GDF 3.0 **(aus [ERTICO 1995])**	**NavTeq** **(aus [NavTeq 2005])**	**TeleAtlas** **(aus [TeleAtlas 2005])**	**OpenStreetMap** **(aus [OpenStreetMap 2008])**
LinkID Identification	**LINK_ID** The unique number used to identify each link in the NAVSTREETS database	**ID** Feature Identification.	**OSM_ID** Object Identification
Name (Text) Street Name	**ST_NAME (Text)** The NAVTEQ attributes Feature Base Name, Street Type, Prefix, and Suffix, are combined to form the full Street Name (Alle Buchstaben großgeschrieben)	**NAME (Text)** Official Street Name or Route Number (Erste Buchstaben im Wort großgeschrieben).	**Name (Text)** Street Name (Erste Buchstaben im Wort großgeschrieben)
Direction of Traffic Flow (DF) • '1' - Allowed in both directions • '2' - Closed in positive direction • '3' - Closed in negative direction • '4' - Closed in both directions	**DIR_TRAVEL** A code used to indicate the direction of traffic flow on a navigable link. This value/code is categorized under the Reference Class BEARING. Applicable values are: • 'F' - Direction of Travel is one way from the reference end of the street to the nonreference	**ONEWAY** Direction of Traffic Flow • Blank - Open in Both Directions (default) • 'FT' - Open in Positive Direction • 'N' - Closed in Both Directions • 'TF' - Open in Negative Direction	**Oneway** To indicate whether a street is oneway • 'true' or 'yes' Closed in positive direction • 'yes' Allowed in both directions **Access**

	end. • 'T' - Direction of Travel is one way from the non-reference end of the street to the reference end. • 'B' - Travel is allowed in both directions.		Further Drving Restriction • 'no' or 'false' Closed in both directions
Ferry Type (FT) • '0' - Default • '1' - Operated by ship or hovercraft • '2' - Operated by train	**FERRY_TYPE** The Ferry Type attribute indicates if the street is part of a boat or ferry route. This field is used for display purposes only. The applicable values are: • 'H' - Street Route. • 'B' - Boat Ferry Route. • 'R' - Rail Ferry Route.	**FT** Ferry Type • '0' - No Ferry (default) • '1' - Ferry Operated by Ship or Hovercraft • '2' - Ferry Operated by Train	-
Form of Way (FW) • '1' - Part of a motorway • '2' - Part of a multiple carriageway which is not a Motorway • '3' - Part of a single carriageway • '4' - Part of a roundabout • '5' - Part of a traffic square • '6' - Part of an Enclosed Traffic Area: parking place • '7' - Part of an Enclosed Traffic Area: parking building • '8' - Part of an Enclosed Traffic	This attribute is not specified in the NAVSTREETS database. It is devided into different attributes, e.g. • **MULTIDIGIT** (2 & 3) • **ROUNDABOUT** (4) • **UNDEFTRAFF** (5) • **RAMP** (10) • **SPECTRFIG** (17) • **INDESCRIB** (20)	**FOW** Form of Way • '-1' - Not Applicable • '1' - Part of Motorway • '2' - Part of Multi Carriageway which is Not a Motorway • '3' - Part of a Single Carriageway (default) • '4' - Part of a Roundabout • '6' - Part of an ETA: Parking Place • '7'- Part of an ETA: Parking Garage (Building) • '8' - Part of an ETA:	**Highway** Values of Highway • default • 'motorway' • 'motorway_link' • 'primary' • 'primary_link' • 'secondary' • 'cycleway' • 'service' • 'footway' • 'living_street' • 'path' • 'pedestrian'

