Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 853

Birgit Kieler

Schema-Matching in räumlichen Datensätzen durch Zuordnung von Objektinstanzen

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 | ISBN 978-3-7696-5265-9

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover
ISSN 0174-1454, Nr. 359, Hannover 2020

Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 853

Schema-Matching in räumlichen Datensätzen durch Zuordnung von Objektinstanzen

Von der Fakultät für Bauingenieurwesen und Geodäsie

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

Vorgelegt von

Dipl.-Ing. Birgit Kieler

Geboren am 24.01.1979 in Berlin

München 2020

Verlag der Bayerischen Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5265-9

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Universität Hannover
ISSN 0174-1454, Nr. 359, Hannover 2020

Adresse der DGK:

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften (DGK)

Alfons-Goppel-Straße 11 ● D – 80 539 München

Telefon +49 – 331 – 288 1685 ● Telefax +49 – 331 – 288 1759

E-Mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission:

Vorsitzender: Prof. Dr.-Ing. habil. Jürgen Müller

Referentin: Prof. Dr.-Ing. habil. Monika Sester

Korreferenten: Prof. Dr.-Ing. habil. Jan-Henrik Haunert (Universität Bonn)
Prof. Dr.-Ing. habil. Christian Heipke

Tag der mündlichen Prüfung: 21.02.2020

© 2020 Bayerische Akademie der Wissenschaften, München

Alle Rechte vorbehalten. Ohne Genehmigung der Herausgeber ist es auch nicht gestattet,
die Veröffentlichung oder Teile daraus auf photomechanischem Wege (Photokopie, Mikrokopie) zu vervielfältigen

ISSN 0065-5325 ISBN 978-3-7696-5265-9

Kurzfassung

Datenintegration, Datenaustausch und Datenaktualisierungen sind im wissenschaftlichen Umfeld nach wie vor aktuelle Themen. Geoinformationen werden zur Unterstützung in vielen Entscheidungen in Politik, Wirtschaft, Verwaltung, aber auch im Alltag benötigt. Im Datenintegrationsprozess verursacht das Kombinieren verschiedener Datensätze aufgrund struktureller, geometrischer und semantischer Unterschiede erhebliche Probleme.

In der vorliegenden Arbeit wurde ein zweistufiges Zuordnungsverfahren entwickelt, das einen Beitrag zur geometrischen und semantischen Datenintegration leistet. Im ersten Schritt wurde ein Objektzuordnungsverfahren entwickelt, das Objektkorrespondenzen durch die geometrische Überlagerung zweier Datensätze mit Polygonobjekten bestimmen kann. Mit Hilfe eines Gesamtähnlichkeitsmaß, das sich aus geometrischen und semantischen Parametern zusammensetzt, werden die vom Verfahren identifizierten Objektrelationen bewertet. Das Verfahren kann sowohl einfache (1:1) als auch komplexe (1:n, n:1, n:m) Relationen durch Aggregation von Nachbarobjekten bestimmen.

Das Objektzuordnungsverfahren wurde in drei verschiedenen Testgebieten mit vier unterschiedlichen Datensätzen getestet. Es kann Datensätze mit verschiedenen Maßstäben, z.B. 1:1.000 bis 1:25.000 berücksichtigen. Es wurden Datensätze untersucht, die semantisch vergleichbare, aber auch verschiedenartige Objekte besitzen. Eine Besonderheit ist, dass Datensätze nicht als Partition vorliegen müssen, sondern auch Objektüberlagerungen innerhalb der Datensätze zugelassen sind, auch wenn dies den Suchraum vergrößert.

Durch den Vergleich mit manuell erstellten Referenzzuordnungen zeigt sich im Ergebnis, dass die im Objektzuordnungsverfahren verwendeten sehr einfachen Ähnlichkeitsmaße und festgelegten Schwellwerte in allen Testgebieten sehr hohe Zuordnungsqualitäten erzielen können, in zwei Fällen sogar über 92 %. Und je ähnlicher sich die Maßstäbe der Datensätze sind, desto vollständiger sind auch die richtigen Zuordnungen. Das bedeutet, dass das Verfahren nur sehr wenige falsche Relationen identifiziert und somit für unterschiedliche Datensätze einsetzbar ist.

Im zweiten Schritt des Verfahrens erfolgt die Zuordnung der Objektklassen auf Schemaebene. Dazu werden aus den Ergebnissen der Objektzuordung Häufigkeitswerte abgeleitet und pro Testgebiet in vier unterschiedlichen Matrizen zusammengefasst. Neben Relationsanteilen wurden auch datensatzbezogene prozentuale Flächenanteile berechnet, die besonders bei großen Maßstabsunterschieden Vorteile bieten. Aufgrund verschiedener Sichtweisen können Häufigkeitsmatrizen auch als bipartite Graphen betrachtet werden.

Für die Zuordnung der Objektklassen wurden existierende Graphalgorithmen verwendet. Neben dem Ansatz des Maximalen Matchings, das nur 1:1 -Schemarelationen in quadratischen Matrizen bestimmen kann, wurde ein heuristisches Verfahren entwickelt, das den Ansatz des Minimalen-2-Schnitts rekursiv anwendet. Mit dem Heuristischen Verfahren können zusätzlich einseitig komplexe Schemarelationen bestimmt werden. Als Ergebnis entstehen in der Zuordnungsmatrix rechteckige Cluster, die sich nicht schneiden und nur pro Zeile und Spalte eine Zuordnung zulassen.

Die Zuordnung der Objektklassen entspricht einer Unterteilung eines Graphen in k Teile, was ein $\mathcal{NP}$ -vollständiges Problem beschreibt, für das nach heutigem Kenntnisstand kein Algorithmus mit polynomieller Laufzeit existiert. Um die nicht garantiert optimalen Ergebnisse des Näherungsverfahrens bewerten zu können, wurden verschiedene ganzzahlige lineare Optimierungsverfahren entwickelt und mit dem existierenden Optimierer (IBM ILOG CPLEX Interactive Optimizer 12.5.1.0) gelöst. Das primäre Optimierungsziel ist die Maximierung der Häufigkeiten innerhalb der Cluster, um die Zuordnung der Objektklassen durch die identifizierten Objektrelationen zu bestätigen. Zusätzlich wurde das Erzeugen von ausgewogenen Clustern hinsichtlich der Zellenanzahl pro Cluster als zweites Optimierungsziel eingefügt. Beide Ziele wurden kombiniert, einerseits gleichgewichtet und andererseits durch Einführung eines Ziels als harte Bedingung. Die Lösung mit der größten Durchschnittshäufigkeit pro Zelle wird als beste Lösung ausgewählt.

In den Ergebnissen zeigt sich, dass das Heuristische Verfahren beim Vergleich in den Kategorien Rechenzeit, Schnittmenge zum Matrixgesamtinhalt und zur Referenzzuordnung gegenüber den anderen Verfahren den ersten Platz belegt. Die Optimierungsverfahren belegen aufgrund der sehr langen Rechenzeiten den zweiten Platz. Der Vergleich der Heuristischen Lösungen mit den besten Optimierungslösungen zeigt, dass in allen Testgebieten die optimalen Lösungen mehr als 91% der Heuristischen Lösungen bestätigen. Das Näherungsverfahren stellt somit einen effizienten Ansatz dar, um das Problem der Objektklassenzuordnung auf Schemaebene automatisch zu lösen.

Schlagworte:
Data-Matching, Schema-Matching, Optimierung, ganzzahlige lineare Programmierung

Abstract

Data integration, exchange and updates are still challenging subjects in the current scientific context. Spatial data are necessary to assist in political, economical, administrative, or common everyday life's decisions. The combination of different datasets causes significant problems due to structural, geometric, and semantic differences during the data integration process.

In this thesis a two-step matching approach has been developed to make a contribution to geometric and semantic data integration. The first step consists of a data matching process in order to derive object correspondences by the geometrical overlay of two polygon object datasets. Using a total similarity measure, composed from geometric and semantic parameters, object relations identified by the matching approach are being evaluated. It is possible to determine as well simple (1:1) as complex (1:n, n:1, n:m) relations by aggregation of neighbouring objects in the proposed approach.

The data matching approach has been tested on three test areas consisting of four different datasets. It is possible to handle datasets with different scales, ranging from 1:1.000 to 1:25.000. The examined datasets consist as well of semantic similar as dissimilar objects. As a special feature datasets do not have to be presented as a partition of the plane but object overlays within the datasets are allowed, even though the search area is expanded by this.

The usage of simple similarity measures and specified thresholds shows in comparison to manually created reference data very good matching results in all three test areas, in two cases even higher than 92%. The more similar the scales of the datasets are, the more complete are the correct matches. That means, the presented approach does identify only few wrong relations and thus can be applied to different datasets.

The second step comprises the matching of object classes on schema level. To achieve this frequency values are derived from the results of the data matching and subsequently summarized in four different matrices per test area. Beside object relation values, also area based percentage values were calculated to offer benefits when dealing with large scale differences. These frequency matrices can also be considered as bipartite graphs.

For the matching of the object classes existing graph algorithms were used. Besides the maximum matching approach that can only process square matrices and identifies simple 1:1 object relations, a heuristical approach, that uses the minimum-2-cut approach recursively has been developed. This heuristic approach allows to calculate additionally one-side complex 1:n/n:1 schema relations. As a result rectangular clusters that do not intersect and allow only one matching per row and column are created in the frequency matrices.

The matching of object classes corresponds with the subdivision of a graph in k parts. That describes an $\mathcal{NP}$-complete problem that can not be solved with an algorithm in polynomial time. To evaluate the not guaranteed optimal results of the approximation process different integer linear optimization approaches have been developed and solved with the existing optimizer (IBM ILOG CPLEX Interactive Optimizer 12.5.1.0). The primary optimization objective is the maximization of frequencies within the clusters to confirm the matching of the object classes with the identified object relations. Additionally a second objective has been defined to generate balanced clusters concerning the number of cells. Both objectives have been combined, on the one hand with a cost function and on the other hand by introducing one criteria as a hard constraint. The solution with the highest average frequency per cell is selected as best solution.

The results show the heuristic approach as the best concerning computing time, intersection to total content of the matrix and to the manually created reference matchings in comparison to the other approaches. The optimization approaches are second place due to long computing times. Comparing the heuristic solutions with the best solutions of the optimization approaches more than 91% of the heuristic results are confirmed by the results of the optimization approaches in all test areas. The approximation approach represents an efficient approach to solve the matching of object classes on schema level automatically.

Keywords:
data-matching, schema-matching, optimization, integer linear programming

Inhaltsverzeichnis

1 Einleitung **9**

1.1 Motivation . . . 9
1.2 Zielsetzung . . . 11
1.3 Gliederung . . . 12

2 Verwandte Arbeiten **13**

2.1 Grundbegriffe . . . 13
2.1.1 Raumbezogene Objekte . . . 13
2.1.2 Ähnlichkeit . . . 13
2.1.3 Relation . . . 14
2.1.4 Schema . . . 14
2.2 Data-Matching . . . 15
2.2.1 Klassifikation von Zuordnungsverfahren auf Objektebene . . . 15
2.2.2 Herausforderungen bei der Objektzuordnung . . . 16
2.2.3 Ausgewählte, merkmalsbasierte Verfahren . . . 18
2.2.4 Ausgewählte, relationale Verfahren . . . 20
2.3 Schema-Matching . . . 23
2.3.1 Klassifikation von Zuordnungsverfahren auf Schemaebene . . . 24
2.3.2 Herausforderungen bei der Zuordnung auf Schemaebene . . . 25
2.3.3 Ausgewählte Schema-Matching-Verfahren im geographischen Kontext . . . 26

3 Grundlagen **29**

3.1 Ähnlichkeitsmaße . . . 29
3.1.1 Geometrische Ähnlichkeit . . . 29
3.1.2 Topologische Ähnlichkeit . . . 32
3.1.3 Semantische Ähnlichkeit . . . 34
3.2 Relationstypen . . . 35
3.2.1 Relationen auf Objektebene . . . 35
3.2.2 Relationen auf Schemaebene . . . 36
3.3 Graphentheorie . . . 36
3.3.1 Graph-Definitionen . . . 37
3.3.2 Graph-Matching . . . 37
3.3.3 Graph-Partitionierung / Graph-Cut . . . 39
3.4 Ganzzahlige lineare Programmierung . . . 41

4 Entwicklung von Data-Matching-Verfahren für verschiedene Objektgeometrien **45**

4.1 Zuordnung von Polygonobjekten . . . 45
4.1.1 Geometrischer Parameter . . . 45
4.1.2 Heterogenitätsparameter . . . 47
4.1.3 Erzeugung eines kombinierten Ergebnisses für das Schema-Matching . . . 48
4.2 Zuordnung von unterschiedlichen Objektgeometrien . . . 50

5 Entwicklung von Schema-Matching-Verfahren basierend auf Instanzdaten 53
5.1 Formale Problemdefinition . . . 53
5.1.1 Synthetisches Beispiel . . . 53
5.2 Einfache Lösungsverfahren . . . 54
5.2.1 Beschränkung auf 1:1-Zuordnungen (Max-Match) . . . 54
5.2.2 Beschränkung auf zwei Cluster (Min-Cut) . . . 55
5.3 Einsatz von Heuristiken . . . 56
5.4 Einsatz der ganzzahligen linearen Programmierung . . . 57
5.4.1 Optimierungsziele und Bedingungen . . . 57
5.4.2 Kombination von Optimierungszielen . . . 60
5.4.3 Einführung einer festen Clustergröße (MaxScoreHardConstraintFixedSize) . . . 64
5.4.4 Optimale Lösung ohne Nullcluster (MaxScoreHardConstraintFixedSizeNonEmpty) . . . 65
5.4.5 Vereinfachtes Programm (MaxScoreHardConstraintFixedSizeUnique) . . . 65

6 Experimente mit Realdaten und Untersuchungsergebnisse 67
6.1 Datenquellen und Datenvorverarbeitung . . . 67
6.1.1 Datenquellen . . . 67
6.1.2 Testgebiete . . . 68
6.1.3 Datenvorverarbeitung . . . 72
6.2 Ergebnisse des Data-Matching . . . 72
6.2.1 Testgebiet A: ALKIS - OSM in Hannover . . . 72
6.2.2 Testgebiet B: ALKIS - ATKIS in Hameln . . . 76
6.2.3 Testgebiet C: ATKIS - GDF in Hannover-Wedemark . . . 80
6.2.4 Zusammenfassung der Data-Matching-Ergebnisse . . . 83
6.3 Ergebnisse des Schema-Matching . . . 86
6.3.1 Testgebiet B: ALKIS - ATKIS in Hameln . . . 86
6.3.2 Testgebiet A: ALKIS - OSM in Hannover . . . 94
6.3.3 Testgebiet C: ATKIS - GDF in Hannover-Wedemark . . . 99
6.3.4 Zusammenfassung aller Schema-Matching-Ergebnisse . . . 103

7 Zusammenfassung und Ausblick 105

A Testgebiet A: ALKIS - OSM in Hannover 109
A.1 Semantik der Datensätze . . . 109
A.2 Häufigkeitsmatrizen . . . 111
A.3 Ergebnis des Heuristischen Verfahrens für H_R . . . 113

B Testgebiet B: ALKIS - ATKIS in Hameln 115
B.1 Semantik der Datensätze . . . 115
B.2 Häufigkeitsmatrizen . . . 116

C Testgebiet C: ATKIS - GDF in Hannover-Wedemark 117
C.1 Semantik der Datensätze . . . 117
C.2 Häufigkeitsmatrizen . . . 119

Abbildungsverzeichnis 120

Tabellenverzeichnis 124

Literaturverzeichnis 128

Danksagung **135**

Lebenslauf **136**

1 Einleitung

1.1 Motivation

So unterschiedlich wie die Kulturen und Landschaften auf der Erde, so vielfältig und zahlreich sind auch die zur Verfügung stehenden Geoinformationen, die diese Objektinformationen mit Ortsbezug repräsentieren. Als Beispiele sind Gebäude- und Grundstücksinformationen, Verkehrsnetze, Schutzgebiete, Gewässer oder auch Klima- und Geologiedaten zu nennen. Auf den ersten Blick scheint die Fülle der raumbezogenen Daten nur Vorteile zu bieten, da ein Datensatz allein niemals eine vollständige und genaue Repräsentation der sich ständig wandelnden Welt darstellen kann. Auf den zweiten Blick birgt diese Datenfülle viele Gefahren. Beispielsweise, dass relevante Informationen nicht entdeckt werden, weil Datensätze bzw. Dateninhalte keine Metadaten besitzen, oder nicht kombiniert werden können, weil unterschiedliche Modellierungen zugrunde liegen.

Geoinformationen werden zur Unterstützung in vielen Entscheidungen in Politik, Wirtschaft, Verwaltung, aber auch im Alltag benötigt. Für länderübergreifende Themen, wie beispielsweise die Umweltberichterstattung zur Verbesserung des Naturschutzes und das Wassermanagement, aber auch für die Planung von Verkehrsnetzen ist die Nutzung von Geoinformationen essentiell. Datenaustausch, Datenintegration und die Verschmelzung von Daten werden immer wichtiger, um Entscheidungen auf Basis aller zur Verfügung stehenden Informationen treffen zu können.

Das Kombinieren bereits erfasster Daten liegt nahe und ermöglicht das Ableiten neuer Datensätze, aus denen sich wiederum neues Wissen generieren lässt. Ein Teil der sehr zeit- und kostenintensiven Neuerfassung von Geodaten lässt sich somit vermeiden. Gefördert wird dies zunehmend durch den vereinfachten Zugang zu Geodaten. Immer mehr Geodaten werden in digitalen geographischen Datenbanken vorgehalten und über standardisierte Webservices im Internet zur Verfügung gestellt, teilweise sogar ohne Zugriffsbeschränkungen. Dennoch bereitet die Kombination der Datensätze Probleme, besonders wenn die semantischen Datenbeschreibungen fehlen oder deren Aussagekraft unzureichend ist.

Damit in Europa eine länderübergreifende Kombination der vorhandenen amtlichen Geodaten möglich wird, trat am 15. Mai 2007 die INSPIRE[1]-Richtlinie 2007/2/EG (Europäische Union, 2007) in Kraft. Ziel ist, eine europäische Geodateninfrastruktur zu errichten, die relevante, harmonisierte und hochwertige Geoinformationen für alle europäischen Mitgliedsstaaten bis zum Jahre 2020 zur Verfügung stellt. In jedem Mitgliedsstaat werden Geodatensätze identifiziert, die in elektronischer Form vorliegen, sich in Verwendung befinden, d.h. noch nicht archiviert wurden und mindestens einem von 34 definierten INSPIRE -Themenbereichen zugeordnet werden können. Diese Datensätze müssen in vordefinierte und themenspezifische Zielstrukturen überführt, einheitlich mit Metadaten beschrieben und halbjährlich aktualisiert über standardisierte Webdienste für die Suche, die Visualisierung und den Download bereitgestellt werden. Erst dann ist in Europa, über Ländergrenzen hinweg, eine vereinfachte Kombination der amtlichen Geodaten möglich.

Die Überführung der Datensätze vom originären Datenmodell, dem sogenannten Quellmodell, in ein definiertes Zielmodell wird als Transformationsprozess bezeichnet und ist auf verschiedene Art und Weise realisierbar. Bei einer Modelltransformation wird das originäre Datenmodell dahingehend angepasst, dass alle notwendigen Informationen des Zielmodells in das Quellmodell aufgenommen werden. Bei einer Schematransformation ist hingegen keine Veränderung des originären Datenmodells notwendig. Hier werden vielmehr Regeln definiert, die die semantischen Korrespondenzen zwischen dem Quell- und Zielmodell beschreiben. Durch die explizite Anwendung der Transformationsregeln, die beispielsweise ein Zusammenfassen und Umbenennen zweier Quell-Attribute in ein Ziel-Attribut erlauben, wird aus dem Quelldatensatz ein Datensatz abgeleitet, der dem Zielmodell entspricht. Unabhängig davon, welcher Transformationsprozess gewählt wird, muss zuallererst eine Analyse der Datenmodelle erfolgen. Für die zunehmende Anzahl an Daten aus unterschiedlichen Datenquellen ist die semantische Analyse, die hauptsächlich von Experten durchgeführt wird, in Zukunft eine ernorme Herausforderung.

Auch in Deutschland werden Geoobjekte mehrfach von verschiedenen öffentlichen Behörden, privatwirtschaftlichen Institutionen und zunehmend von Laien mit jeweils eigenen Spezifikationen (Li und Goodchild, 2010) freiwillig erfasst und beschrieben. Straßendaten werden beispielsweise sowohl von den Vermessungsverwaltungen

[1] INfrastructure for SPatial InfoRmation in Europe

der Länder für die Erstellung topographischer Karten erfasst als auch von Herstellern von Navigationssystemen. Für die Fahrzeugnavigation sind, neben der als Knoten- und Kantenmodell beschriebenen Geometrie, Straßennamen und Zusatzinformationen, wie z.B. Einbahnstraßenregelungen, Ampelpositionen, Abbiegeverbote, Tempolimits und die Art des Straßenbelags wichtig. In Abhängigkeit vom Anwendungszweck werden unterschiedliche Datenmodelle und Erfassungsmaßstäbe definiert, für die Objekte mit verschiedenen Aufnahmeverfahren und Messinstrumenten mit unterschiedlichen Messgenauigkeiten erfasst werden. Dementsprechend sind für ein und dasselbe Real-Welt-Objekt mehrere Repräsentationsformen möglich, die verschiedene Eigenschaften bzw. Aspekte des Objekts wiedergeben und sich hinsichtlich der strukturellen, geometrischen und semantischen Eigenschaften deutlich unterscheiden können. Diese Unterschiede verursachen in einem Datenintegrationsprozess erhebliche Probleme.

Strukturelle Differenzen, wie unterschiedliche Datenformate oder Koordinatensysteme, lassen sich laut Bishr (1997) durch die Verwendung von standardisierten Datenformaten (z.B. ISO[2], OGC[3]) bzw. durch Konvertierungen und Koordinatentransformationen überwinden. Im Rahmen der INSPIRE-Richtlinie ist die Verwendung des Europäischen Referenzsystems ETRS89[4] vorgeschrieben.

Geometrische Differenzen können erst analysiert und bewertet werden, wenn korrespondierende Objekte in unterschiedlichen Datensätzen des gleichen räumlichen Gebiets mit Hilfe von Objektzuordnungsverfahren, sogenannten Data-Matching-Verfahren, identifiziert werden. Dazu werden insbesondere die räumliche Koinzidenz, die Objektgeometrie bzw. -form analysiert und gegebenenfalls auf Nachbarobjekte ausgedehnt, aber auch Attributwerte und Objektklassenzugehörigkeiten untersucht. In der Vergangenheit wurden Data-Matching-Verfahren entwickelt, die hauptsächlich auf eine Zuordnung von Objekten ähnlicher Gruppen, z.B. von Straßen- oder Gebäudeobjekten, ausgerichtet waren. Das Wissen, dass Objekte semantisch ähnlichen Gruppen angehören, verringert den Suchraum, weil die Objektauswahl begrenzt werden kann. Diese Information ist für den Zuordnungsprozess besonders hilfreich, wenn zwischen den Datensätzen große Maßstabsunterschiede bestehen und unterschiedliche Geometriedimensionen zu erwarten sind. In kleinmaßstäbigen Datensätzen, die mit Hilfe von Generalisierungsregeln zunehmend automatisiert aus detaillierten Datensätzen abgeleitet werden, erfolgt oftmals eine Reduktion der geometrischen Objektdimension. Beispielsweise kann ein Fluss in einer großmaßstäbigen Karte als Fläche und in einer kleinmaßstäbigen Karte als Linie repräsentiert sein (Haunert und Sester, 2008). Folglich erhöhen sich die Anforderungen an das Zuordnungsverfahren, wenn Objekte mit unterschiedlichen Geometriedimensionen einander zugeordnet werden müssen.

Das größte Hindernis für Data-Matching-Verfahren sind jedoch die häufig unbekannten semantischen Korrespondenzen zwischen Datensätzen. Die Beschleunigung der Objektzuordnung durch die Verringerung des Suchraums ist ohne diese Information kaum möglich. Die Identifizierung von bestimmten Informationen in Datensätzen verschiedener Wissensgebiete, die für eine komplexe Fragestellung wichtig sein könnten, ist schwierig, gerade wenn Objekte nicht mit eindeutigen Bezeichnern, wie z.B. Städtenamen, annotiert sind. Für die Bezeichnung von Objekten werden oftmals Klassifikationsschemas genutzt, welche Begriffe aus unterschiedlichen Sprachen, fachspezifisches Vokabular oder alphanumerische Codes verwenden.

Lösungsansätze bieten sogenannte Schema-Matching-Verfahren, mit deren Hilfe explizit semantische Korrespondenzen zwischen Objektgruppen verschiedener Datenmodelle, sogenannte Schemas, bestimmt werden können. Die Mehrzahl der existierenden Verfahren richtet ihre Aufmerksamkeit nur auf die Schemas selbst und ist nicht automatisiert. Rückschlüsse werden hauptsächlich aus linguistischen Vergleichen der formalen Datenbeschreibungen, d.h. Objektklassen- und Attributbeschreibungen bzw. -werte oder durch die Analyse der Struktur des Klassifizierungsschemas gewonnen. Diese Herangehensweise wird von vielen geodatenhaltenden Stellen praktiziert, die ihre Datensätze allein mit Hilfe der zur Verfügung stehenden Modellbeschreibungen in die von INSPIRE vorgegebenen Datenmodelle transformieren müssen.

Existieren zu Datenmodellen auch Objektinformationen, können Objektklassenkorrespondenzen explizit aus den Instanzen der Datensätze gewonnen werden, um unabhängig von den semantischen Informationen zu bleiben, die Fehlinterpretationen hervorrufen können. Aus den räumlichen und geometrischen Objekteigenschaften können neue Informationen abgeleitet werden, die den tatsächlichen Inhalt und die Bedeutung von Objektklassen besser ausdrücken als dies durch eine formale Datenbeschreibung, z.B. durch einen Namen oder eine begrenzte Anzahl von Attributen möglich ist (Duckham und Worboys, 2005). Demzufolge kann ein Data-Matching-Verfahren auch für die Identifizierung von semantischen Korrespondenzen auf Schemaebene oder für die Verifizierung vorhandener Korrespondenzen eingesetzt werden.

[2]Internationale Organisation für Normung (engl. International Organization for Standardization)
[3]Open Geospatial Consortium
[4]Europäisches Terrestrisches Referenzsystem 1989

1.2 Zielsetzung

Ziel der Arbeit ist es, ein Verfahren zu entwickeln, das auf Schemaebene automatisch semantische Korrespondenzen zwischen Objektklassen identifiziert, die von zwei verschiedenen geographischen Datensätzen im gleichen räumlichen Gebiet stammen. Die Zuordnung erfolgt hauptsächlich durch die geometrische Analyse der zugrunde liegenden Objektinstanzen. Solch ein datengetriebenes Verfahren ermöglicht in Zukunft die Aufdeckung semantischer Korrespondenzen zwischen vielen unterschiedlichen Datensätzen, unabhängig davon, ob Schemainformationen zu den Datensätzen vorhanden sind oder Experten für die Interpretation der Schemas zur Verfügung stehen. Die Ergebnisse können für die Erstellung von allgemein gültigen Transformationsregeln zwischen Datenmodellen genutzt und regelmäßig aktualisiert werden.

Im Rahmen der vorliegenden Arbeit wird ein Beitrag zur geometrischen und semantischen Datenintegration geleistet. Abbildung 1.1 stellt den Ablauf des zweistufigen Verfahrens schematisch dar. Zuerst werden zwei verschiedene Data-Matching-Verfahren vorgestellt, die in jeweils zwei Datensätzen automatisch Objektkorrespondenzen zwischen verschiedenen Repräsentationen gleicher Real-Welt-Objekte bestimmen. Ausführlich wird auf die Zuordnung von Polygonobjekten eingegangen. Zusätzlich wird ein Verfahren für die Zuordnung von unterschiedlichen Objektgeometrien vorgestellt. Die Objektklassenzugehörigkeiten und die identifizierten Objektrelationen werden verwendet, um Häufigkeitsmatrizen zu erstellen, die für das anschließende Schema-Matching-Verfahren als Eingabe dienen. In jeder Matrix werden Cluster gebildet, die als semantische Übereinstimmungen zwischen den Objektklassen der verschiedenen Datenmodelle interpretiert werden.

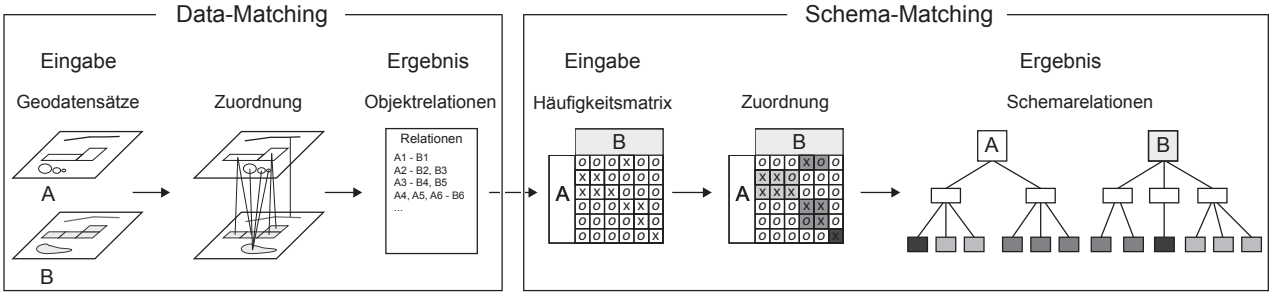


Abbildung 1.1: Schematischer Verfahrensablauf für die automatische Zuordnung von Objektklassen zweier Datensätze A und B auf Basis von geometrischen Objektzuordnungen.

In den Data-Matching-Verfahren stützt sich die Objektzuordnung auf die Auswertung der räumlichen und geometrischen Merkmale der datensatzbeschreibenden Objekte. Für die Zuordnung von Polygonobjekten werden vier Testdatensätze untersucht, die von unterschiedlichen Organisationen für verschiedene Zwecke erfasst werden und sich hinsichtlich ihrer Modellierung und ihres Maßstabs unterscheiden. Neben amtlichen Geodaten werden Navigationsdaten und von Laien erfasste Geoinformationen verwendet. Bei der Zuordnung werden alle Objekte geometrisch überlagert. Um automatisiert eine Zuordnungsentscheidung herbeizuführen, werden objektspezifische Ähnlichkeitsmaße, wie z.B. Überlagerungsverhältnisse und Objektausrichtungen, bestimmt. Basierend auf den Ähnlichkeitsmaßen wird für jede Objektrelation ein kombiniertes Qualitätsmaß abgeleitet.

Das Zuordnungsverfahren für Polygonobjekte kann sowohl Datensätze verarbeiten, die überlagerungsfrei als Partition modelliert sind als auch Datensätze, in denen sich Objekte innerhalb eines Datensatzes geometrisch überdecken, wie z.B. die Überlagerung eines See -Objekts mit einer Freizeitanlage. Des Weiteren können aufgrund der Unterschiedlichkeit der Daten sowohl einfache als auch komplexe Objektrelationen bestimmt werden. In Bezug auf die Objektzuordnung ist das Ziel der Arbeit, zuverlässige Relationen aufzudecken. Es geht weniger um eine vollständige oder garantiert optimale Zuordnung aller Objekte in den Datensätzen. In Hinblick auf die spätere Zuordnung der Objektklassen werden semantisch klare Objektrelationen bevorzugt, an denen nur wenige Objektklassen beteiligt sind. Denn je größer die Vielfalt der Objektklassen in einer Relation, desto unspezifischer ist die semantische Korrespondenz.

Für die Zuordnung der Objektklassen auf Schemaebene werden die Objektrelationen aus dem Data-Matching-Prozess verwendet. Damit ist es möglich, einerseits unabhängig von Schemainformationen zu bleiben und andererseits die in den Objektrelationen verborgenen, semantischen Korrespondenzen zwischen den Objektklassen aufzudecken. Die Zuordnung von Objekten der Klasse See mit Objekten der Klasse Lake aus einem anderen Datensatz lässt sich ohne linguistische Analyse, die erkennt, dass die Objekte ähnliche Sachverhalte beschreiben, nicht ohne Weiteres automatisieren. Werden im Data-Matching-Verfahren allerdings ausschließlich Relationen zwischen den beiden Klassen identifiziert, drückt dies eine eindeutige semantische Übereinstimmung beider

Objektklassen aus. Diese semantische Ähnlichkeitsbeziehung kann durch die Auswertung der Objektrelationen abgeleitet werden.

Das Zuordnen der Objektklassen stellt ein kombinatorisches Problem dar, in dem alle möglichen Kombinationen berücksichtigt werden müssen. Mit steigender Objektklassenanzahl kann dies hohe Rechenzeiten verursachen. Das primäre Ziel ist, die Anzahl der Objektrelationen zu maximieren, um die Zuordnung auf Schemaebene durch die Objektzuordnung zu bestätigen. Aus den Objektrelationen werden verschiedene Häufigkeitwerte bestimmt und in Matrizen zusammengefasst. Jede Häufigkeitsmatrix kann als Graph interpretiert werden, was den Einsatz von existierenden Graphalgorithmen ermöglicht. Einfache Lösungsverfahren können allerdings Einschränkungen im Zuordnungsergebnis verursachen. Im Rahmen der Arbeit werden neben zwei einfachen Lösungsverfahren auch ein neues Heuristisches Schema-Matching-Verfahren und verschiedene neue Optimierungsverfahren entwickelt und vorgestellt, die einfache und komplexe Relationen zwischen mehreren Objektklassen bestimmen. Einige der entwickelten Optimierungsverfahren liefern garantiert optimale Lösungen des Zuordnungsproblems und können als Referenzlösungen für die Bewertung der anderen Verfahren eingesetzt werden.

1.3 Gliederung

In dieser Arbeit werden zunächst verwandte Arbeiten vorgestellt und Grundlagen vermittelt, die für die Entwicklung der Zuordnungsverfahren sowohl auf Objektebene als auch auf Schemaebene benötigt werden. Kapitel 2 erläutert zu Beginn Grundbegriffe, die allgemein für das Verständnis der Arbeit notwendig sind. Anschließend werden im Abschnitt Data-Matching merkmalsbasierte und relationale Zuordnungsverfahren auf Objektebene vorgestellt und Herausforderungen bei der Zuordnung beschrieben. Es werden ausgewählte Arbeiten präsentiert, die Schwierigkeiten, aber auch Lösungsansätze bei der Zuordnung von geographischen Daten verdeutlichen. Im Abschnitt Schema-Matching wird zunächst ein Klassifizierungsschema vorgestellt, das einen Überblick über die Vielfalt der bisher entwickelten Zuordnungsverfahren auf Schemaebene gibt. Neben Problemen und Herausforderungen werden auch hierfür ausgewählte Arbeiten präsentiert, die sich speziell mit geographischen Daten befassen.

Kapitel 3 vermittelt sowohl mathematische als auch algorithmische Grundlagen. Es werden verschiedene Ähnlichkeitsmaße vorgestellt und unterschiedliche Relationsbeziehungen definiert, die auf Objektebene und Schemaebene auftreten. Zusätzlich werden Grundlagen der Graphentheorie präsentiert, da die Zuordnung auf Schemaebene als Graphzuordnungsproblem betrachtet wird, deren Lösung mit existierenden Graphalgorithmen möglich ist. Für die Bestimmung einer optimalen Lösung ist die Formulierung als spezielles Optimierungsproblem notwendig. Dazu werden abschließend Grundlagen der linearen Optimierung erläutert.

In Kapitel 4 werden zwei Ansätze für die Zuordnung verschiedener Objektgeometrien vorgestellt, die hauptsächlich geometrische und topologische Objekteigenschaften nutzen. Ausführlich befasst sich die Arbeit mit der Zuordnung von Polygonobjekten. Die automatisierte Zuordnungsentscheidung wird auf Basis von einzelnen Ähnlichkeitsmaßen getroffen, die zunächst einzeln definiert und dann in einem Gesamtähnlichkeitsmaß zusammengefasst werden. Als Ergebnis werden einfache und komplexe Objektrelationen identifiziert, die anschließend in Zuordnungsmatrizen kombiniert werden und als Eingabedaten für die Schema-Matching-Verfahren dienen.

Kapitel 5 beschreibt das Zuordnungsproblem auf Schemaebene zunächst formal, bevor verschiedene Lösungsansätze mit Hilfe eines synthetischen Beispiels vorgestellt werden. Dazu werden als erstes zwei einfache Lösungsverfahren präsentiert, die allerdings Kompromisse in Bezug auf die Ergebnisse erfordern. Darauf aufbauend wird ein Heuristisches Schema-Matching-Verfahren entwickelt, das eines der einfachen Verfahren rekursiv anwendet. Anschließend werden Optimierungsverfahren beschrieben, die in Abhängigkeit der formulierten Bedingungen optimale Lösungen des Zuordnungsproblems erzielen. Um die Korrektheit und Leistungsfähigkeit der entwickelten Verfahren zu bewerten, werden manuell erstellte Referenzlösungen und Optimierungslösungen verwendet.

In Kapitel 6 werden zuerst die verwendeten Datensätze mit ihren Besonderheiten in Hinblick auf die geometrische und semantische Modellierung erläutert, bevor drei ausgewählte Testgebiete vorgestellt werden. Im Anschluss daran werden die erzielten Ergebnisse des Objektzuordnungsverfahrens für Polygonobjekte für jedes Testgebiet präsentiert und mit manuell erstellten Referenzdaten evaluiert. Aus den Objektzuordnungen werden für jedes Testgebiet vier verschiedene Zuordnungsmatrizen abgeleitet, die als Eingabedaten für die in Kapitel 5 vorgestellten Schema-Matching-Verfahren dienen. Entsprechend werden die Ergebnisse für jedes Testgebiet und für jedes Schema-Matching-Verfahren vorgestellt und mit Referenzlösungen verglichen, die aus linguistischen Analysen der Objektklassennamen abgeleitet wurden. Abschließend wird eine Empfehlung gegeben, welches Schema-Matching-Verfahren sich hinsichtlich der Korrektheit und Leistungsfähigkeit am besten in der Praxis für die automatisierte Objektklassenzuordnung eignet. Kapitel 7 gibt eine Zusammenfassung und einen Ausblick.

2 Verwandte Arbeiten

In diesem Kapitel werden die Themen Data-Matching und Schema-Matching vorgestellt, die eine Zuordnung auf Objekt- bzw. Schemaebene ermöglichen. Dabei wird besonders auf bestehende Probleme und Herausforderungen eingegangen. Für das Verständnis der Arbeit werden zuerst zentrale Grundbegriffe in Abschnitt 2.1 eingeführt und anhand von Beispielen näher erläutert. Abschnitt 2.2 behandelt das Thema Data-Matching speziell für raumbezogene Vektordaten. In zahlreichen Forschungsarbeiten wurden verschiedene Verfahren entwickelt, die unterschiedliche Lösungsansätze für eine Objektzuordnung bieten. Im Anschluss wird in Abschnitt 2.3 das Thema Schema-Matching vorgestellt, das in Bezug auf Geodaten, trotz seiner zunehmenden Bedeutung, weniger untersucht ist und in Zukunft weiteren Forschungsbedarf hat.

2.1 Grundbegriffe

2.1.1 Raumbezogene Objekte

In der vorliegenden Arbeit stellen zweidimensionale topographische Objekte in vektorisierter Form die verwendeten, raumbezogenen Daten dar. Dazu zählen künstliche und natürliche Objekte, die für die Beschreibung einer Landschaft notwendig sind und in topographischen Karten dargestellt werden, beispielsweise Straßen, Wege, Gebäude, Flüsse, Seen oder Flächen bestimmter Nutzungen. Im folgenden Text werden diese topographischen Objekte kurz mit dem Begriff *Objekt* bezeichnet.

Für das im Rahmen der Arbeit entwickelte Zuordnungsverfahren sind pro Objekt lediglich drei Attribute notwendig: *Koordinaten*, die die Position und die Geometriedimension widerspiegeln und aus denen sich die Form ableiten lässt, ein *Identifikator*, um jedes Objekt eindeutig zu identifizieren, sowie die Zuordnung zu einer bestimmten *Objektklasse*, die durch einen numerischen Code und einen Namen festgelegt ist.