Area: unstructured traffic square • ‘9’ - Part of another type of Enclosed Traffic Area • ‘10’ - Part of a slip road • ‘11’ - Part of a service road • ‘12’ - Entrance/exit to/from a car park • ‘13’ - Entrance/exit to/from a service • ‘14’ - Part of a pedestrian zone • ‘15’ - Part of a walkway not passable for vehicles		Unstructured Traffic Square • ‘10’ - Part of a Slip Road • ‘11’ - Part of a Service Road • ‘12’ - Entrance / Exit to / from a Car Park • ‘14’ - Part of a Pedestrian Zone • ‘15’ - Part of a Walkway • ‘17’ - Special Traffic Figures • ‘20’ - Road for Authorities	• ‘residential’ • ‘steps’ • ‘tertiary • ‘track’ • ‘unclassified’
Functional Road Class (FC) • ‘0’ - Main road • ‘1’ - First class road • ‘2’ - Second class road • ‘3’ - Third class road • ‘4’ - Fourth class road • ‘5’ - Fifth class road • ‘6’ - Sixth class road • ‘7’ - Seventh class road • ‘8’ - Eighth class road • ‘9’ - Ninth class road	**FUNC_CLASS** Functional Class defines the network used to determine a logical and efficient route for a traveller. The Streets layer uses the following Functional Class Levels: • ‘1’ - Roads with very few, if any speed changes, typically controlled access, and provide high volume, maximum speed movement between and through major metropolitan areas. • ‘2’ - Roads with very few, if any speed changes, and provide high volume, high speed traffic movement. Typically used to channel traffic to (and from) Level 1 roads.	**FRC** Functional Road Class • ‘-1’ - Not Applicable (for FeatTyp 4165) • ‘0’ - Motorways • ‘1’ - Roads not belonging to ‘Main Road’ Major Importance • ‘2’ - Other Major Roads • ‘3’ - Secondary Roads • ‘4’ - Local Connecting Roads • ‘5’ - Local Roads of High Importance • ‘6’ - Local Roads • ‘7’ - Local Roads of Minor Importance • ‘8’ - Others	-

	• '3' - Roads which interconnect Level 2 roads and provide a high volume of traffic movement at a lower level of mobility than Level 2 roads. • '4' - Roads that provide for a high volume of traffic movement at moderate speeds between neighborhoods. • '5' - All other roads. • 'NA' - Not Applicable'		
Ownership (OW) • '1' - Publicly owned • '2' - Privately owned	**PRIVATE** Indicates if the street is private. • 'Y' - means the street is private. • 'N' - means the street is not private.	**PRIVATERD** Private Road • '0' - No Special Restriction (default) • '2' - Not Publicly Accessible	**Access** • 'destination' 1 • 'no' 1 • 'private' 2 • 'psv' 1
Road Number (RN) (Text)	**ST_NAME** The NAVTEQ attributes Feature Base Name, Street Type, Prefix, and Suffix, are combined to form the full Street Name (Alle Buchstaben großgeschrieben)	**ROUTENUM** Route Number (Text)	-
Toll Road (TR) • '0' - Not a Toll Road • '1' - Toll Road	**TOLLWAY** Indicates if the street is a tollway. • 'Y' - means the street is a tollway. • 'N' - means the street is not a tollway.	**TOLLRD** Toll Road • 'Blank' - No Toll Road (default) • 'B' - Toll Road in Both Directions • 'FT' - Toll Road in Positive	-