2.1.2 Ähnlichkeit

Bruns und Egenhofer (1996) beschreiben den Begriff der *Ähnlichkeit* sehr treffend mit den folgenden Worten: *„Similarity is the assessment of deviation from equivalence“*, was besagt, dass die Ähnlichkeit als Abweichung von der Gleichheit zu beurteilen ist. Das bedeutet, die Ähnlichkeit ist umso größer, je kleiner die Abweichung ist. Doch wie genau lässt sich die Abweichung bestimmen und wie groß darf diese sein, um Elemente als einander ähnlich einzustufen? Dies ist von vielen verschiedenen Faktoren abhängig, z.B. dem Fachgebiet, dem Anwendungszweck oder der Objektausprägung.

Die folgenden Beispiele zeigen, wie unterschiedlich Entscheidungen bezüglich der Ähnlichkeit getroffen werden können. Im ersten Beispiel werden jeweils zwei Begriffe miteinander verglichen. Der Vergleich der Begriffspaare 1) Bank und Sparkasse gegenüber 2) Bank und Bank lässt zwei Schlussfolgerungen zu. Aus linguistischer Sicht sind sich die Begriffe aus Beispiel 2 ähnlicher, da sie aus den gleichen Buchstaben bestehen und formal identisch sind. Aus semantischer Sicht sind sich die Begriffe aus Beispiel 1 ähnlicher, da beide Begriffe Geldinstitute beschreiben, während in Beispiel 2 der Begriff Bank sowohl für ein Sitzmöbel als auch für ein Geldinstitut verwendet wird.

Im zweiten Beispiel sollen die in Abbildung 2.1 dargestellten Objekte auf ihre Ähnlichkeit zueinander untersucht werden. Auf den ersten Blick wird ein und dasselbe Gebäude zweimal abgebildet. Doch sind sich diese Abbildungen wirklich ähnlich, obwohl Unterschiede bezüglich der Größe, der Orientierung und der Farbgebung vorhanden sind? Oder sind diese Veränderungen zu vernachlässigen, da die Gemeinsamkeiten überwiegen, wie z.B. die Dachform, die Dachneigung sowie die Anzahl und die relative Position der Fenster. Tversky (1977) schlägt ein Kontrastmodell vor, wobei die Ähnlichkeit zweier Objekte als Linearkombination aus Maßen der gleichen und unterschiedlichen Eigenschaften beschrieben wird.

Beide Beispiele verdeutlichen, dass für einen Vergleich das zur Anwendung kommende Ähnlichkeitsmaß von entscheidender Bedeutung ist. Es muss die Gemeinsamkeiten und Unterschiede zwischen den Objekten entsprechend der Fragestellung richtig herausstellen, um zuverlässige Entscheidungen ableiten zu können.

Im Kontext der Arbeit sind verschiedene Ähnlichkeitsmaße notwendig, um sowohl die Zuordnung der raumbezogenen Objekte als auch die Zuordnung der Objektklassen zu erreichen. Bei der Objektzuordnung spielen

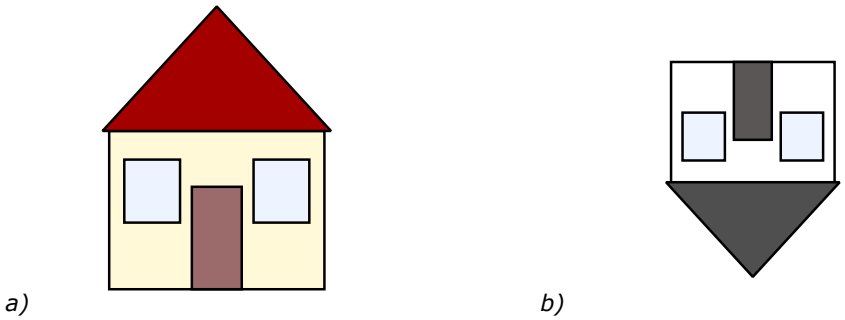


Abbildung 2.1: Zwei Darstellungen eines Gebäudeobjekts.

die räumliche Position, die geometrische Objektform und Beziehungen zu Nachbarobjekten eine entscheidende Rolle. Die Zuordnung der Objektklassen wird auf Basis der Objektzuordnungen durchgeführt und soll daher eine semantische Ähnlichkeit der Objektklassen widerspiegeln. In Abschnitt 3.1 wird daher eine Unterteilung in *geometrische*, *topologische* und *semantische Ähnlichkeit* vorgenommen.

2.1.3 Relation

Eine *Relation r* ist allgemein formuliert eine Beziehung zwischen mehreren Elementen. In der vorliegenden Arbeit sind sowohl raumbezogene Objekte als auch Objektklassen Vertreter dieser Elemente. Wenn eine Relation genau zwischen zwei Elementen besteht, kennzeichnet dies eine einfache Eins-zu-Eins-Äquivalenz-Relation, die im Folgenden als 1:1 -Relation bezeichnet oder mit 1:1 abgekürzt wird. Sind mehr als zwei Elemente an einer Relation beteiligt, wird von einer komplexen Aggregationsrelation gesprochen. Hier wird zwischen einseitig (1:n/n:1) und beidseitig zusammengefassten Relationen (n:m) unterschieden. Eine 1:n -Relation besagt, dass ein Element mehreren anderen Elementen gegenübersteht, während eine n:m -Relation die Beziehung einer Gruppe von Elementen zu einer anderen Gruppe beschreibt.

Neben der Anzahl der beteiligten Elemente, den sogenannten Kardinalitäten, kennzeichnet auch der Kontext, in dem die Elemente vorkommen, die Art der Relation. In Abbildung 2.2 werden am Beispiel von raumbezogenen Objekten zwei unterschiedliche Relationen vorgestellt. Eine Ähnlichkeitsbeziehung kann zwischen Objekten unterschiedlicher Datensätze bestehen, z.B. zwischen den Objekten Park und City Park (1:1) oder zwischen den Objekten Platz mit Grünfläche und Square (2:1). Aufgrund der räumlichen Lage und vergleichbarer Flächeninhalte können sie als ähnlich betrachtet werden. Im Gegensatz dazu besteht zwischen den Objekten City Park und Square bzw. Park und Platz, die jeweils aus einem Datensatz stammen und deren Objekträndер sich berühren, eine Nachbarschaftsbeziehung. In Abschnitt 3.2 werden alle Relationstypen, die in der vorliegenden Arbeit untersucht werden, ausführlich vorgestellt.

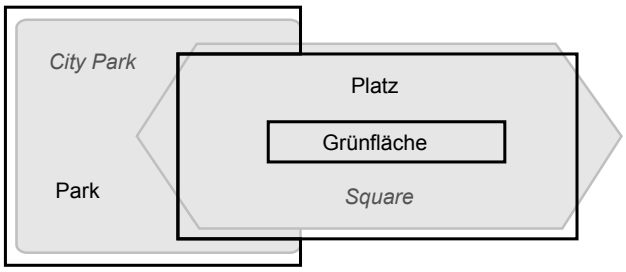


Abbildung 2.2: Verschiedene Relationsarten: Eine Ähnlichkeitsbeziehung besteht zwischen den Objekten Platz mit Grünfläche aus Datensatz A (schwarz) und dem Square-Objekt aus Datensatz B (grau), während eine Nachbarschaftsbeziehung zwischen den Objekten innerhalb eines Datensatzes vorliegt, z.B. City Park und Square bzw. Park und Platz.

2.1.4 Schema

Ein *Schema* ist eine formale Struktur, die eine vereinfachte und abstrahierte Sicht auf zu repräsentierende Daten beschreibt. Die Modellierung und Strukturierung der Daten kann auf unterschiedliche Weise erfolgen: entweder

einfach und flach als Entity-Relationship-Modell, als Hierarchie in einem Baum oder einer XML[1]-Struktur, als objektorientiertes Modell oder als gerichteter Graph, um nur die wichtigsten zu nennen. Gleichermaßen entwickelten sich auch in den verschiedenen Fachgebieten unterschiedliche Bezeichnungen für diese formale Struktur. Am häufigsten sind in der Literatur die Begriffe Schema, Konzept und Ontologie zu finden.

Im Rahmen dieser Arbeit wird der Begriff Schema verwendet und als eine Menge von Elementen definiert, die durch eine Struktur miteinander verbunden sind. Die Schemaelemente entsprechen den Objektklassen der Datensätze, die eine Menge von Objekten mit ähnlichen Eigenschaften zusammenfassen.

Abbildung 2.3 zeigt ein hierarchisches Schema, bei dem nur den untersten Objektklassen Objekte zugeordnet sind. Abstrakte Klassen besitzen keine eigenen Instanzen und sind hier mit rot gekennzeichnet. Objekteigenschaften werden durch verschiedene Attribute beschrieben und sind fester Bestandteil der Objektklassendefinition. Neben geometrischen Attributen, wie z.B. Koordinaten der einzelnen Objektpunkte, sind auch semantische Attribute, wie der geographische Name oder eine eventuelle Fließgeschwindigkeit definiert. Alle Pfeile in diesem Schema repräsentieren semantische is-a-Relationen. Diese Beziehungen entsprechen Generalisierungsrelationen zwischen genau zwei Objektklassen, wobei der Pfeil von der spezifischen auf die allgemeinere Objektklasse zeigt. Für das Beispiel in Abbildung 2.3 stellt die Objektklasse Fluss eine Spezialisierung der abstrakten Klassen Fließgewässer und Gewässer dar, die die Attribute der Oberklassen erbt.

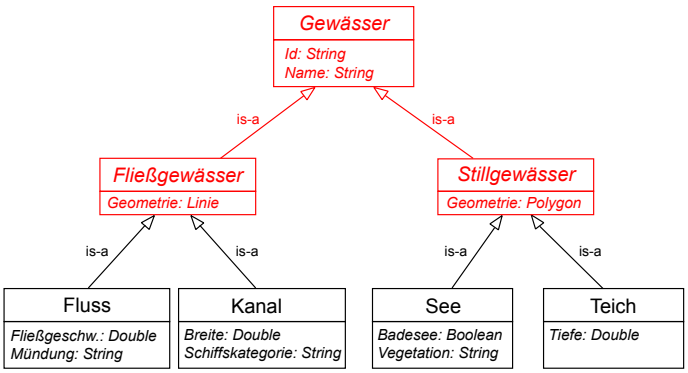


Abbildung 2.3: Einfaches, hierarchisches Schema zur Beschreibung von Gewässerdaten. Die Objektklassen Fluss, Kanal, See und Teich besitzen im Gegensatz zu den rot gekennzeichneten, abstrakten Objektklassen eigene Instanzen mit identischen Attributen, wie z.B: Id und Name, aber auch unterschiedliche Attribute, um sich von anderen Klassen abzugrenzen.

2.2 Data-Matching

Im Bereich der Geoinformatik wird der Begriff *Data-Matching* für die Identifizierung von georeferenzierten Objekten verwendet, die das gleiche Real-Welt-Objekt repräsentieren, aber in verschiedenen Datenbanken abgelegt sind. Es existieren viele digitale, geographische Datenbanken, in denen große Teile der Erde mehrfach erfasst sind. Die Verfügbarkeit solcher GIS -Datenbanken wird weiter ansteigen.

Durch die rasante Entwicklung des Internets ist eine neue Art der Erzeugung von raumbezogenen Daten entstanden, bei denen Freiwillige zeitnah und kostengünstig geographische Informationen erheben und der Öffentlichkeit zur Verfügung stellen, z.B. im Rahmen des OpenStreetMap-Projekts (Yang u. a., 2013). Um alle erfassten raumbezogenen Daten nutzen zu können, sind Zuordnungsverfahren notwendig, um beispielsweise Objektveränderungen zu erkennen, Daten zu aktualisieren, Objekt- und Attributinformationen für komplexe Analysen auszutauschen bzw. zu teilen, Datensätze miteinander zu verschmelzen oder Objekte in einen anderen Datensatz zu integrieren (Fan u. a., 2014). Das Ergebnis von Zuordnungsverfahren sind Verbindungen, sogenannte Links, zwischen korrespondierenden Objekten. Das explizite Abspeichern dieser Links in Datenbanken ermöglicht einen einfachen Zugriff. Besonders für Aktualisierungsprozesse ist dies vorteilhaft, da so jedes Update eines Objekts nachvollziehbar ist und bleibt (Harrie und Hellström, 1999; Dunkars, 2004).

2.2.1 Klassifikation von Zuordnungsverfahren auf Objektebene

Bei der Zuordnung von Objekten wird zwischen merkmalsbasierten und relationalen Verfahren unterschieden. Merkmalsbasierte Verfahren (engl. Feature-Based Matching) ordnen Objekte aufgrund von geometrischen und

[1] erweiterbare Auszeichnungssprache (engl. Extensible Markup Language)

thematischen Objektinformationen einander zu. Die zugrunde liegende Idee ist, dass bei räumlicher Koinzidenz zweier Objekte mit ähnlichen geometrischen Eigenschaften, wie z.B. Größe, Form und Ausrichtung oder identischen Bezeichnungen bzw. Attributwerten, diese einander zugeordnet werden können. Allerdings besitzen korrespondierende Objekte gewöhnlich keine eindeutigen Namen oder Attributwerte, mit Ausnahme einzigartiger Namen von Städten oder Straßen. Demzufolge wird der Grad der Ähnlichkeit hauptsächlich anhand der räumlichen Position und der geometrischen Objekteigenschaften bestimmt. Diese Verfahren sind besonders für klar abgrenzbare Objekte, wie z.B. Gebäude oder Seen, geeignet.

Im Gegensatz dazu berücksichtigen relationale Verfahren (engl. Relational Matching) neben den bereits genannten Objektinformationen auch noch Eigenschaften und Beziehungen zu Nachbarobjekten. Das Ausnutzen von topologischen Eigenschaften ist besonders für Netzwerke, wie Straßen, Flüsse oder Eisenbahnstrecken, geeignet, da Kontextinformationen hinzufügt werden. Eine Zuordnung auf Basis eines großen Ähnlichkeitswertes ist mit einem rein geometrischen Vergleich der einzelnen Netzwerkobjekte, unabhängig von ihren Nachbarobjekten, nicht immer möglich. Bei merkmalsbasierten Verfahren verstärken sich die Zuordnungsprobleme, je größer der Maßstabsunterschied zwischen den Datensätzen ist. Im Vergleich zu großmaßstäbigen Objekten besitzen Objekte in kleinmaßstäbigen Datensätzen meist eine geometrisch vereinfachte Form. Topologische Beziehungen bleiben während einer Generalisierung überwiegend erhalten und können somit in relationalen Zuordnungsverfahren genutzt werden.

Im Rahmen der vorliegenden Arbeit werden Zuordnungsverfahren entwickelt, die neben geometrischen Objektinformationen zusätzlich topologische Beziehungen zu Nachbarobjekten berücksichtigen und dementsprechend den relationalen Verfahren zuzuordnen sind. Bevor in den Abschnitten 2.2.3 und 2.2.4 ausgewählte Arbeiten für merkmalsbasierte und relationale Verfahren vorgestellt werden, werden zunächst die größten Herausforderungen bei einer Objektzuordnung benannt.

2.2.2 Herausforderungen bei der Objektzuordnung

Die Identifizierung von korrespondierenden räumlichen Objekten wurde in der Vergangenheit oftmals manuell oder mit halb-automatisierten Verfahren durchgeführt. Dieses Vorgehen ist besonders für sehr große Datensätze kosten- und zeitintensiv. Daher lag der Fokus der letzten Jahre auf der Automatisierung dieses Prozesses. Für spezielle Zuordnungsprobleme wurden automatische Verfahren entwickelt, die allerdings nur eingeschränkt auf andere Datensätze bzw. Probleme anwendbar sind. Gerade die Einschränkungen in den Algorithmen, aber auch die Mehrdeutigkeit der Daten, stellen Probleme bei der Automatisierung dar. Einschränkungen der Algorithmen können teilweise durch Eingriffe bei der Bearbeitung kompensiert werden, z.B. indem falsche Zuordnungen gelöscht oder Objekte ohne Zuordnung manuell korrigiert werden.

In den meisten Arbeiten zum Data-Matching wird die Zuordnung der Schemaelemente, d.h. der Objektklassen, vor der eigentlichen Objektzuordnung vorausgesetzt. Das bedeutet, dass Objekte von semantisch ähnlichen Objektklassen einander zugeordnet werden, z.B. nur Straßen- oder Gebäudeobjekte (Walter, 1997; van Wijngaarden u. a., 1997; Uitermark u. a., 1999; Walter und Fritsch, 1999; Chen u. a., 2006; Volz, 2006; Kieler u. a., 2007; Lüscher u. a., 2007; Diez u. a., 2008; Mustière und Devogele, 2008; Kieler u. a., 2009b; Zhang, 2009; Koukoletsos u. a., 2012; Luan, 2012; Yang u. a., 2013; Fan u. a., 2014; Zhang u. a., 2014; Abdolmajidi u. a., 2015; Fan u. a., 2016). Die Vielzahl der Arbeiten zeigt, dass das Wissen, welche Objektklassen semantisch ähnlich sind, wichtig ist, um den Suchraum für Matching-Kandidaten im Zuordnungsprozess zu reduzieren. Da genau diese Korrespondenzen auf Schemaebene vorab nicht immer zur Verfügung stehen bzw. nur schwierig abzuleiten sind (siehe Abschnitt 2.3.1), können dafür ebenfalls Data-Matching-Verfahren eingesetzt werden. In einer eigenen früheren Arbeit wurde am Beispiel zweier Polygondatensätze gezeigt, dass eine geometrische Überlagerung der Datensätze, gefolgt von einer Häufigkeitsanalyse der Objektrelationen, eine Zuordnung der Objektklassen durchaus möglich macht (Kieler u. a., 2007). Die dort vorgestellten Überlegungen werden im Rahmen dieser Arbeit erweitert.

Ein weiteres Problem besteht, wenn die zu untersuchenden Datensätze unterschiedliche Kartenprojektionen besitzen und die Transformation zwischen den Koordinatensystemen unbekannt ist. Das fehlende gemeinsame Referenzsystem erschwert die Zuordnung, da implizite Verbindungen, d.h. die gleiche räumliche Position, zwischen den Objekten fehlen. Chen u. a. (2006), Diez u. a. (2008) und Luan (2012) untersuchten dieses Problem bei der Zuordnung von zwei Straßennetzwerken. Während Chen u. a. (2006) eine mögliche Skalierung und Verschiebung beider Datensätze berücksichtigen, beziehen Diez u. a. (2008) zusätzlich eine mögliche Rotation der Datensätze zueinander in ihr Verfahren ein. Luan (2012) präsentiert ein Verfahren, das nur die wichtigsten Kreuzungen zuordnet, um daraus Transformationsparameter für das gesamte Straßennetzwerk abzuleiten. Dafür wird das Problem als Graphzuordnungsproblem formuliert und aus den Straßennetzwerken werden Graphen

erzeugt und miteinander verglichen. Die Korrespondenzen werden mit einem Algorithmus zur Berechnung eines größten gemeinsamen Teilgraphen identifiziert.

Letztendlich hängt der Erfolg der Objektzuordnung sehr stark von den beteiligten Daten ab. Je unterschiedlicher Datenbestände und die ihnen zugrunde liegenden Schemas sind, desto größer sind die Schwierigkeiten. Dies ist besonders bei Daten zu beobachten, die zu unterschiedlichen Zeitpunkten erfasst wurden, aus verschiedenen Fachgebieten stammen oder unterschiedliche Maßstäbe aufweisen.

Anhand eines Beispiels lässt sich die Problematik verdeutlichen: Gewässerobjekte sind für viele unterschiedliche Fachgruppen (Kartographen, Landesumweltämter oder Bundesschifffahrtsamt) von besonderem Interesse. Daher erfolgt die Erfassung der Objekte mehrfach, entsprechend der Anforderungskriterien der erfassenden Institutionen und mit den ihr zur Verfügung stehenden Erfassungsmethoden. Daraus entstehen nicht nur Ungenauigkeiten bezüglich der Position, sondern auch Unterschiede hinsichtlich der Geometrie (Objektform und -größe) und der Attributinformationen (Abb. 2.4 a). Für die Erstellung einer topographischen Karte sind z.B. die Position, Ausdehnung und der Name eines Gewässers von Bedeutung, nicht aber die chemische Gewässerbeschaffenheit oder die Länge des durch Binnenschiffe befahrbaren Flusslaufs. Dementsprechend entwickelt jedes Fachgebiet ein individuelles Schema mit unterschiedlichen Strukturelementen, was das Auffinden von identischen Objekten, z.B. anhand eines Namens oder eines Attributs, im Schema erschwert.

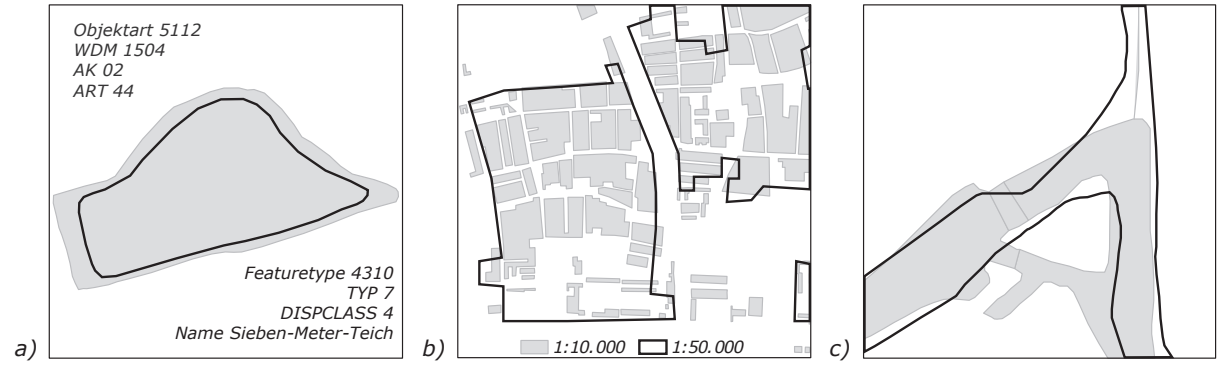


Abbildung 2.4: Probleme bei der Objektzuordnung: a) geometrische Unterschiede hinsichtlich Position, Objektform und -größe, aber auch bzgl. der Attributinformationen, b) unterschiedliche Maßstäbe können zu komplexen Objektrelationen (1:n/n:1 bzw. n:m) führen und c) unterschiedliche Geometriedimensionen (Polygon- und Linienobjekte)

Werden für die Zuordnung statt der thematischen Objektinformation geometrische Repräsentationsformen verwendet, ist der Erfassungsmaßstab bedeutend, da bei der Objekterfassung für verschiedene Maßstäbe unterschiedliche Mindestobjektgrößen oder Geometriedimensionen gelten. Ein Gebäudeobjekt kann für einen großen Maßstab detailliert als Polygonobjekt erfasst werden, während eine zunehmende Verkleinerung des Maßstabes eine Zusammenfassung mit Nachbarobjekten oder gar eine Nichterfassung bewirkt (Abb. 2.4 b). Bei einem Flussobjekt kann die Darstellung von einem Polygon- zu einem Linienobjekt wechseln, wenn die zulässige Mindestbreite für einen bestimmten Maßstab unterschritten wird (Abb. 2.4 c). Solch ein Geometrietypwechsel innerhalb eines Objektes, aber auch die Gegenüberstellung unterschiedlicher Geometrietypen oder zusammengefasster Objekte zu einem einzelnen Objekt, sind große Herausforderungen, die im Zuordnungsprozess gelöst werden müssen.

Aus diesem Grund werden verschiedene Data-Matching-Verfahren entwickelt, die speziell auf die Eingangsdaten angepasst sind. Einige Zuordnungsverfahren arbeiten nur mit Datensätzen gleichen bzw. ähnlichen Maßstabs, wohingegen die Thematiken voneinander abweichen können. Walter und Fritsch (1999), Zhang und Meng (2007) und Kieler u. a. (2007) präsentieren Beispiele für diese Arbeiten, in denen Objekte eines topographischen Datensatzes den Objekten eines Datensatzes für die Fahrzeugnavigation zugeordnet werden. Während in den beiden ersten Arbeiten der Fokus auf Straßenobjekten lag, d.h. semantisch ähnlichen Objektklassen, verwendeten Kieler u. a. (2007) Objekte verschiedener Objektklassen, mit dem Ziel die verborgenen semantischen Ähnlichkeiten zwischen den Schemas aufzudecken.

Dagegen behandeln andere Verfahren speziell Datensätze mit unterschiedlichen Maßstäben. In den Untersuchungen von Lüscher u. a. (2007) und Mustière und Devogele (2008) werden ausschließlich linienförmige Straßenobjekte verwendet, während bei Uitermark u. a. (1999) das Straßennetzwerk durch Polygonobjekte repräsentiert war. Allerdings wurde das Polygonnetzwerk in einem Vorverarbeitungsschritt mit Hilfe eines Skelettierungsalgorithmus in ein Liniennetzwerk umgewandelt. Kieler u. a. (2009b) leiteten nach gleichem Vorbild ein geschlossenes linienförmiges Gewässernetz ab, das im Gegensatz zu Uitermark u. a. (1999) sowohl Polygon- als auch Linie-

nobjekte beinhaltete. Die Schwierigkeit besteht darin, die Topologie zwischen den Objekten der verschiedenen Geometriedimensionen zu bewahren. Das Objektzuordnungsverfahren wird in Abschnitt 4.2 vorgestellt.

Datensätze mit unterschiedlichen Maßstäben können durch die Anwendung von Generalisierungsregeln auf einen Datensatz vergleichbar gemacht werden, ohne die Objektdimension zu verändern. In einer eigenen früheren Arbeit werden flächenhafte Objekte einander zugeordnet, um semantische Informationen, d.h. Namen von aktuellen Siedlungsgebieten (1:50.000) auf geometrisch genauere Gebäude (1:10.000) zu übertragen (Kieler u. a., 2007) (Abb. 2.4 b). Maßstabsunterschiede werden mit Hilfe einer distanzbasierten Aggregation der Gebäudeobjekte ausgeglichen. Zhang u. a. (2014) bestimmen Korrespondenzen zwischen Gebäuden, die aus zwei voneinander abgeleiteten Datensätzen der Maßstäbe 1:10.000 und 1:50.000 stammen. In dem iterativen Relaxation-Labeling-Verfahren werden für die Objektzuordnung kontextbezogene Informationen, d.h. die relative Position, Orientierung, Größe und Form benachbarter Objekte herangezogen.

Das Data-Matching ist für Objekte mit gleichen Geometriedimensionen am besten erforscht. In der Vergangenheit standen dabei vor allem linienförmige Objekte und speziell Straßenobjekte im Fokus der Forschung. Dies zeigt auch die Entwicklung des Standard-Matching-Tools *RoadMatcher* (Vivid Solutions, 2005), welches die Zuordnung und Verschmelzung von Liniennetzwerken erlaubt. RoadMatcher wird im Rahmen dieser Arbeit nicht eingesetzt, da einerseits nur Linien- und keine Polygonobjekte verwendet werden können und andererseits Zuordnungen auf 1:1 -Relationen beschränkt sind. Viele Forschungsarbeiten zeigen, dass Zuordnungsverfahren, die Datensätze verschiedener Maßstäbe berücksichtigen, in der Lage sein müssen, auch komplexe Objektrelationen (1:n/n:1 bzw. n:m) zu identifizieren. Fan u. a. (2014) identifizieren mit einem Data-Matching-Verfahren verschiedene Objektrelationen und nutzen diese für die Qualitätsbewertung von OpenStreetMap -Daten. Die Zuordnung von Objekten unterschiedlicher Geometrietypen mit komplexen Objektrelationen (1:n/n:1 bzw. n:m) stellt die Forschung weiterhin vor Probleme.

Besonders die Verwendung von komplexen geometrischen Analyseoperationen, bei der Zuordnung von raumbezogenen Objekten, kann im Vergleich zur Zuordnung reiner Buchstaben-Zahlen-Kombinationen, z.B. bei ISBN[2]-Nummern in verschiedenen Produktkatalogen, viel Rechenzeit beanspruchen. Ein Brute-Force-Ansatz, der alle Objekte eines Datensatzes mit allen Objekten eines anderen Datensatzes geometrisch, möglicherweise sequentiell, vergleicht, ist in der Praxis nicht nutzbar. Um lange Rechenzeiten zu verringern, muss der Suchbereich verkleinert und der Suchprozess optimiert werden. Li und Goodchild (2010) vergleichen zu diesem Zweck zwei Suchstrategien, mit denen sowohl eine effektive als auch effiziente Suche nach Matching-Kandidaten möglich ist. Neben einem Greedy-Algorithmus, der nacheinander alle Objekte paarweise zuordnet und der Lösungsmenge hinzufügt, aber niemals eine Fehlzuordnung aus dieser Menge entfernen kann, berücksichtigt ein Optimierungsansatz alle möglichen Matching-Kandidaten gleichzeitig, mit dem Ziel, die Gesamtähnlichkeit zwischen allen Objekten zu maximieren. Die Untersuchungen mit hypothetischen Punktdaten sowie mit realen linienhaften Straßendaten haben den deutlichen Qualitätsvorteil des Optimierungsansatzes gezeigt. Das Zuordnungsproblem wurde dabei unter der Bedingung formuliert, dass nur 1:1 -Korrespondenzen möglich sind. Li und Goodchild (2011) lockerten auch diese Einschränkung.

2.2.3 Ausgewählte, merkmalsbasierte Verfahren

In diesem Abschnitt werden beispielhaft zwei merkmalsbasierte Zuordnungsverfahren vorgestellt. Während in der Arbeit von Van Wijngaarden u. a. (1997) das Hauptaugenmerk auf der Zuordnung von flächenhaften Gebäudeobjekten auf Basis geometrischer Eigenschaften liegt, werden in der Arbeit von Li und Goodchild (2011) linienförmige Straßenobjekte unter Verwendung geometrischer und semantischer Informationen, wie z.B. Objektnamen, zugeordnet. Der wichtigste Unterschied ist, dass Li und Goodchild die Zuordnung als Optimierungsproblem formulieren und somit für das gesamte Untersuchungsgebiet die bestmögliche Zuordnung erhalten. Dagegen bestimmen Van Wijngaarden u.a. für jedes Objekt lokal den besten Matching-Kandidaten.

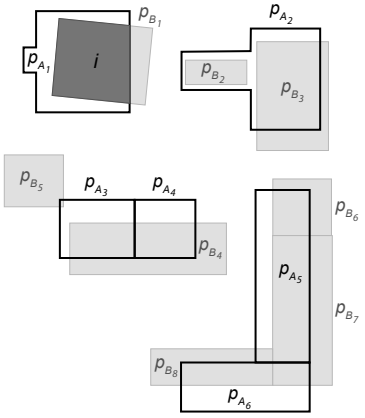
Lokale Zuordnung von Gebäudeobjekten unterschiedlichen Maßstabs

Van Wijngaarden u. a. (1997) entwickelten für die Übertragung von Updates zwischen ähnlichen Objekten aus unterschiedlichen Datenbanken einen *Map Integrator*, um die manuelle Bearbeitung im Update-Prozess zu reduzieren. Der Prozess umfasst ingesamt sechs Stufen. An dieser Stelle wird nur die erste Stufe, die Bestimmung von korrespondierenden Objekten, erläutert.

[2]Internationale Standard-Buch-Nummer

Für die Objektzuordnung werden flächenhafte Gebäudeobjekte zweier Datensätze unterschiedlichen Maßstabs (GBKN[3] 1:1.000 und TOP10vector[4] 1:10.000) überlagert. Es gilt die Annahme, dass zwei Polygone eine gewisse Überlappung haben müssen, wenn sie das selbe Real-Welt-Objekt repräsentieren. Ist die Überlagerung geringer als ein definierter Schwellwert, werden die Gebäude nicht als korrespondierend betrachtet.

Abbildung 2.5 erläutert die Zuordnung im Detail. Als erstes werden alle Objekte der Datensätze A und B überlagert und *Überlagerungsverhältnisse* s_{i_j} zwischen der *Schnittfläche* i (engl. Intersection Area) und den Objektflächen p_j mit $j = A, B$ bestimmt. Anhand der Ergebnisse wird die Relationsliste 1 (L1) erstellt. Für mindestens ein Überlagerungsverhältnis muss der Wert größer sein als ein definierter *Schwellwert* t (engl. Threshold). Die sechste Relation $p_{A_3} \rightarrow p_{B_5}$ fällt heraus, da beide Verhältnisse mit 4 % kleiner sind als der definierte Schwellwert von 5 %. Anschließend wird die Relationsliste L2 für die komplexen n:m -Relationen erstellt, indem alle Relationen von L1 selektiert werden, bei denen beide Matching-Kandidaten auch an anderen Relationen beteiligt sind. Die entsprechenden Objekte sind fett hervorgehoben. Zu dieser n:m -Liste werden alle Relationen aus L1 hinzugefügt, bei denen sich bereits ein Objekt in L2 befindet. Als nächstes wird L3 mit den 1:n - bzw. n:1 -Relationen erzeugt, indem entweder ein Objekt von Datensatz A oder Datensatz B mehrmals vorkommt. Die übrigen Relationen werden in L4 mit den 1:1 -Relationen zusammengefasst.



Nr.		s_{i_A} [%]	s_{i_B} [%]	L1 ($s_{i_A} \vee s_{i_B} > t$)	L2 (n:m)	L3 (1:n/n:1)	L4 (1:1)
1	$p_{A_1} \rightarrow p_{B_1}$	70	80	$p_{A_1} \rightarrow p_{B_1}$			$p_{A_1} \rightarrow p_{B_1}$
2	$p_{A_2} \rightarrow p_{B_2}$	40	100	$p_{A_2} \rightarrow p_{B_2}$		$\boldsymbol{p_{A_2}} \rightarrow p_{B_2}$	
3	$p_{A_2} \rightarrow p_{B_3}$	50	80	$p_{A_2} \rightarrow p_{B_3}$		$\boldsymbol{p_{A_2}} \rightarrow p_{B_3}$	
4	$p_{A_3} \rightarrow p_{B_4}$	60	30	$p_{A_3} \rightarrow p_{B_4}$		$p_{A_3} \rightarrow \boldsymbol{p_{B_4}}$	
5	$p_{A_4} \rightarrow p_{B_4}$	60	30	$p_{A_4} \rightarrow p_{B_4}$		$p_{A_4} \rightarrow \boldsymbol{p_{B_4}}$	
6	$p_{A_3} \rightarrow p_{B_5}$	4	4				
7	$p_{A_5} \rightarrow p_{B_6}$	30	50	$p_{A_5} \rightarrow p_{B_6}$	$p_{A_5} \rightarrow p_{B_6}$		
8	$p_{A_5} \rightarrow p_{B_7}$	50	70	$p_{A_5} \rightarrow p_{B_7}$	$\boldsymbol{p_{A_5}} \rightarrow \boldsymbol{p_{B_7}}$		
9	$p_{A_6} \rightarrow p_{B_7}$	20	20	$p_{A_6} \rightarrow p_{B_7}$	$\boldsymbol{p_{A_6}} \rightarrow \boldsymbol{p_{B_7}}$		
10	$p_{A_6} \rightarrow p_{B_8}$	40	40	$p_{A_6} \rightarrow p_{B_8}$	$p_{A_6} \rightarrow p_{B_8}$		

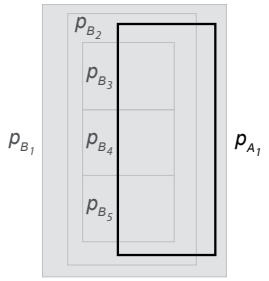
Abbildung 2.5: Objektzuordnung nach Van Wijngaarden u. a. (1997). Durch die geometrische Überlagerung der schwarzen und grauen Objekte werden Schnittflächen bestimmt und Überlagerungsverhältnisse s_{i_j} mit $j = A, B$ abgeleitet. Basierend auf den Verhältnissen werden einfache (L4) und komplexe Objektrelationen (L2 und L3) bestimmt, die letztendlich die Objektzuordnungen $p_{A_1} \rightarrow p_{B_1}$, $p_{A_2} \rightarrow \{p_{B_2}, p_{B_3}\}$, $\{p_{A_3}, p_{A_4}\} \rightarrow p_{B_4}$ und $\{p_{A_5}, p_{A_6}\} \rightarrow \{p_{B_6}, p_{B_7}, p_{B_8}\}$ widerspiegeln.

Das Prinzip der Objektzuordnung ist nicht anwendbar, wenn sich mehrere Objekte innerhalb eines Datensatzes überlagern. In Abbildung 2.6 ist solch eine Situation dargestellt: Gebäude (p_{B_3}, p_{B_4} und p_{B_5}) überlagern eine Wohnbaufläche (p_{B_2}), die wiederum ein Verwaltungsbezirk (p_{B_1}) überlagert. Sogenannte Container-Objekte, die mehrere Objekte des anderen Datensatzes umfassen, vergrößern die Anzahl der potentiellen Matching-Kandidaten. Die Frage ist, ob in diesem Beispiel Objekt p_{A_1} ein Gebäude darstellt und eine komplexe Relation $p_{A_1} \rightarrow \{p_{B_3}, p_{B_4}, p_{B_5}\}$ beschreibt oder eine Wohnbaufläche repräsentiert und $p_{A_1} \rightarrow p_{B_2}$ wahrscheinlicher ist. Die Bestimmung des besten Matching-Kandidaten bzw. das Ableiten der richtigen Relationen durch die Auswertung der beiden Überlagerungsverhältnisse nach Vorbild von van Wijngaarden u. a. (1997) ist hier nicht möglich. Als Ergebnis entsteht folgende Relation $p_{A_1} \rightarrow \{p_{B_1}, p_{B_2}, p_{B_3}, p_{B_4}, p_{B_5}\}$, die von einem Experten als nicht korrekt interpretiert wird. Das im Rahmen der vorliegenden Arbeit entwickelte Verfahren für Polygonobjekte erlaubt die Überlagerung von mehreren Objekten innerhalb eines Datensatzes und berücksichtigt dies bei der Zuordnung.

Globale Zuordnung von Straßenobjekten

Li und Goodchild (2011) stellten für die Zusammenführung verschiedener Datensätze einen Ansatz für die Zuordnung von linienhaften Objekten gleichen Maßstabs vor. Das Ziel ist nicht, für jedes Objekt den besten Matching-Kandidaten zu finden, sondern vielmehr global die beste Lösung über alle möglichen Objektkombinationen zu bestimmen. Sie entwickelten dazu ein Optimierungsmodell, das anhand eines Ähnlichkeitsmaßes die Gesamtähnlichkeit zwischen allen zugeordneten Objektpaaren maximiert.

[3]Großmaßstäbige/Detaillierte Topographische Standardkarte der Niederlande im Maßstab 1:1.000
[4]Kleinmaßstäbige Topographische Karte der Niederlande im Maßstab 1:10.000



Nr.		s_{i_A} [%]	s_{i_B} [%]	L1 ($s_{i_A} \vee s_{i_B} > t$)	L2 (n:m)	L3 (1:n/n:1)	L4 (1:1)
1	$p_{A_1} \rightarrow p_{B_1}$	100	70	$p_{A_1} \rightarrow p_{B_1}$		$\boldsymbol{p_{A_1}} \rightarrow p_{B_1}$	
2	$p_{A_1} \rightarrow p_{B_2}$	80	70	$p_{A_1} \rightarrow p_{B_2}$		$\boldsymbol{p_{A_1}} \rightarrow p_{B_2}$	
3	$p_{A_1} \rightarrow p_{B_3}$	20	70	$p_{A_1} \rightarrow p_{B_3}$		$\boldsymbol{p_{A_1}} \rightarrow p_{B_3}$	
4	$p_{A_1} \rightarrow p_{B_4}$	20	70	$p_{A_1} \rightarrow p_{B_4}$		$\boldsymbol{p_{A_1}} \rightarrow p_{B_4}$	
5	$p_{A_1} \rightarrow p_{B_5}$	20	70	$p_{A_1} \rightarrow p_{B_5}$		$\boldsymbol{p_{A_1}} \rightarrow p_{B_5}$	

Abbildung 2.6: Die Objektzuordnung nach dem Vorbild von van Wijngaarden u. a. (1997) erzielt bei Datensätzen, in denen mehrere Objekte das gleiche Gebiet überlagern, keine zuverlässige Lösung.

Das Ähnlichkeitsmaß setzt sich aus geometrischen und semantischen Informationen zusammen. Die Autoren nutzen ein Winkelmaß und die Hausdorff-Distanz, um Objekte in räumlicher Nachbarschaft zuzuordnen sowie die Hamming-Distanz (Hamming, 1950), um Objektnamen zu vergleichen.

Damit zwischen Linienobjekten auch komplexe Relationen (1:n/n:1) identifiziert werden können, wird die Asymmetrieeigenschaft der Hausdorff-Distanz genutzt. Das heißt, es werden die gerichteten Hausdorff-Distanzen bestimmt, die sich in Abhängigkeit vom Ausgangsobjekt unterscheiden. In Abschnitt 3.1.1 wird dies an Beispielen verdeutlicht. Die Identifikation von komplexen n:m -Relationen ist in diesem Verfahren allerdings nicht vorgesehen.

Verfügen Objekte über Objektnamen, wird die Hamming-Distanz d_{Ham} bestimmt. Sie drückt die Unterschiedlichkeit der Buchstaben zweier Zeichenketten aus, indem die Stellen gezählt werden, die nicht übereinstimmen. Für die beiden Flussnamen Leine und Seine ist $d_{Ham}(\text{Leine}, \text{Seine}) = 1$, da sich lediglich die Anfangsbuchstaben unterscheiden. Ursprünglich wurde die Hamming-Distanz für Zeichenketten gleicher Länge entwickelt. Damit das Distanzmaß auch auf Objektnamen unterschiedlicher Längen anwendbar ist, modifizierten die Autoren das Maß.

Experimente haben die Leistungsfähigkeit des Optimierungsmodells in Kombination mit dem entwickelten Ähnlichkeitsmaß gezeigt. In allen Untersuchungsgebieten, sowohl im städtischen als auch ländlichen Bereich, werden hohe Qualitäten mit mehr als 90 % korrekter Zuordnungen identifiziert. Für den Fall, dass zwischen den Datensätzen eine globale Verzerrung existiert, ist zur Verbesserung der Ergebnisse eine Affintransformation in das Optimierungsmodell integriert.

Der größte Nachteil dieses Verfahrens bleibt allerdings die Rechenzeit: Für ein kleines Testgebiet mit wenigen Objekten (666 Objekte in zwei Datensätzen) waren bereits mehr als 2 Stunden Rechenzeit notwendig, ohne an dieser Stelle auf detaillierte Rechnerspezifikationen einzugehen. Obwohl der Lösungsansatz optimal ist, ist dessen Anwendung in der vorgestellten Art und Weise in der Praxis nicht umsetzbar, da der Suchraum bei steigender Objektanzahl und der Berücksichtigung von zusätzlichen semantischen oder topologischen Objektinformationen exponentiell wächst. Um die Berechnung zu beschleunigen, schlagen die Autoren eine Partitionierung der Daten in kleine Gebiete und eine Parallelisierung der Berechnungen vor.

Im Rahmen dieser Arbeit wird die Idee des Optimierungsansatzes aufgegriffen, jedoch nicht für das Data-Matching, sondern für das sich daran anschließende Schema-Matching, mit dem Ziel, die semantische Gesamtähnlichkeit zwischen den Objektklassen zu maximieren.

2.2.4 Ausgewählte, relationale Verfahren

In diesem Abschnitt werden stellvertretend für eine Reihe von Arbeiten zwei relationale Zuordnungsverfahren vorgestellt, die neben geometrischen Objektinformationen auch topologische Eigenschaften zu Nachbarobjekten analysieren. Mustière und Devogele (2008) entwickelten das Verfahren *NetMatcher*, das eine lokale Zuordnung von verschiedenen Netzwerken unterschiedlicher Maßstäbe ermöglicht. Im Gegensatz dazu entwickelte Walter (1997) einen Ansatz für die Zuordnung von linienhaften Straßendaten vergleichbarer Maßstäbe, der eine global optimale Lösung bestimmt, die außerdem komplexe n:m -Relationen beinhalten kann.

Lokale Zuordnung von Linienobjekten unterschiedlichen Maßstabs

Mustière und Devogele (2008) entwickelten zusammen den *NetMatcher*, um eine automatische Nachführung von Aktualisierungen des großmaßstäbigen BD TOPO® (1:25.000) zu BD CARTO® (1:100.000), einem klein-

maßstäbigen Datensatz zu ermöglichen bzw. Unstimmigkeiten zwischen den Datenbanken aufzudecken. Dafür wurden natürliche und anthropogene linienförmige Objekte der Kategorien Straßen, Flüsse, Elektrische Leitungen, Schienen und Wanderrouten untersucht, die sich hinsichtlich der Anzahl der enthaltenen Objekte und ihrer Struktur unterschieden.

Insgesamt umfasst der Zuordnungsprozess sechs Schritte. Im ersten Schritt werden die geographischen Objekte in eine Graphstruktur, bestehend aus Knoten und Kanten, überführt. Ein *Knoten* v repräsentiert einen Punkt, eine *Kante* e entsprechend eine Linie mit Anfangs- und Endpunkt. Die Graphelemente können anhand von thematischen Informationen qualifiziert werden. Knoten, die einen Kreisverkehr oder Kanten, die eine Autobahn repräsentieren, können ein höheres Gewicht bekommen als eine Vier-Wege-Kreuzung oder innerstädtische Hauptstraßen.

Im zweiten und dritten Verfahrensschritt werden potentielle Matching-Kandidaten v_T und e_T für jedes Element des großmaßstäbigen Datensatzes BD TOPO® bestimmt. Die Vorauswahl wird mit Hilfe von einfachen Distanzkriterien in Verbindung mit Schwellwerten getroffen, die abhängig von Zusatzinformationen veränderbar sind. Auf Basis dieser Ergebnisse wird zuerst das Knoten-Matching (vierter Schritt) und anschließend das Kanten-Matching (fünfter Schritt) für jedes Element v_C und e_C des kleinmaßstäbigen Datensatzes BD CARTO® durchgeführt. Das Knoten-Matching ist der wichtigste Schritt im Verfahren, da das Kanten-Matching stark vom Erfolg des Knoten-Matchings abhängig ist.

Zuerst wird jeder schwarze Knoten v_T des detaillierten Datensatzes einer der folgenden drei Kategorien zugeteilt: vollständig, unvollständig und unmöglich. Als vollständige Matching-Kandidaten werden Knoten bezeichnet, wenn unter der Vorauswahl Verbindungen zwischen einigen ihrer Kanten und allen Kanten des kleinmaßstäbigen Datensatzes zu finden sind (vgl. Abb. 2.7 a) und b)). Zusätzlich muss ein Drehkriterium erfüllt sein, das besagt, dass die Zuordnung der einzelnen Kanten in der gleichen Reihenfolge, z.B. im Uhrzeigersinn, erfolgt. In Abbildung 2.7 a) kann die Klassifizierung von v_T einfach nachvollzogen werden, da alle drei Kanten von v_C mit unterschiedlichen Kanten von v_T in der gleichen Reihenfolge zugeordnet wurden. Als unvollständige Matching-Kandidaten werden Knoten bezeichnet, die mindestens eine Verbindung zwischen jeweils einer Kante beider Datensätze besitzen. Dazu zählt der in Abbildung 2.7 c) dargestellte Fall. Obwohl jede Kante einen Matching-Kandidaten hat, wird hier das Drehkriterium nicht erfüllt und der Knoten v_T wird als unvollständig markiert. Eine Zuordnung ist nicht möglich, wenn wie in Abbildung 2.7 d) gezeigt, keine Korrespondenz zwischen den Kanten besteht.

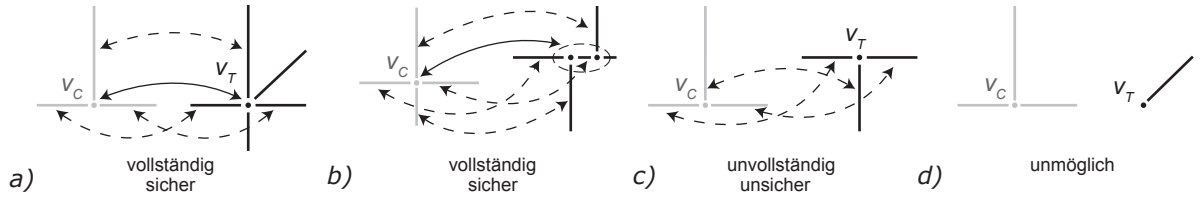


Abbildung 2.7: Knoten-Matching nach Mustière und Devogele (2008). Die schwarzen Knoten v_T des detaillierten Datensatzes werden als vollständige (a, b), unvollständige (c) und unmögliche (d) Matching-Kandidaten zu den grauen Knoten v_C annotiert. Darauf aufbauend werden die Relationen zwischen den Knoten als sicher (a, b) bzw. unsicher (c) klassifiziert. Zwischen Knoten beider Datensätze wird in a) eine 1:1 -Relation und in b) eine 1:n -Relation identifiziert. Die gezeigten Beispiele wurden aus Mustière und Devogele (2008) entnommen und abgewandelt.

Anschließend wird jeder Knoten v_C einzeln betrachtet und die Verbindungen zum detaillierten Datensatz mittels der Kategorien sicher bzw. unsicher annotiert. Wenn zu v_C nur ein Kandidat existiert, der als vollständig gekennzeichnet ist, wird die Verbindung zwischen beiden Knoten als sicher bezeichnet. Wenn dagegen nur ein unvollständiger Knoten existiert, wird die Verbindung als unsicher beschrieben. Werden zu einem Knoten mehrere Kandidaten bestimmt, müssen 1:n -Relationen untersucht werden. Dafür werden die entsprechenden Knoten-Kandidaten zu verbundenen Teilgraphen zusammengefügt und als sogenannte Überknoten betrachtet und nach den gleichen Prinzipien ausgewertet und annotiert.

Beim anschließenden Kanten-Matching wird jede Kante e_C einzeln betrachtet. Es wird im großmaßstäbigen Datensatz nach einem Pfad gesucht, der einerseits aus Matching-Kandidaten besteht, bereits zugeordnete Knoten umfasst und sich möglichst nah an der Kante e_C befindet. Dafür wird ein Pfad bestimmt, der die Fläche zwischen den potentiellen Matching-Kandidaten minimiert. Kanten, deren Anfangs- und Enpunkte ohne Fehler zugeordnet werden, werden als fehlerfrei, alle anderen als fehlerbehaftet, markiert.

Im letzten Schritt erfolgt eine globale Auswertung, bei der Entscheidungen, die auf Basis lokaler Kriterien getroffen wurden, hinsichtlich Mehrfachzuordnungen überprüft werden. Hierbei gilt, dass Elemente des großmaßstä-

bigen Datensatzes nicht mehreren, unterschiedlichen Elementen des kleinmaßstäbigen Datensatzes zugeordnet sein dürfen.

Umfangreiche Experimente haben die Leistungsfähigkeit des *NetMatchers* gezeigt, indem hohe Zuordnungsqualitäten von mehr als 90 % erzielt werden. Ein wichtiger Vorteil des Verfahrens ist, dass einseitig zusammengefasste 1:n -Relationen bestimmt werden können, sei es zwischen einer einfachen Kreuzung und einem Kreisverkehr oder zwischen einer einzelnen Straße und mehreren Straßenabschnitten.

Globale Zuordnung von Straßenobjekten

Walter (1997) entwickelte ein global optimierendes Zuordnungsverfahren, um Attribute, wie z.B. Straßennamen zwischen linienhaften Objekten der Datensätze ATKIS[5] und GDF[6] auszutauschen, die aus unterschiedlichen Fachdisziplinen stammen. Die optimale Zuordnung stellt ein kombinatorisches Problem dar. Die Minimierung des Suchraums ist besonders wichtig, um ein Ergebnis in Polynomialzeit zu erhalten.

In diesem Fall stellt der Raumbezug die natürliche Einschränkung des Suchbereichs dar. Objekte, die sich an völlig unterschiedlichen Positionen befinden, brauchen nicht miteinander verglichen werden. Der Suchraum vergrößert sich wieder, wenn neben einfachen 1:1 -Relationen auch komplexe n:m -Relationen ausgewertet werden müssen. Für die Auswertung von komplexen Relationen wird das *Buffer Growing* eingesetzt. Dieses Verfahren errichtet um jedes Objekt ein Pufferpolygon, um darin Matching-Kandidaten des anderen Datensatzes zu finden. Abbildung 2.8 erläutert die Vorgehensweise anhand eines Beispiels.

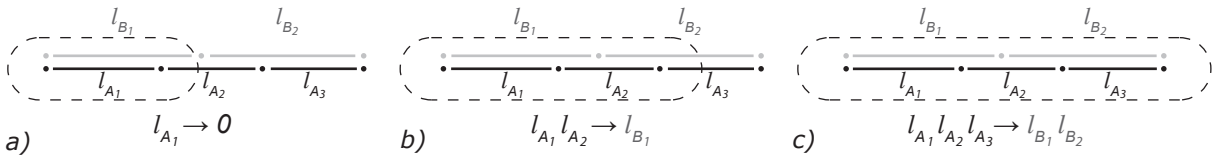


Abbildung 2.8: Buffer Growing nach Walter (1997) für die Identifizierung von komplexen n:m -Relationen zwischen Linienobjekten der Datensätze A (schwarz) und B (grau).

Zunächst wird um das Linienobjekt l_{A_1} aus Datensatz A ein Puffer gelegt (siehe Abb. 2.8 a). Es wird keine Linie aus Datensatz B gefunden, die vollständig innerhalb des gestrichelten Pufferpolygons liegt. Anschließend werden die Objekte l_{A_1} und l_{A_2} zusammengefasst betrachtet und der Puffer nach rechts erweitert (siehe Abb. 2.8 b). Die Linie l_{B_1} befindet sich vollständig innerhalb des Puffers. Ein Vergleich der Linienlängen stellt jedoch für diese Zuordnung ein deutliche Differenz fest. Eine erneute Puffererweiterung (siehe Abb. 2.8 c) identifiziert letztendlich die vorliegende komplexe 3:2 -Relation mit einem geringen Längenunterschied.

Durch die Anwendung des Buffer Growing kann eine Liste mit allen potentiellen Zuordnungspaaren zwischen den beiden Datensätzen aufgestellt werden. Je länger die Liste ist, desto größer ist der Aufwand, die optimale und eindeutige Lösung zu finden. Aus diesem Grund muss die Liste im Wachstum begrenzt werden. Dies kann durch verschiedene geometrische Bedingungen erreicht werden. Beispielsweise dürfen benachbarte Objekte nur dann zusammengefasst werden, wenn ein bestimmtes Winkelkriterium erfüllt ist. Weiterhin kann das Wachsen des Puffers nur in eine bestimmte Richtung erlaubt bzw. auf eine Maximalgröße festgelegt werden. Zudem sind Zuordnungspaare sehr unwahrscheinlich, deren Objekte sich in einem Winkel von ca. 90° schneiden oder deren Längen bzw. deren Form stark voneinander abweichen.

Die Definition der geometrischen Beschränkungen wie Puffergröße, Winkelkriterium und zulässiger Längenunterschied hat einen hohen Einfluss auf die Größe des Suchraums. Im Rahmen der Arbeit wurden Schwellwerte bzw. Parameter anhand von statistischen Analysen von zuvor zugeordneten Daten festgelegt. Dazu wurde eingangs, wie bei Li und Goodchild (2011), der globale Fehler zwischen den beiden Datensätzen durch die Messung von Passpunkten bestimmt und mit Hilfe einer Transformation eliminiert.

Die Bewertung der einzelnen Zuordnungen erfolgt mit Hilfe eines informationstheoretischen Maßes – der gegenseitigen Information. Das Maß gibt an, wie viele Informationen ein Element über ein anderes Element besitzt. Walter (1997) bestimmt für jedes Attribut und jede Relation eine individuelle, gegenseitige Information. Anschließend werden diese Maße kombiniert, um die Ähnlichkeit aller Attribute und Relationen gemeinsam widerzuspiegeln. Die gegenseitige Information ist eine Leistungsfunktion und hat gegenüber einer Kostenfunktion

[5] Topographische Geobasisdaten des Amtlichen Topographisch-Kartographischen Informationssystems von Deutschland
[6] Geodaten der Firma TomTom im Geographic Data Files-Format (GDF)

den Vorteil, dass 1:0 -Objektzuordnungen keine negativen Einflüsse haben. Für die Bestimmung der gegenseitigen Information wird die bedingte Wahrscheinlichkeit benötigt, die ebenfalls aus den statistischen Analysen der Datensätze abgeleitet werden kann. Dabei ist es wichtig, dass eine Zuordnung seltener Attributwerte bzw. Relationen gegenüber häufigen Relationen ein höheres Gewicht erhält. Abschließend wird die Summe aller Informationsmaße mit Hilfe des Bergsteiger-Algorithmus, einem Baumsuchverfahren, maximiert.

Die von Walter (1997) durchgeführten Experimente haben die Leistungsfähigkeit des Zuordnungsverfahrens in Kombination mit dem entwickelten Ähnlichkeitsmaß gezeigt. In drei von vier Testgebieten mit unterschiedlichen Straßendichten werden hohe Zuordnungsqualitäten von mehr als 90 % korrekter Zuordnungen erzielt. Im vierten Testgebiet ist der Unterschied zwischen den Datensätzen allerdings sehr stark, so dass signifikant mehr Fehlzuordnungen gefunden werden. Da im Rahmen des Verfahrens für jede Zuordnung ein Qualitätsmaß bestimmt wird, können Fehlzuordnungen automatisch identifiziert werden, um sie gegebenenfalls manuell zu überprüfen oder ganz zu entfernen. Die hohe Rechenzeit des Programms stellt jedoch eine große Einschränkung für eine effiziente Anwendung in der Praxis dar.

2.3 Schema-Matching

In der Forschung wird das Thema *Schema-Matching* als Notwendigkeit für viele verschiedene Anwendungen identifiziert und seither in der Literatur als eigene Thematik umfassend diskutiert (Rahm und Bernstein, 2001; Rahm und Peukert, 2018b). Schema-Matching bezeichnet die Identifikation von Korrespondenzen zwischen Elementen zweier Schemas. Eine Korrespondenz beschreibt wiederum eine Relation auf Schemaebene zwischen einem oder mehreren Elementen des einen Schemas mit einem oder mehreren Elementen des anderen Schemas (vgl. Abschnitt 3.2.2). Korrespondenzen spiegeln semantische Verknüpfungen wider.

Eine 1:1 -Relation zwischen zwei Schemaelementen entspricht einer semantischen Übereinstimmung. Dieser Rückschluss ist nur bedingt korrekt, wie das Einführungsbeispiel in Abbildung 2.9 verdeutlicht. Ein Algorithmus, der Wörter zuordnet (engl. Name-Matcher), identifiziert in beiden Schemas das Wort Bank, allerdings mit unterschiedlichen Bedeutungen. Während das Wort Bank in Schema A eine Universalbank beschreibt, ist es in Schema B ein Sitzmöbel. Homonyme, die gleich geschrieben werden, aber unterschiedliche Bedeutungen besitzen, verfälschen das Ergebnis eines Algorithmus stark.

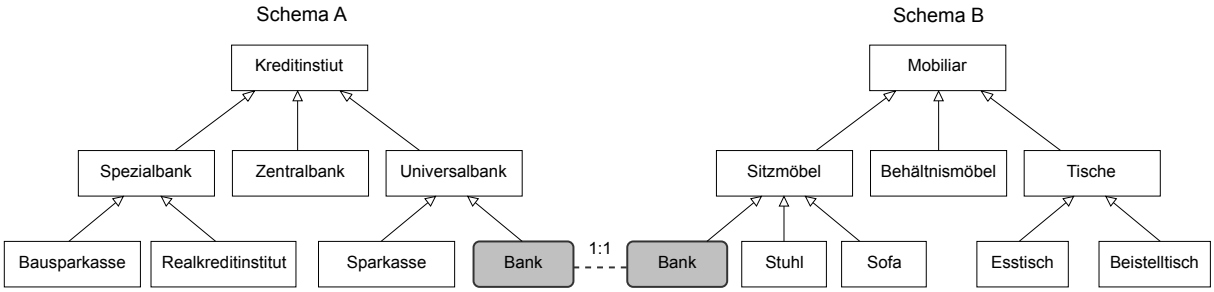


Abbildung 2.9: Schema-Matching zwischen Schema A und Schema B. Die gestrichelte Verbindungslinie zwischen den grau hervorgehobenen Schemaelementen Bank kennzeichnet eine 1:1 -Schemarelation.

Ergebnisse von Schema-Matchings sind Voraussetzung für viele verschiedene Aufgaben. Beispielsweise können aus semantischen Relationen Regeln für einen Datenaustausch, für alle Arten der Datenintegration oder für die Entwicklung neuer Schemas abgeleitet werden, damit Daten aus verschiedenen Disziplinen zusammengeführt werden können.

Neben dem Begriff Schema-Matching sind auch die Bezeichnungen *Ontology Alignment* (Noy und Musen, 2003) oder *Model Matching* (Falleri u. a., 2008) üblich. Im weiteren Verlauf dieser Arbeit wird der Begriff Schema-Matching verwendet.

Techniken, die beim Schema-Matching angewendet werden, um semantische Übereinstimmungen zu entdecken, die verschlüsselt im Schema vorhanden sind, werden im folgenden Abschnitt beleuchtet. Ebenso werden Probleme und Herausforderungen benannt, die bei der Entwicklung und Verwendung bereits existierender Verfahren berücksichtigt werden müssen. Für einen ausführlicheren Überblick sei an dieser Stelle bereits auf die Bücher *Schema Matching und Mapping* (Bellahsene u. a., 2011) und *Ontology Matching* (Euzenat und Shvaiko, 2007, 2013) sowie auf die Übersichten von Rahm und Bernstein (2001), Kalfoglou und Schorlemmer (2003), Shvaiko und Euzenat (2005), Bernstein u. a. (2011), Rahm und Peukert (2018a) und Rahm und Peukert (2018b) verwiesen.

2.3.1 Klassifikation von Zuordnungsverfahren auf Schemaebene

Techniken, die beim Schema-Matching zum Einsatz kommen, werden anhand des Klassifikationsschemas (Abbildung 2.10) von Rahm und Bernstein (2001) vorgestellt. Die Taxonomie wird in der Forschung häufig als ein Standard genutzt, um neu entwickelte Zuordnungsverfahren auf Schemaebene einzuordnen.

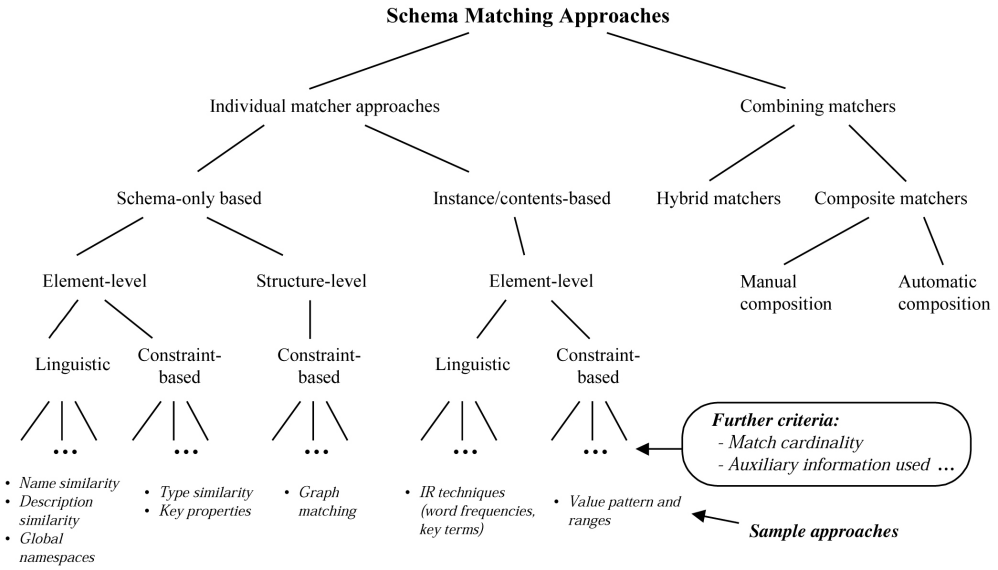


Abbildung 2.10: Klassifikation von Schema-Matching-Ansätzen (aus Rahm und Bernstein (2001)).

Abbildung 2.10 zeigt, dass ein großes Spektrum an Schema-Matching-Techniken existiert. Unterschieden wird zwischen individuellen und kombinierten Verfahren, sogenannten Individual und Combining Matchers. Die Kombination mehrerer Matcher wird empfohlen, um die beste Zuordnung zu finden, die ein Verfahren allein in der Regel nicht erzielen kann. Bei kombinierten Matchern wird zwischen Hybrid und Composite Matcher unterschieden. Während Hybrid Matcher verschiedene Techniken entweder simultan oder in einer bestimmten Reihenfolge ausführen und Ergebnisse auf vorhergehende Matcher aufbauen, kombinieren Composite Matcher die Ergebnisse von unabhängig ausgeführten Verfahren, was sie bei der Auswahl der einzelnen Techniken flexibler macht. COMA++ (Aumueller u. a., 2005) zählt zu den kombinierten Verfahren, bei dem 15 verschiedene Zuordnungsalgorithmen auswählbar sind.

Individuelle Verfahren nutzen entweder die Metadaten der Schema (Schema-Based) oder ihre Objekte (Instance-Based). Bei schemabasierten Ansätzen werden nur Informationen des Schemas verwendet, entweder Eigenschaften der Schemaelemente, wie z.B. Name, Beschreibung, Datentyp, Restriktionen bzgl. des Wertebereichs analysiert oder strukturelle Eigenschaften des Schemas. Die Zuordnung erfolgt zwischen einzelnen Elementen (Element-level) oder zwischen komplexen Schemastrukturen, einer Kombination aus mehreren Schemaelementen (Structure-level). Beim zuletzt genannten Ansatz wird zusätzlich zwischen einer vollständigen oder partiellen Zuordnung unterschieden.

Bei instanzbasierten Ansätzen wird die Korrespondenz der Schemaelemente allein aus der Ähnlichkeit der Objekte abgeleitet. Das heißt, je ähnlicher sich einzelne Objekte sind, desto häufiger werden sie einander zugeordnet und umso wahrscheinlicher und eindeutiger ist die semantische Beziehung ihrer Schemaelemente. Für die Ähnlichkeitsbewertung ist die Definition eines Ähnlichkeitsmaßes notwendig, das für die Daten geeignet ist, damit Fehlentscheidungen, wie im Einführungsbeispiel (Abb. 2.9), nicht entstehen. Beim Vergleich von Artikelnummern in einem Produktkatalog müssen andere Kriterien berücksichtigt werden als bei raumbezogenen Objekten, die durch Koordinaten repräsentiert werden.

Objekte stellen eine wertvolle Quelle dar, da sie wichtige Einblicke in den tatsächlichen Inhalt bzw. die Bedeutung eines Schemaelements gewähren, besser als es sich durch einen Namen oder eine kurze Beschreibung ausdrücken lässt (Duckham und Worboys, 2005). Diese Informationen sind umso wichtiger, wenn nützliche Informationen zum Schema fehlen oder überhaupt kein Schema vorhanden ist. Nachteilig ist, dass für die Identifikation hochqualitativer Korrespondenzen genügend sowie geeignete Objektdaten für alle Schemaelemente benötigt werden, die nicht immer zur Verfügung stehen. Demzufolge wird der instanzbasierte Ansatz häufig nur als Ergänzung zum schemabasierten Ansatz verwendet.

Bei schema- und instanzbasierten Ansätzen werden hauptsächlich sprachwissenschaftliche Techniken (Linguistic-Matcher) verwendet, um semantisch ähnliche Elemente zu finden. Der Einfachheit halber wird vielfach angenommen, dass die Ähnlichkeit zweier Namen die Ähnlichkeit der Semantik ausdrückt, was für Schemas mit einfacher Semantik durchaus zulässig ist (Kokla und Kavouras, 2005). Hierbei werden einzelne Wörter (Name-Matcher), wie z.B. der Name eines Schemaelements oder eines Produkts, oder ganze Texte (Description-Matcher) analysiert. Die Gleichheit bzw. Ähnlichkeit wird ganz unterschiedlich bestimmt. Wörter können als gleich angesehen werden, wenn die komplette Zeichenfolge identisch ist oder mindestens eine Teilfolge (z.B. bei Auto und Automobil) abbildet. Des Weiteren werden Wörter als ähnlich bezeichnet, wenn es sich um Synonyme (z.B. Gebäude ≅ Haus), Hyperonyme (z.B. Gewässer als Oberbegriff von Teich), Akronyme (z.B. S-Bahn als Kurzwort für Schnellbahn oder Stadtbahn) oder Übersetzungen (z.B. See, Lake, Lac) handelt. Für die Identifikation solcher Beziehungen sind Wortschatzsammlungen (Thesauren), Lexika oder Wörterbücher notwendig. Beim Description-Matching werden Schlüsselwörter oder auch komplexe, semantisch äquivalente Ausdrücke auf Basis der relativen Häufigkeit der Wörter oder Wortkombinationen aus Beschreibungen extrahiert und mittels der bereits vorgestellten linguistischen Techniken analysiert. Diese linguistische Charakterisierung erfolgt in Form von Techniken der Informationsrückgewinnung (engl. Information Retrieval) (Rahm und Bernstein, 2001).

Beinhalten Schemaelementbezeichnungen Sonderzeichen, wie z.B. in *„qh/W“*, ist eine Ähnlichkeitsbestimmung basierend auf der Gleichheit der Namen nicht mehr direkt möglich. Andere Kriterien müssen herangezogen werden, um Korrespondenzen aufzudecken. Die Unterstützung von Experten ist in Erwägung zu ziehen. Eine andere Möglichkeit ist, sogenannte Constraint-Based Matcher einzusetzen. Damit werden definierte Restriktionen überprüft, z.B. bei verwendeten Datentypen (z.B. String vs. Integer), Wertebereiche (z.B. positive Werte zwischen 1000 und 9999), Beziehungen (z.B. part-of, is-a) oder Kardinalitäten. Die alleinige Verwendung der Informationen über bestehende Bedingungen führt allerdings häufig zu fehlerhaften n:m -Zuordnungen, da mehrere Schemaelemente oft vergleichbare Bedingungen besitzen.

Die Auswahl bzw. die Entwicklung eines geeigneten Schema-Matching-Verfahrens bleibt schwierig, da es zur Anwendung und zum Schematyp passen muss. Im Rahmen der Arbeit wird ein instanzbasierter Ansatz entwickelt, um möglichst unabhängig von den Informationen der Schemaelemente zu bleiben, die nicht bei jedem Datensatz zur Verfügung stehen. Welche Herausforderungen im Allgemeinen bei der Zuordnung auf Schemaebene auftreten, wird im folgenden Abschnitt vorgestellt.

2.3.2 Herausforderungen bei der Zuordnung auf Schemaebene

Obwohl Experten die Fehlzuordnung zwischen Bank als Kreditinstitut und Bank als Mobiliar aus dem Einführungsbeispiel einfach erkennen können, ist ein manuelles Schema-Matching nicht geeignet. Die wiederkehrende manuelle Zuordnung durch Experten, die die Terminologie der erfassten Datenbestände kennen, ist sehr zeit- und kostenintensiv und lediglich für eine begrenzte Anzahl von Datensätzen möglich. Der Prozess ist zudem sehr fehleranfällig, insbesondere wenn Schemas an Komplexität zunehmen. Aus diesem Grund liegt der Fokus der Forschung auf der Entwicklung von automatisierten Verfahren (Volz, 2006), um einerseits die Zuordnung immer größer werdender Schemas (engl. Large-Scale Schema-Matching) bewältigen und andererseits das Matching von mehr als zwei Schemas gleichzeitig (engl. Holistic Schema-Matching) ermöglichen zu können. Gegenwärtig existieren viele Verfahren, bei denen eine manuelle Unterstützung notwendig ist. Es werden dem Anwender die wahrscheinlichsten Matching-Kandidaten präsentiert, die ein Zuordnungsverfahren ermittelt hat und anhand derer er die finale Entscheidung treffen kann.

Besonders große und unübersichtliche Schemas mit vielen Elementen, Attributen und tiefen Verschachtelungen stellen Schema-Matching-Verfahren bezüglich der Zuordnungsqualität und Leistungsfähigkeit zunehmend vor große Herausforderungen. Um hohe Zuordnungsqualitäten zu erreichen, müssen die identifizierten semantischen Korrespondenzen korrekt und vollständig sein. Je größer der Suchraum, desto schwieriger ist die Aufgabe. Für die paarweise 1:1 -Zuordnung wächst der Suchraum mindestens quadratisch mit der Anzahl der Schemaelemente. Somit erhöhen sich die Wahrscheinlichkeit der Falschzuordnungen und die Ausführungszeit. Die individuelle Leistungsfähigkeit jedes Matching-Algorithmus ist demzufolge sehr stark von der Art beeinflusst, wie die Eingangsinformationen verarbeitet werden. Element-Level-Techniken, die nur einzelne Schemaelemente in Isolation betrachten, wie z.B. der Name-Matcher, sind beispielsweise einfacher und schneller auszuführen als Structure-Level-Techniken.

In der Regel werden mehrere Matcher kombiniert, um die Zuordnungsqualitäten zu erhöhen. Die größte Herausforderung besteht allerdings darin, geeignete Matcher auszuwählen und deren Ausführungsreihenfolge festzulegen. Meta Matching Systeme (Shvaiko und Euzenat, 2005; Ehrig u. a., 2005; Lee u. a., 2007; Duchateau u. a., 2008; Peukert u. a., 2010) sollen genau diesen Modellierungsprozess unterstützen. Dafür ist ein detailliertes

Wissen über die Fülle der existierenden Matching-Techniken notwendig, um sie gewinnbringend für die individuelle Anwendung einsetzen zu können. Für eine weitere Optimierung werden neue spezielle Matching-Verfahren entwickelt, die sowohl den Eingangsdaten entsprechen als auch den eigenen Anforderungen genügen.

Je mehr Matching-Techniken an einem Zuordnungsprozess beteiligt sind, desto höher ist die Komplexität des Verfahrens und desto geringer die Leistungsfähigkeit. Viele Matching-Systeme haben gerade bei der Zuordnung von großen Schemas Probleme mit der Leistungsfähigkeit. Der Fokus dieser Systeme lag bisher auf der Qualität und nicht auf der Leistungsfähigkeit (Peukert u. a., 2010). Die Herausforderung besteht darin, einen Kompromiss zwischen Zuordnungsqualität und Leistungsfähigkeit zu finden.

Die Performance eines Algorithmus kann erhöht werden, wenn der Suchraum verringert wird (Rahm und Peukert, 2018b). Zu Beginn kann beispielsweise ein schneller Matcher genutzt werden, um unwahrscheinliche Matching-Kandidaten frühzeitig zu eliminieren. Im Anschluss kann dann ein exakter Matcher die kleinere Anzahl an Kandidaten auswerten (Ehrig und Staab, 2004; Peukert u. a., 2010). Nachteilig ist, dass einmal aussortierte Kandidaten nicht wieder zurück in den Suchprozess eingebracht werden können.

Sehr große Schemas können mit der Divide-and-Conquer-Strategie manuell oder automatisch in Fragmente (Aumueller u. a., 2005; Do und Rahm, 2007), Blöcke (Hu und Qu, 2006; Hu u. a., 2008), Partitionen (Abadi u. a., 2007; Paulheim, 2008) oder Cluster (Smiljanic u. a., 2006; Saleem u. a., 2008) zerlegt (engl. Partition-Based Matching) werden. Die Zuordnungsalgorithmen werden dann auf die jeweils reduzierten Suchräume angewendet. Allerdings kann sich dadurch die Qualität der Gesamtlösung verschlechtern. Des Weiteren ist ein paralleles Matching möglich, d.h. entweder können verschiedene Matcher oder mehrere Partitionen gleichzeitig berechnet werden. Damit lässt sich die Laufzeit von Zuordnungsverfahren verbessern. Erste Prototypen sind Gomma (Gross u. a., 2010), ein sehr schneller Name-Matcher und SPHeRe (Amin u. a., 2016). Das vorherige Filtern der Schemainformationen in relevante und nicht relevante Informationen kann ebenfalls den Suchraum reduzieren. Der Einsatz von speziellen Datenstrukturen, wie z.B. Indexe oder Hash-Tabellen, kann die Suchprozesse verbessern.

Derzeitige Schema-Matching-Ansätze liefern hauptsächlich paarweise 1:1 -Zuordnungen, z.B. COMA++ (Aumueller u. a., 2005), AgreementMaker (F. Cruz u. a., 2009) oder Falcon (Hu und Qu, 2008). Verfahren müssen zukünftig in der Lage sein, auch komplexe Zuordnungen, wie z.B. 1:n - oder n:m -Relationen, zu identifizieren. Besonders wenn zwei Schemas stark unterschiedliche Strukturen und Terminologien aufweisen, werden komplexe Relationen immer wahrscheinlicher. Eine Zuordnung zwischen einem eher kleinen Datenbankschema, das nur Schemaelementnamen besitzt und einem Thesaurus-Schema, bestehend aus vielen hunderttausend Einträgen mit einer sehr reichen Semantik, ist allein mit 1:1 -Schemarelationen nicht denkbar. Arnold und Rahm (2014) entwickelten den STROMA-Algorithmus, der neben Äquivalenzbeziehungen auch is-a und part-of Relationen identifizieren kann.

Komplizierter wird es, wenn Schemaelemente aus unbekannten Synonymen oder kryptischen Elementen bestehen und eine sprachwissenschaftliche Analyse durchgeführt werden soll. Die lexikalische Problematik wird zumindest bei der Verwendung von instanzbasierten Verfahren umgangen. Schemaelemente werden nur in der Art und Weise zugeordnet, wie sie tatsächlich durch die Instanzen repräsentiert werden. Der instanzbasierte Ansatz ist somit auch resistent gegenüber Fehlern, die durch manuelle Annotationen verursacht worden sind.

Zukünftige Herausforderungen für das Schema-Matching sind die qualitative Bewertung der automatisch bestimmten Ergebnisse sowie die Aufstellung von Vergleichskriterien (engl. Benchmarking) für die entwickelten Zuordnungsalgorithmen, um Verfahrensschwächen aufzudecken. Rahm und Peukert (2018b) empfehlen, Zuordnungsergebnisse von Menschen prüfen zu lassen und gegebenfalls anzupassen, d.h. falsche Korrespondenzen zu löschen bzw. fehlende Korrespondenzen hinzuzufügen. Dazu werden allerdings noch Schnittstellen zur Benutzerbeteiligung benötigt.

Die meisten Arbeiten des Schema-Matchings beschäftigen sich mit Textdokumenten, in denen Wörter die Objekte der Schemaelemente widerspiegeln. Beispielsweise können in Webanwendungen Produktkataloge, im Gesundheitswesen Patientenakten und andere medizinische Reports und im Bereich des elektronischen Handels (engl. E-Commerce) Nachrichtenformate für verschiedene Geschäftsdokumente wie Bestellungen und Rechnungen abgeglichen werden. Im Rahmen dieser Arbeit liegt der Fokus auf geographischen Daten. Aus diesem Grund werden im nächsten Abschnitt Schema-Matching-Verfahren speziell für den geographischen Kontext vorgestellt.

2.3.3 Ausgewählte Schema-Matching-Verfahren im geographischen Kontext

In diesem Abschnitt werden beispielhaft ein schemabasiertes und zwei instanzbasierte Zuordnungsverfahren vorgestellt. Während in der Arbeit von Hess u. a. (2007) das Hauptaugenmerk auf die Analyse der Schemaelemente

hinsichtlich ihrer Gemeinsamkeiten und Unterschiede gelegt wird, werden in den Arbeiten von Duckham und Worboys (2005) und Volz (2006) direkt geographische Objekte der verwendeten Datensätze für die Zuordnung der Objektklassen zweier Schemas genutzt.