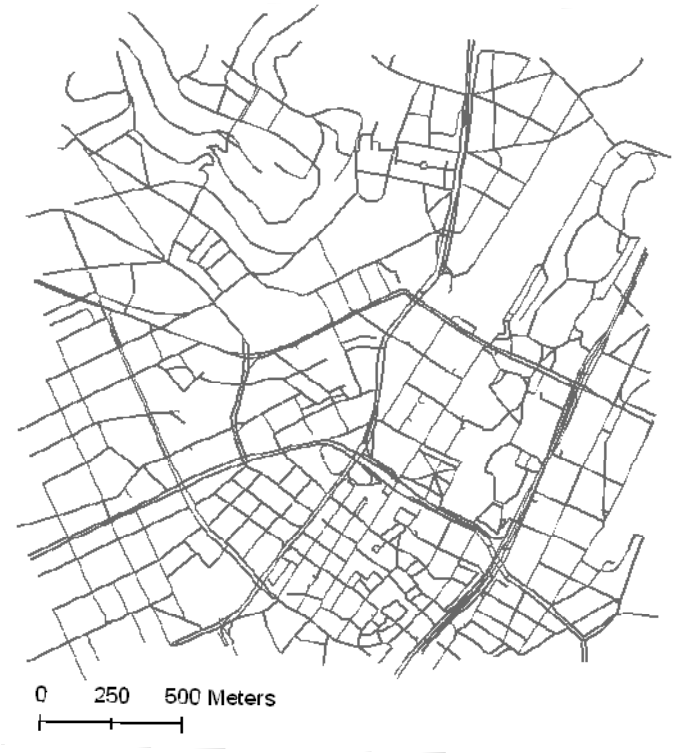
		Direction • 'TF'- Toll Road in Negative Direction	
F_JNCID From Junction Identification	**REF_IN_ID** The unique Intersection ID (NAVTEQ NodeID) used to identify the Reference End of the Intersection.	**F_JNCTID** From Junction Identification	**F_ID** From Node Identification
T_JNCID To Junction Identification	**NREF_IN_ID** The unique Intersection ID (NAVTEQ NodeID) used to identify the Non-Reference End of theIntersection.	**T_JNCTID** To Junction Identification	**T_ID** To Node Identification
Left House Number Structure (L_Struct) • '1' - Odd address range • '2' - Even address range • '3' - Mixed address range • '4' - Undefined address range	**L_ADDRSCH** The numbering scheme for the left address range. This value/code is categorized under the Reference class ADDRSCH. Applicable values are: • 'M' - Mixed address range. • 'O' - Odd address range. • 'E' - Even address range. • ' ' - Undefined address range	**GC.L_Struct** Left House Number Structure • '0' - Not Applicable (default) • '1' - No House Numbers at All • '2' - Even • '3' - Odd • '4' - Mixed • '5' - Irregular House Number Structure	-
First House Number Left (LS)	**L_REFADDR** The left side reference address.	**GC.L_F_Add** Left First Base House Number	-
First House Number Right (RS)	**L_NREFADDR** The left side non-reference address.	**GC.L_T_Add** Left Last Base House Number	
Reft House Number Structure	**R_ADDRSCH**	**GC.R_Struct**	-

(R_Struct) • '1' - Odd address range • '2' - Even address range • '3' - Mixed address range • '4' - Undefined address range	The numbering scheme for the left address range. This value/code is categorized under the Reference class ADDRSCH. Applicable values are: • 'M' - Mixed address range. • 'O' - Odd address range. • 'E' - Even address range. • ' ' - Undefined address range	Right House Number Structure • '0' - Not Applicable (default) • '1' - No House Numbers at All • '2' - Even • '3' - Odd • '4' - Mixed • '5' - Irregular House Number Structure	
Last House Number Left (LE)	**R_REFADDR** The right side reference address.	**GC.R_F_Add** Right First Base House Number	-
Last House Number Right (RE)	**R_NREFADDR** The right side non-reference address.	**GC.R_T_Add** Right Last Base House Number	-
Speed Category (SPEED_CAT) • '0'- Not Applicable (default) • '1' - Greater than 130 kph / 80 mph • '2' - 101-130 kph / 65-80 mph • '3' - 91-100 kph / 55-64 mph • '4' - 71-90 kph / 41-54 mph • '5' - 51-70 kph / 31-40 mph • '6' - 31-50 kph / 21-30 mph • '7' - 11-30 kph / 6-20 mph • '8' - Less than 11 kph / 6 mph	**SPEED_CAT** A code that classifies the speed of the road based on posted or legal speed and is used to enhance route calculation. The applicable values are: • '1' - Greater than 130 kph / 80 mph • '2' - 101-130 kph / 65-80 mph • '3' - 91-100 kph / 55-64 mph • '4' - 71-90 kph / 41-54 mph • '5' - 51-70 kph / 31-40 mph • '6' - 31-50 kph / 21-30 mph • '7' - 11-30 kph / 6-20 mph	**KPH** Calculated Average Speed (kilometers per hour)	**MaxSpeed** Maximal Allowed Speed **Minspeed** Minimal Allowed Speed

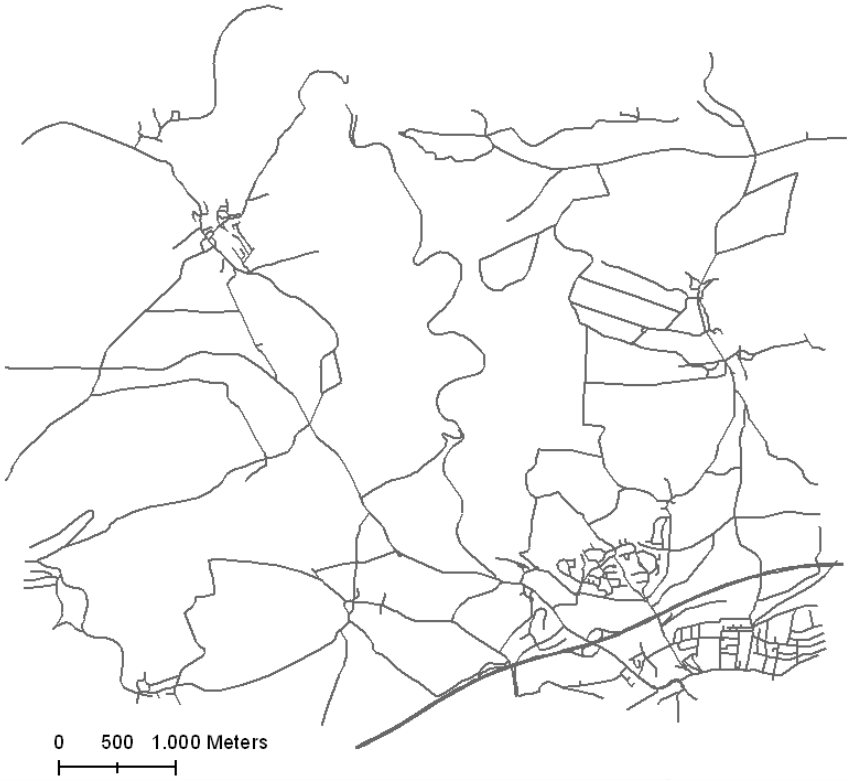
	• ‘8’ - Less than 11 kph / 6 mph • 'NA' - Not Applicable'		
Maximum Height Allowed (MH)	-	-	**MaxHeight** Maximal Allowed Height
Maximum Total Weight Allowed (MT)	-	-	**MaxWeight** Maximal Allowed Weight

Anhang B Graphische Übersicht der Testgebiete

Testgebiet I (NavTeq Q1/2005)



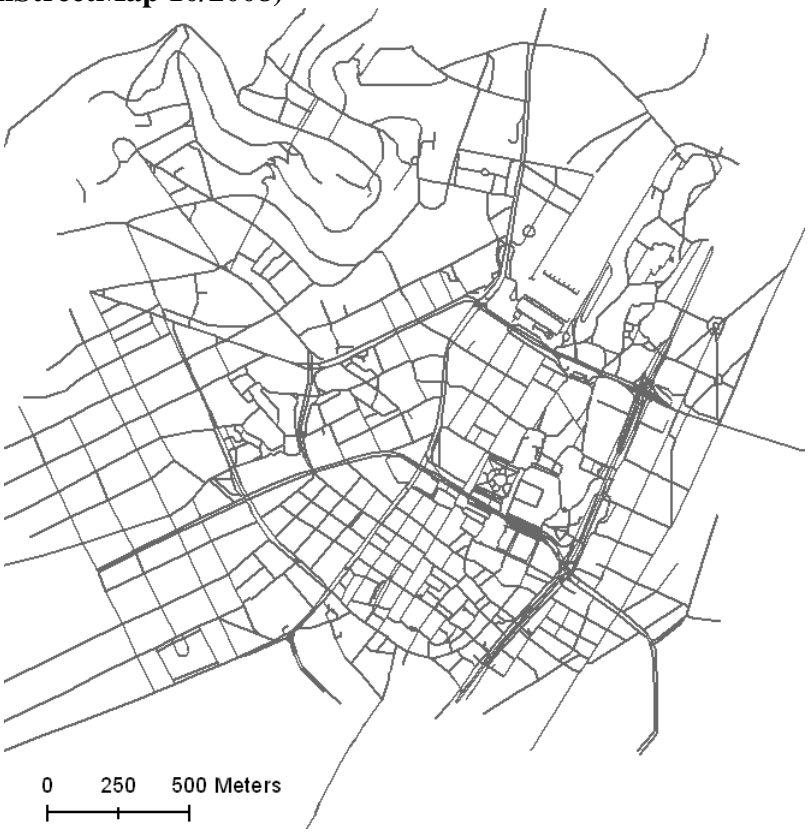
Testgebiet II (NavTeq Q1/2005)