Schemabasierte Zuordnung

Hess u. a. (2007) entwickelten einen schemabasierten Algorithmus mit dem Namen *G-Match*, der die Integration von geographischen Ontologien möglich macht. Für das Schema-Matching werden sowohl Gemeinsamkeiten als auch Unterschiede zwischen den Schemaelementen, hier als Konzepte bezeichnet, gemessen und in verschiedenen Ähnlichkeitsmaßen ausgedrückt.

G-Match ist ein dreistufiges Verfahren. Zuerst werden die Ähnlichkeit der Konzeptnamen und deren Attribute bestimmt. Dazu wird WordNet® (Miller, 1995), eine umfangreiche, lexikalische Datenbank der englischen Sprache verwendet, um Synonyme, Homonyme, aber auch verwandte Terme zu finden. Aufbauend auf diesen Ergebnissen wird im zweiten Schritt die Ähnlichkeit der Taxonomien und die Ähnlichkeit der konventionellen und topologischen Relationen zwischen den Konzepten bestimmt. Ob sich zwei ähnliche Konzepte auf der gleichen Hierarchieebene befinden, wird zum einen aus der Anzahl der Sub-Konzepte und zum anderen aus der Position in der Taxonomie abgeleitet. Die Ähnlichkeit der konventionellen Relationen wird durch das Zählen der gemeinsamen und unterschiedlichen is-a-Beziehungen (z.B. Museum is-a Attraction) der ähnlichen Konzepte bestimmt.

Im Gegensatz dazu sind topologische Relationen, die räumliche Beziehungen zwischen zwei Geometrien ausdrücken und sich formal in Kategorien des 9-Intersection-Modells von Egenhofer (1989) einteilen lassen, explizit durch Bezeichner, wie z.B. overlap, inside, cross charakterisiert. Für die Ähnlichkeitsbestimmung werden sowohl die Relationsnamen verglichen als auch die Geometriedimensionen der daran beteiligten Konzepte berücksichtigt. In Belussi u. a. (2005) wird ausführlich beschrieben, dass bestimmte topologische Beziehungen in Abhängigkeit ihrer Geometriedimension gleiche Bedeutungen haben. Beispielsweise ist overlap als Überlappungsrelation nur zwischen Paaren von Polygon- oder Linienobjekten definiert und nicht zwischen einem Polygon und einer Linie. In diesem Fall ist die Relation cross gleichbedeutend mit der Relation overlap.

Abschließend werden im dritten Schritt alle Ähnlichkeitsmaße in einer gewichteten Summe zu einem Gesamtmaß kombiniert. Die optimale Kombination der einzelnen Ähnlichkeitsmaße ist sehr stark von den Charakteristika der Eingangsontologien abhängig und muss daher noch manuell unterstützt werden.

Instanzbasiertes Schema-Matching mittels Techniken der Verbandstheorie

Duckham und Worboys (2005) realisierten einen automatisierten Schema-Fusionierungsprozess, bei dem raumbezogene Objektdaten für die Identifikation von ähnlichen Schemaelementen verwendet werden. Sie präsentieren ein hypothetisches Beispiel, bei dem zwei Landnutzungsdatensätze, die als Partitionen bestehend aus Polygonobjekten mit dazugehörigen Schemas modelliert vorliegen, geometrisch überlagert werden.

Als Ergebnis der Überlagerung entsteht eine fusionierte Karte mit teilweise neuen geometrischen Objekten, annotiert mit neuen Objektklassen, die durch die Zusammenfassung aller an der Relation beteiligten Objektklassen entstanden sind. Basierend auf den Relationen in den Ausgangsschemas und den neuen Klassifikationsergebnissen wird ein integriertes Schema nach Regeln der Verbandstheorie konstruiert. Hierfür spielen weder die Art noch die Häufigkeit der identifizierten Objektrelationen eine Rolle. Existieren vorab Informationen über bestimmte Objektklassenverbindungen, sogenannte intensionale Informationen, können diese als Bedingung in den Konstruktionsprozess mit eingebracht werden. Eine ausführliche Einführung in die Thematik der Verbandstheorie, die sich allgemein mit der Untersuchung von Strukturen beschäftigt, ist in Grätzer (1978) zu finden.

Im Idealfall sind im integrierten Schema neue bzw. bis dahin unbekannte Schemarelationen auffindbar, die eine Verbesserung gegenüber der konventionellen Überlagerung darstellen. Allerdings ist dies stark von der Qualität der Daten, den sogenannten extensionalen Informationen abhängig. Unsicherheiten in den Eingangsdaten, wie z.B. Unrichtigkeiten oder Ungenauigkeiten, können erhebliche Auswirkungen auf die Schlussfolgerung haben. Bestehen zwischen den Objekten geometrische Abweichungen, führt dies zur Bildung von zusätzlichen Objekten mit meist kleiner Fläche und zu neuen, weiteren Objektklassen, die sich schwer integrieren lassen. Um der Ungenauigkeit entgegenzuwirken, schlagen die Autoren die Einführung von Schwellwerten für neu gebildete Objekte vor, um so sehr kleine Flächen eher unklassifiziert zu lassen und im Schemakonstruktionsprozess zu vernachlässigen.

Die Festlegung eines Schwellwerts kann ein Gleichgewicht zwischen der Qualität der Objekt- und der Objektklassenzuordnung im fusionierten Datensatz unter Ungenauigkeiten schaffen. Eine hohe Qualität bei der Objektzuordnung mit geringen, unklassifizierten Regionen kann nur durch einen kleinen Schwellwert erreicht

werden, was wiederum zu Lasten geringer, intensionaler Informationen geht und demzufolge nur eine schlechte Integration der Eingangsschemas ermöglicht.

Instanzbasiertes Schema-Matching basierend auf statistischen Korrelationswerten

Einen wichtigen Beitrag zum instanzbasierten Schema-Matching im Bereich Geoinformatik lieferte Volz (2006). Für zwei linienförmige Straßenverkehrsdatensätze (ATKIS und GDF) entwickelte Volz ein semi-automatisches Data-Matching-Verfahren, das Objektzuordnungen auf Basis von geometrischen und topologischen Eigenschaften, vereint in einem Gesamtähnlichkeitsmaß, ermöglicht. Aus den Zuordnungsergebnissen können vollautomatisch eindeutige Klassenzuordnungen der Kardinalität 1:1, aber auch Klassenkombinationen der Kardinalitäten 1:n und n:m ausgewertet werden.

Im Gegensatz zu Duckham und Worboys (2005) spielen an dieser Stelle die Häufigkeiten und die Art der Objektrelationen eine wichtige Rolle, weil daraus statistische Korrelationsmaße für die beteiligten Objektklassen abgeleitet und als prozentuales Maß angegeben werden.

Das nachfolgende Beispiel, das ausschnittsweise aus der Arbeit von Volz (2006) entnommen wurde, soll den Ansatz kurz erläutern. Aus den in Tabelle 2.1 dargestellten Zuordnungsergebnissen für die GDF-Objektklasse RoadElement und die ATKIS-Objektklassen Straße und Weg können die in Tabelle 2.2 angegebenen Korrelationswerte bestimmt werden.

Tabelle 2.1: Auszug aus den Ergebnissen der Objektzuordnung für die GDF-Objektklasse RoadElement *für das zweite Testgebiet aus der Arbeit von Volz (2006).*

	Straße					Straße-Weg					Weg				
	1:1	1:n	n:1	n:m	Σ	1:1	1:n	n:1	n:m	Σ	1:1	1:n	n:1	n:m	Σ
RoadElement	333	42	32	22	429	0	0	6	7	13	59	24	13	11	107

Tabelle 2.2: Korrelationswerte für die GDF-Objektklasse RoadElement *aus den in Tabelle 2.1 angegebenen Zuordnungsergebnissen.*

Korrespondenzen	Anzahl	Gesamt	Prozent
RoadElement → Straße	429	435,5	79,3
RoadElement → Weg	107	113,5	20,7
RoadElement → Straße, Weg	13		

Die Auswertung der Klassenzuordnungen, an denen auf einer Seite mehrere unterschiedliche Objektklassen beteiligt sind, ist besonders. Diese Relationen werden einfach auf die eindeutigen Klassenzuordnungen unter Annahme der Gleichverteilung aufgeteilt. Das bedeutet, dass die 13 Relationen RoadElement → {Straße, Weg} jeweils zur Hälfte auf die Klassenkombinationen RoadElement → Straße und RoadElement → Weg addiert werden. Die neuen Gesamtrelationswerte 435,5 und 113,5 werden für die Bestimmung der prozentualen Korrelationswerte genutzt.

Beim Ansatz der Gleichverteilung haben eine 1:2-Objektrelation mit RoadElement → {Straße, Weg} und eine 1:4-Relation mit RoadElement → {Straße, Straße, Straße, Weg} den gleichen Einfluss auf das Ergebnis, obwohl die letztgenannte eher die Zuordnung der Klassen RoadElement → Straße empfehlen würde.

Die umfangreichen Experimente von Volz (2006) haben die Funktionsfähigkeit des Verfahrens bestätigt. Sowohl die vollautomatische Bestimmung von Klassenkorrespondenzen als auch die Einbeziehung von Attributen waren für die untersuchten Datensätze sehr gut möglich und entsprachen den Erkenntnissen, die auch ein Experte manuell zwischen den Objektklassen beider Schemas ermitteln würde. Die Korrelationswerte für die Objektklassen konnten nochmals gesteigert werden, indem nur Objektrelationen mit einem hohen Gesamtähnlichkeitsmaß verwendet wurden.

Ein Nachteil des Verfahrens ist die fehlende Interpretation der Ergebnisse. Diese muss der Anwender selbst analysieren. Liefern die Prozentwerte keine eindeutigen Ergebnisse, wie z.B. dass Objekte der Klasse Weg ausschließlich Objekten der Klasse RoadElement zugeordnet werden, gibt der Algorithmus keine Empfehlung, welche Objektklassen am besten korrespondieren.

Im Rahmen der vorliegenden Arbeit werden einerseits die gemischten Objektrelationen genauer analysiert, um die vorhandenen Informationen, wie sie in der oben vorgestellten 1:4-Relation zur Verfügung stehen, besser für die Ableitung der Klassenkorrespondenzen zu nutzen. Anderseits bestimmt das entwickelte Schema-Matching-Verfahren für jede Objektklasse beider Schemas einen konkreten Zuordnungspartner.

3 Grundlagen

In diesem Kapitel werden mathematische und algorithmische Grundlagen beschrieben, die für das Verständnis der Arbeit notwendig sind. In Abschnitt 3.1 werden Ähnlichkeitsmaße präsentiert, die entweder für die Zuordnung räumlicher Objekte oder von Objektklassen auf Schemaebene einsetzbar sind. Anschließend werden in Abschnitt 3.2 verschiedene Relationstypen erläutert, die auf Objekt- und Schemaebene auftreten. Abschnitt 3.3 stellt Grundlagen der Graphentheorie vor, da die Zuordnung der Objektklassen u.a. mit existierenden Graphalgorithmen durchgeführt wird. Die ganzzahlige lineare Programmierung wird als Optimierungstechnik in Abschnitt 3.4 erläutert. Sie wird genutzt, um garantiert optimale Lösungen für das Zuordnungsproblem zu bestimmen.

3.1 Ähnlichkeitsmaße

In Abschnitt 2.1.2 wurde die Schwierigkeit der Ähnlichkeitsbewertung anhand von Beispielen verdeutlicht. Für einen Vergleich sind Ähnlichkeitsmaße erforderlich, die helfen, Gemeinsamkeiten und Unterschiede zwischen Objekten entsprechend der Fragestellung richtig zu interpretieren. Im Kontext der Arbeit wird einerseits die Ähnlichkeit von raumbezogenen Objekten untersucht. Hierbei spielen vor allem die räumliche Position, die geometrische Objektform und Beziehungen zu Nachbarobjekten eine entscheidende Rolle. Daher wird eine Unterscheidung in *geometrische* und *topologische Ähnlichkeit* vorgenommen. Zum anderen wird die Ähnlichkeit zwischen den Objektklassen auf Schemaebene bewertet und im Folgenden als *semantische Ähnlichkeit* bezeichnet. Dazu können entweder Schemainformationen, wie z.B. Objektklassennamen oder Attribute analysiert oder Korrelationswerte aus den Objektrelationshäufigkeiten bestimmt werden. In den nachfolgenden Abschnitten werden verschiedene Ähnlichkeitsmaße für diese Kategorien vorgestellt, von denen viele getestet wurden, aber nicht alle abschließend in die Verfahren der vorliegenden Arbeit eingeflossen sind. Viele der hier vorgestellten Ähnlichkeitsmaße wurden jedoch in einer früheren Arbeit verwendet, in der die Maße für verschiedene Gebäude bestimmt wurden. Mittels Methoden der unüberwachten Klassifikation wurden Objekte in Cluster mit gleichen geometrischen Eigenschaften zusammengefasst und anschließend semantisch angereichert (Werder u. a., 2010).

3.1.1 Geometrische Ähnlichkeit

Den entscheidenden Hinweis für die Zuordnung von raumbezogenen Objekten liefert die geographische Position. Alle Repräsentationen eines Real-Welt-Objektes müssen sich an der gleichen Position oder zumindest in unmittelbarer Nachbarschaft befinden. Eine Lageabweichung kann durch verschiedene Gründe hervorgerufen werden, z.B. durch die Erfassung der Datensätze mit unterschiedlichen Genauigkeiten und Methoden oder durch die Modellierung der Objekte in verschiedenen Geometriedimensionen. Daher stellt die Nachbarschaftsbeziehung ein bedeutendes Ähnlichkeitsmaß dar, das durch Distanzmaße ausgedrückt werden kann (Li und Goodchild, 2012). Je geringer die Distanz zwischen zwei Objekten, desto größer ist die Wahrscheinlichkeit, dass diese Objekte das gleiche Real-Welt-Objekt beschreiben.

Distanz

Es gibt eine Vielzahl von Distanzmaßen, die es in Abhängigkeit von den zu vergleichenden Geometrietypen zu wählen gilt. Aufgrund der Fokussierung der Arbeit auf Objekte mit linien- und flächenhafter Ausprägung werden im Folgenden geeignete Maße für diese beiden Objektdimensionen beschrieben.

Für die Zuordnung von Linienobjekten eignet sich die Hausdorff-Distanz (Yuan und Tao, 1999), obwohl sie ursprünglich als Distanzmaß für Punktmengen entwickelt wurde. Wie Abbildung 3.1 a) darstellt, beschreibt die Hausdorff-Distanz d_{Haus} die maximale Distanz aller gefundenen gerichteten minimalen Distanzen $\vec{d}_{Haus}$, weil $\vec{d}_{Haus}(L_A, L_B) \neq \vec{d}_{Haus}(L_B, L_A)$ ist:

$$d_{Haus}(L_A, L_B) = \max\{\vec{d}_{Haus}(L_A, L_B), \vec{d}_{Haus}(L_B, L_A)\} \text{ mit} \tag{3.1}$$

$$\vec{d}_{Haus}(L_A, L_B) = \max_{l_A \in L_A} \{ \min_{l_B \in L_B} \{d(l_A, l_B)\}\}, \tag{3.2}$$

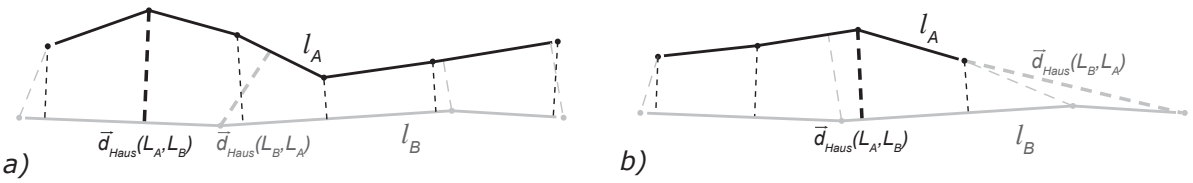


Abbildung 3.1: Hausdorff-Distanz $d_{Haus}(L_A, L_B)$ zwischen a) Linien etwa gleicher Länge und b) Linien unterschiedlicher Länge. Die dünn gestrichelten Linien repräsentieren die minimalen Abstände in den Stützpunkten zwischen beiden Linien, wohingegen die dick gestrichelten Linien die Maximalwerte widerspiegeln. Daraus leitet sich in a) $d_{Haus}(L_A, L_B) = \vec{d}_{Haus}(L_A, L_B)$ und in b) $d_{Haus}(L_A, L_B) = \vec{d}_{Haus}(L_B, L_A)$ ab.

wobei L_A und L_B die Punktmengen der Linien und $d(l_A, l_B)$ eine Abstandsfunktion zwischen zwei Punkten beschreiben. Stehen sich dagegen, wie in Abbildung 3.1 b), zwei Linien mit stark unterschiedlichen Längen gegenüber, sollte die gerichtete Hausdorff-Distanz von der kürzeren zur längeren Linie verwendet werden, da das Ergebnis sonst falsch interpretiert wird. Werden hier die Unterschiede in den Linienlängen nicht berücksichtigt, entspricht die Hausdorff-Distanz d_{Haus} für Beispiel b) $\vec{d}_{Haus}(L_B, L_A) = 332{,}3\,\text{m}$ und wäre fast 150% länger als $\vec{d}_{Haus}(L_A, L_B) = 132{,}4\,\text{m}$.

Für den Vergleich von sinusartigen Linien oder Kurven, wie z.B. Küstenlinien, eignet sich die Fréchet-Distanz (Fréchet, 1906). Sie berücksichtigt im Gegensatz zur Hausdorff-Distanz die Position und die Ordnung der Punkte entlang der Kurven (Mascret u. a., 2006). Die Fréchet-Distanz kann am besten anhand eines Beispiels erklärt werden: Ein Mann geht mit seinem Hund an der Leine Gassi. Beide Beteiligten bewegen sich auf ihren vorgegebenen Wegen, bei denen sie die Geschwindigkeit selbst bestimmen können, aber niemals umkehren dürfen. Die Fréchet-Distanz beschreibt hierbei die kürzeste mögliche Hundeleine. Für die Berechnung der Fréchet-Distanz können sogenannte Freiraumdiagramme genutzt werden. Eine ausführlichere Beschreibung findet sich in Alt und Godau (1995).

Die Distanz zweier Polygonobjekte p_A und p_B kann über den Abstand der Objektschwerpunkte d_S oder den geringsten Abstand zwischen den Objekträndern d_R bestimmt werden. Bei konkaven Polygonen besteht die Gefahr, dass sich der Schwerpunkt außerhalb des Polygons befindet. In Abhängigkeit von der Objektlage und -ausdehnung ergeben sich, wie Abbildung 3.2 a) zeigt, Distanzen mit deutlichem Unterschied.

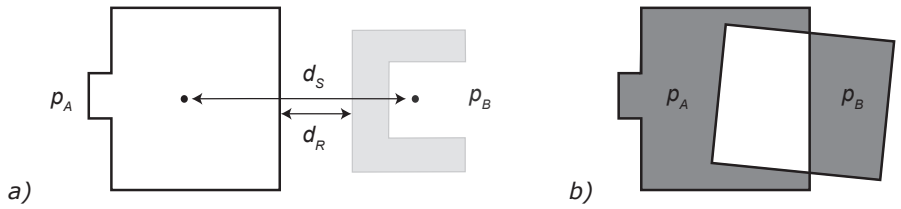


Abbildung 3.2: a) Distanzmaße für Polygone: Distanz zwischen den Objektschwerpunkten $d_S(p_A, p_B)$ im Vergleich zur Distanz zwischen den Polygonrändern $d_R(p_A, p_B)$. b) Symmetrische Differenz $p_A \Delta p_B = (p_A \backslash p_B) \cup (p_B \backslash p_A)$.

Überdeckung

Die Distanz d_R ist 0, wenn sich die Polygone wie in Abbildung 3.2 b) überlagern. In diesem Fall müssen weitere Kriterien untersucht werden, um auf eine Ähnlichkeitsbeziehung zu schließen. Es gibt verschiedene Maße, die die Größen und den Grad der Überdeckung beider Polygone in einem Wert zwischen 0 und 1 wiedergeben. Die symmetrische Differenz $p_A \Delta p_B = (p_A \backslash p_B) \cup (p_B \backslash p_A)$ ist umso geringer, je größer die Überlagerung der Polygone und je ähnlicher die Objektgrößen sind. Im gleichen Fall ist der Wert Intersection over Union $\mathsf{IuO} = p_A \cap p_B / p_A \cup p_B$ umso größer. Vorwiegend wird dieses Maß im Teilbereich Deep Learning des Maschinellen Lernens angewendet.

Größe

Die Größe einer Linie wird als Länge l_L, d.h. als Distanz zwischen den Endpunkten beschrieben. Bei Polygonen beschreiben der Flächeninhalt F und der Umfang U die Größe des Objekts.

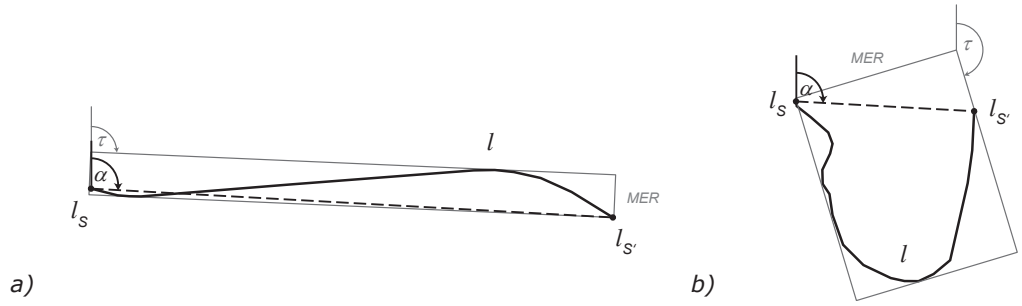


Abbildung 3.3: Vereinfachte Bestimmung der Ausrichtung von langgestreckten und gekrümmten Linienobjekten. Der Richtungswinkel α beschreibt die Orientierung der geradlinigen, gestrichelten Verbindung zwischen Anfangs- und Endpunkt einer Linie und τ beschreibt die Orientierung des mit grau gekennzeichneten, minimal umschließenden Rechtecks.

Ausrichtung

Die Definition und Bestimmung der Ausrichtung bzw. Orientierung eines Objekts sind hingegen nicht eindeutig. Für Linienobjekte kann, wie Abbildung 3.3 zeigt, vereinfacht der Richtungswinkel α der geradlinigen Verbindung beider Endpunkte l_S und $l_{S'}$ bestimmt werden. Diese Vereinfachung führt allerdings dazu, dass für unterschiedlich gekrümmte Linien gleiche Richtungswinkel α bestimmt werden.

Aus diesem Grund wird die Verwendung des Richtungswinkels τ zur Längsseite des minimal umschließenden Rechtecks minimaler Breite *MER* empfohlen. Während bei geradlinigen, langgestrecken Linien die Winkel α und τ etwa gleich groß sind, ist bei stark gekrümmten Linien ein deutlicher Größenunterschied sichtbar.

Die Ausrichtung von Polygonobjekten kann wie Abbildung 3.4 a) verdeutlicht, ebenfalls aus dem minimal umschließenden Rechteck abgeleitet werden.

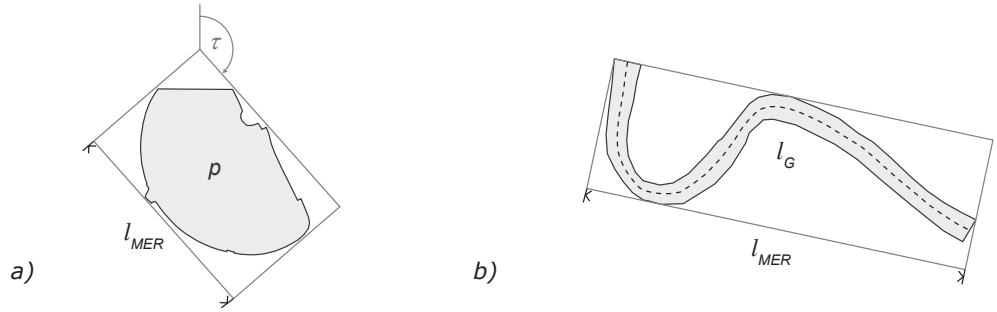


Abbildung 3.4: Bestimmung der Ausrichtung und Länge von Polygonobjekten. a) Die Objektausrichtung kann durch den Richtungswinkel τ zur Längsseite des MERs beschrieben werden. Die Objektlänge wird durch die Ausdehnung der Längsseite l_{MER} angenähert. b) Für mäandernde Objekte ist die Länge des Hauptskeletts l_G zuverlässiger.

Länge

Die Länge einer Linie berechnet sich aus der Summe der Längen der Einzelsegmente zu einer Gesamtlänge. Die Definition und Bestimmung einer Polygonlänge ist nicht trivial und wird im Rahmen dieser Arbeit als Ausdehnung der Längsseite des minimal umschließenden Rechtecks l_{MER} angesetzt, um den Berechnungsaufwand gering zu halten. Bei mäandernden Objekten, wie dem Fluss in Abbildung 3.4 b), ist die Längsseite des minimal umschließenden Rechtecks, aber deutlich kürzer als die tatsächliche Länge. Die Länge der Hauptskelettlinie l_G, die mit Hilfe eines Skelettierungsalgorithmus gewonnen werden kann, nähert die tatsächliche Objektlänge in diesem Fall besser an als l_{MER}.

Form

Die tatsächliche Form eines Objekts kann mit Hilfe von verschiedenen Merkmalen beschrieben werden. Für Polygone können z.B. die in Abbildung 3.5 dargestellten Formparameter berechnet werden:

– die Kompaktheit als Verhältnis des Umfangs zur Fläche eines Quadrates $C_S = \frac{U^2}{4^2 \cdot F}$ oder eines Kreises $C_C = \frac{4\pi \cdot F}{U^2}$, mit dem Flächeninhalt und dem Umfang des Polygons,

– die Langgestrecktheit $El = \frac{l_{MER}}{b_{MER}}$ als Verhältnis der beiden Hauptachsen zueinander, mit der Länge l_{MER} und der Breite b_{MER} des minimal umschließenden Rechtecks,

– die Rechtwinkligkeit $Re = 1 - \frac{\sum_{i=1}^{x} \left| |\Delta\beta_i| - \frac{\pi}{2} \right|}{x \cdot \frac{\pi}{2}}$ als Maß der Winkeländerungen entlang der Objektkontur, mit der Anzahl der Polygonpunkte x und dem Winkel β.

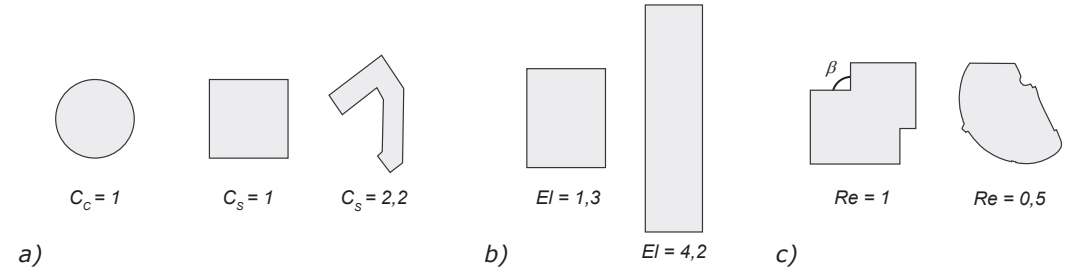


Abbildung 3.5: Geometrische Formmerkmale für Polygonobjekte: a) Kompaktheitsmaß bezogen auf einen Kreis C_C bzw. auf ein Quadrat C_S, b) Langgestrecktheit El und c) Rechtwinkligkeit Re.

Neben diesen Maßen können weitere Hüllen, wie z.B. das achsparallele Rechteck oder die konvexe Hülle wichtige Entscheidungshilfen für eine Objektzuordnung sein. Sind die vorgestellten, eher groben Merkmale für eine hinreichende Zuordnung nicht ausreichend, kann die Kontur jedes Objekts mittels Fourier-Deskriptoren exakt beschrieben werden. Der Vorteil ist, dass die Fourier-Deskriptoren invariant bzgl. Rotation, Translation, Skalierung und gewähltem Anfangspunkt sind, so dass sich Unterschiede in der Position, Größe und Ausrichtung nicht gravierend auswirken.

3.1.2 Topologische Ähnlichkeit

Für die topologische Ähnlichkeit ist nicht die Ausrichtung bzw. die Form der einzelnen Objekte entscheidend, sondern die Lage im Raum und die daraus resultierenden Beziehungen zu benachbarten Objekten. Dafür können entweder Adjazenz- bzw. Inzidenzbeziehungen oder räumliche Relationen zwischen Objekten untersucht und miteinander verglichen werden. Diese Beziehungen sind an keine metrischen oder formgebenden Eigenschaften gebunden und bleiben daher auch bei geometrischen Abweichungen erhalten, die zwischen unterschiedlichen Datensätzen auftreten können.

Knotengrad

Bei der Untersuchung von Linienobjekten hinsichtlich ihrer topologischen Ähnlichkeit werden beispielsweise für die Endpunkte der Linien die Knotengrade überprüft, ungeachtet der räumlichen Distanz der Linien zueinander, deren Ausrichtung oder Form. Der Knotengrad $deg(l_S)$ gibt die Anzahl der inzidenten (hineinfallenden) Kanten an. Im Beispiel in Abbildung 3.6 stimmen für zwei Linien die Knotengrade der Anfangspunkte mit $deg(l_S) = 3$ und die der Endpunkte mit $deg(l_{S'}) = 4$ überein, obwohl die Formen voneinander abweichen. Dieses Maß der topologischen Ähnlichkeit kann genutzt werden, um die Menge der möglichen Matching-Kandidaten einzugrenzen.

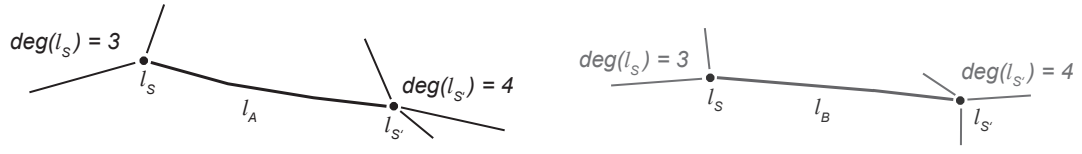


Abbildung 3.6: Der Knotengrad $deg(l_S)$ gibt als topologisches Ähnlichkeitsmaß die Anzahl der inzidenten Kanten wieder.

Links-Rechts-Relation

Schulze u. a. (2014) entwickelten ein Verfahren, das für die Zuordnung von Linienobjekten zusätzlich die Semantik der angrenzenden Objektflächen in Form von Objektklassenzugehörigkeiten berücksichtigt. In Abbildung 3.7 repräsentieren die unterschiedlichen Flächentexturen verschiedene Objektklassen. Für Linie l_A werden als Ergebnis einer geometrischen Analyse bezüglich Entfernung und Ausrichtung die Linien l_{B_1} und l_{B_2} als potentielle

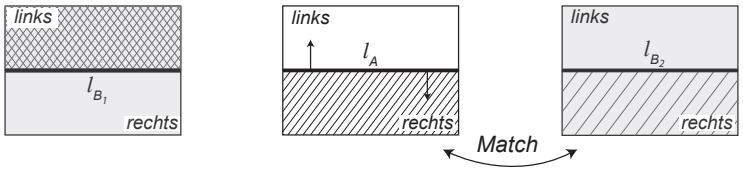


Abbildung 3.7: Die Rechts-Links-Relation für die Zuordnung von Linien unter Berücksichtigung der semantischen Informationen der angrenzenden Flächenobjekte.

Matching-Kandidaten identifiziert. Im nächsten Schritt werden jeweils die linken und rechten Objektnachbarn verglichen. Hierbei muss die Ausrichtung der Linien beachtet werden. Wenn Linien entgegengesetzt verlaufen, müssen die Nachbarflächen getauscht werden. Im konkreten Beispiel haben die Linien die gleiche Ausrichtung. Die Relation $l_A \rightarrow l_{B_2}$ wird als beste Zuordnung ausgewählt, da die semantische Differenz zwischen den linken und rechten Objektflächen insgesamt kleiner ist als bei $l_A \rightarrow l_{B_1}$. Während die Objektklassen der linken Nachbarn im erstgenannten Fall identisch sind, unterscheiden sich die rechten nur geringfügig. Im Vergleich dazu sind die Unterschiede auf beiden Seiten der anderen Relation größer. Damit dieser Rückschluss möglich ist, muss vorab ein semantisches Ähnlichkeitsmaß, z.B. in Form einer Distanz, definiert werden, um die Bedeutung der Objektklassen miteinander zu vergleichen. In diesem Beispiel sind sich die beiden Linienschraffuren mit unterschiedlichen Abständen ähnlicher als eine Linien- zu einer Kreuzschraffur. Abschnitt 3.1.3 stellt verschiedene semantische Distanzmaße vor.

Topologische Relationen

Eine weitere Möglichkeit ist, sogenannte topologische Relationen zwischen Objekten zu bestimmen, wie sie Egenhofer in zahlreichen Publikationen für die zweidimensionalen Geometrietypen Polygon, Linie und Punkt definiert hat (Egenhofer, 1989; Egenhofer und Herring, 1991). Die Beziehung zwischen zwei einfachen Polygonobjekten kann in acht topologische Relationen unterteilt werden, die schematisch in Abbildung 3.8 dargestellt sind.

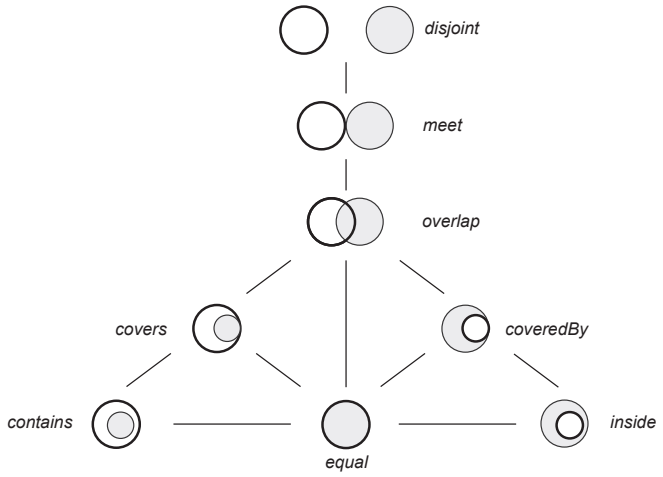


Abbildung 3.8: Konzeptionelles Nachbarschaftmodell der topologischen Relationen für einfache flächenhafte Objekte nach Bruns und Egenhofer (1996).

Zwei Objekte werden als disjoint bezeichnet, wenn sie sich weder überdecken noch berühren. Im Gegensatz dazu werden zwei identische Objekte equal benannt. Berühren sich die Ränder der Objekte, muss geprüft werden, ob außerdem eine Überlagerung vorliegt oder nicht. Die Relationen covers und coveredBy bestehen, wenn neben der Berührung auch eine Überlagerung gegeben ist und meet, wenn dies nicht der Fall ist. Die Überlagerung zweier Objekte ohne eine Berührung im Rand wird als overlap bezeichnet. Befindet sich ein Objekt vollständig im anderen Objekt, ist, in Abhängigkeit des zu betrachtenden Objekts, die Relation contains bzw. inside zu wählen.

Diese topologischen Relationen können als Basis für eine topologische Ähnlichkeitsbewertung gesehen werden. Liegen zwei ähnliche räumliche Szenen vor, kann aus den topologischen Unterschieden ein Ähnlichkeitsmaß für

die Szene abgeleitet werden, indem die unterschiedlich vorkommenden Relationen gezählt werden. Allerdings kann eine geringe Änderung in der Szene eine enorme Auswirkung auf die Art der vorkommenden Relationen bewirken.

Gradual Change

Eine andere Möglichkeit ist, eine Szene stufenweise in die andere Szene zu überführen. Dieses Konzept wird als Gradual Change bezeichnet. Dazu werden die Schritte gezählt, die für die Transformation nötig sind. Bruns und Egenhofer (1996) erstellten dafür ein konzeptionelles Nachbarschaftmodell (Abb. 3.8), woraus die Ähnlichkeit der topologischen Relationen ablesbar ist. Beispielsweise ist die Änderung der Relation meet in overlap in einem Transformationsschritt möglich, da sich die Relationen laut Nachbarschaftsmodell ähnlicher sind als die Relationen meet und contain. Das bedeutet, dass sich räumliche Szenarien umso ähnlicher sind, je weniger Änderungen beim Wechsel der Relationen ineinander nötig sind.

3.1.3 Semantische Ähnlichkeit

Wenn zwei Objekte einander geometrisch zugeordnet werden, wie die beiden Gewässer in Abbildung 2.4 a), liegt die Vermutung nahe, dass eine semantische Ähnlichkeit im Objektklassennamen oder einem der Attribute erkennbar ist. Diese Korrespondenz ist nicht immer festzustellen, gerade weil die Informationen, die in den Objektklassenbeschreibungen formalisiert werden, stark von der individuellen Betrachtungsweise abhängen. Allein die Einteilung in unterschiedliche Klassen ist von vornherein verlustbehaftet, weil nicht alle Details formalisiert werden können. Hinzu kommt, dass ein Sachverhalt in einem Schema als Objektattribut und in einem anderen Schema als eine eigene Objektklasse beschrieben sein kann. Identische Bezeichnungen für Objektklassen, wie Bank aus dem Einführungsbeispiel, können unterschiedliche Bedeutungen besitzen. Der umgekehrte Fall ist ebenfalls möglich, wenn Bezeichnungen voneinander abweichen, aber semantisch übereinstimmen, z.B. Wassertiefe gegenüber Wasserhöhe. Die feinen Unterschiede können größtenteils nur mit Hilfe von Experten gefunden werden, die die Terminologie der erfassten Datenbestände kennen und in der Lage sind, Kontextinformationen im manuellen Zuordnungsprozess zu berücksichtigen. Allerdings ist diese Arbeit sehr zeitaufwendig und kostenintensiv und kann nur für eine begrenzte Anzahl von Datensätzen durchgeführt werden.

Semantische Distanz

Die semantische Ähnlichkeit kann als Distanzmaß automatisch aus zur Verfügung stehenden, numerischen Attributen abgeleitet werden, indem die absolute Differenz der Attributwerte berechnet wird. Je geringer der Differenzwert, desto größer ist die Ähnlichkeit der Objekte. In der vorliegenden Arbeit werden für die Bildung von komplexen Objektrelationen zuerst die Objekte innerhalb eines Datensatzes zusammengefasst, die sich semantisch am ähnlichsten sind. Dafür wird die Zugehörigkeit zu den Objektklassen analysiert. Beispielsweise werden beim ATKIS-Datensatz die Objektklassen in Objektgruppen wie z.B. Siedlung, Verkehr und Vegetation zusammengefasst und zusätzlich mit Zahlencodes 2000, 3000 und 4000 annotiert. Durch die Differenzbildung der Zahlencodes wird die Ähnlichkeit zweier Objektklassen widergespiegelt. Objekte der Klassen 2111 und 2221 sind sich ähnlicher als die der Klassen 2111 und 4101, da die Differenz kleiner ist. Ein ausführliches Beispiel wird in Abschnitt 4.1.1 präsentiert.

Für alphanumerische Werte kann bei Wörtern gleicher Länge die Hamming-Distanz und bei unterschiedlichen Längen die Levenshtein-Distanz, auch bekannt als Editierdistanz, verwendet werden. Alle Distanzen bestimmen die minimale Anzahl der Operationen, z.B. Einfügen, Löschen und Ersetzen, die notwendig sind, um eine Zeichenfolge in eine andere Zeichenfolge zu transformieren.

Semantische Ähnlichkeit durch Häufigkeiten

Im Rahmen dieser Arbeit wird die semantische Ähnlichkeit zwischen Objektklassen unterschiedlicher Datensätze durch Häufigkeitswerte ausgedrückt, die aus den Objektzuordnungen im Data-Matching-Prozess abgeleitet werden. Das heißt, je mehr Zuordnungen zwischen Objekten zweier Klassen existieren, desto größer ist ihre semantische Ähnlichkeit. Allerdings sagt die Anzahl der Objekte wenig darüber aus, ob nur eins von hundert Objekten oder nur das eine vorhandene Objekt der Objektklasse Teil der Zuordnung ist. Aus diesem Grund werden neben den absoluten Häufigkeiten auch datensatzbezogene, prozentuale Flächenanteile bestimmt und für das Schema-Matching als Eingabewerte verwendet. In Abschnitt 6.3.1 wird das Verfahren ausführlich beschrieben.