Testgebiet I (TeleAtlas Q1/2006)

Testgebiet II (TeleAtlas Q1/2006)

Testgebiet I (OpenStreetMap 10/2008)



Testgebiet II (OpenStreetMap 10/2008)

Anhang C Manuelle Bewertung der Zuordnungspaare

Im Folgenden werden die Kriterien und die Ergebnisse der manuellen Bewertung der Zuordnungspaare vorgestellt.

Kriterien der manuellen Bewertung

Bei der manuellen Zuordnung werden die Zuordnungspaare nach Gesichtspunkten der topologischen Ähnlichkeit (Anzahl der Knoten und Konnektivität), geometrischen Ähnlichkeit (Lage, Verlauf und geometrische Modellierung) und thematischen Ähnlichkeit (Straßennamen) bewertet. Die Zuordnungspaare werden in folgende fünf Gruppen eingeteilt:

- „*Identisch*“: Das Zuordnungspaar ist mit der Relation *1:1* erfasst und die Anfangs- und Endpositionen liegen weniger als zehn Meter voneinander entfernt. Der Verlauf der Linien ist sehr ähnlich und der Längenunterschied ist kleiner als fünf Meter. Die Konnektivität mit anderen Kanten ist in beiden Datensätzen identisch. Darüber hinaus ist der Straßenname in beiden Datensätzen identisch.

- „*Sehr Ähnlich*“: Das Zuordnungspaar ist nicht mit der Relation *1:1* zugeordnet. Allerdings betragen die Abweichungen von Anfangs- und Endpositionen weniger als zehn Meter. Der Verlauf der Linien ist identisch und der Längenunterschied ist kleiner als zehn Meter. Die Konnektivität mit anderen Kanten ist in beiden Datensätzen identisch. Außerdem ist der Straßenname in beiden Datensätzen identisch.

- „*Ähnlich*“: Die Abweichungen von Anfangs- und Endpositionen in beiden Datensätzen betragen weniger als zwanzig Meter und die geometrische Modellierung in beiden Datensätzen ist ähnlich. Die Konnektivität mit anderen Kanten ist in beiden Datensätzen identisch.

- „*Wenig Ähnlich*“: Die Abweichungen von Anfangs- und Endpositionen sind größer als zwanzig Meter. Außerdem ist die geometrische Modellierung oder der Straßenname in beiden Datensätzen unterschiedlich. Die Konnektivität mit anderen Kanten kann auch in beiden Datensätzen unterschiedlich sein.

- „*Unähnlich*“: Hierbei sind zwei Fälle zu unterscheiden. Im ersten Fall sind die Abweichungen der Anfangs- und Endpositionen größer als zwanzig Meter und die Straßennamen in beiden Datensätzen unterschiedlich. Beim zweiten Fall handelt es sich um unterschiedliche geometrische Modellierungen (Zuordnung zwischen Knoten und Kanten).