3.2 Relationstypen

Der Begriff Relation wurde bereits in Abschnitt 2.1.3 als Beziehung zwischen mehreren Elementen eingeführt. Außerdem wurde die Einteilung in einfache (1:1) und komplexe (1:n, n:1, n:m) Relationen vorgestellt, die von der Anzahl der beteiligten Elemente abhängig ist. Für die Entwicklung eines Ähnlichkeitsmaßes ist es wichtig, ob Relationen zwischen Elementen eines Datensatzes als sogenannte interne Relationen oder datensatzübergreifend als externe Relationen untersucht werden. Die Nachbarschaftsbeziehung in Abbildung 2.2 beschreibt zwischen den Objekten City Park und Square eine interne, und die Ähnlichkeitsbeziehung zwischen Park und City Park eine externe Relation, für deren Analyse sich unterschiedliche Ähnlichkeitsmaße eignen. Für die vorliegende Arbeit ist die Menge der externen Objektrelationen R_o von besonderem Interesse, da daraus die Menge der Schemarelationen R_s abgeleitet wird.

In den folgenden Abschnitten werden weitere Unterteilungen der Relationen vorgenommen und anhand von Beispielen vorgestellt. Auf Objektebene werden zusätzlich geometrisch reine und geometrisch gemischte Relationen unterschieden, während auf der Schemaebene nur eine Unterscheidung zwischen homogenen und heterogenen Relationen vollzogen wird.

3.2.1 Relationen auf Objektebene

Den Idealfall bei der Objektzuordnung stellen 1:1-Objektrelationen dar, bei denen genau ein Objekt eines Datensatzes genau einem einzelnen Objekt eines anderen Datensatzes zugeordnet wird. Die Überprüfung auf Richtigkeit und die eindeutige Interpretation des Ergebnisses ist somit einfach möglich. Bei der Untersuchung von Objekten unterschiedlicher Geometriedimensionen (Punkt pk, Linie l, Polygon p) sind bereits sechs verschiedene Geometrietypkombinationen denkbar, die die Entscheidung über eine 1:1-Relation erschweren. Aus diesem Grund wird eine Unterteilung in geometrisch reine und geometrisch gemischte Relationen vorgestellt. Als eine geometrisch reine 1:1-Relation wird eine Beziehung zwischen Objekten vom gleichen Geometrietyp bezeichnet, wie $pk \rightarrow pk$, $l \rightarrow l$ oder $p \rightarrow p$. Im Gegensatz dazu definiert eine geometrisch gemischte 1:1-Relation eine Beziehung zwischen Objekten unterschiedlicher Geometrietypen, wie z.B. $pk \rightarrow l$, $pk \rightarrow p$ oder $l \rightarrow p$. In Abbildung 3.9 a) und b) sind für die Geometriedimensionen Linie und Polygon Beispiele sowohl für geometrisch reine als auch für gemischte Relationen dargestellt.

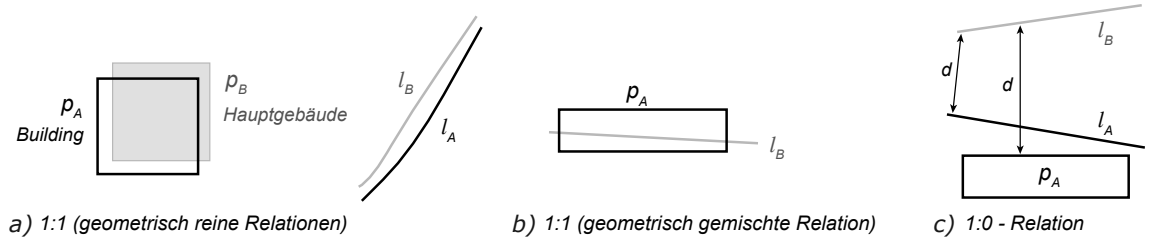


Abbildung 3.9: Einfache Objektrelationen zwischen Objekten zweier Datensätze A und B: a) geometrisch reine 1:1-Relationen zwischen Objekten gleicher Geometriedimension, b) geometrisch gemischte 1:1-Relation zwischen Objekten unterschiedlicher Geometriedimensionen, c) 1:0-Relation aufgrund zu großer Mindestabstände.

Neben der 1:1-Relation stellt die 1:0- bzw. die 0:1-Relation eine weitere einfache Relation dar. Hierbei kann einem Objekt kein anderes Objekt eindeutig zugewiesen werden. Ein möglicher Grund dafür ist, dass sich kein anderes Objekt an derselben Position befindet, z.B. weil die Datensätze von unterschiedlicher Aktualität sind und Objekte zum Zeitpunkt der Erfassung noch nicht bzw. nicht mehr vorhanden waren. Die Entscheidung für eine 1:0-Relation kann trotz potentieller Matching-Kandidaten getroffen werden, wenn das definierte Ähnlichkeitsmaß nicht ausreichend ist, z.B. wenn die Distanz zwischen den Objekten, wie in Abbildung 3.9 c) dargestellt, zu groß ist.

Komplexe Objektrelationen treten vorrangig beim Vergleich von Datensätzen unterschiedlichen Maßstabs auf, weil die Datenerfassung für verschiedene Maßstäbe oft unterschiedliche Objektmindestgrößen verlangt und der Detailgrad voneinander abweicht. Ebenso führt die anwendungsbezogene Wahrnehmung einer Situation in verschiedenen Fachgebieten dazu, dass ein Real-Welt-Objekt unterschiedlich modelliert wird. In Abbildung 3.10 a) ist eine 1:2-Objektrelation dargestellt, bei der ein Polygon der Objektklasse Building zwei Polygonen der Klassen Nebengebäude und Hauptgebäude zugeordnet wird. Offensichtlich ist, dass die Objekte von Datensatz B nicht nur geometrisch, sondern auch semantisch detaillierter modelliert sind, weil hier eine Unterscheidung zwischen Haupt- und Nebengebäuden erfolgte.

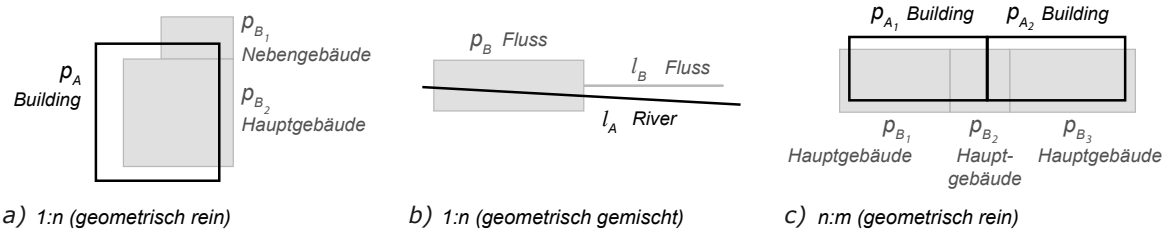


Abbildung 3.10: Komplexe Objektrelationen zwischen Objekten zweier Datensätze A und B: a) geometrisch reine 1:n-Relation, b) geometrisch gemischte 1:n-Relation, c) geometrisch reine n:m-Relation

Im Gegensatz zur geometrisch reinen Relation wird in Abbildung 3.10 b) eine geometrisch gemischte 1:n-Relation zwischen Flussobjekten gezeigt. Bei der Objektbildung erfolgt in Datensatz B ein Geometrietypwechsel aufgrund der Modellierungsbedingung, die besagt, dass Flüsse bis zu einer Breite von 12 m als Linie und darüber hinaus als Polygon darzustellen sind. Weitere geometrisch gemischte 1:n-Relationen sind denkbar, die allein aus der Betrachtung unterschiedlicher Geometriedimensionen abgeleitet werden können, z.B. $p \rightarrow \{l, l\}$, $p \rightarrow \{l, p\}$, $l \rightarrow \{p, p\}$ oder $l \rightarrow \{l, p\}$.

Durch die Beteiligung von mehreren Objekten auf beiden Seiten entstehen gleichermaßen geometrisch reine und geometrisch gemischte n:m-Relationen. Abbildung 3.10 c) zeigt ein Beispiel für eine geometrisch reine n:m-Relation. Hierbei wird ein Gebäudekomplex in Datensatz A in zwei Objekte untergliedert, während in Datensatz B drei einzelne Objekte vorhanden sind.

3.2.2 Relationen auf Schemaebene

Analog zu den Objektrelationen werden bei 1:1-Schemarelationen jeweils zwei Schemaelemente, d.h. zwei Objektklassen, einander zugeordnet. Wird in Abbildung 3.9 a) davon ausgegangen, dass die Objekte p_A und p_B die einzigen Vertreter ihrer Klassen sind und einander zugeordnet werden, besteht zwischen den Objektklassen Building und Hauptgebäude folglich eine Äquivalenzbeziehung.

In Abbildung 3.10 c) wird eine komplexe Objektrelation erkannt, die im Rahmen dieser Arbeit als *homogene Schemarelation* bezeichnet wird, weil insgesamt nur zwei unterschiedliche Objektklassen beteiligt sind. Am Beispiel bedeutet dies, dass aus der 2:3-Objektrelation mit den Objektklassen {Building, Building} → {Hauptgebäude, Hauptgebäude, Hauptgebäude} ein Rückschluss auf die homogene 1:1-Schemarelation Building→ Hauptgebäude möglich ist.

Im Gegensatz dazu wird eine Schemarelation als *heterogen* bezeichnet, wenn mehr als zwei unterschiedliche Objektklassen an der Relation beteiligt sind. Im Beispiel in Abbildung 3.10 a) wird aus der 1:2-Objektrelation anhand der Objektklassenbeteiligung die heterogene 1:2-Schemarelation Building → {Hauptgebäude, Nebengebäude} abgeleitet. Daraus folgt, dass heterogene Schemarelationen nicht nur semantisch ähnliche Objektklassen zusammenfassen, sondern gleichzeitig auch komplexe Objektrelationen widerspiegeln.

Theoretisch sind auf Schemaebene ebenfalls 1:0- bzw. 0:1-Relationen möglich. Die Nichtzuordnung einer Objektklasse ist gerechtfertigt, wenn keinerlei Objekte im Data-Matching-Prozess zugeordnet werden können oder der Anteil der zugeordneten Objekte bezogen auf die Gesamtzahl zu gering ist. Die im Rahmen dieser Arbeit entwickelten Verfahren bestimmen jedoch für jede Objektklasse einen Zuordnungspartner.

3.3 Graphentheorie

In diesem Abschnitt werden graphentheoretische Grundlagen präsentiert, weil die Ergebnisse der Objektzuordnung in Häufigkeitsmatrizen zusammengefasst werden und als Graphen interpretierbar sind. Die Zuordnung von Objektklassen wird als Graphzuordnungsproblem betrachtet, dessen Lösung mit existierenden Graphalgorithmen möglich ist. Dazu werden neben den wesentlichen Begriffen auch die im Rahmen der Arbeit verwendeten Algorithmen vorgestellt. Für eine ausführlichere Einführung und weiterführende Informationen zum Thema der Graphentheorie wird auf folgende Grundlagenliteratur verwiesen: Diestel (2000) sowie Ottmann und Widmayer (2002).

3.3.1 Graph-Definitionen

Ein *Graph* $G(V, E)$ ist definiert durch eine Menge *Knoten* V und eine Menge *Kanten* E. Eine Kante $e \in E$ besitzt zwei verschiedene Knoten $u, v \in V$, die als Endpunkte einer Kante $e = \{u, v\}$ bezeichnet werden. Wird für jede Kante ein Anfangsknoten u und ein Endknoten v festgelegt, wird aus dem ungerichteten Graphen ein gerichteter Graph. Dies ist von besonderer Bedeutung, wenn z.B. ein Flussnetzwerk/Gewässernetz mit Beschränkungen in der Fließrichtung als Graph dargestellt werden soll. Durch die Einführung von Gewichten bzw. Kapazitäten $w(e \in E)$ zu den Kanten wird aus einem ungewichteten Graphen ein gewichteter Graph. Im weiteren Verlauf dieser Arbeit wird immer von nicht-negativen Gewichten $w(e) \geq 0$ ausgegangen. In Abbildung 3.11 sind drei Beispielgraphen dargestellt, deren Knoten durch Punkte und deren Kanten durch Linien, die die Knoten miteinander verbinden, repräsentiert werden. Die Darstellung der Kanten durch gerade Linien ist nicht zwingend erforderlich. Pfeile an den Kanten markieren deren Richtung, Zahlen stehen für Kantengewichte.

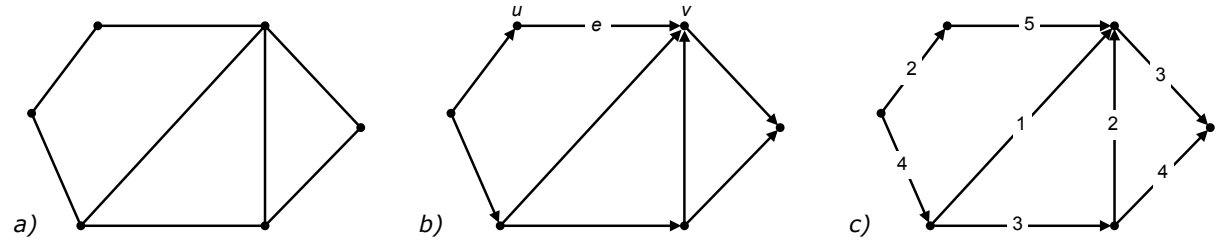


Abbildung 3.11: Verschiedene Graphtypen: a) ungerichteter und ungewichteter Graph, b) gerichteter Graph und c) gerichteter und gewichteter Graph.

Für die Zuordnung von Objektklassen zweier Datensätze wird ein spezieller Graph, ein sogenannter bipartiter oder paarer Graph benötigt. Ein Graph heißt bipartit, wenn sich die Knotenmenge V in zwei disjunkte Teilmengen A und B ($V = A \cup B$ und $A \cap B = \emptyset$), z.B. in die Objektklassen der Datensätze, zerlegen lässt, die Kanten E ungerichtet sind sowie nur zwischen den Teilmengen A und B und nicht innerhalb einer Teilmenge verlaufen. Besteht zwischen jedem Knoten $a_i \in A$ mit $i = 1, ..., n$ und jedem Knoten $b_j \in B$ mit $j = 1, ..., m$ eine Kante, dann ist der Graph vollständig bipartit (Abb. 3.12). Die Häufigkeit einer Klassenkombination, die aus der Objektzuordnung abgeleitet wird, kann als Gewicht w_{ij} der Kante zwischen den Klassen im Graphen betrachtet werden.

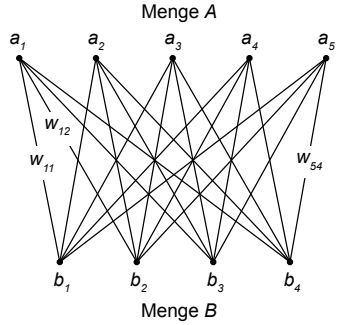


Abbildung 3.12: Vollständig bipartiter Graph mit Gewichten w_{ij} mit $i = 1, ..., n$ und $j = 1, ..., m$.

Die Verwendung von bipartiten anstatt allgemeinen Graphen bedeutet für viele Fragestellungen eine Reduktion der Komplexität bzw. der Berechnungen, da ein vollständig verbundener Graph mit vielen Knoten in der Praxis schwer lösbar sein kann. Das Ziel, Zuordnungen zwischen den einzelnen Objektklassen herzustellen, kann mit Methoden des Graph-Matchings oder des Graph-Cuts erreicht werden, die in den folgenden Abschnitten näher beschrieben werden.

3.3.2 Graph-Matching

In einem bipartiten Graph, beispielhaft in Abbildung 3.13 a) gezeigt, wird eine Zuordnung zwischen zwei Knoten der Teilmengen A und B als Paarung M (engl. Matching) bezeichnet und durch starke Kanten hervorgehoben. Eine Paarung stellt eine Teilmenge von Kanten $M \subseteq E$ dar, mit der Bedingung, dass jeder Knoten mit maximal einer Kante aus M inzident sein darf. Anders ausgedrückt, zwei Kanten dürfen nicht den gleichen Endpunkt haben. Ein Matching ist nicht erweiterbar (engl. Maximal Matching), wenn es keine Kante $e \in E \backslash M$ gibt, die

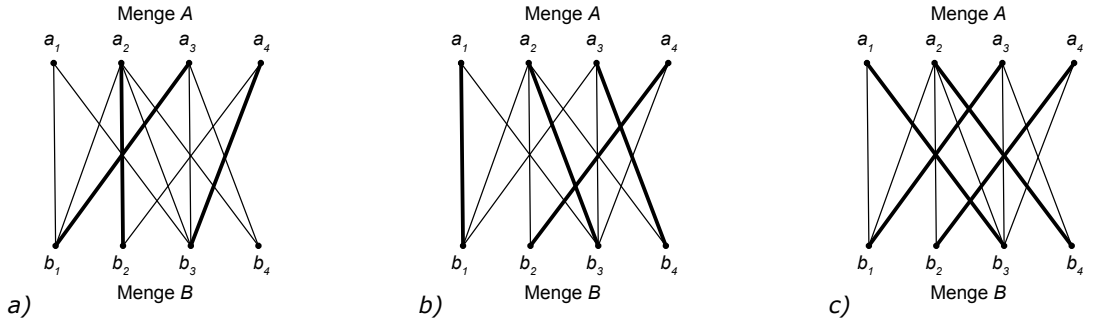


Abbildung 3.13: Für einen bipartiten Graph werden verschiedene Matchingtypen dargestellt: a) Nicht erweiterbares Matching mit $M = 3$, b) und c) maximale und perfekte Matchings mit $M = 4$. Die starken Kanten kennzeichnen die Zuordnung.

mit den Kanten aus M zusammen ein größeres Matching bilden kann. Abbildung 3.13 a) zeigt solch ein nicht erweiterbares Matching mit $M = 3$, da zwischen den freien Knoten a_1 und b_4 keine Kante existiert, die das Matching ergänzen würde. Ein maximales Matching (engl. Maximum Matching) liegt vor, wenn es kein weiteres Matching M' gibt, das mehr Kanten $|M'| > |M|$ beinhaltet. Eine vollständige Paarung bzw. ein perfektes Matching besteht, wenn alle Knoten beider Teilmengen am Matching beteiligt sind, d.h. jeder Knoten in einer Kante von M vorkommt. In Abbildung 3.13 b) und c) werden zwei maximale Matchings der Kardinalität 4 gezeigt, die gleichzeitig auch perfekte Matchings darstellen.

Wenn sich die beiden Teilmengen A und B in der Anzahl der Elemente stark unterscheiden, kann das maximale Matching jedoch nur $\min(|A|, |B|)$ groß sein. Im Allgemeinen ist das maximale Matching nicht eindeutig. So sind z.B. in einem ungewichteten, vollständig bipartiten Graph mit $\min(|A|, |B|) = 4$ theoretisch bereits $4! = 24$ maximale Matchings möglich.

Die derzeit effizienteste Bestimmung des maximalen Matchings in ungewichteten bipartiten Graphen ist mit Hilfe des Algorithmus von Hopcroft und Karp (1973) möglich. Hierbei werden anhand eines bestehenden Matchings M simultan mehrere verbessernde bzw. augmentierende Pfade gesucht. In Abbildung 3.14 wird die Entstehung solch eines Pfads gezeigt. Der Pfad P heißt verbessernd, wenn er in einem freien Knoten, d.h. einem Knoten, der zu keiner Kante in M inzident ist, beginnt und endet und dazwischen alternierend Kanten aus M und aus $E \backslash M$ enthält. Für das Beispiel ergibt sich folgender verbessernder Pfad: $(a_1, b_1, a_3, b_3, a_4, b_2, a_2, b_4)$. Aus diesem Pfad $P \subseteq E$ wird ein neues Matching M' abgeleitet, indem die zu M gehörenden Kanten des Pfades aus dem Matching entfernt und die nicht zu M gehörenden stattdessen hinzugefügt werden (siehe Abb. 3.14 c). Die Laufzeit ist $\mathcal{O}(\sqrt{V}E) = \mathcal{O}(n^{\frac{5}{2}})$.

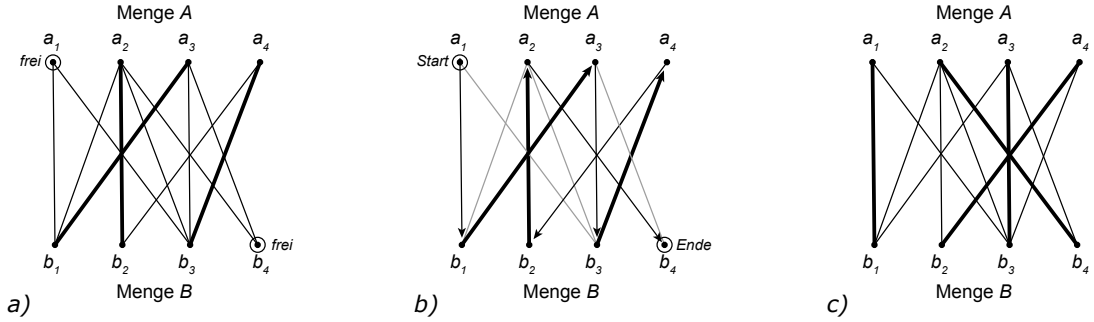


Abbildung 3.14: Bestimmung eines maximalen Matchings mittels augmentierender Pfade: Ausgangspunkt ist das in a) präsentierte, nicht erweiterbare Matching $M = 3$, anschließend wird in b) ein verbessernder, alternierender Pfad $(a_1, b_1, a_3, b_3, a_4, b_2, a_2, b_4)$ bestimmt und in c) daraus ein verbessertes, maximales und perfektes Matching $M = 4$ abgeleitet.

Das in der vorliegenden Arbeit untersuchte Zuordnungsproblem beschreibt dagegen einen gewichteten bipartiten Graphen. Somit ist ein gewichtetes maximales Matching ein Matching mit maximalem Gewicht:

$$w(M) := \sum_{e \in M} w(e) = \max . \tag{3.3}$$

Dieses Problem kann mit Hilfe der Ungarischen Methode als ein Spezialfall der linearen Optimierung gelöst werden und ein optimales Ergebnis liefern. Die Methode wird auch als Kuhn-Munkres-Algorithmus bezeichnet, da von Kuhn (1955) entwickelt und anschließend von Munkres (1957) hinsichtlich der Laufzeit von $O(n^4)$ auf $O(n^3)$ verbessert. Der Ansatz basiert ebenfalls auf dem Konzept von augmentierenden Pfaden bzgl. M. Ein Nachteil der Ungarischen Methode ist die Voraussetzung eines bipartiten Graphen mit gleicher Anzahl von Knoten für beide Teilmengen. Wenn sich die Anzahl der Objektklassen unterscheidet, werden sogenannte Dummy-Elemente eingeführt, um eine Angleichung zu erzielen. Bei sehr großen Unterschieden müssen viele Dummy-Elemente eingeführt werden, die wiederum einen hohen Rechenaufwand verursachen können. In Abbildung 3.15 wird für ein Beispiel mit unterschiedlicher Anzahl Knoten für jede Teilmenge das maximale Matching mit maximalem Gewicht präsentiert. In diesem Fall wird der Knoten a_5 dem Dummy-Knoten zugeordnet.

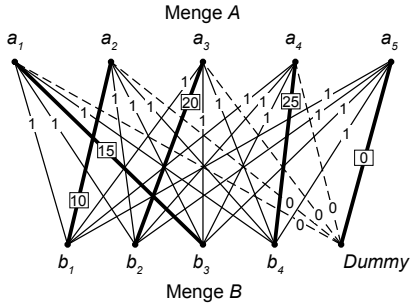


Abbildung 3.15: Beispiel für ein gewichtetes maximales Matching in einem bipartiten Graphen mit den Mengen $A = \{a_1, a_2, a_3, a_4, a_5\}$ und $B = \{b_1, b_2, b_3, b_4\}$. Die starken Kanten kennzeichnen das Ergebnis des maximalen Matchings mit $\sum = 70$. Die gestrichelten Kanten markieren die Kanten zum Dummy-Element mit dem Gewicht 0.

Mit der Methode des Graph-Matchings ist es jedoch nur möglich, 1:1 -Relationen zwischen Elementen der Teilmengen A und B abzuleiten. Um das Problem der reinen 1:1 -Zuordnung dahingehend aufzulösen, dass auch mehrere Elemente zu Clustern zusammengefasst und dann einander zugeordnet werden, können Methoden der Graph-Partitionierung verwendet werden. Laut Goldschmidt und Hochbaum (1988) kann das Cluster-Problem vereinfacht als Partitionierungsproblem interpretiert werden, da beide Probleme identisch sind. Im folgenden Abschnitt wird das Problem formal beschrieben und anhand von Beispielen verdeutlicht.

3.3.3 Graph-Partitionierung / Graph-Cut

Ein einfacher, beliebiger Schnitt (engl. Cut) unterteilt die Knotenmenge V eines Graphen G in zwei nicht leere und disjunkte Teilmengen V_1 und V_2 ($V = V_1 \cup V_2$). Außerdem wird eine Kantenmenge $C \subseteq E$ gebildet, die jede Kante enthält, die ein Element aus V_1 mit einem Element aus V_2 verbindet und damit alle Kanten umfasst, die zwischen den beiden Mengen verlaufen.

Minimaler-2-Schnitt

Für die Bestimmung des minimalen Schnitts (engl. Min-Cut) muss die kleinste Anzahl Kanten C gefunden werden, die zwei Partitionen miteinander verbindet. Abbildung 3.16 a) zeigt ein Beispiel, bei dem mindestens zwei Kanten entfernt werden müssen, um den Graph in genau zwei Teile zu schneiden. An dieser Stelle sind drei minimale Schnitte mit jeweils zwei Kanten ($|C| = 2$) möglich, was deutlich macht, dass der Min-Cut nicht immer eindeutig ist.

In einem gewichteten Graphen müssen dagegen die Kanten mit den geringsten Gewichten zwischen zwei Partitionen geschnitten werden, um die Summe der entfernten Kantengewichte zu minimieren:

$$w(C) := \sum_{e \in C} w(e) = \min . \tag{3.4}$$

In Abbildung 3.16 b) repräsentieren die numerischen Werte an den Kanten die Kantengewichte. In diesem Beispiel sind zwei Lösungen mit gleichem Kantengewicht, nämlich $w(C) = 6$, aber unterschiedlicher Kantenanzahl $|C| = 2$ und $|C| = 3$ möglich. Durch den minimalen Schnitt werden Teilmengen voneinander getrennt, die sich am unähnlichsten sind, ausgedrückt durch verbindende Kanten mit geringem Gewicht.

In einem bipartiten Graphen darf ein Schnitt die ursprünglichen Teilmengen A und B nicht wieder voneinander trennen. Vielmehr muss jeweils eine Partition dieser Teilmengen gebildet werden, d.h. P_A von A und P_B von B.

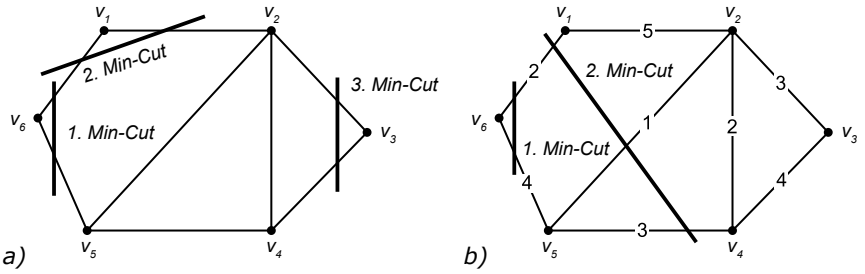


Abbildung 3.16: Minimaler-2-Schnitt in a) ungewichteten und b) gewichteten Graphen. In a) gibt es drei verschiedene minimale Schnitte mit der gleichen Kantenanzahl $|C| = 2$: 1. min-Cut: $V_1 = \{v_6\}$ und $V_2 = \{v_1, v_2, v_3, v_4, v_5\}$, 2. min-Cut: $V_1 = \{v_1\}$ und $V_2 = \{v_2, v_3, v_4, v_5, v_6\}$ und 3. min-Cut: $V_1 = \{v_3\}$ und $V_2 = \{v_1, v_2, v_4, v_5, v_6\}$. In b) sind zwei minimale Schnitte mit $w(C) = 6$ bei $|C| = 2$ und $|C| = 3$ möglich, wobei der 1. min-Cut mit dem ersten Schnitt aus a) identisch ist und der 2. min-Cut andere Kanten schneidet: $V_1 = \{v_5, v_6\}$ und $V_2 = \{v_1, v_2, v_3, v_4\}$.

Wenn die Anzahl der Elemente in den neuen Teilmengen der Partitionen größer als eins ist, entspricht dies einer Zusammenfassung bzw. Clusterbildung von Elementen. In Abbildung 3.17 wird der minimale Schnitt für das bereits eingeführte Beispiel aus Abbildung 3.15, allerdings ohne den Dummy-Knoten, präsentiert. Die Strichpunktlinien mit $w(e) = 1$ werden entfernt und es entstehen die Partitionen: $P_A = \{\{a_1, a_2, a_4, a_5\}, \{a_3\}\}$ und $P_B = \{\{b_1, b_3, b_4\}, \{b_2\}\}$, die folgende 1:1-Zuordnung zwischen den Partitionen: $\{a_1, a_2, a_4, a_5\} \rightarrow \{b_1, b_3, b_4\}$ und $a_3 \rightarrow b_2$ erlauben.

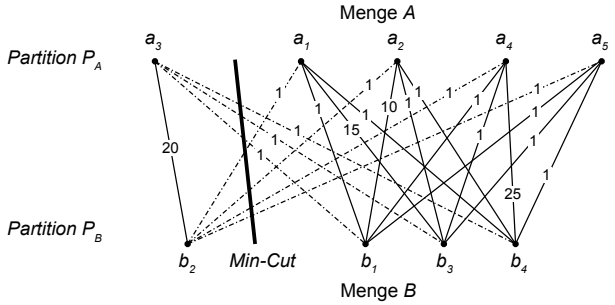


Abbildung 3.17: Minimaler Schnitt für den gewichteten Graphen aus Abb. 3.15. Der Graph wird in die Partitionen $P_A = \{\{a_1, a_2, a_4, a_5\}, \{a_3\}\}$ und $P_B = \{\{b_1, b_3, b_4\}, \{b_2\}\}$ geteilt. Der Schnitt ist durch eine starke Linie und die geschnittenen Kanten mit Strichpunktlinien gekennzeichnet. Der minimale Schnitt ermöglicht folgende 1:1-Zuordnung zwischen den Partitionen: $\{a_1, a_2, a_4, a_5\} \rightarrow \{b_1, b_3, b_4\}$ und $a_3 \rightarrow b_2$. Die Summe der geschnittenen Kanten beträgt $w(C) = 7$.

Für die Partitionierung eines Graphens in zwei Teile existieren eine Reihe von effizienten Algorithmen, die eine polynomielle Laufzeit haben. Traditionell werden die minimalen Schnitte mittels maximalen Fluss-Algorithmen berechnet. Das bedeutet, dass das Min-Cut-Problem auf das Maximale-Fluss-Problem (engl. Max-Flow) transformiert wird. Dafür wird das bekannte Max-Flow-Min-Cut-Theorem von Ford und Fulkerson (1958) verwendet, welches besagt, dass es eine Dualität zwischen den Problemen gibt. In einem Netzwerk, bei dem zwei spezielle Knoten (so: Source und ta: Target) separiert werden sollen, entspricht der maximale Fluss dem Wert des minimalen Schnitts (engl. Min-so-ta-Cut). King u. a. (1994) entwickelten dafür den bekanntesten, schnellsten Algorithmus. Um jedoch den minimalen Schnitt zu finden, ohne vorher spezielle Knoten festzulegen, die getrennt werden sollen, können für einen festgelegten Knoten *so* alle Min-so-ta-Cuts mit $ta \in V \backslash so$ bestimmt werden. Aus den Resultaten kann anschließend der Schnitt mit dem geringsten Gewicht ausgewählt werden. In den neunziger Jahren wurden für das Min-Cut-Problem viele effiziente Algorithmen entwickelt, die entweder deterministischen (Nagamochi und Ibaraki, 1992; Hao und Orlin, 1994; Stoer und Wagner, 1997) oder randomisierten (Karger und Stein, 1993, 1996; Karger, 2000) Algorithmen zuzuordnen sind.

Minimaler-k-Schnitt

Die Zerlegung des Graphen in lediglich zwei Teile ist für die zu untersuchende Aufgabenstellung, sprich die Zuordnung von vielen, sich gegenüberstehenden Objektklassen, nicht ausreichend. Der Graph muss in k beliebige Teile teilbar sein. Ein k-Schnitt (engl. k-Cut) ist definiert durch eine Partition der Knotenmenge $V = \{V_1, V_2, ..., V_k\}$ in k Teilmengen und die Schnittkantenmenge $E' \subseteq E$. Das Entfernen von E' zerteilt den Graphen in exakt k nichtleere Komponenten.

Ein minimaler-k-Schnitt (engl. Min-k-Cut) findet die Kantenmenge E' mit der kleinsten Summe der Kantengewichte, die zwischen den Partitionen verlaufen. Dieses Problem ist eine Erweiterung des Min-Cut-Problems. Goldschmidt und Hochbaum (1994) haben bewiesen, dass bei einem willkürlich gewählten k das Problem $\mathcal{NP}$-vollständig ist. Dies ergibt sich aus der Reduktion des Cliquenproblems. Für ein fest vorgegebenes k konnten Goldschmidt und Hochbaum (1994) einen deterministischen Polynomialzeitalgorithmus entwickeln, der eine Divide-and-Conquer-Methode nutzt. Das Problem der exponentiellen Laufzeitabhängigkeit von k bleibt allerdings bestehen, es sei denn $\mathcal{NP} = \mathcal{P}$.

Nagamochi u. a. (1999) präsentieren speziell für $k = 5$ und $k = 6$ einen schnellen deterministischen Algorithmus. Einige Jahre später entwickelten Kamidoi u. a. (2007) für das Min-k-Cut-Problem mit fest vorgegebenem k einen schnelleren deterministischen Algorithmus, der wie der Algorithmus von Goldschmidt und Hochbaum (1994) auf Max-Flow-Berechnungen basiert, allerdings eine komplexere Rekursion verwendet. Im Gegensatz dazu erzielt Thorup (2008) mit einem Tree-Packing-Ansatz für $k > 6$ eine quadratische Laufzeitverbesserung gegenüber dem Algorithmus von Kamidoi u. a. (2007).

Durch das Festhalten von k vorher bestimmten Knoten, sogenannten *terminals*, welche vergleichbar sind mit so: Source und ta: Target, wird der Min-so-ta-Cut zum Min-k-terminal-Cut erweitert (Kamidoi u. a., 2007). Obwohl sich die Probleme sehr ähneln, ist das Min-k-terminal-Cut-Problem bereits für $k \geq 3$ $\mathcal{NP}$-schwer (Dahlhaus u. a., 1992). Das im Rahmen dieser Arbeit zu lösende Zuordnungsproblem zwischen Objektklassen zweier Datensätze entspricht dem Min-k-terminal-Cut-Problem, da in der Knotenmenge, die durch die Partitionierung entsteht, mindestens eine Objektklasse aus jedem Datensatz festgehalten werden muss.

Für $\mathcal{NP}$-schwere Probleme ist es sehr unwahrscheinlich, Algorithmen mit polynomieller Laufzeit zu finden. Probleme dieser Art können entweder durch exakte Algorithmen oder durch Näherungsverfahren gelöst werden. Während die erstgenannten immer das beste Ergebnis liefern, wenn teilweise auch nur sehr langsam, können Näherungsalgorithmen niemals das beste Ergebnis garantieren, aber dafür effizient sein. Damit die Lösung eines Näherungsalgorithmus bewertet werden kann, ist zum Vergleich eine exakte Lösung notwendig. Aus diesem Grund wird im folgenden Abschnitt auf die ganzzahlige lineare Programmierung eingegangen, die es ermöglicht, mit existierender Software, z.B. dem IBM ILOG CPLEX Interactive Optimizer, der bereits fortgeschrittene Algorithmen beinhaltet, ein $\mathcal{NP}$-schweres Problem durch die Überführung in ein mathematisches Modell zu lösen.

3.4 Ganzzahlige lineare Programmierung

Die Zuordnung semantisch ähnlicher Objektklassen zweier Datensätze ist ein spezielles Optimierungsproblem, das mit Verfahren der linearen Optimierung gelöst werden kann. Ziel ist es, möglichst viele der vorkommenden Objektklassen einander zuzuordnen und dabei die Anzahl der Objektrelationen, die die Zuordnung bestätigen, zu maximieren. Das Maximierungsproblem ist von bestimmten Eigenschaften und Bedingungen (Restriktionen) abhängig. Das Problem wird als Funktion in Abhängigkeit der Variablen modelliert, für die es unter Einhaltung der Restriktionen bei gleichzeitiger Maximierung der Zielfunktion gilt, die beste Lösung zu finden.

Die allgemeine mathematische Formulierung eines Optimierungsproblems lautet wie folgt:

$$\begin{aligned} \text{Maximiere} \quad & f(x) \\ \text{gemäß den Restriktionen} \quad & g_i(x) \geq 0, i \in \{1, ..., m\} && (3.5) \\ & h_i(x) = 0, i \in \{1, ..., p\} && (3.6) \\ \text{mit} \quad & x \in \mathbb{R}^n \end{aligned}$$

wobei f die Zielfunktion, x den Vektor der Variablen, die die Lösungen des Problems repräsentieren sowie h und g als Restriktionsfunktionen, entweder in Form von Ungleichungen (3.5) oder Gleichungen (3.6), die die Lösungsmenge der Variablen einschränken, kennzeichnen. Zusätzlich können auch Hilfsvariablen definiert werden, für die mit Hilfe von Restriktionen bestimmte Sollwerte erzwungen werden können, um sie auch in der Zielfunktion verwenden zu können.

Abhängig von der Form der Zielfunktion, der Restriktionsfunktionen und den Variablen ergeben sich verschiedene Klassen von Optimierungsproblemen. Sind f, g und h lineare Funktionen von x wird von linearer Optimierung gesprochen. Ein lineares Optimierungsproblem heißt auch lineares Programm (LP – engl. Linear Program) und ist in der Regel effizient lösbar. Umfasst im Gegensatz dazu die Zielfunktion quadratische Terme, wird von einem quadratischen Optimierungsproblem (QP – engl. Quadratic Program) gesprochen. Sind auch die Restriktionsfunktionen quadratische Funktionen, wird das Problem als quadratisches Optimierungsproblem mit

quadratischen Nebenbedingungen (QCP – engl. Quadratically Constrained Quadratic Optimization Problem) bezeichnet und die Komplexität gegenüber einem LP nimmt zu.

Die Art der Variablen, die in f, g und h vorkommen, haben ebenfalls Einfluss auf die Lösung des Problems. In einem LP sind alle Variablen kontinuierlich, was die Lösung in polynomieller Zeit ermöglicht. Werden die Variablen dahingehend beschränkt, dass sie nur diskrete Werte, d.h. Integer annehmen dürfen, wird von einem ganzzahligen Optimierungsproblem (IP – engl. Integer Program) bzw. diskreter Optimierung gesprochen. Diese Einschränkung steigert die Komplexität im Vergleich zu einem LP deutlich und lässt das Problem sogar $\mathcal{NP}$-schwer sein. Wenn die Variablen sowohl kontinuierlich als auch diskret sind, ist es ein gemischt-ganzzahliges Programm (MIP – engl. Mixed-Integer Program). Für einen ausführlicheren Überblick sei an dieser Stelle auf die Bücher *Numerical Optimization* (Nocedal und Wright, 2006), *Combinatorial Optimization: Algorithms and Complexity* (Papadimitriou und Steiglitz, 1982) und das Buchkapitel *Räumliche Analyse durch kombinatorische Optimierung* (Haunert und Wolff, 2016) verwiesen.

Die Notwendigkeit, ganzzahlige Optimierungsprobleme zu formulieren, wird durch viele praxisrelevante Beispiele gezeigt, bei denen keine gestückelten Lösungen erwünscht sind, z.B. bei der Produktion von Artikeln oder der Tourenplanung, für die keine halben Fahrzeuge eingeteilt werden können. Die Objektklassenzuordnung, die im Rahmen dieser Arbeit untersucht wird, erfordert ebenfalls die Formulierung als ganzzahliges Optimierungsproblem, da die Objektklassen nicht teilweise, sondern nur vollständig zugeordnet werden sollen. Mit Hilfe eines einfachen Beispiels werden die Unterschiede zwischen den Lösungen eines LPs gegenüber denen eines IP graphisch veranschaulicht. Folgendes Problem ist dafür gegeben:

$$\begin{aligned} \text{Maximiere} \quad & f(x) = 3x_1 + 4x_2 \\ \text{gemäß den Restriktionen} \quad & g_1(x) : -4x_1 + 5x_2 \leq 5 \\ & g_2(x) : 6x_1 + 4x_2 \leq 24 \\ \text{mit} \quad & x_1 \geq 0 \\ & x_2 \geq 0 \,. \end{aligned} \tag{3.7}$$

In Abbildung 3.18 werden für das Beispiel zwei Geraden für die Restriktionen g_1 und g_2 in das Koordinatensystem eingezeichnet. Sie bilden mit der x_1- und x_2-Achse den zulässigen Lösungsraum als zweidimensionales, konvexes Polytop. Wird das Beispiel als LP gelöst, mit den x-Variablen als kontinuierliche Werte, sind alle Eckpunkte des Lösungsraums ($(0;0)$, $(4;0)$, $(2{,}17;2{,}74)$ und $(0;1)$) potentielle Lösungen. Der Maximalwert kann durch Testen aller Eckpunkte bestimmt werden. In diesem Beispiel entsteht der Maximalwert im Punkt $(2{,}17;2{,}74)$ und beträgt $f_{\mathsf{LP}}(x) = 17{,}48$. Alle Eckpunkte zu testen ist nicht sehr effizient, besonders wenn sich die Anzahl der Variablen und Restriktionen erhöht und sich somit auch die Dimension des Polytops vergrößert.

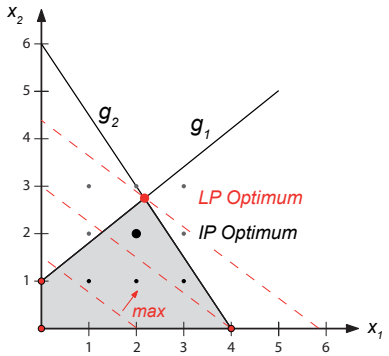


Abbildung 3.18: Graphische Lösung eines LP *und eines* IP. *Das zweidimensionale, konvexe Polytop (graue Fläche) stellt den Lösungsraum dar und wird durch die Restriktionsgleichungen g_1, g_2 und die Koordinatenachsen x_1, x_2 begrenzt. Die rot markierten Eckpunkte sind mögliche Lösungspunkte des* LPs *und die schwarzen Punkte innerhalb des Polytops die des* IPs. *Der vergrößerte rote Punkt kennzeichnet das* LP*-Optimum und der schwarze das* IP*-Optimum. Die rot gestrichelten Linien sind Linien gleicher Kosten und erhöhen sich mit Abstand vom Koordinatenursprung.*

Stattdessen wird für die Lösung von linearen Programmen häufig der weit verbreitete Simplex Algorithmus (Dantzig, 1963) verwendet. Dieser Algorithmus basiert auf der Idee, auf der Suche nach dem besten Eckpunkt die Kosten bei der Bewegung von Punkt zu Punkt des Polytops stetig zu verbessern. Die Bewegung erfolgt

dabei entlang des Polytoprandes immer zum nächsten Nachbarpunkt. Je größer die Dimension des Polytops wird, desto wahrscheinlicher ist auch hier, dass die Anzahl der Berechnungsschritte exponentiell wächst.

Wird im Gegensatz dazu das Beispiel als IP gelöst, dürfen die x-Variablen nur ganzzahlige Werte annehmen. Es sind insgesamt neun Lösungen zulässig, die sich entweder auf dem Rand oder im Inneren des Polytops befinden: $(0;0)$, $(1;0)$, $(2;0)$, $(3;0)$, $(4;0)$, $(0;1)$, $(1;1)$, $(2;1)$, $(3;1)$ und $(2;2)$. Das globale Optimum entsteht im Punkt $(2;2)$, weil dort der Maximalwert von $f_{\mathsf{IP}}(x) = 14$ innerhalb des Polytops erreicht wird, der allerdings 20 % geringer als das LP-Optimum ist.

Eine weitere Möglichkeit zur Lösung von IPs ist die Lockerung der Restriktion von ganzzahligen x-Variablen, so dass auch kontinuierliche Werte erlaubt sind. In diesem Fall ist eine effiziente Lösung mit dem Simplex-Verfahren möglich. Werden tatsächlich ganzzahlige Lösungswerte erzielt, entspricht dies gleichzeitig der optimalen IP-Lösung. Allerdings kann das globale IP-Optimum nur in seltenen Fällen mit solch einem LP-Lösungsverfahren bestimmt werden. Gleichermaßen führt das Auf- bzw. Abrunden der IP-Lösung auf den nächst größeren bzw. kleineren ganzzahligen Wert nicht in jedem Fall direkt zur optimalen Lösung des IPs. Sowohl $\lfloor f_{\mathsf{LP}}(x) \rfloor = 17$ als auch $\lceil f_{\mathsf{LP}}(x) \rceil = 18$ sind Ergebnisse, die sich außerhalb des zulässigen Lösungsbereichs befinden. Trotzdem wird die Methode der Auf- oder Abrundung oft verwendet, um das zu lösende Problem in zwei Teilprobleme zu trennen.

Um ganzzahlige Optimierungsprobleme zu lösen, können entweder exakte Verfahren, wie beispielsweise Branch-and-Bound (dt. Verzweigen und Begrenzen) und Schnittebenenverfahren (engl. Cutting Plane Algorithm) oder Heuristiken eingesetzt werden, da es keinen bekannten praktikablen Algorithmus gibt, der große IP-Probleme lösen kann. Übermäßige Zeitanforderungen machen ihn auch für kleine Instanzen praktisch undurchführbar.

4 Entwicklung von Data-Matching-Verfahren für verschiedene Objektgeometrien

Die Zuordnung von raumbezogenen Daten erfolgt im Rahmen der vorliegenden Arbeit mit geometrischen, topologischen und semantischen Objekteigenschaften. Es werden einzelne Ähnlichkeitsmaße definiert und zu einem Gesamtähnlichkeitsmaß zusammengefasst. Der Schwerpunkt der Arbeit liegt auf der Zuordnung von Polygonobjekten, die in Abschnitt 4.1 präsentiert wird. Hierzu werden Datensätze mit vergleichbaren, aber auch unterschiedlichen Maßstäben betrachtet. Als Ergebnis werden einfache und komplexe Objektrelationen identifiziert. Die Kombination verschiedener Relationsarten in einer Häufigkeitsmatrix zeigt Abschnitt 4.1.3. Die Matrix dient als Eingabe für das Schema-Matching. Abschnitt 4.2 stellt eine eigene frühere Arbeit vor, die die Zuordnung von Objekten unterschiedlicher Geometriedimensionen erlaubt (Kieler u. a., 2009b). Ein Datensatz mit Polygon- und Linienobjekten wird in einen reinen Liniendatensatz überführt und einem anderen Liniendatensatz zugeordnet.

4.1 Zuordnung von Polygonobjekten

Bei der Zuordnung von Polygonobjekten wird in dieser Arbeit hauptsächlich auf einen hohen Überdeckungsgrad vertraut. Wenn im einfachsten Fall zwei Polygone verschiedener Datensätze das gleiche Real-Welt-Objekt repräsentieren, müssen sie sich mit hoher Wahrscheinlichkeit in der gleichen räumlichen Lage befinden und ähnliche geometrische Eigenschaften aufweisen. Die Erfüllung beider Bedingungen spiegelt sich in einer deutlichen Überlagerung beider Objekte wider.

Abbildung 4.1 stellt die Vorgehensweise für die Objektzuordnung schematisch dar. Als erstes werden alle Objekte der Datensätze A und B geometrisch überlagert. Auf Basis eines geometrischen Parameters (siehe Abschnitt 4.1.1), der die Überlagerungsflächen und die Objektausrichtungen zueinander auswertet, werden einfache und komplexe Objektrelationen identifiziert und in eine Liste mit potentiellen Zuordnungsrelationen aufgenommen. In dieser Liste kann theoretisch jedes Objekt mehrfach vorkommen, wenn beispielsweise ein Objekt p_{A_1} selbst Teil unterschiedlicher Relationen ist oder verschiedenen, sich überlagernden Objekten des anderen Datensatzes zugeordnet wird: $p_{A_1} \rightarrow p_{B_1}$ und $p_{A_1} \rightarrow \{p_{B_1}, p_{B_2}\}$. Überlagerungsobjekte sind beispielsweise Brücken, die über Gewässer führen oder Stadtteile, die in Ortslagen liegen (siehe Abbildung 4.4). Um komplexe Relationen zu bevorzugen, bei denen vorrangig Objekte gleicher Objektklassen zusammengefasst werden, wird ein Heterogenitätsparameter (siehe Abschnitt 4.1.2) berücksichtigt.

Anschließend wird für jedes Objekt die Relation mit dem höchsten Gesamtparameter ausgewählt. Bei Übernahme der Objektrelationen in die Ergebnisliste wird eine Prüfung auf doppelte Objekteinträge durchgeführt. Es werden alle Relationen entfernt, in denen doppelte Objekteinträge vorkommen. Aus der finalen Liste wird eine Häufigkeitsmatrix abgeleitet, die als Eingabe für das Schema-Matching dient.

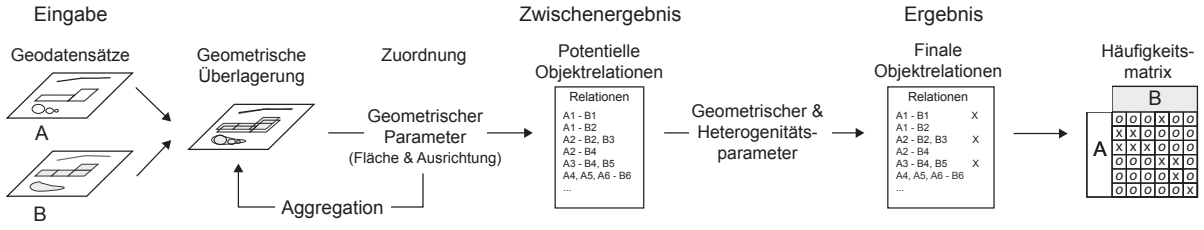


Abbildung 4.1: Schematische Darstellung des Zuordnungsprozesses von Polygonobjekten. Die als Ergebnis identifizierten finalen Objektrelationen werden in einer Häufigkeitsmatrix zusammengefasst und dienen als Eingabe für das Schema-Matching.

4.1.1 Geometrischer Parameter

Für die Bestimmung des geometrischen Parameters werden Polygonobjekte hinsichtlich ihrer geometrischen und topologischen Eigenschaften analysiert und miteinander verglichen. Dafür werden einerseits die in Abschnitt 2.2.3 eingeführten zwei Überlagerungsverhältnisse s_{i_j} mit $j = A, B$, folgend als Flächenparameter

bezeichnet, und andererseits ein Ausrichtungsparameter s_a definiert und zu gleichen Anteilen in einem geometrischen Gesamtmaß s_g zusammengefasst:

$$s_g = s_{i_A} + s_{i_B} + s_a. \tag{4.1}$$

Jeder geometrische Parameter kann maximal den Wert Eins annehmen. Daraus ergibt sich für s_g ein Maximalwert von Drei.

Flächenparameter

Für die beiden Flächenparameter werden die Objekte geometrisch überlagert und ihre Schnittfläche i bestimmt (siehe Abb. 4.2). Anschließend wird das Verhältnis der Schnittfläche zu beiden Polygonflächen p_j mit $j = A, B$ ermittelt und als s_{i_j} definiert:

$$s_{i_j} = \frac{i}{p_j} \quad \text{mit } j = A, B. \tag{4.2}$$

Je größer der Wert, desto größer ist die eigene Fläche, die sich im gegenüberliegenden Objekt befindet.

Für das in Abbildung 4.2 a) dargestellte Beispiel ergeben sich folgende Parameterwerte: $s_{i_A} = 415\,\mathrm{m}^2/624\,\mathrm{m}^2 = 0{,}66$ und $s_{i_B} = 415\,\mathrm{m}^2/460\,\mathrm{m}^2 = 0{,}90$. Die deutliche Differenz zwischen den Werten entsteht einerseits durch die unterschiedlichen Objektgrößen, weil Polygon p_B mehr als $^1/_4$ kleiner als Polygon p_A ist und andererseits durch die abweichende Position, was zu einer kleineren Schnittfäche führt.

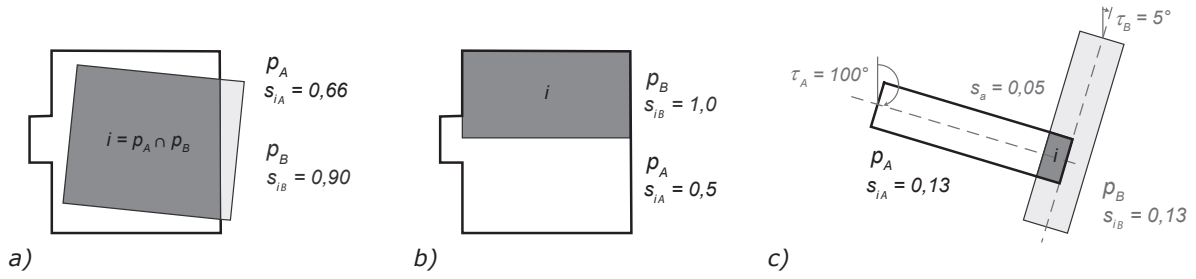


Abbildung 4.2: Bei der Überlagerung von verschiedenen Polygonobjekten p_A und p_B werden unterschiedliche Werte für die Flächenparameter s_{i_j} und den Ausrichtungsparameter s_a bestimmt.

Eine Differenzbildung beider Flächenparameter und die Vermutung, dass ein kleiner Wert für die Zuordnung besser ist, ist nicht immer richtig. Das Beispiel in Abbildung 4.2 c) zeigt, dass die Differenz beider Flächenparameter sogar 0 ist, die Werte jedoch zu klein sind, um eine Zuordnung zuzulassen. An dieser Stelle wird der Zuordnungsversuch vorzeitig abgebrochen, weil kein zuverlässiges Ergebnis zu erwarten ist. Aufgrund dieser Beobachtung wird für die Flächenparameter ein Schwellwert t eingeführt. Die Größenordnung des Schwellwerts ist stark von den verwendeten Datensätzen abhängig und kann empirisch aus Testdaten ermittelt werden. Um eine Fehlentscheidung aufgrund zu kleiner Flächenparameter zu vermeiden, die beispielsweise durch einen systematischen Versatz beider Datensätze zueinander hervorgerufen werden können, wird ein zusätzlicher geometrischer Parameter eingeführt, der die Entscheidung bestätigen soll.

Ausrichtungsparameter

Der Vergleich der Objektausrichtung kann das Eliminieren falscher Matching-Kandidaten unterstützen. Für die Bestimmung der Polygonausrichtungen werden, wie zuvor in Abschnitt 3.1.1 vorgestellt, vereinfacht die Richtungswinkel τ_j mit $j = A, B$ zur Längsseite des jeweiligen minimal umschließenden Rechtecks bestimmt. Für das Beispiel in Abbildung 4.2 c) stimmen die minimal umschließenden Rechtecke mit der jeweiligen Objektform überein. Der Ausrichtungsparameter s_a wird mit Hilfe der Winkeldifferenz $\Delta\tau = |\tau_A - \tau_B|$ bestimmt und gibt die Abweichung vom rechten Winkel an:

$$s_a(\Delta\tau) = \begin{cases} 1 - \frac{\Delta\tau}{90^\circ}, & \text{für } \Delta\tau \leq 90^\circ \\ \frac{\Delta\tau}{90^\circ} - 1, & \text{für } \Delta\tau > 90^\circ. \end{cases} \tag{4.3}$$

Daraus folgt, je größer s_a ist, umso ähnlicher ist die Ausrichtung der Polygone.

Bildung komplexer Objektrelationen durch Aggregation

Grundsätzlich gilt, je größer die einzelnen Parameter sind, umso größer wird der geometrische Parameter s_g und damit die Wahrscheinlichkeit der Objektzuordnung. Wenn, wie in Abbildung 4.2 b), der Flächenparameter s_{i_B} den Maximalwert besitzt und s_{i_A} deutlich kleiner ist, dann liegt p_B vollständig in p_A. Dieser Hinweis ist nützlich, um die Analyse auf die Objektnachbarschaft auszudehnen.

Das Aggregieren von Nachbarobjekten, ähnlich dem Buffer Growing Prinzip von Walter (1997), kann dazu führen, dass komplexe Objektrelationen mit verbesserten geometrischen Parametern bestimmt werden. Abbildung 4.3 erläutert die Vorgehensweise des Zusammenfassens anhand eines Beispiels.

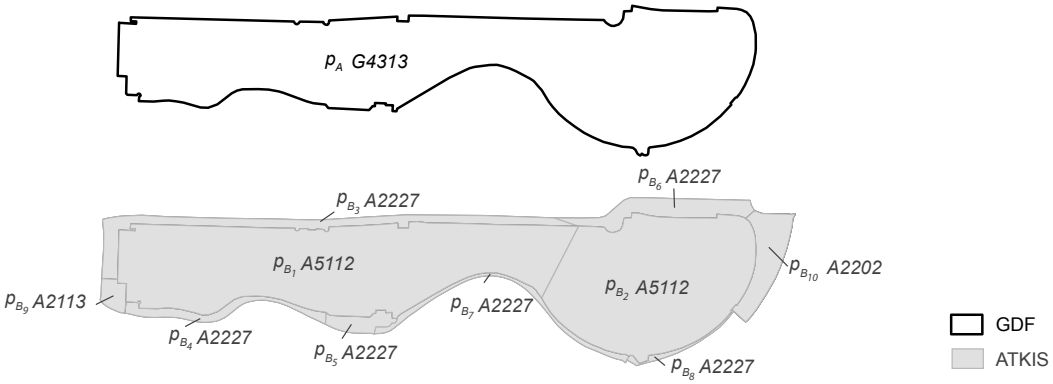


Abbildung 4.3: Bildung einer komplexen Objektrelation zwischen dem schwarz umrandeten GDF-Objekt von Datensatz A und den grauen ATKIS-Objekten aus Datensatz B durch die Zusammenfassung von Nachbarobjekten unter Berücksichtigung ihrer Objektklassen. Zur Bewahrung der Übersichtlichkeit, werden die sich überlagernden Objekte beider Datensätze räumlich getrennt dargestellt. Erläuterungen zu den vorliegenden Objektklassen sind im Anhang in der Übersichten C.1 und C.2 zu finden.

Das schwarz umrandete GDF-Objekt aus Datensatz A und die grauen ATKIS-Objekte aus Datensatz B werden hinsichtlich ihrer Korrespondenz untersucht. Als erstes werden die Objekte p_A und p_{B_1} überlagert und beide Flächenparameter, der Ausrichtungsparameter und das daraus resultierende geometrische Gesamtmaß berechnet: $s_{i_A} = 0{,}99$, $s_{i_B} = 0{,}59$, $s_a = 0{,}98$ und $s_g = 2{,}56$. Die Größenordnungen der Flächenparameter verdeutlichen, dass p_{B_1} nur etwa halb so groß wie p_A ist und somit eine Zusammenfassung von p_{B_1} mit den Nachbarobjekten zu prüfen ist. Jedoch hat p_{B_1} insgesamt sechs Nachbarobjekte mit unterschiedlichen Objektklassen. Im Rahmen dieser Arbeit besitzen alle Objektklassen zusätzlich zu dem Objektklassennamen einen numerischen Code, der hier für die Bestimmung der semantischen Ähnlichkeit der Objektklassen verwendet wird. Dazu wird die in Abschnitt 3.1.3 vorgestellte semantische Distanz berechnet. Für die Aggregation werden als erstes Objekte gleicher Objektklassen ausgewählt. Möglicherweise könnten Objekte aufgrund von Modellierungskriterien bei der Datenerfassung unterteilt worden sein, obwohl sie eine Einheit bilden. Besitzt keines der Nachbarobjekte die gleiche Objektklasse wird eine Rangliste genutzt, die durch Differenzbildung der Klassencodes erstellt wird. Die Rangfolge für p_{B_1} mit der Objektklasse A5112 (Binnensee, Stausee, Teich) sieht wie folgt aus: als erstes wird das Objekt p_{B_2} mit identischer Objektklasse aggregiert, bevor die Objekte p_{B_3}, p_{B_4} und p_{B_7} der Objektklasse A2227 (Grünanlage) oder p_{B_5} mit A2201 (Sportanlage) oder p_{B_9} mit A2113 (Fläche gemischter Nutzung) getestet werden. Das Zusammenfassen von p_{B_1} und p_{B_2} erhöht den geometrischen Parameter gegenüber der 1:1-Objektrelation $p_A \rightarrow p_{B_1}$ von $s_g = 2{,}56$ auf $s_g = 2{,}99$. Dies entspricht gleichzeitig dem Maximalwert aller möglichen Kombinationen.

4.1.2 Heterogenitätsparameter

Die Prüfung aller Objekte eines Datensatzes mit allen Objekten des anderen Datensatzes erzeugt eine Liste mit vielen potentiellen Objektrelationen, in der ein Objekt auch mehrfach vorhanden sein kann. Abbildung 4.4 zeigt zwei unterschiedliche Zuordnungsrelationen für das schwarze p_A-Objekt, die beide gültig sind.

Unter alleiniger Berücksichtigung geometrischer Aspekte ist als beste Zuordnung die Lösung mit dem höchsten geometrischen Parameter zu wählen. In diesem Fall hat die komplexe Objektrelation aus Abbildung 4.4 b) mit $s_g = 2{,}95$ einen größeren Parameterwert als die einfache Objektrelation mit $s_g = 2{,}89$ aus Abbildung 4.4 a). In Hinblick auf das sich anschließende Schema-Matching ist die einfache Objektrelation, bei der sich nur zwei

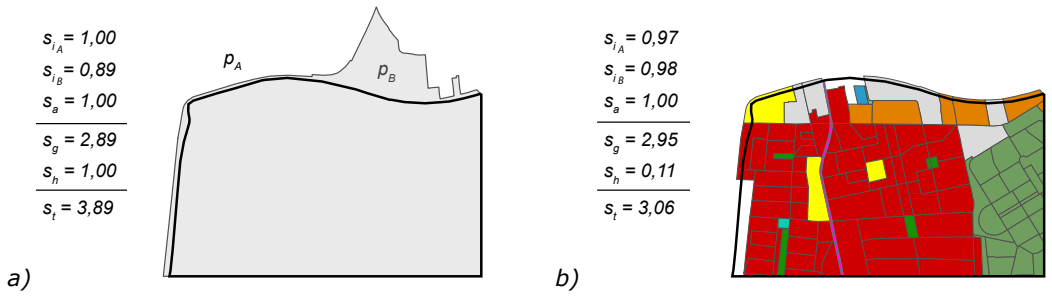


Abbildung 4.4: Die Einführung des Heterogenitätsparameter s_h kann die in a) dargestellte einfache Objektrelation mit geringerem geometrischen Parameter gegenüber der in b) dargestellten komplexen Objektrelation für die endgültige Relationsauswahl bevorzugen.

Objektklassen gegenüberstehen und die semantische Korrespondenz eindeutig ist, besser geeignet. Bei der komplexen Objektrelation wird p_A mehreren Objekten mit sehr unterschiedlichen Objektklassen, gekennzeichnet durch verschiedene Farben, zugeordnet. Die Zuordnung auf Schemaebene ist in diesem Fall eher unspezifisch. Um klare Zuordnungen zu stärken, werden alle Objektrelationen auf die Heterogenität der beteiligten Objektklassen hin überprüft. Dafür wird ein Heterogenitätsparameter s_h eingeführt:

$$s_h = \frac{(|A| + |B|) - (|A|_r + |B|_r)}{(|A| + |B|) - 2} \,. \tag{4.4}$$

Hierbei entsprechen $|A|$ und $|B|$ der Gesamtanzahl der in Datensatz A bzw. B vorkommenden Objektklassen und $|A|_r$ und $|B|_r$ der Anzahl der Objektklassen je Datensatz pro Objektrelation. Für das Beispiel ergeben sich für den Heterogenitätsparameter aus Gleichung 4.4 bei einer Gesamtzahl von 11 Objektklassen in beiden Datensätzen stark unterschiedliche Werte. Während die 1:1-Relation in a): Stadtteil → Ortslage mit $s_h = 1{,}00$ dem Idealfall entspricht, wird bei b) nur ein Wert von $s_h = 0{,}11$ erreicht, da fast alle vorhandenen Objektklassen in der Relation vertreten sind: Stadtteil → {Wohnbaufläche, Industriefläche, Fläche gemischter Nutzung, Fläche besonderer Prägung, Sportanlage, Freizeitfläche, Friedhof, Grünanlage, Fläche unbekannt}.

Durch Addition des Heterogenitätsparameters zum geometrischen Parameter entsteht das Gesamtähnlichkeitsmaß s_t :

$$s_t = s_g + s_h, \tag{4.5}$$

das die Auswahl der besten Matching-Entscheidungen zugunsten homogener und einfach nachvollziehbarer Relationen zwischen Objektklassen beeinflusst. Für s_t ergibt sich ein Maximalwert von Vier.

4.1.3 Erzeugung eines kombinierten Ergebnisses für das Schema-Matching

Für das Schema-Matching werden sowohl einfache als auch komplexe Objektrelationen in einer Häufigkeitsmatrix H zusammengefasst, die exemplarisch in Tabelle 4.1 dargestellt ist.

H		Menge B				
		b_1	b_2	b_3	...	b_m
Menge A	a_1	h_{11}	h_{12}	h_{13}	...	h_{1m}
	a_2	h_{21}	h_{22}	h_{23}	...	h_{2m}
	...	...	...	...	...	...
	a_n	h_{n1}	h_{n2}	h_{n3}	...	h_{nm}

Tabelle 4.1: Allgemeine Häufigkeitsmatrix H als Ergebnis des Data-Matchings zwischen Menge A (Objektklassen in Datensatz A) und Menge B (Objektklassen in Datensatz B). Jede Zelle der Häufigkeitsmatrix repräsentiert eine Klassenkombination und beinhaltet eine Trefferzahl $h_{ij} = H(a_i, b_j)$ mit $i = 1, ..., n$ und $j = 1, ..., m$.

Menge A umfasst alle Objektklassen von Datensatz A und Menge B die von Datensatz B. Für jedes Element aus A und B beinhaltet die Matrix einen Häufigkeitswert, also ist $H\colon A \times B \to \mathbb{R}^+$. $H(a, b)$ steht für die Anzahl der Relationen, die einer Objektklasse $a \in A$ und $b \in B$ gleichzeitig zugeordnet sind.

Für jede 1:1-Objektrelation ($1{\cdot}a \to 1{\cdot}b$) wird der Wert $H(a, b)$ in der Häufigkeitsmatrix um 1 erhöht. Bei komplexen Objektrelationen werden die beteiligten Objektklassen untersucht. Abbildungen 4.5 a) und b) zeigen zwei

verschiedene 1:2-Objektrelationen, die sich hinsichtlich der beteiligten Objektklassen, gekennzeichnet durch unterschiedliche Grautöne, voneinander unterscheiden. Während in a) eine homogene Schemarelation O2 → {A2121, A2121} mit $s_h = 1{,}00$ präsentiert wird, stellt b) eine heterogene Schemarelation O2 → {A0931, A0932} mit $s_h < 1$ dar. Homogene Schemarelationen werden im weiteren Verlauf der Arbeit wie 1:1-Relationen behandelt. Der Wert $H(a_{\text{O2}}, b_{\text{A2121}})$ wird um 1 erhöht. Bei heterogenen Schemarelationen werden Relationsanteile für jede beteiligte Objektklasse bestimmt, um die mehrdeutige Zuordnung zu erhalten.

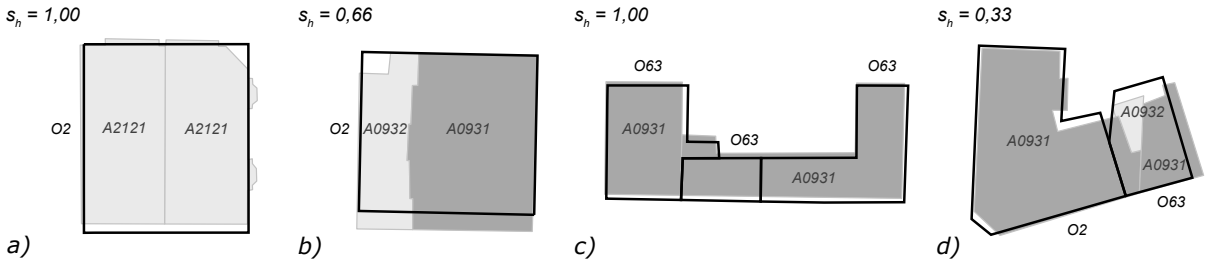


Abbildung 4.5: Auswertung der komplexen Objektrelationen für die Objektklassenzuordnung. Die Objektrelationen in a) und c) sind Vertreter von homogenen Schemarelationen und werden wie 1:1-Relationen berücksichtigt. Im Gegensatz dazu repräsentieren b) und d) heterogene Schemarelationen, erkennbar an den unterschiedlichen Grautönen und den Heterogenitätswerten $s_h < 1$. Die Bestimmung von s_h basiert in diesem Beispiel auf $|A| = 2$ und $|B| = 3$.

Volz (2006) wählte einen Ansatz, der die beteiligten Objektklassen zu jeweils gleichen Teilen berücksichtigt. Für das Beispiel in 4.5 b) bedeutet das, dass sich in der Häufigkeitsmatrix die Werte $H(a_{\text{O2}}, b_{\text{A0931}})$ und $H(a_{\text{O2}}, b_{\text{A0932}})$ um jeweils 0,5 erhöhen würden. Der hier vorgestellte Ansatz geht darüber hinaus, indem die einzelnen Flächengrößen der Objekte berücksichtigt werden. Ein kleines Objekt hat gegenüber einem größeren Objekt einen geringeren Einfluss auf die endgültige Objektklassenzuordnung. Es werden prozentuale Anteile der Einzelflächen an der Gesamtfläche berechnet. Der Gebäudeteil mit der Objektklasse A0931 ist etwa doppelt so groß wie der Teil der Objektklasse A0932 und somit erhöht sich $H(a_{\text{O2}}, b_{\text{A0931}})$ um 0,72, während die Steigerung bei $H(a_{\text{O2}}, b_{\text{A0932}})$ mit 0,28 deutlich geringer ausfällt. Alle Anteile einer Relation ergeben immer den Wert Eins.

Für die in Abbildung 4.5 d) dargestellte 2:3-Objektrelation {O2, O63} → {A0931, A0931, A0932} verändert sich die Häufigkeitsmatrix gleich an vier Stellen. In Datensatz A ist das Objekt der Objektklasse O2 drei Mal größer als das O63-Objekt. In Datensatz B umfassen beide Objekte der Klasse A0931 96 % der gemeinsamen Fläche. Daraus ergeben sich folgende Relationsanteile $H(a_{\text{O2}}, b_{\text{A0931}}) = 0{,}75 \cdot 0{,}96 = 0{,}72$, $H(a_{\text{O2}}, b_{\text{A0932}}) = 0{,}75 \cdot 0{,}04 = 0{,}03$, $H(a_{\text{O63}}, b_{\text{A0931}}) = 0{,}25 \cdot 0{,}96 = 0{,}24$ und $H(a_{\text{O63}}, b_{\text{A0932}}) = 0{,}25 \cdot 0{,}04 = 0{,}01$. Tabelle 4.2 fasst die Relationsanteile der vier Objektrelationen aus Abbildung 4.5 a) bis d) zusammen.

H		Menge B		
		A0931	A0932	A2121
Menge A	O2	$0{,}72_{b)} + 0{,}72_{d)} = \mathbf{1{,}44}$ [36 %]	$0{,}28_{b)} + 0{,}03_{d)} = \mathbf{0{,}31}$ [7,75 %]	$\mathbf{1}_{a)}$ [25 %]
	O63	$1{,}00_{c)} + 0{,}24_{d)} = \mathbf{1{,}24}$ [31 %]	$\mathbf{0{,}01}_{d)}$ [0,25 %]	0

Tabelle 4.2: Häufigkeitsmatrix H für die in Abbildung 4.5 präsentierten Objektrelationen. Neben den einzelnen Relationsanteilen, die mit dem Index des Beispiels gekennzeichnet sind, werden in fett die Gesamtrelationsanteile und in [ ] die prozentualen Anteile angegeben.

Aufgrund der unterschiedlichen Klassenanzahl ist eine eindeutige Zuordnung nicht möglich. Es werden drei bedeutende Relationsanteile zwischen vier der sechs Objektklassen identifiziert. Der größte Anteil wird zwischen den Objektklassen O2 (Appartments) → A0931 (Wohngebäude) mit $H(a_{\text{O2}}, b_{\text{A0931}}) = 1{,}44$ bestimmt, was einem prozentualen Anteil von 36 % entspricht. Der zweitgrößte Relationsanteil wird zwischen O63 (Offices) → A0931 (Wohngebäude) mit $H(a_{\text{O63}}, b_{\text{A0931}}) = 1{,}24$ (31 %) berechnet. Mit einem Anteil von 25 % darf auch der semantische Zusammenhang zwischen O2 (Appartments) → A2121 (Wohngebäude mit Handel und Dienstleistungen) nicht vernachlässigt werden. Das Zusammenfassen der vier beteiligten Objektklassen {O2, O63} → {A0931, A2121} kann als Zuordnungsergebnis durch 92 % der im Data-Matching festgestellten Objektrelationen bestätigt werden. Diese Zuordnung bedeutet auch, dass die Relation O63 (Offices) → A2121 (Wohngebäude mit Handel und Dienstleistungen) mit einen Anteil von Null in das Zuordnungsergebnis eingeht.

Die alleinige Maximierung der Summe führt dazu, dass alle Objektklassen von A allen Objektklassen von B zugeordnet werden und somit keine semantische Differenzierung mehr möglich ist. Dieses Kriterium muss demzufolge ergänzt werden, um semantisch nachvollziehbare Korrespondenzen zwischen Objektklassen aufzudecken. Bevor

in Kapitel 5 verschiedene Kriterien vorgestellt werden, die bei der Auswertung der Häufigkeitsmatrix angewendet werden, wird im nächsten Abschnitt ein Matching-Verfahren beschrieben, das die Zuordnung unterschiedlicher Objektgeometrien ermöglicht.

4.2 Zuordnung von unterschiedlichen Objektgeometrien

Im Gegensatz zur Zuordnung von Polygonobjekten kann bei der Zuordnung von unterschiedlichen Objektgeometrien, wie z.B. Linien- zu Polygonobjekten oder Linien- zu Linienobjekten, nicht auf einen hohen Überdeckungsgrad vertraut werden. Eine teilweise Überlagerung bzw. ein Schnitt der Geometrien ist zwar möglich, aber nicht vorauszusetzen. Allerding muss die räumliche Nähe gegeben sein. In diesem Abschnitt wird ein Ansatz vorgestellt, der die Überführung von Objekten in eine andere Geometriedimension vorschlägt, weil eine Zuordnung zwischen gleichartigen Objekten einfacher ist.

In einer früheren Arbeit (Kieler u. a., 2009b) wurde ein Zuordnungsverfahren entwickelt, das es ermöglicht, zwei Gewässernetzwerke mit unterschiedlichen Maßstäben, die sowohl Linien- als auch Polygonobjekte beinhalten, durch die Überführung in reine Liniennetzwerke einander zuzuordnen. Der Zuordnungsprozess umfasst drei Schritte und ist schematisch in Abbildung 4.6 dargestellt. Das Überführen der Objektrelationen in die Häufigkeitsmatrix wird aufgrund von fehlenden Objektklasseninformationen an dieser Stelle nicht weiter ausgeführt.

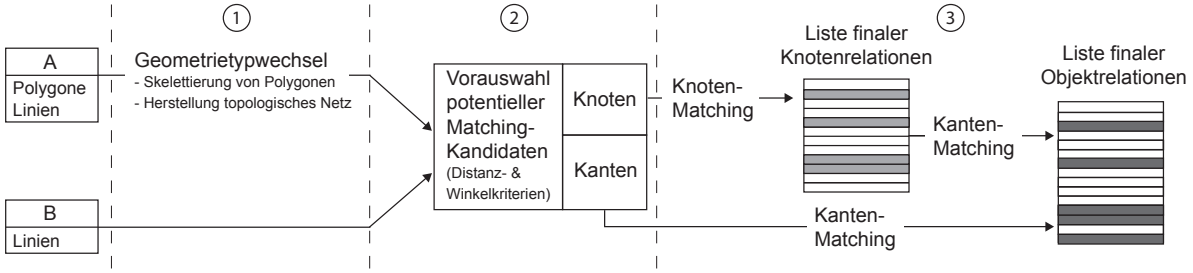


Abbildung 4.6: Schematische Darstellung des Zuordnungsprozesses von Polygon- und Linienobjekten.

Im ersten Schritt, einem Vorverarbeitungsschritt, werden automatisch topologisch korrekte Liniennetzwerke erzeugt. Dazu werden zunächst die Polygonobjekte mit Hilfe eines Skelettierungsalgorithmus auf ihre Mittellinien kollabiert. Auf die Erzeugung von exakten Skeletten, wie der Medialen Achse, wird aufgrund des zu hohen Rechenaufwands verzichtet. Stattdessen wird mit Hilfe der bedingten Delaunay Triangulation im Inneren des Polygons ein einfaches Skelett bestimmt. Penninga u. a. (2005) stellen die Methode im Detail vor. Die Bewahrung der korrekten Gewässernetztopologie ist in Folge des Geometrietypwechsels umso wichtiger. Damit die Skelettlinien der einzelnen Polygone nicht mühsam zusammengeführt werden müssen bzw. um die Bildung von Artefakten bei zwei aneinander grenzenden Polygonen zu vermeiden, werden im Vorfeld alle benachbarten Polygone verschmolzen. Anschließend kann für alle Polygonobjekte ein gemeinsames Skelett bestimmt werden. Abbildung 4.7 a) und b) zeigen die unterschiedlichen Ergebnisse.

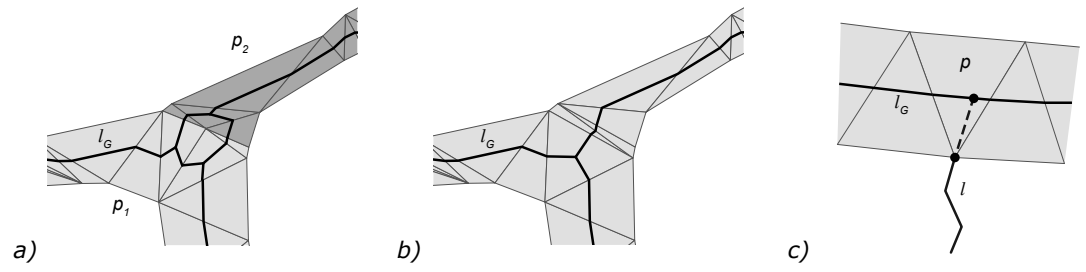


Abbildung 4.7: Die unterschiedliche Skelettbildung bei a) zwei benachbarten Polygonen p_1 und p_2 und b) bei einem aggregierten Polygon. In c) wird die Verlängerung des linienhaften Objektes l bis hin zur Skelettlinie des Polygons l_G gezeigt, die notwendig ist, um die topologische Verbindung zu erhalten.