Ergebnisse der manuellen Bewertung

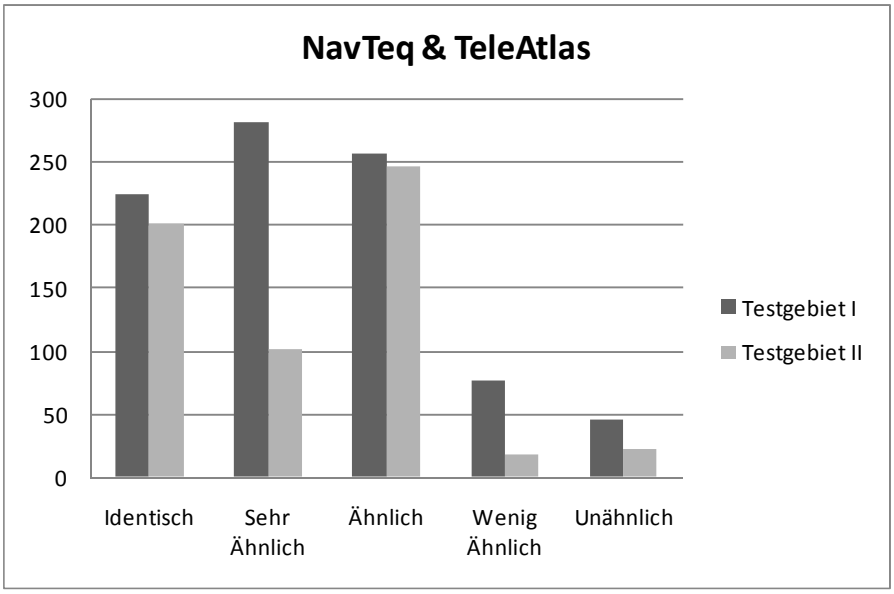


Abbildung C.1: Ergebnisse der manuellen Bewertung der Zuordnungspaare zwischen NavTeq und TeleAtlas

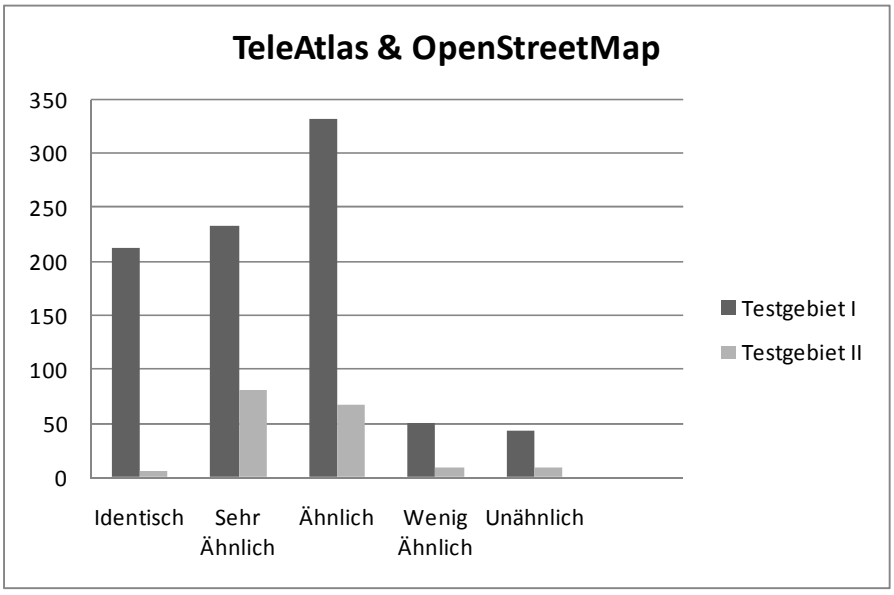


Abbildung C.2: Ergebnisse der manuellen Bewertung der Zuordnungspaare zwischen TeleAtlas und OpenStreetMap