Des Weiteren ist die Verbindung zwischen einem schmalen Fluss (Linie) und einem weiten Fluss (Polygon) auch nach der Skelettbildung sicherzustellen. Abbildung 4.7 c) zeigt dafür eine Möglichkeit, indem die Linie in ihrer Originalrichtung bis hin zum Skelett verlängert wird. Diese einfache Konstruktion kann unter bestimm-

ten Voraussetzungen jedoch eine Topologieänderung bewirken. In Kieler u. a. (2009b) werden diese Probleme ausführlicher dargestellt und Lösungsvorschläge präsentiert.

Der zweite Schritt des Zuordnungsverfahrens entspricht einer Vorauswahl, die getrennt für Knoten und Kanten der Liniennetzwerke geeignete Matching-Kandidaten identifiziert. Dafür werden neben Distanzkriterien auch Winkelkriterien verwendet, um die Qualität der potentiellen Kandidaten zu bewerten.

Für jeden Knoten v_B des kleinmaßstäbigen Datensatzes wird eine Menge V'_{v_B} von Knoten bestimmt, die möglicherweise mit v_B korrespondieren. Zu diesem Zweck wird um jeden Knoten v_B ein Puffer mit der Distanz δ gelegt. Jeder Knoten v_A des großmaßstäbigen Datensatzes, der innerhalb des Puffers liegt, wird der Menge V'_{v_B} als Matching-Kandidat hinzugefügt. Hierbei wird die Distanz zwischen den Knoten als Qualitätsmaß betrachtet. Je kleiner die Distanz, desto größer ist die Qualität.

Entsprechend wird für jede Kante e_B des kleinmaßstäbigen Datensatzes eine Menge E'_{e_B} von Kanten bestimmt, die mit e_B übereinstimmen könnten. Ähnlich wie bei den Knoten wird um jede Kante e_B ein Pufferobjekt mit einer Distanz ϵ gebildet (siehe Abb. 4.8 a). Jede Kante e_A des großmaßstäbigen Datensatzes, die vollständig innerhalb des Pufferobjekts liegt oder den Rand schneidet, wird in die Menge E'_{e_B} als Matching-Kanditat aufgenommen. Die Qualität der Kandidaten wird aus dem Vergleich der Ausrichtungen abgeleitet. Da die Kanten aufgrund der unterschiedlichen Maßstäbe verschiedene Längen haben können, wird der Vergleich der Kantenausrichtungen nur lokal im Schnittpolygon beider Pufferobjekte durchgeführt. Je kleiner die Winkeldifferenz, desto höher ist die Qualität. Mit der Vorauswahl wird eine Rangliste potentieller Matching-Kandidaten und eine Reduktion des Suchraums für die endgültige Zuordnung erreicht.

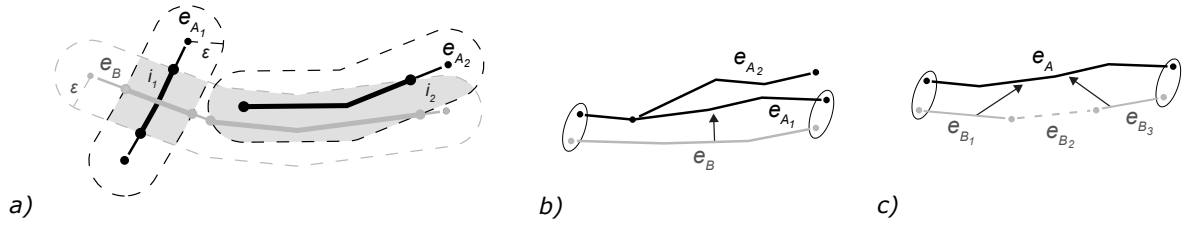


Abbildung 4.8: In a) wird die Bestimmung des besten Matching-Kandidaten für die graue Kante e_B des kleinmaßstäbigen Maßstabs gezeigt. Beide schwarzen Kanten e_{A_1} und e_{A_2} des großmaßstäbigen Datensatzes sind Matching-Kandidaten, da jede Kante das Pufferpolygon von e_B schneidet. Innerhalb der grauen Schnittpolygone werden die fett markierten Kantenteile auf die Winkeldifferenz hin überprüft. Aufgrund der geringeren Winkeldifferenz besitzt e_{A_2} eine höhere Match-Qualität als e_{A_1}. In b) und c) werden auf Basis der zugeordneten Knoten, gekennzeichnet durch Ellipsen, Beispiele für die Zuordnung der Kanten präsentiert. Während in b) beide Endpunkte von e_B einen Matching-Kandidaten besitzen, hat in c) entweder nur ein Endpunkt oder gar kein Endpunkt einen Matching-Kandidaten.

Im letzten Schritt der Zuordnung wird ein finales Knoten-Matching durchgeführt. Für die Entscheidung, welche Knoten am besten korrespondieren, wird zusätzlich zur Distanz auch der Knotengrad der inzidenten Kanten (siehe Abschnitt 3.1.2) verglichen. Matching-Kandidaten, die einen größeren Knotengrad als v_B besitzen, werden im weiteren Verlauf ignoriert, da angenommen wird, dass kleinmaßstäbige Datensätze weniger detailliert sind und somit im Vergleich zum detaillierten Datensatz keine zusätzlichen Kanten besitzen. Sind die Knotengrade identisch, wird außerdem die Ausrichtung der inzidenten Kanten untersucht. Anhand dieser Ergebnisse wird für jeden Knoten der jeweils beste Matching-Partner bestimmt. Eine Nichtzuordnung ist ebenfalls möglich.

Das finale Kanten-Matching stützt sich auf das Knoten-Matching und unterscheidet drei Fälle:

1. beide Endpunkte der Kante e_B des kleinmaßstäbigen Datensatzes haben einen Matching-Kandidaten im großmaßstäbigen Datensatz (siehe Abbildung 4.8 b),
2. nur ein Endpunkt von e_B hat einen Matching-Kandidaten oder
3. kein Endpunkt von e_B hat einen Matching-Kandidaten (siehe Abb. 4.8 c).

Für Fall 1) wird der Pfad mit minimalen Kosten zwischen den korrespondierenden Knoten gesucht, der aus potentiellen Matching-Kandidaten besteht. Dieser Ansatz wurde von Mustière und Devogele (2008) vorgeschlagen. Die Kosten berechnen sich aus den Winkeldifferenzen gewichtet mit den Kantenlängen. Für das Beispiel in Abbildung 4.8 b) liefert e_{A_1} den Pfad minimaler Kosten. Für Fall 2), in dem nur ein Knoten der Kante zugeordnet wird, wird die nächstgelegende Kante ausgewählt, die eine vorgegebene Winkeldifferenz unterschreitet. In diesem Fall wird die Kante e_A sowohl für e_{B_1} als auch für e_{B_3} als Matching-Kandidat ausgewählt. Bei Fall 3) erhält e_{B_2} keinen Matching-Partner aus dem anderen Datensatz und ist aus diesem Grund gestrichelt dargestellt.

Mit dem beschriebenen Matching-Verfahren wurden im Rahmen von Testuntersuchungen zwischen Datensätzen mit unterschiedlichen Maßstäben (1:50.000 bzw. 1:250.00) sehr gute Zuordnungsergebnisse erzielt. Das Verfahren hat 90,5 % der Objektkorrespondenzen identifiziert, obwohl große Teile des Gewässernetzwerks im großmaßstäbigen Datensatz durch Polygonobjekte und im kleinmaßstäbigen Datensatz durch Linienobjekte repräsentiert waren. Ziel der Testuntersuchung war es, den Austausch von Attributen zwischen den Datensätzen zu ermöglichen.

Das Ableiten von Objektklassenkorrespondenzen ist nicht möglich, weil Informationen bezüglich einer Objektklassenzugehörigkeit nicht zur Verfügung stehen. Mit Kenntnis dieser Zusatzinformationen kann nach gleichem Vorbild wie in Abschnitt 4.1.3 beschrieben, eine Häufigkeitsmatrix für das Schema-Matching aus den Objektrelationen abgeleitet werden. Für die komplexen Objektrelationen muss, anders als bei den Polygonobjekten, der jeweilige Relationsanteil bezogen auf die Linienlängen statt auf die Flächen bestimmt werden. Somit hat ein kürzeres Linienobjekt auch einen geringeren Einfluss auf die endgültige Objektklassenzuordnung.

5 Entwicklung von Schema-Matching-Verfahren basierend auf Instanzdaten

In diesem Kapitel werden verschiedene Verfahren vorgestellt, mit denen es möglich ist, aus Ergebnissen der Objektzuordnung automatisch semantische Korrespondenzen zwischen Objektklassen abzuleiten. Identifizierte Objektrelationen werden dazu in Häufigkeiten umgeformt und als Eingangsdaten verwendet. Sie können aufgrund verschiedener Sichtweisen entweder als Häufigkeitsmatrix oder als bipartiter Graph betrachtet werden. In Abschnitt 4.1.3 wurde die Erzeugung der Häufigkeitsmatrix erläutert. Zunächst wird das vorliegende Zuordnungsproblem formal beschrieben, bevor unterschiedliche Lösungsansätze anhand eines synthetischen Beispiels erläutert werden.

5.1 Formale Problemdefinition

Gegeben ist eine Menge A der Objektklassen in Datensatz A und eine Menge B der Objektklassen in Datensatz B sowie eine Häufigkeitsmatrix H, die für jedes Element $a_i \in A$ und $b_j \in B$ mit $i = 1, ..., n$ und $j = 1, ..., m$ einen Häufigkeitswert $h_{ij} = H(a_i, b_j)$ beinhaltet und das Ergebnis der Objektzuordnung widerspiegelt.

Ein Häufigkeitswert 0 gibt an, dass für diese Klassenkombination keine Objektrelation vorliegt. Ein hoher Wert deutet auf eine bedeutsame Klassenkombination hin und repräsentiert eine starke semantische Ähnlichkeit. Ein geringer Wert sagt jedoch nichts darüber aus, ob eine Kombination unbedeutend oder gar fehlerhaft ist. Geringe Werte treten auf, wenn Objektklassen nur wenige oder sehr spezielle Objekte beinhalten. Beispielsweise ist die Anzahl von Wohngebäuden in einem Gebäudedatensatz um ein Vielfaches höher als die Anzahl vorhandener Schulen. Dies ist bereits durch den Zweck bzw. die unterschiedliche Notwendigkeit von Wohngebäuden und Schulen begründet. Die Schemarelation für Schulgebäude ist relevant und soll identifiziert werden. Die Größe und der Inhalt der Häufigkeitsmatrix sind sehr stark von den verwendeten Datensätzen und dem untersuchten Testgebiet beeinflusst.

Gesucht wird eine Zuordnung der Objektklassen aus beiden Datensätzen. Im Idealfall sind dies ausschließlich 1:1 -Schemarelationen, damit eine einfache Interpretation der Ergebnisse möglich ist. Dieses Resultat kann in den seltensten Fällen erzielt werden, da sich Datensätze hinsichtlich ihrer Objektklassenanzahl oft unterscheiden. Um die Bedingung zu lockern, nur 1:1 -Relationen zwischen den einzelnen Objektklassen zuzulassen, wird die Zusammenfassung von Objektklassen zu Clustern ermöglicht. Gesucht wird für die Zuordnung eine Partition P_A von A und eine Partition P_B von B in k Teilmengen. Die Anzahl k der Teilmengen kann dabei entweder vorgegeben werden, z.B. $k = 2$, oder das Problem wird mehrfach über $k = 1, ..., c$ gelöst und die verschiedenen Lösungen werden anhand eines Qualitätsmaßes bewertet.

Zusammenfassend wird eine 1:1 -Zuordnung der Elemente aus P_A und P_B gesucht, wobei $g(p)$ das Element in P_B bezeichnet, dem $p \in P_A$ zugeordnet ist. Die Summe der Häufigkeiten soll über die zugeordneten Partitionen maximiert werden:

$$\text{Maximiere} \quad \sum_{q_B \in g(p)} \sum_{q_A \in p} \sum_{p \in P_A} H(q_A, q_B), \tag{5.1}$$

wobei q_A für die Objektklassen in $p \in P_A$ und q_B für die Objektklassen in $g(p) \in P_B$ stehen.

Durch die Problemdefinition in Gleichung 5.1 wird erzwungen, dass in der Häufigkeitsmatrix pro Zeile bzw. Spalte nur eine Zuordnung enthalten sein darf. Allerdings können sich Zuordnungen über mehrere Zeilen bzw. Spalten erstrecken. Das bedeutet, dass in der Matrix rechteckige Cluster gesucht werden, die sich nicht schneiden dürfen.

5.1.1 Synthetisches Beispiel

Die Zuordnung von Partitionen wird mit Hilfe eines synthetisch generierten Beispiels verdeutlicht. Menge A umfasst sechs Objektklassen Wald (Wa), Acker (Ac), Siedlung (Si), Fluss (Fl), Straße (St), Bahn (Ba), während Menge B sieben Objektklassen Bruchwald (Bw), Nadelwald (Nw), Laubwald (Lw), Acker (Ac), Gewässer (Ge), Verkehr (Ve), Schiene (Sc) besitzt. Tabelle 5.1 zeigt die Häufigkeitsmatrix für das Beispiel. Zur Wahrung der

Übersichtlichkeit werden im weiteren Verlauf der Arbeit nur die Abkürzungen der Objektklassennamen verwendet. Die Häufigkeitsmatrix ist cirka zu einem Drittel mit Werten belegt, was einem realen Untersuchungsszenario entspricht. Die Gesamthäufigkeit beträgt $H_{\text{total}} = 600$.

Tabelle 5.1: Synthetisches Beispiel: Die Gesamthäufigkeit der 6×7 Matrix beträgt $H_{\text{total}} = 600$.

H		Menge B Bruchwald (Bw)	Nadelwald (Nw)	Laubwald (Lw)	Acker (Ac)	Gewässer (Ge)	Verkehr (Ve)	Schiene (Sc)	$\sum$
Menge A	Wald (Wa)	74	74	74	0	1	0	0	223
	Acker (Ac)	2	1	0	74	0	0	0	77
	Siedlung (Si)	0	0	0	74	0	1	0	75
	Fluss (Fl)	0	1	1	0	74	0	0	76
	Straße (St)	0	0	0	0	0	74	0	74
	Bahn (Ba)	0	0	0	1	0	0	74	75
	$\sum$	76	76	75	149	75	75	74	600

5.2 Einfache Lösungsverfahren

Das vorgestellte Zuordnungsverfahren ist $\mathcal{NP}$-schwer (siehe Abschnitt 3.3) und somit ist kein Algorithmus mit polynomieller Laufzeit zu erwarten, der ein exakt optimales Ergebnis liefert. Demnach können entweder Exponentialzeitalgorithmen (z.B. ganzzahlige lineare Programme) oder Verfahren, die garantiert eine Lösung in polynomieller Zeit finden, aber Kompromisse in Bezug auf das Ergebnis verursachen, eingesetzt werden. Welche Einschränkungen bzw. Vereinfachungen vorgenommen werden können, wird in den folgenden Abschnitten zusammengefasst und hinsichtlich der tatsächlichen Anwendbarkeit diskutiert.

5.2.1 Beschränkung auf 1:1 -Zuordnungen (Max-Match)

Für die Interpretation von semantischen Korrespondenzen zwischen Objektklassen verschiedener Datensätze sind für den Anwender 1:1 -Relationen am besten geeignet. Die Ergebnisse lassen sich einfach und schnell auf Korrektheit prüfen und für die Definition von Transformationsregeln zwischen zwei Datensätzen verwenden. Allerdings sollten sich dafür zwei Schemas gegenüberstehen, die Objekte mit semantisch ähnlichen Objektklassen beschreiben, z.B. nur Straßen- oder Gebäudeobjekte. Erst dann kann die Vereinfachung des Problems durch die Beschränkung auf 1:1 -Schemarelationen sinnvoll und gerechtfertigt sein. Bei semantisch abweichenden Objektklassen werden voraussichtlich nur sehr wenige Objektrelationen identifiziert. Das Ziel, die Zuordnung auf Schemaebene durch die Objektzuordnung zu bestätigen, kann damit verfehlt werden.

Um 1:1 -Zuordnungen in gewichteten bipartiten Graphen, wie sie im Rahmen dieser Arbeit vorliegen, effizient zu bestimmen, eignen sich Algorithmen für das maximal gewichtete Matching. Im weiteren Verlauf der Arbeit wird der Lösungsansatz mit Max-Match bezeichnet. Dem Anwender müssen die Auswirkungen dieses Kompromisses auf das Ergebnis bewusst sein. Nur wenn die Mengen A und B die gleiche Anzahl Elemente enthalten, kann für jede Objektklasse ein Matching-Kandidat bestimmt werden. Bei unterschiedlichen Objektklassenanzahlen werden nur $\min(|A|, |B|)$ Zuordnungen identifiziert. Für $||A| - |B||$ Objektklassen gibt es somit keine Matching-Kandidaten.

Die Ungarische Methode eignet sich zur Bestimmung eines gewichteten maximalen Matchings und wurde in Abschnitt 3.3.2 erläutert. Der Algorithmus ist so angelegt, dass auf der Suche nach der maximalen Anzahl von Paaren das Kostenminimum gesucht wird. Für die Bestimmung des Maximums muss jeder Häufigkeitswert in der Matrix H durch die Differenz aus Maximalwert und Häufigkeitswert ersetzt werden:

$$h'_{ij} = \max(h_{ij}) - h_{ij}. \tag{5.2}$$

Der Algorithmus ist an quadratische Matrizen gebunden, was unter Umständen die Einführung von Hilfsklassen, sogenannten Dummy-Elementen, notwendig macht. Für das synthetische Beispiel ist die Einführung solch einer Hilfsklasse in der Menge A notwendig. Tabelle 5.2 zeigt das optimale Max-Match -Ergebnis in der Häufigkeitsmatrix. Die Häufigkeit der Zuordnung beträgt $H_{(k=7)} = 372$. Die Objektklasse Laubwald wird dem Dummy-Element zugeordnet und erhält keine semantische Entsprechung in A. Im Vergleich dazu hätte ein Experte folgende Zuordung getroffen: Wald $\rightarrow$ {Bruchwald, Nadelwald, Laubwald}. Eine Lösung, die auch komplexe Schemarelationen (1:n, n:1 bzw. n:m) zulässt und für jede Objektklasse einen Matching-Kandidaten bestimmt, wird angestrebt.

Tabelle 5.2: Max-Match *-Lösung für das synthetische Beispiel in Matrixdarstellung. Die grau unterlegten Werte kennzeichnen die Zuordnung mit* $H_{(k=7)} = 372$.

H		Menge B						
		Bw	Nw	Lw	Ac	Ge	Ve	Sc
Menge A	Wa	74	74	74	0	1	0	0
	Ac	2	1	0	74	0	0	0
	Si	0	0	0	74	0	1	0
	Fl	0	1	1	0	74	0	0
	St	0	0	0	0	0	74	0
	Ba	0	0	0	1	0	0	74
	Dummy	0	0	0	0	0	0	0

5.2.2 Beschränkung auf zwei Cluster (Min-Cut)

Die Einschränkung auf 1:1 -Schemarelationen kann verhindert werden, indem das Zuordnungsergebnis zwischen allen beteiligten Objektklassen auf genau zwei Cluster begrenzt wird. Komplexe Schemarelationen sind möglich, wenn sich Schemas mit mehr als zwei Objektklassen gegenüberstehen. Bei sehr großen Schemas führt die Beschränkung allerdings zu einer Zusammenfassung vieler verschiedener Objektklassen, was eine semantische Interpretationen unmöglich macht.

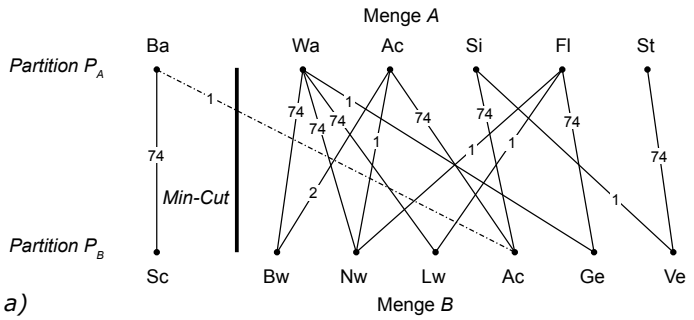
Die Partitionierung eines Graphen in genau zwei Teile kann mit dem Min-Cut-Verfahren erreicht werden. Im weiteren Verlauf der Arbeit wird das Verfahren mit Min-Cut gekennzeichnet. Für diesen Spezialfall existieren effiziente Algorithmen, die zuvor in Abschnitt 3.3.3 beschrieben wurden. Es werden die Kanten mit den geringsten Gewichten geschnitten, bis zwei getrennte Partitionen entstehen. Eine Besonderheit bei bipartiten Graphen ist, dass der Schnitt nicht die Knotenteilmengen A und B voneinander trennt, sondern in jedem Cluster mindestens ein Element pro Teilmenge behält.

Der Vollständigkeit halber wird an dieser Stelle das Ergebnis des Verfahrens für das synthetische Beispiel präsentiert, obwohl es im Rahmen dieser Arbeit nicht als Lösungsverfahren vorgeschlagen wird. In Abbildung 5.1 werden beide Partitionen mit einer Häufigkeit von $H_{(k=2)} = 599$ gezeigt. Für die Zerlegung des Graphen wird nur eine Kante mit dem Gewicht 1 geschnitten. In der Häufigkeitsmatrix repräsentieren die Partitionen zwei rechteckige Cluster, die sich nicht schneiden. Aus den Partitionen:

$$P_A = \{\{\text{Wa, Ac, Si, Fl, St}\}, \text{Ba}\} \quad \text{und}$$
$$P_B = \{\{\text{Bw, Nw, Lw, Ac, Ge, Ve}\}, \text{Sc}\}$$

ergibt sich folgende Zuordnung:

$$\{\text{Wa, Ac, Si, Fl, St}\} \rightarrow \{\text{Bw, Nw, Lw, Ac, Ge, Ve}\} \quad \text{und}$$
$$\text{Ba} \rightarrow \text{Sc}.$$



	Bw	Nw	Lw	Ac	Ge	Ve	Sc
Wa	74	74	74	0	1	0	0
Ac	2	1	0	74	0	0	0
Si	0	0	0	74	0	1	0
Fl	0	1	1	0	74	0	0
St	0	0	0	0	0	74	0
Ba	0	0	0	1	0	0	74

Abbildung 5.1: Min-Cut *-Lösung für das synthetische Beispiel. Darstellung der Partitionierung im a) bipartiten Graph und b) in der Häufigkeitsmatrix:* $P_A = \{\{Wa, Ac, Si, Fl, St\}, Ba\}$ *und* $P_B = \{\{Bw, Nw, Lw, Ac, Ge, Ve\}, Sc\}$.

Für das synthetische Beispiel gibt es jedoch eine weitere Lösung mit der gleichen Häufigkeit. Die Zuordnung ist genauso nachvollziehbar und lautet wie folgt:

$$\{\mathsf{Ba, Wa, Ac, Si, Fl}\} \rightarrow \{\mathsf{Sc, Bw, Nw, Lw, Ac, Ge}\} \quad \text{und}$$
$$\mathsf{St} \rightarrow \mathsf{Ve}.$$

Wenn das einfache Lösungsverfahren in einem heuristischen Ansatz rekursiv angewendet wird, werden verbesserte Zuordnungsergebnisse erwartet.

5.3 Einsatz von Heuristiken

Die rekursive Unterteilung der Teilmengen A und B mit dem effizienten Min-Cut-Algorithmus stellt einen heuristisch hierarchischen Lösungsansatz für das Zuordnungsproblem dar, das nach nahezu optimalen Ergebnissen sucht. In Abbildung 5.2 wird die Verfahrensweise der Rekursion für das synthetische Beispiel gezeigt.

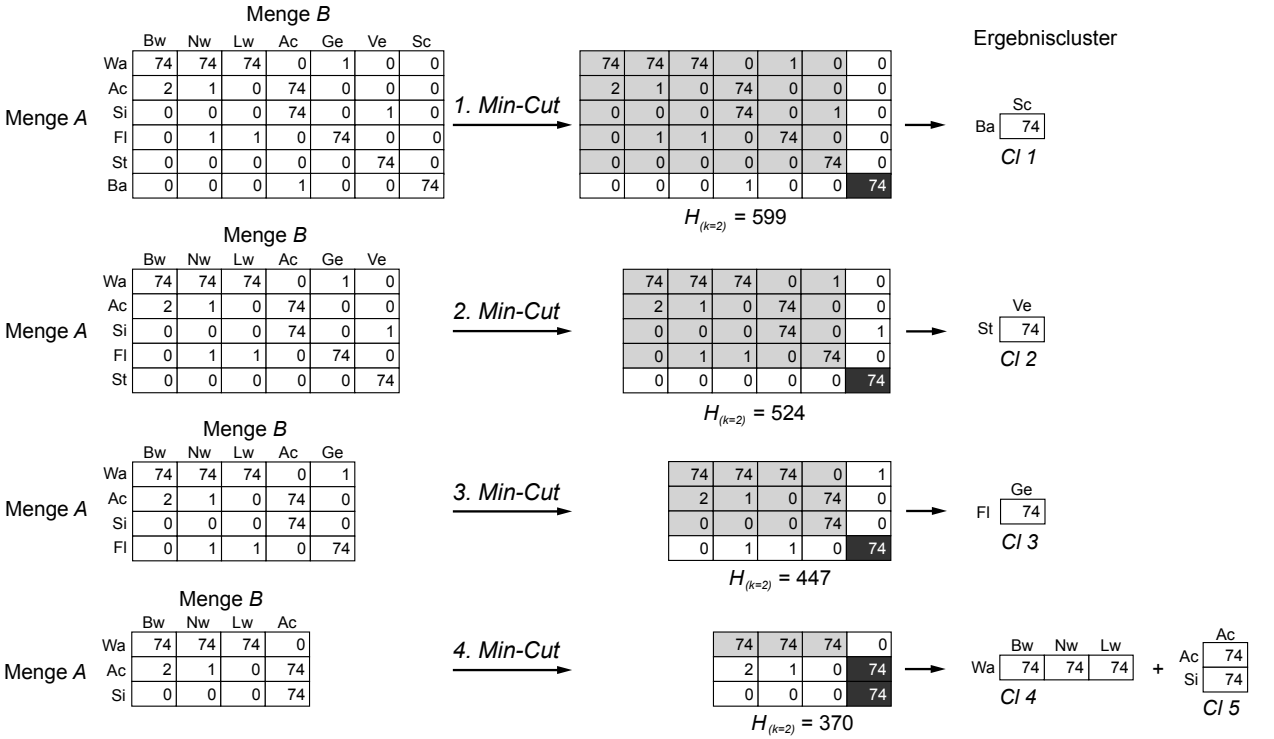


Abbildung 5.2: Rekursive Unterteilung der Häufigkeitsmatrix des synthetischen Beispiels mit dem Min-Cut-Algorithmus. In jedem Verfahrensschritt entstehen zwei Cluster. Das hellgraue Cluster wird weiter unterteilt. Insgesamt sind vier Berechnungen des minimalen Schnitts notwendig. Es entstehen fünf Cluster mit einer Häufigkeit von $H_{(k=5)} = 592$.

Bei der ersten Anwendung des minimalen Schnitts entstehen zwei Cluster unterschiedlicher Größe. Ein dunkelgraues Cluster mit nur einer und ein hellgraues Cluster mit 30 Zellen. Das dunkelgraue Cluster kann nicht weiter geteilt werden. Am hellgrauen Cluster sind beide Datensätze mit mehr als zwei Objektklassen beteiligt, demzufolge ist eine weitere Zerlegung möglich. Die Vorgehensweise wird solange fortgeführt, bis keine weitere Trennung der Cluster mehr möglich ist. Dies ist der Fall, wenn nur noch eine Objektklasse von Datensatz A oder B am Cluster beteiligt ist. Für das synthetische Beispiel sind insgesamt vier Berechnungsschritte notwendig. Es werden fünf Cluster mit acht Zellen und einer Häufigkeit von $H_{(k=5)} = 592$ gebildet. Folgende Objektklassenzuordnungen werden identifiziert, die sowohl 1:1-, 1:n- und n:1-Schemarelationen beinhalten:

$$\mathsf{Ba} \rightarrow \mathsf{Sc},$$
$$\mathsf{St} \rightarrow \mathsf{Ve},$$
$$\mathsf{Fl} \rightarrow \mathsf{Ge},$$
$$\mathsf{Wa} \rightarrow \{\mathsf{Bw, Nw, Lw}\} \quad \text{und}$$
$$\{\mathsf{Ac, Si}\} \rightarrow \mathsf{Ac}.$$

Der vorgestellte heuristisch hierarchische Ansatz kann sowohl in angemessener Zeit als auch mit vertretbarem Aufwand, Ergebnisse bestimmen. Allerdings ist die Lösung weder garantiert optimal, noch liefert sie Informationen zur Qualität der Ergebnisse. Mit dem Verfahren können auch keine komplexen n:m -Relationen bestimmt werden. Dies kann zum Nachteil werden, wenn sich die beteiligten Schemas stark unterscheiden.

Die Korrektheit und Leistungsfähigkeit des Verfahrens kann mit manuell erstellten Referenzlösungen oder mit optimalen Ergebnissen verglichen und bewertet werden. Heuristische Verfahren mit guten Lösungsergebnissen sind für praktische Probleme häufig ausreichend, da sie, im Gegensatz zu Verfahren mit optimalen Ergebnissen, viel effizienter sind. Da manuell erstellte Referenzlösungen selten sind, werden im nächsten Abschnitt Optimierungsverfahren vorgestellt.

5.4 Einsatz der ganzzahligen linearen Programmierung

Manche Optimierungsverfahren können garantiert optimale Ergebnisse für die Objektklassenzuordnung liefern. Durch die Formulierung des Zuordnungsproblems als ganzzahliges lineares Programm (ILP – engl. Integer Linear Program) ist dies möglich. Dafür ist ein Modell aufzustellen, das aus Zielfunktion, Variablen und Bedingungen besteht.

5.4.1 Optimierungsziele und Bedingungen

Das Zuordnungsproblem erfordert eine optimale Aufteilung der Knotenteilmengen A und B und der Kantenmenge E in k Partitionen. Knoten, die durch Kanten mit hohen Gewichten verbunden sind, sollen zu Clustern zusammengefasst werden. Knoten, die durch Kanten mit geringen Gewichten verbunden sind, sollen sich dagegen in unterschiedlichen Clustern befinden. Daraus folgt, dass die Summe der Kantengewichte innerhalb der Cluster maximal und außerhalb der Cluster minimal wird. Die Maximierung der Häufigkeiten stellt das primäre Optimierungsziel dar und wird daher im folgenden Abschnitt ausführlich vorgestellt.

Maximierung der Häufigkeiten (MaxScore)

Für die Modellierung des Problems wird für jeden Knoten eine Variable $x_{k,v} \in \{0,1\}$ definiert: $x_{k,v} = 1$ bedeutet, dass der Knoten $v \in V$ ($V = A \cup B$) der Partition $k = 1, ..., c$ zugeordnet ist. Für jede Kante wird eine Variable $y_{k,e} \in \{0,1\}$ definiert, mit $y_{k,e} = 1$, wenn beide Endpunkte der Kante $e = \{u, v\}$ der gleichen Partition k zugeordnet sind. Um eine Zuordnung mit größtmöglichem Gesamtkantengewicht zu erreichen, wird die folgende lineare Zielfunktion aufgestellt:

$$\text{Maximiere} \quad H(y) = \sum_{k=1}^{c} \sum_{e \in E} h_{ij}(e) \cdot y_{k,e}. \tag{5.3}$$

Hierbei müssen folgende Restriktionen gelten:

$$\forall v \in V: \quad \sum_{k=1}^{c} x_{k,v} = 1, \tag{5.4}$$

$$\forall k \in \{1, ..., c\}: \quad \sum_{v \in A} x_{k,v} \geq 1, \quad \sum_{v \in B} x_{k,v} \geq 1, \tag{5.5}$$

$$\forall e = \{u, v\} \in E, \forall k \in \{1, ..., c\}: \quad y_{k,e} \leq x_{k,u}, \quad y_{k,e} \leq x_{k,v} \quad \text{und} \tag{5.6}$$

$$\forall e = \{u, v\} \in E, \forall k \in \{1, ..., c\}: \quad y_{k,e} \geq x_{k,u} + x_{k,v} - 1. \tag{5.7}$$

Restriktion (5.4) stellt sicher, dass jeder Knoten in genau einer Partition vorhanden sein muss, d.h. alle Objektklassen müssen zugeordnet werden. Die Restriktionen (5.5) garantieren, dass jede Partition mindestens je einen Knoten aus A und B enthält. Restriktion (5.6) ist notwendig, damit nur Kanten einer Partition zugeordnet werden, wenn auch beide Endpunkte der gleichen Partition zugeordnet sind. Die Definition der Restriktion (5.7) erzwingt, dass $y_{k,e}$ auf 1 gesetzt wird, wenn $x_{k,v} = 1$ und $x_{k,u} = 1$ gelten.

Mit der Zielfunktion $H(y)$ in Gleichung 5.3 ergeben sich sehr unterschiedliche Clustergrößen. Häufig entstehen neben einem sehr großen Cluster (mehrere) kleine Cluster. Die Abspaltung von kleinen Clustern erfolgt aufgrund des absoluten Maximum-Kriteriums, das für die Kanten innerhalb der Partitionen gilt. Je mehr Zellen in einem Cluster zusammengefasst werden, desto mehr Häufigkeitswerte können aufsummiert werden. Im weiteren Verlauf der Arbeit wird dieser Lösungsansatz mit MaxScore bezeichnet. Abbildung 5.3 verdeutlicht dieses Verhalten anhand des synthetischen Beispiels.

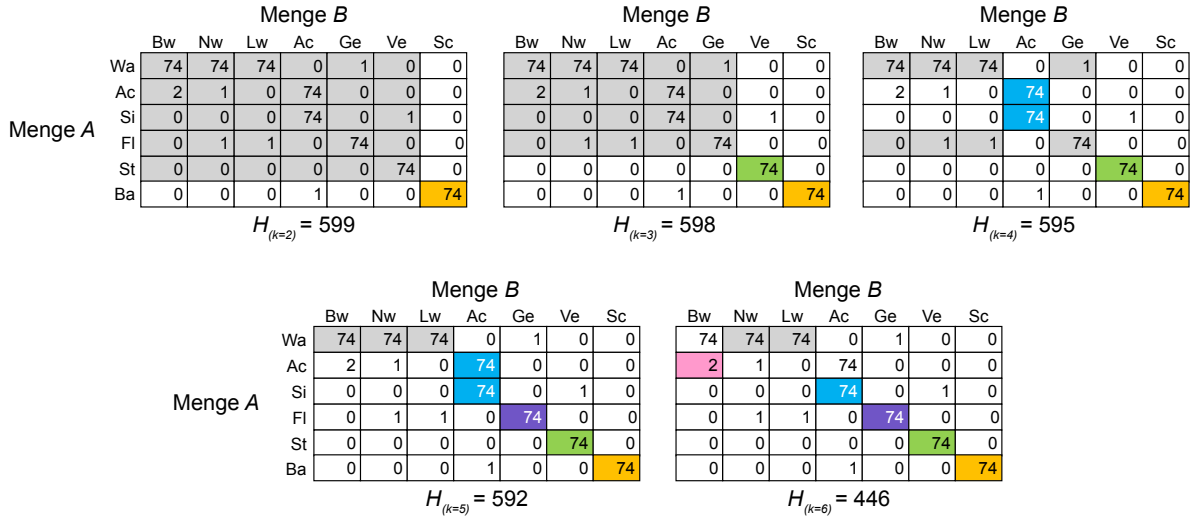


Abbildung 5.3: MaxScore *-Lösung: Clusterbildung unter Maximierung der Häufigkeiten* $H_{(k)}$ *für* $k = 2, 3, 4, 5, 6$. *Zellen, die zu einem Cluster gehören, haben den gleichen farbigen Hintergrund.*

Bei der Partitionierung in zwei bzw. drei Teile sind die Ergebnisse mit den Lösungen der rekursiven Unterteilung identisch. Ab $k = 4$ werden erste Unterschiede deutlich. Während beim Heuristischen Verfahren einmal gebildete Cluster immer Teil der Lösungsmenge bleiben, können sich Cluster beim Optimierungsverfahren mit wechselndem k verändern. Beispielhaft sei auf das blaue Cluster $\{\mathsf{Ac}, \mathsf{Si}\} \rightarrow \mathsf{Ac}$ verwiesen, das bei $k = 4$ entsteht, bei $k = 5$ erhalten bleibt, aber bei $k = 6$ nicht mehr Teil der Lösungsmenge ist. Im Vergleich zur Heuristischen Lösung erzeugt das Optimierungsverfahren ein Ergebniscluster mehr.

Für die semantische Interpretation ist die Bildung von allüberdeckenden Clustern nicht zielführend. Um dem entgegenzuwirken, sollte bei der Bildung der Cluster ebenfalls die Ausgewogenheit der Clustergrößen berücksichtigt werden. Cluster können entweder hinsichtlich der Anzahl der umfassenden Zellen oder gemäß der aufsummierten Häufigkeiten ausgewogen sein, was zu unterschiedlichen Zielfunktionen führt. In den folgenden Abschnitten werden beide Zielfunktionen vorgestellt. Der Vollständigkeit halber werden die Ergebnisse beider Zielfunktionen für das synthetische Beispiel präsentiert, obwohl sie im Rahmen dieser Arbeit nicht als eigenständige Lösungsverfahren vorgeschlagen werden.

Ausgewogene Cluster bezüglich der Zellenanzahl (BalancedSize)

Cluster mit ähnlicher Zellanzahl können erzeugt werden, indem die Größe des größten Clusters minimiert und die des kleinsten Clusters maximiert wird. Daraus leitet sich die Zielfunktion H_a wie folgt ab:

$$\text{Maximiere} \quad H_a = cell_{\text{total}} - clsize_{\max} + clsize_{\min}, \tag{5.8}$$

mit $cell_{\text{total}} = |A| \cdot |B|$ als Gesamtanzahl der Zellen in der Matrix, $clsize_{\min}$ der Anzahl der Zellen im kleinsten Cluster und $clsize_{\max}$ der Anzahl der Zellen im größten Cluster. Die Häufigkeitswerte haben hier keinen Einfluss. Je kleiner der Unterschied zwischen den einzelnen Clustergrößen, desto größer wird der Funktionswert H_a. Um die Werte für die Variablen zu bestimmen, werden Zusatzvariablen eingeführt, die von der Anzahl der Partitionen abhängig sind und die Gesamtzahl der Zellen in jedem Cluster enthalten. Im Modell werden diese Variablen mit $clsize_k$ bezeichnet. Die Bestimmung der Variablen $clsize_k$, $clsize_{\min}$ und $clsize_{\max}$ erfolgt unter Hinzunahme folgender Restriktionen:

$$\forall k \in \{1, ..., c\}: \quad clsize_k = \sum_{e \in E} y_{k,e}, \tag{5.9}$$

$$\begin{aligned} \forall k \in \{1, ..., c\}: \quad & clsize_{\min} \leq clsize_k, \\ & clsize_{\max} \geq clsize_k. \end{aligned} \tag{5.10}$$

Im Verlauf der Arbeit wird dieser Ansatz mit BalancedSize bezeichnet.