Anhang D Verteilung der lokalen geometrischen und topologischen Ähnlichkeit

Verteilung der lokalen geometrischen Ähnlichkeiten

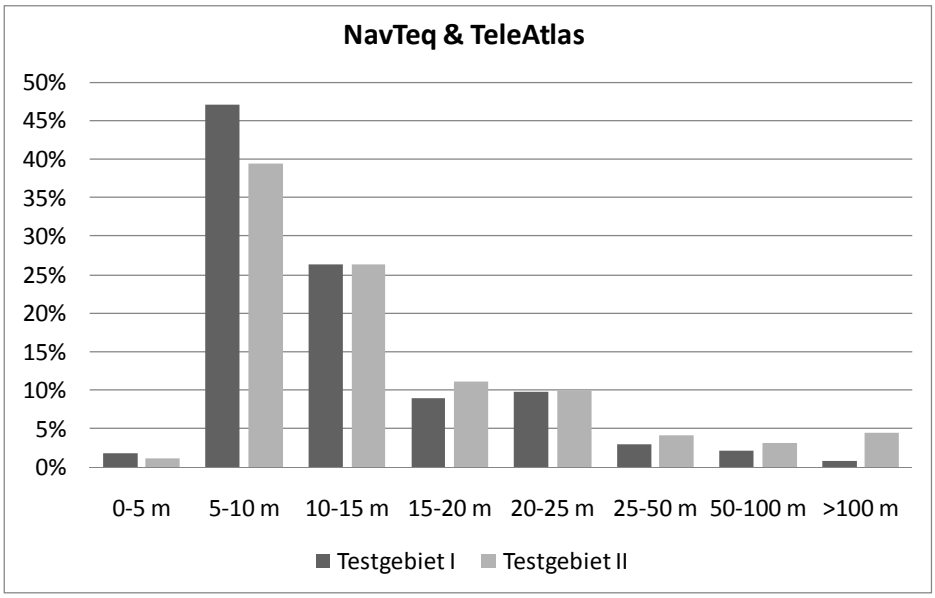


Abbildung D.1: Häufigkeitsverteilung der geometrischen Ähnlichkeiten zwischen NavTeq und TeleAtlas

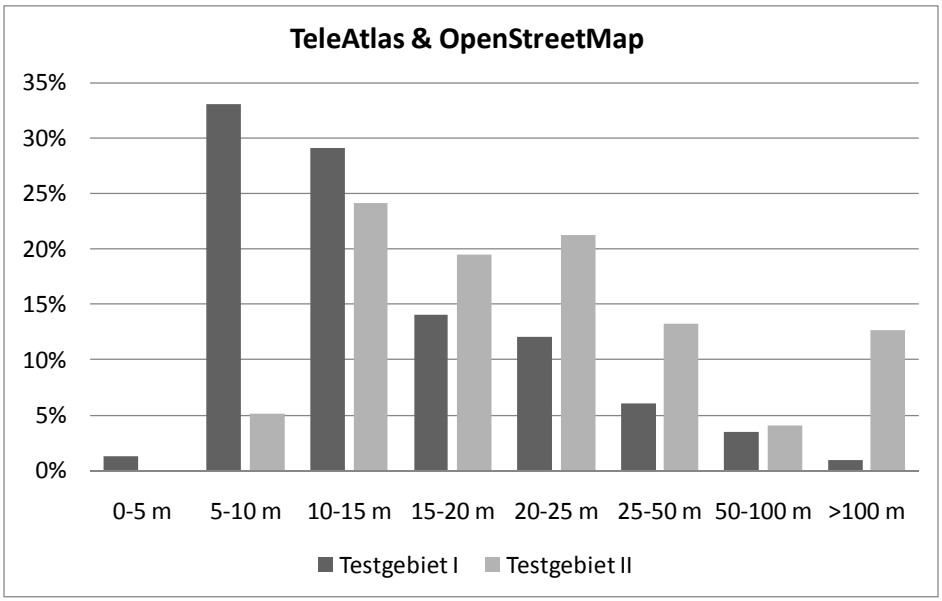


Abbildung D.2: Häufigkeitsverteilung der geometrischen Ähnlichkeiten zwischen TeleAtlas und OpenStreetMap

Verteilung der lokalen topologischen Ähnlichkeiten

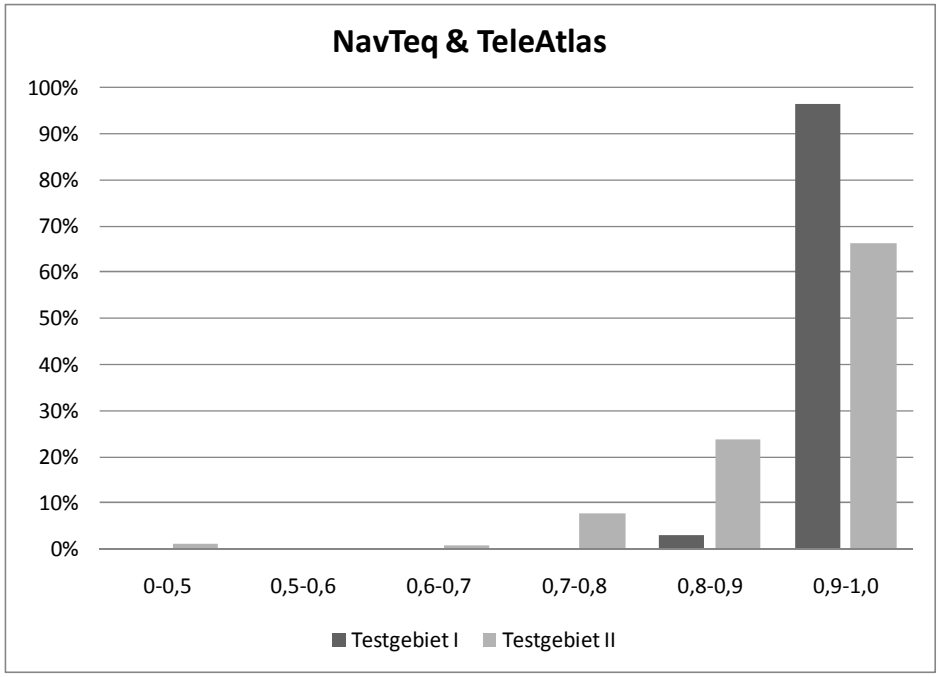


Abbildung D.3: Häufigkeitsverteilung der topologischen Ähnlichkeiten zwischen NavTeq und TeleAtlas

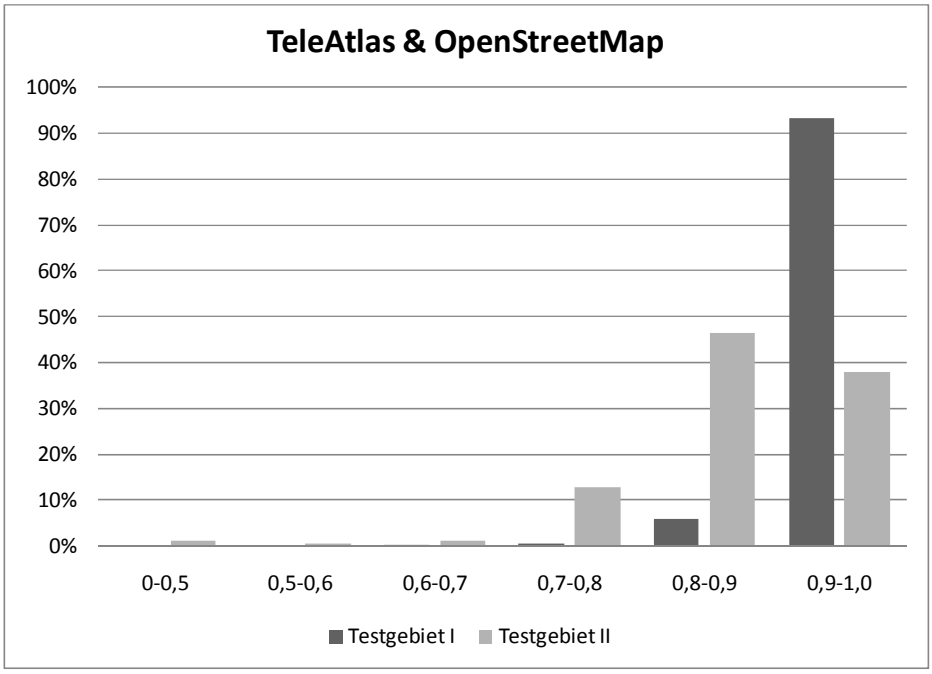


Abbildung D.4: Häufigkeitsverteilung der topologischen Ähnlichkeiten zwischen TeleAtlas und OpenStreetMap

Lebenslauf

Hainan Chen

13. Mai 1979	Geboren in Fujian, China
1986 - 1997	Grundschule und Hauptschule in Fujian, China
1997 - 2001	Bachelorstudium der Geoinformatik und Kartographie an der Universität Wuhan in China
2001 - 2006	Diplomstudium der Geodäsie und Geoinformatik an der Universität Stuttgart
2006 - 2009	Doktorand in der Daimler Forschung und Entwicklung in Zusammenarbeit mit Prof. Dieter Fritsch beim Institut für Photogrammetrie der Universität Stuttgart
2009 - 2010	Doktorand im Institut für Photogrammetrie der Universität Stuttgart
2010 -	Entwicklungsingenieur bei MBtech Group GmbH & Co. KGaA (Daimler AG)