Für das synthetische Beispiel werden in Abbildung 5.4 die Ergebnisse der Clusteraufteilung BalancedSize vorgestellt und in Tabelle 5.3 den Ergebnissen von MaxScore gegenübergestellt. Die Anzahl der Clusterzellen unterscheidet sich stark. Für das Bilden von zwei Clustern waren bei MaxScore 31 Zellen und bei BalancedSize nur

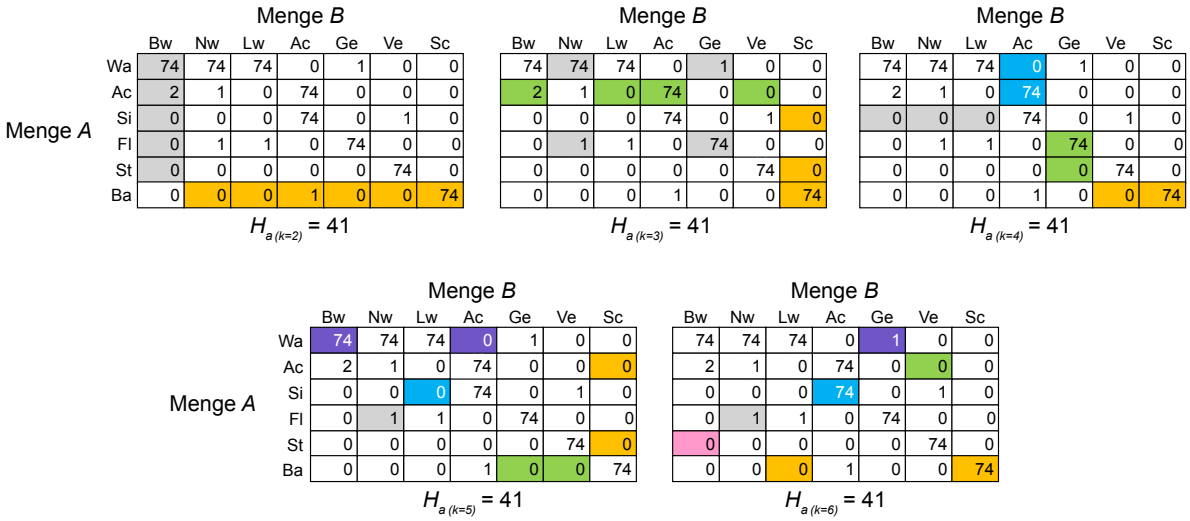


Abbildung 5.4: BalancedSize *-Lösung: Clusterbildung unter Berücksichtigung ausgewogener Clustergrößen bezüglich der Zellenanzahl $H_{a(k)}$ für $k = 2, 3, 4, 5, 6$. Die präsentierte Lösung stellt eine von vielen optimalen Lösungen dar. Zellen, die zu einem Cluster gehören, haben den gleichen farbigen Hintergrund.*

11 Zellen notwendig. Bei $k = 3$ ist immer noch eine Verdoppelung der Zellen zu erkennen. Mit zunehmenden k nähert sich die Anzahl der clusterüberspannenden Zellen weiter an. Bei $k = 6$ sind die Clustergrößen identisch.

Bei allen Partitionierungen wird der gleiche Zielfunktionswert $H_a = 41$ erzielt, da sich jeweils das größte und das kleinste Cluster nur durch eine Zelle unterscheiden. Dieses Verhalten verdeutlicht, dass für die Bestimmung von ausgewogenen Clustern ungeachtet der Häufigkeitswerte viele optimale Lösungen mit unterschiedlichsten semantischen Korrespondenzen möglich sind. Der maximale Zielfunktionswert von $H_a = 42$ kann für dieses Beispiel nicht erreicht werden, da die Anzahl der Elemente in A und B unterschiedlich ist und demzufolge eine Bildung von gleich großen Clustern nicht zulässig ist.

Tabelle 5.3: Vergleich der Clusterbildung unter verschiedenen Bedingungen: MaxScore*;* BalancedSize*;* BalancedScore*. $clsize_k$ gibt die Anzahl der Zellen pro Cluster, $clscore_k$ die einzelnen Clusterhäufigkeiten und $\emptyset H_{Ze}$ die Durchschnittshäufigkeit pro Zelle wieder.*

	MaxScore				BalancedSize		BalancedScore	
k	$H_{(k)}$	$clsize_k$	$clscore_k$	$\emptyset H_{Ze}$	$H_{a(k)}$	$clsize_k$	$H_{b(k)}$	$clscore_k$
2	599	(1|30)	[74|525]	19,32	41	(5|6)	600	[222|222]
3	598	(1|1|20)	[74|74|450]	27,18	41	(3|4|4)	600	[74|74|74]
4	595	(1|1|2|8)	[74|74|148|299]	49,58	41	(2|2|2|3)	600	[0|0|0|0]
5	592	(1|1|1|2|3)	[74|74|74|148|222]	74,00	41	(1|1|2|2|2)	600	[0|0|0|0|0]
6	446	(1|1|1|1|1|2)	[2|74|74|74|74|148]	63,71	41	(1|1|1|1|1|2)	600	[0|0|0|0|0|0]

Ausgewogene Cluster bezüglich der Häufigkeiten (BalancedScore)

Für das Bilden von ausgewogenen Clustern unter Berücksichtigung der Häufigkeitswerte definiert sich die Zielfunktion wie folgt:

$$\text{Maximiere} \quad H_b = H_{\text{total}} - clscore_{\max} + clscore_{\min}, \tag{5.11}$$

mit $H_{\text{total}} = \sum_{e \in E} h_{ij}(e)$ als Gesamthäufigkeitswert der Matrix, $clscore_{\min}$ als kleinstem und $clscore_{\max}$ als größtem Häufigkeitswert aller Cluster. Hierfür werden Variablen eingeführt, die einerseits für jedes Cluster die Häufigkeiten $clscore_k$ aufsummieren und andererseits jeweils das Maximum $clscore_{\max}$ und Minimum $clscore_{\min}$ aller Clusterhäufigkeiten bestimmen:

$$\forall k \in \{1, ..., c\} : \quad clscore_k = \sum_{e \in E} h_{ij}(e) \cdot y_{k,e}, \tag{5.12}$$

$$\begin{aligned} \forall k \in \{1, ..., c\} : \quad & clscore_{\min} \leq clscore_k, \\ & clscore_{\max} \geq clscore_k. \end{aligned} \tag{5.13}$$

Restriktion (5.12) unterscheidet sich von (5.9) allein durch die Berücksichtigung der Häufigkeitswerte bzw. der Kantengewichte $h_{ij}(e)$. Dieser Ansatz wird mit BalancedScore bezeichnet.

Für das synthetische Beispiel sind die optimalen Ergebnisse für BalancedScore in Abbildung 5.5 dargestellt und in Tabelle 5.3 aufgelistet. Für jedes k wurde der maximal mögliche Zielfunktionswert von $H_b = 600$ erzielt. Das bedeutet, dass bei jeder Partitionierung jedes Cluster die gleiche Häufigkeit besitzt. Ab $k = 4$ werden allerdings nur noch Nullcluster gebildet. Das Bilden von ausgewogenen Clustern mit sehr niedrigen Häufigkeiten oder Häufigkeiten 0 ist nicht zielführend für das vorgestellte Zuordnungsproblem. Um ausgewogene Cluster mit hohen Häufigkeiten zu erzeugen, müssen diese Optimierungsziele mit MaxScore kombiniert werden.

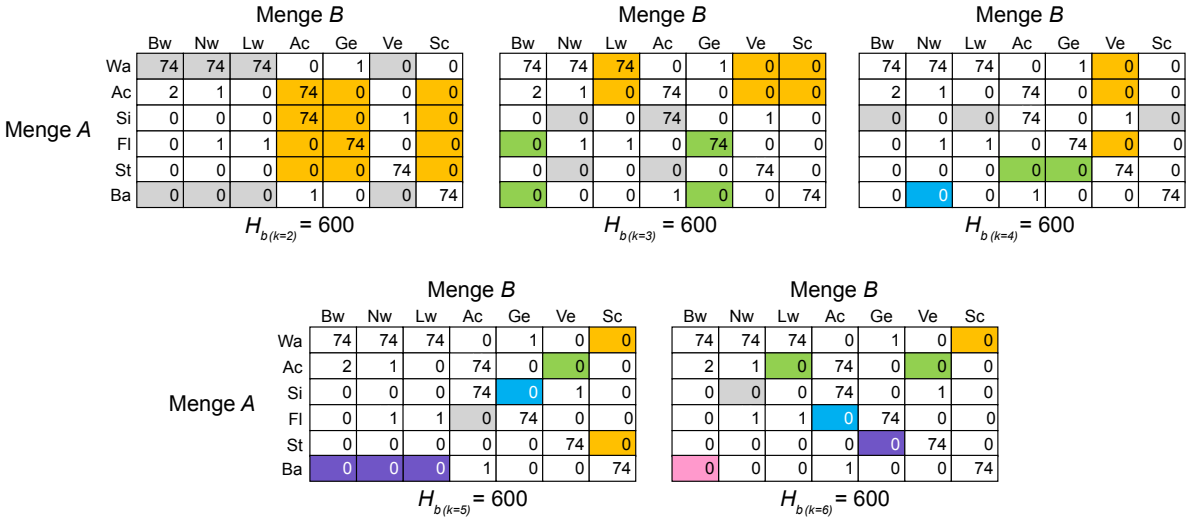


Abbildung 5.5: BalancedScore *-Lösung: Clusterbildung unter Berücksichtigung ausgewogener Clustergrößen bezüglich der Häufigkeiten $H_{b(k)}$ für $k = 2, 3, 4, 5, 6$. Zellen, die zu einem Cluster gehören, haben den gleichen farbigen Hintergrund.*

5.4.2 Kombination von Optimierungszielen

Die Kombination mehrerer Ziele in einem Optimierungsprozess kann auf verschiedene Art und Weise erfolgen: (1) Gewichtung der Ziele gegeneinander oder (2) Einführung einer strikten Bedingung für ein Kriterium. Im Rahmen dieser Arbeit werden beide Möglichkeiten für jeweils zwei der drei genannten Ziele (MaxScore mit BalancedSize bzw. MaxScore mit BalancedScore) untersucht.

Gewichtung der Ziele (WeightedSum)

Die einfachste Kombination von zwei Zielen ist die Addition beider Zielfunktionsterme:

$$\text{Maximiere} \quad H_{Ga} = H + H_a \quad \text{bzw.} \quad H_{Gb} = H + H_b. \tag{5.14}$$

Beide Ziele, das Bilden von Clustern mit maximalen Häufigkeiten und die Erzeugung ausgewogener Cluster, stehen in Konflikt zueinander. Ein Kompromiss zwischen den Zielen kann durch die Einführung eines Gewichtsfaktors $s \in [0, 1]$ erreicht werden. Die Zuordnungsziele ändern sich wie folgt:

$$\text{Maximiere} \quad H_{Ga} = s \cdot H + (1 - s) \cdot H_a \quad \text{bzw.} \quad H_{Gb} = s \cdot H + (1 - s) \cdot H_b. \tag{5.15}$$

Die Gleichungen 5.14 stellen den Spezialfall für $s = 0{,}5$ dar. Die kombinierten Zielfunktionen ergeben sich daraus zu:

$$\text{Maximiere} \quad H_G = s \cdot \sum_{k=1}^{c} \sum_{e \in E} h_{ij}(e) \cdot y_{k,e} + (1 - s) \cdot \begin{cases} (cell_{\text{total}} - clsize_{\max} + clsize_{\min}) & \text{falls } H_a \\ (H_{\text{total}} - clscore_{\max} + clscore_{\min}) & \text{falls } H_b. \end{cases} \tag{5.16}$$

Die Lösungsansätze werden entsprechend mit WeightedSumMaxScoreBalancedSize bzw. WeightedSumMaxScore-BalancedScore bezeichnet.

Abbildung 5.6 a) verdeutlicht geometrisch, wo mögliche optimale Lösungen bei der Kombination von zwei Zielen mittels gewichteter Summe liegen. Jeder schwarze Punkt steht für eine mögliche Lösung. Die Qualität jedes Ziel-

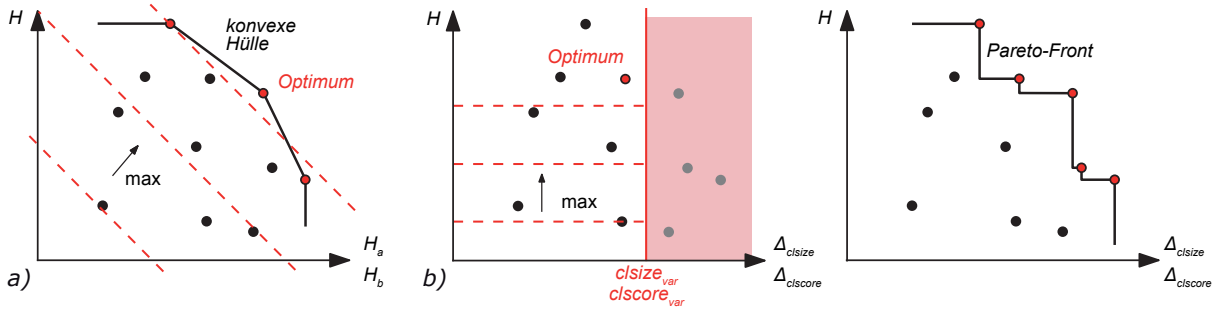


Abbildung 5.6: Geometrische Darstellung der Kombination von zwei Optimierungszielen: a) mittels gewichteter Summe und einem Gewichtsfaktor $s = 0{,}5$, b) durch Maximierung von H bei gleichzeitiger Beschränkung der Clustervariabilität als harte Restriktion, welches entweder die Clustergrößendifferenz Δ_{clsize} bzw. die Clusterhäufigkeitsdifferenz $\Delta_{clscore}$ entspricht. Die roten gestrichelten Linien sind Linien gleicher Qualität.

funktionswertes ist anhand der Position ablesbar. Mit zunehmendem Abstand vom Koordinatenursprung nimmt der Zielfunktionswert in Richtung des Pfeils zu, den es zu maximieren gilt. Für einen beliebigen Gewichtsfaktor stellen alle roten Punkte optimale Lösungen dar, die sich auf der konvexen Hülle befinden, die den Rand des zulässigen Lösungsbereiches beschreibt. Bei der Gleichgewichtung beider Ziele mit $s = 0{,}5$ bildet der rote Punkt mit der Markierung Optimum das beste Ergebnis.

In Tabelle 5.4 und Abbildung 5.7 sind die Ergebnisse für das synthetische Beispiel mit dem Gewichtsfaktor $s = 0{,}5$ zusammengefasst. Bei WeightedSumMaxScoreBalancedSize sind beide Zielfunktionsterme H und H_a unterschiedlich groß. Der MaxScore-Term ist etwa fünfzehn Mal größer als der BalancedSize-Term und wird somit vorrangig optimiert. Im Gegensatz dazu sind bei WeightedSumMaxScoreBalancedScore die Zielfunktionsterme H und H_b bis auf $k = 6$ etwa gleich groß. Das entspricht eher einer Gleichgewichtung.

Tabelle 5.4: Auflistung der einzelnen Zielfunktionsterme für die gleichgewichtete Kombination zweier Optimierungsziele nach Gleichung 5.16 mit $s = 0{,}5$ für WeightedSumMaxScoreBalancedSize *und* WeightedSumMaxScoreBalancedScore*. $clsize$ repräsentiert die Anzahl aller Zellen in den Clustern und $\varnothing H_{Ze} = H/clsize$ gibt die Durchschnittshäufigkeit pro Zelle an.*

	WeightedSumMaxScoreBalancedSize					WeightedSumMaxScoreBalancedScore				
k	H_{Ga}	$=$	$s \cdot H + (1-s) \cdot H_a$	$clsize$	$\varnothing H_{Ze}$	H_{Gb}	$=$	$s \cdot H + (1-s) \cdot H_b$	$clsize$	$\varnothing H_{Ze}$
2	317,5	$=$	$0{,}5 \cdot 597 + 0{,}5 \cdot 38$	20	29,85	598,0	$=$	$0{,}5 \cdot 597 + 0{,}5 \cdot 599$	20	29,85
3	316,0	$=$	$0{,}5 \cdot 593 + 0{,}5 \cdot 39$	13	45,62	559,0	$=$	$0{,}5 \cdot 592 + 0{,}5 \cdot 526$	14	42,29
4	315,5	$=$	$0{,}5 \cdot 592 + 0{,}5 \cdot 39$	10	59,20	523,0	$=$	$0{,}5 \cdot 519 + 0{,}5 \cdot 527$	10	51,90
5	316,0	$=$	$0{,}5 \cdot 592 + 0{,}5 \cdot 40$	8	74,00	522,5	$=$	$0{,}5 \cdot 519 + 0{,}5 \cdot 526$	8	64,88
6	243,5	$=$	$0{,}5 \cdot 446 + 0{,}5 \cdot 41$	7	63,71	451,0	$=$	$0{,}5 \cdot 373 + 0{,}5 \cdot 529$	7	53,29

Durch die Veränderung des Gewichtsfaktors kann der Anwender Einfluss auf das zu optimierende Ziel nehmen. Die Wahl des richtigen Gewichtsfaktors ist schwierig, weil er stark von den Häufigkeitswerten in der Matrix beeinflusst ist und für jeden Datensatz neu ermittelt werden muss. Dies ist ein entscheidener Nachteil für diesen Lösungsansatz. Damit H und H_a bzw. H_b gleich groß sind und so den gleichen Einfluss auf das Ergebnis haben, können normierte Mittelwerte als Gewichtsfaktoren eingesetzt werden. Gleichung 5.15 verändert sich somit zu:

$$\begin{aligned} &\text{Maximiere} \quad H_{Ga_{Mean}} = \frac{m_{H_a}}{(m_H + m_{H_a})} \cdot H + \frac{m_H}{(m_H + m_{H_a})}) \cdot H_a \quad \text{bzw.} \\ &\text{Maximiere} \quad H_{Gb_{Mean}} = \frac{m_{H_b}}{(m_H + m_{H_b})} \cdot H + \frac{m_H}{(m_H + m_{H_b})}) \cdot H_b. \end{aligned} \tag{5.17}$$

mit m_H als Mittelwert aller Einzelwerte von H über alle k und entsprechend m_{H_a} als Mittelwert aller Einzelwerte von H_a und m_{H_b} von H_b.

Bei stark unterschiedlichen Häufigkeitswerten ist die Maximierung der Gesamthäufigkeit unter Berücksichtigung ausgewogener Cluster hinsichtlich der Zellenanzahl einfacher als hinsichtlich der Häufigkeiten, da sowohl Nullcluster als auch Cluster mit großen Häufigkeitswerten zulässig sind. Dagegen können Cluster mit ausgewogenen Häufigkeiten nur erzeugt werden, wenn entweder große Häufigkeitswerte vernachlässigt oder viele Häufigkeitswerte zusammengefasst werden.

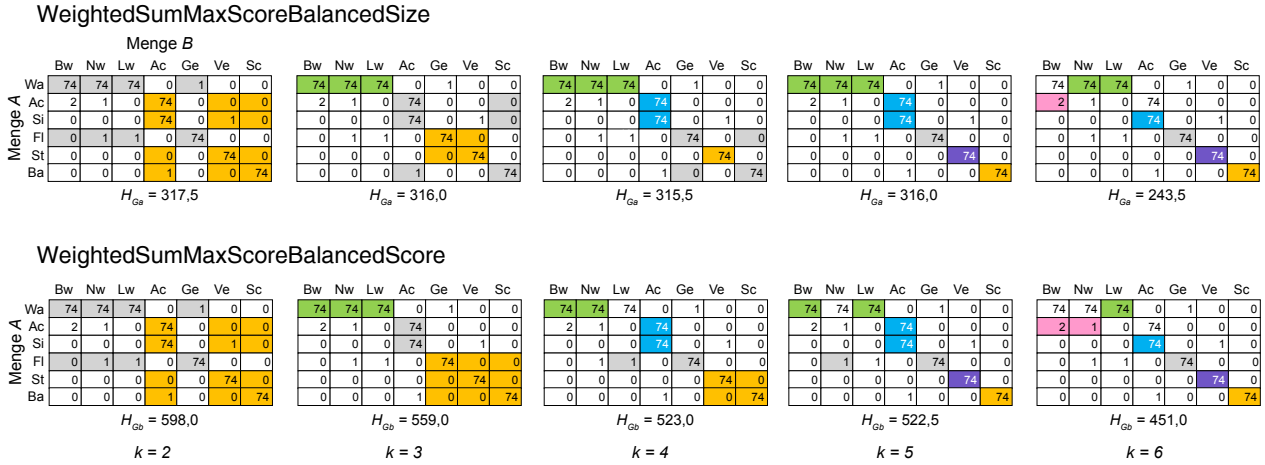


Abbildung 5.7: Clusterbildung bei gleichgewichteter Kombination von zwei Optimierungszielen mit dem Gewichtsfaktor $s = 0{,}5$*:* WeightedSumMaxScoreBalancedSize *(oben) und* WeightedSumMaxScoreBalancedScore *(unten).*

Für die Bewertung der Ergebnisse bestimmte ein Experte für das synthetische Beispiel eine Referenzzuordnung mit fünf Clustern. Diese Zuordnung stimmt mit der WeightedSumMaxScoreBalancedSize-Lösung bei $k = 5$ überein. Um diese Zuordnung automatisch als beste aus allen optimalen Lösung zu erkennen, wird im nächsten Abschnitt ein übergeordnetes Optimierungsziel vorgestellt.

Übergeordnetes Optimierungsziel

Für die bessere Beurteilung der erzielten Ergebnisse wird die Durchschnittshäufigkeit pro Zelle $\varnothing H_{Ze}$ als übergeordnetes Optimierungsziel eingeführt:

$$\varnothing H_{Ze} = H/clsize. \tag{5.18}$$

Das Verhältnis zwischen der Gesamthäufigkeit aller Cluster H und der Anzahl der von den Clustern überdeckten Zellen $clsize$ entspricht einer Normierung. Es wird ein hoher Wert erzielt, wenn wenige Clusterzellen mit großen Häufigkeitswerten als Zuordnungslösung zusammengefasst werden. Allüberdeckende Cluster wirken sich dagegen negativ aus.

In Tabelle 5.4 sind alle notwendigen Kenngrößen für das synthetische Beispiel aufgelistet. Sowohl bei WeightedSumMaxScoreBalancedSize als auch bei WeightedSumMaxScoreBalancedScore werden die maximalen Durchschnittshäufigkeiten pro Zelle bei den Lösungen mit fünf Clustern erzielt. Beide Zuordnungslösungen stimmen in sieben von acht Zellen überein (siehe Abb. 5.7). In zwei Clustern ergeben sich Unterschiede: Wa $\rightarrow$ {Bw, Nw, Lw} und Fl $\rightarrow$ Ge bei WeightedSumMaxScoreBalancedSize gegenüber Wa $\rightarrow$ {Bw, Lw} und Fl $\rightarrow$ {Nw, Ge} bei WeightedSumMaxScoreBalancedScore, die auf den Einfluss des zweiten Optimierungsziels zurückzuführen sind. Durch die verschiedenartige Zuordnung der Klasse Nadelwald wird das Kriterium der Clusterausgewogenheit geändert.

Harte Bedingung (HardConstraint)

Die zweite Möglichkeit, Optimierungsziele zu kombinieren, ist, für eines der Ziele eine harte Bedingung einzuführen. Das Optimierungsproblem verändert sich zu: Maximiere die Gesamthäufigkeit unter der Bedingung, dass (a) die Differenz der Zellanzahl des größten und kleinsten Clusters $\Delta_{clsize} = clsize_{\max} - clsize_{\min}$ oder (b) die Differenz der Clusterhäufigkeiten des größten und kleinsten Clusters $\Delta_{clscore} = clscore_{\max} - clscore_{\min}$ kleiner sein muss als ein bestimmter Wert, der hier als Clustervariabilität $clsize_{var}$ für (a) und $clscore_{var}$ für (b) bezeichnet wird. Diese Lösungsansätze werden entsprechend mit MaxScoreHardConstraintVariableSize bzw. MaxScoreHardConstraintVariableScore bezeichnet.

Abbildung 5.6 b) zeigt die geometrische Interpretation des Optimierungsproblems. Im linken Teil der Abbildung sind alle Lösungen gültig, die sich links der roten Schranke befinden, die die festgesetzte Clustervariabilität markiert. Der rote Punkt repräsentiert das optimale Ergebnis. Für eine beliebige Wahl des Variabilitätsparameters können im Gegensatz zur gewichteten Summe verschiedene Lösungen erreicht werden. Im rechten Teil der Abbildung sind alle rot markierten Punkte sogenannte nicht dominierte Punkte, die hinsichtlich der beiden Kriterien nicht übertroffen werden können. Diese Punkte bilden die Pareto-Front und die Lösungen werden als Pareto-optimale Lösungen bezeichnet (Ehrgott, 2005). Dieser Parameter ist im Vergleich zum Gewichtsfaktor s leichter zu interpretieren und für den Optimierungsprozess von einem Anwender einfacher festzulegen.

Für die Lösung der beiden Optimierungsprobleme ist die Zielfunktion aus Gleichung 5.3 weiterhin gültig. Zusätzlich bleiben die Restriktionen (5.4) bis (5.7) erhalten. Ausgewogene Cluster hinsichtlich der Zellgrößen erfordern die Restriktionen (5.9), (5.10) und (5.19). Für ausgewogene Cluster bezüglich der Häufigkeiten sind (5.12), (5.13) und (5.20) zu berücksichtigen:

$$clsize_{\max} - clsize_{\min} \leq clsize_{var} \quad \text{bzw.} \tag{5.19}$$

$$clscore_{\max} - clscore_{\min} \leq clscore_{var}. \tag{5.20}$$

Für das synthetische Beispiel werden im oberen Teil der Tabelle 5.5 die Ergebnisse für MaxScoreHardConstraintVariableSize gezeigt.

Tabelle 5.5: Ergebnisse für MaxScoreHardConstraintVariableSize *(oben) und* MaxScoreHardConstraintVariableScore *(unten) für das synthetische Beispiel. Die Clustervariabilität $clsize_{var}$ bzw. $clscore_{var}$ wurde als harte Bedingung eingefügt. Die mit $*$ gekennzeichneten Zielfunktionswerte repräsentieren die* MaxScore*-Lösung. Die grau hervorgehobenen Lösungen repräsentieren die optimalen Lösungen mit der höchsten Durchschnittshäufigkeit pro Zelle $\varnothing H_{Ze}$.*

	k									
	2		3		4		5		6	
$clsize_{var}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$
29	$*599$	19,32								
28	598	24,92								
19			$*598$	27,18						
18			596	39,73						
15	597	29,85								
7					$*595$	49,58				
6			595	42,50	593	53,91				
5			593	45,62						
4					592	59,20				
3	595	28,33								
2	523	29,06	523	37,36	519	51,90	$*592$	74,00		
1	372	33,82	519	47,18	518	57,56	519	64,88	$*446$	63,71

	k									
	2		3		4		5		6	
$clscore_{var}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$	H	$\varnothing H_{Ze}$
451	$*599$	19,32								
450	598	24,92								
376			$*598$	27,18						
375			596	27,09						
373			596	39,73						
301	597	29,85								
225					$*595$	49,58				
224			595	42,50	593	53,91				
150			593	45,62						
148					592	59,20	$*592$	74,00		
147					520	47,27	520	57,78		
146									$*446$	63,71
145									373	53,29
75					519	47,18	519	64,75		
74			592	42,29	519	51,90				
73			520	40,00			374	46,75		
72			449	32,07	302	33,56			373	53,29
70									7	1,00
1			448	32,00			373	46,63	5	0,71

Ausgehend von den mit $*$ gekennzeichneten MaxScore-Lösungen, werden für jedes k die Schrankenwerte $clsize_{var}$ bestimmt. Bei $k = 2$ liefert die MaxScore-Optimierung Cluster der Größe (1|30). Somit beträgt die Clusterdifferenz $\Delta_{clsize} = 30 - 1 = 29$. Im nächsten Berechnungsschritt wird $clsize_{var}$ auf 28 begrenzt, woraufhin Cluster der Größe (4|20) entstehen. Die Lösung $H = 598$ ist bis einschließlich $clsize_{var} = 16$ gültig. Die nächste Schranke wird auf $clsize_{var} = 15$ verringert. Diese Vorgehensweise wird für alle k und alle notwendigen $clsize_{var}$

durchgeführt. Insgesamt sind hier 20 Berechnungsschritte notwendig. Die grau hinterlegte Lösung besitzt bei $k = 5$ und $clsize_{var} = 2$ den höchsten Wert für $\varnothing H_{Ze}$ und entspricht gleichzeitig der Referenzlösung.

Im unteren Teil der Tabelle 5.5 sind die Ergebnisse für MaxScoreHardConstraintVariableScore gezeigt. Die Vorgehensweise zur Berechnung ist ähnlich. Ausgehend von der MaxScore-Lösung wird die Schranke $clscore_{var}$ schrittweise um 1 verringert, da in der Häufigkeitsmatrix des synthetischen Beispiels nur ganzzahlige Werte vorhanden sind. Bei Fließkommazahlen muss die Reduzierung der Schranke auf die Genauigkeit der verwendeten Zahlen angepasst werden. Die Anzahl der notwendigen Berechnungen erhöht sich dadurch. Im Vergleich zur MaxScoreHardConstraintVariableSize-Optimierung waren 45 % mehr Berechnungen notwendig. Die größte Durchschnittshäufigkeit pro Zelle wird bei $k = 5$ und $clscore_{var} = 148$ erzielt.

Das Verfahren MaxScoreHardConstraintVariableScore wird für die Anwendung in der Praxis nicht empfohlen, da alle Häufigkeitsmatrizen der realen Testgebiete Fließkommazahlen beinhalten und dadurch noch mehr Berechnungsschritte zu erwarten sind. Allerdings wurde auch die Rechenzeit bei der Anwendung von MaxScoreHardConstraintVariableSize auf reale Testdaten zum Problem. Die variable Clustergröße $clsize_{var}$ verursachte mitunter hohe Rechenzeiten bzw. verhindert sogar eine garantiert optimale Lösbarkeit des Problems. Somit schränkt das Rechenzeitproblem den Einsatz des Algorithmus in der Praxis stark ein. Demzufolge wird im folgenden Abschnitt eine Vereinfachung des Lösungsverfahren vorgeschlagen.

5.4.3 Einführung einer festen Clustergröße (MaxScoreHardConstraintFixedSize)

Das Verfahren MaxScoreHardConstraintVariableSize wird dahingehend verändert, dass die in Gleichung (5.19) eingeführte Clustervariabilität durch einen fest vorgegebenen Wert ersetzt wird. Somit wird nicht mehr der Unterschied zwischen dem größten und kleinsten Cluster, sondern die Gesamtanzahl $clsize$ der in allen Clustern verwendeten Zellen beschränkt. Die Bedingungen (5.9) und (5.10) müssen dafür entfernt und (5.19) durch folgende Bedingung ersetzt werden:

$$\forall k \in \{1, ..., c\} : \quad \sum_{e \in E} y_{k,e} \leq clsize \ . \tag{5.21}$$

Das Optimierungsproblem wird mit MaxScoreHardConstraintFixedSize bezeichnet. Für die Identifikation der besten Lösung müssen viele Kombinationen für die Parameter k und $clsize$ berechnet werden, wobei nicht alle $clsize$-Werte sinnvoll sind. Aus der Matrix-Dimension und der Clusteranzahl k lassen sich obere und untere Schrankenwerte für jedes k festlegen, um Rechenschritte zu begrenzen. Die Schranken bestimmen sich wie folgt:

$$\begin{aligned} \forall k \in \{1, ..., c\} : \quad & clsize_{o_k} = (|A| - k + 1) \cdot (|B| - k + 1) + (k - 1) \\ & clsize_{u_k} = |A| + |B| - k \ . \end{aligned} \tag{5.22}$$

In Tabelle 5.6 sind die Werte $clsize_o$ und $clsize_u$ für das synthetische Beispiel zusammengefasst. Es sind maximal 45 Berechnungen notwendig.

Tabelle 5.6: Schranken für $clsize_o$ und $clsize_u$ für das synthetische Beispiel.

k	$clsize_o$	$clsize_u$	$\Delta clsize_o clsize_u + 1$
2	31	11	21
3	22	10	13
4	15	9	7
5	10	8	3
6	7	7	1
			$\sum$ 45

Tabelle 5.7 zeigt für das synthetische Beispiel die Ergebnisse für MaxScoreHardConstraintFixedSize. Für die Bestimmung der besten Lösung waren 21 Berechnungsschritte notwendig. Durch die Einführung der festen Clustergesamtgröße ($clsize$) gegenüber der variablen Clustergröße ($clsize_{var}$) konnte die Rechenzeit für das synthetische Beispiel um 30 % reduziert werden.

Während der Untersuchungen wurde festgestellt, dass viele Lösungen mit Nullcluster bestimmt werden. Für die Interpretation der Ergebnisse wird angenommen, dass eine Zuordnungslösung ohne Nullcluster besser geeignet ist, da so sichergestellt wird, dass jede Objektklasse einen Zuordnungskandidaten besitzt, dessen Korrespondenz gleichzeitig durch Objektrelationen bestätigt wird. Um dieser Forderung gerecht zu werden, wird das Verfahren MaxScoreHardConstraintFixedSize erweitert und im nächsten Abschnitt vorgestellt.

Tabelle 5.7: Ergebnisse für MaxScoreHardConstraintFixedSize *für das synthetische Beispiel. Die mit ∗ gekennzeichneten Zielfunktionswerte repräsentieren die* MaxScore*-Lösung und clsize ist die maximale Anzahl der Zellen, die zur Clusterbildung erlaubt sind. Die grau hervorgehobene Lösung repräsentiert die optimale Lösung mit der höchsten Durchschnittshäufigkeit pro Zelle $\emptyset H_{Ze}$.*

	k									
	2		3		4		5		6	
$clsize$	H	$\emptyset H_{Ze}$	H	$\emptyset H_{Ze}$	H	$\emptyset H_{Ze}$	H	$\emptyset H_{Ze}$	H	$\emptyset H_{Ze}$
31	∗599	19,32								
30	598	24,92								
23	597	29,85								
22			∗598	27,18						
21			596	39,73						
19	523	29,06								
17	448	28,00								
15	447	29,80			∗595	39,67				
14	372	33,82	595	42,50	595	49,58				
13			593	45,62						
12			520	43,33						
11			519	47,18	593	53,91				
10			446	44,60	592	59,20	∗592	74,00		
9					519	57,67				
7									∗446	63,71

5.4.4 Optimale Lösung ohne Nullcluster (MaxScoreHardConstraintFixedSizeNonEmpty)

Um Nullcluster in Lösungen zu unterbinden, wird folgende Bedingung hinzugefügt, die in jedem Cluster eine Summe der Häufigkeiten größer Null fordert:

$$\forall k \in \{1, ..., c\}, \forall y_e \cdot w(e) > 0 : \quad \sum_{e \in E} y_{k,e} \geq 1 \ . \tag{5.23}$$

Das Optimierungsproblem wird mit MaxScoreHardConstraintFixedSizeNonEmpty bezeichnet. Für das synthetische Beispiel ist diese zusätzliche Bedingung nicht notwendig, da bereits bei MaxScoreHardConstraintFixedSize kein Nullcluster Bestandteil der Lösung ist.

In den Experimenten mit realen Daten wurde hingegen festgestellt, dass die eingeführte Restriktion sehr hohe Rechenzeiten verursacht. Um Rechenzeit zu reduzieren, könnten Rechenschritte an anderer Stelle eingespart werden. Aus Tabelle 5.6 wird deutlich, dass für verschiedene k identische $clsize$-Werte gültig sind und daher mehrfach zu bestimmen sind. Die Berechnung von $clsize = 10$ ist für $k = 3$, $k = 4$ und $k = 5$ erforderlich. Insgesamt werden vier Schrankenwerte ($clsize = 15, 14, 11, 10$) mehrmals für verschiedene k berechnet. Ein Programm, das alle $clsize$-Werte unabhängig von k nur einmal testet, ist gerade in Hinblick auf große Häufigkeitsmatrizen effizienter.

Im Rahmen dieser Arbeit wird für diesen Fall ein vereinfachtes, ganzzahliges lineares Programm (MaxScoreHardConstraintFixedSizeUnique) entwickelt, das jedoch geringfügig veränderte Variablen und Bedingungen beinhaltet. Die Ergebnisse von (MaxScoreHardConstraintFixedSizeUnique) sind mit denen von (MaxScoreHardConstraintFixedSize) identisch. Das Verfahren wird im nächsten Abschnitt vorgestellt.

5.4.5 Vereinfachtes Programm (MaxScoreHardConstraintFixedSizeUnique)

Für die vereinfachte Modellierung des Problems wird, statt der Knoten und Kanten wie in Abschnitt 5.4.1, eine Variable $x_{a,b} \in \{0, 1\}$ für jede Zelle der Häufigkeitsmatrix H definiert. $x_{a,b} = 1$ bedeutet, dass die Zelle Bestandteil der Lösungscluster ist. Es wird aber nicht explizit modelliert, zu welchem Cluster eine Zelle gehört. Damit ist es nicht möglich, Restriktionen an die Eigenschaften der Cluster (Größe und Variabilität) zu knüpfen. Die lineare Zielfunktion ändert sich zu:

$$\text{Maximiere} \quad H(x) = \sum_{a \in A} \sum_{b \in B} w(a, b) \cdot x_{a,b} \quad \forall w(a, b) \cdot x_{a,b} > 0 \ . \tag{5.24}$$

Außerdem müssen folgende Restriktionen gelten:

$$\sum_{a \in A} \sum_{b \in B} x_{a,b} \leq clsize \tag{5.25}$$

$$\forall a \in \{1, ..., n\} : \quad \sum_{b \in B} x_{a,b} \geq 1, \quad \forall b \in \{1, ..., m\} : \quad \sum_{a \in A} x_{a,b} \geq 1 \quad \text{und} \tag{5.26}$$

$$\forall a \in A, \forall b \in B, \forall u \in A, \forall v \in B : \quad x_{a,b} + x_{u,v} + x_{a,v} - x_{u,b} \leq 2 \ . \tag{5.27}$$

Restriktion (5.25) garantiert, dass die Anzahl der Zellen im Lösungscluster nicht den fest vorgegebenen $clsize$-Wert übersteigt. Die Bedingungen (5.26) gewährleisten, dass jede Objektklasse der Mengen A und B mindestens einmal zugeordnet wird. Die Definition der Restriktion (5.27) stellt sicher, dass auch komplexe n:m-Zuordnungen möglich sind und in Form von rechteckigen Clustern gebildet werden, d.h. wenn die Zellen x_{a_1,b_1}, x_{a_1,b_2} und x_{a_2,b_1} gewählt wurden, muss auch x_{a_2,b_2} Teil des Clusters sein.

Algorithmus 1 zeigt den genauen Berechnungsablauf in Form von Pseudocode. Zunächst wird $clsize$ auf die maximale Zellenanzahl gesetzt, die sich aus dem Produkt der Zeilen $|A|$ und Spalten $|B|$ der Häufigkeitsmatrix H ergibt. Damit wird das oben beschriebene vereinfachte MaxScoreHardConstraintFixedSizeUnique-Verfahren gelöst. Die erzielte optimale Lösung umfasst Cluster mit einer Zellenanzahl z und einen Gesamtzielfunktionswert $score$. Daraus wird die Durchschnittshäufigkeit pro Zelle $\emptyset H_{Ze}$ abgeleitet. In der nächsten Iteration entspricht $clsize$ dem um 1 verringerten Wert von z. Diese Vorgehensweise wird solange durchgeführt, bis $clsize$ dem Maximalwert der Zeilen bzw. Spalten entspricht. Die Lösung mit dem größten $\emptyset H_{Ze}$ wird als optimale Lösung für das Zuordnungsproblem ausgewählt. Als Ergebnis wird für jede Zelle angegeben, ob sie Teil der Zuordnungslösung ist oder nicht. Zusätzlich werden der beste Gesamtzielfunktionswert $bestscore$ und die dazugehörige Zellanzahl $bestz$ aufgeführt.

Algorithmus 1 Vereinfachtes Programm MaxScoreHardConstraintFixedSizeUnique

```
1: |A| ← Anzahl der Zeilen von H
2: |B| ← Anzahl der Spalten von H
3: clsize ← Anzahl der Zellen, die die Lösungscluster beschreiben
4: clsize ← |A| · |B|
5: bestØH_Ze ← H/clsize
6: bestz ← clsize
7: bestscore ← H
8: while clsize ≥ max(|A|, |B|) do
9:     solve MaxScoreHardConstraintFixedSizeUnique (H(A, B), clsize) return score, z
10:    ØH_Ze ← score/z
11:    if ØH_Ze > bestØH_Ze then
12:        bestØH_Ze ← ØH_Ze
13:        bestz ← z
14:        bestscore ← score
15:    end if
16:    clsize ← z − 1
17: end while
18: return bestz, bestscore, bestØH_Ze, x_{a,b}
```

Experimente zeigten, dass sich das vereinfachte Programm MaxScoreHardConstraintFixedSizeUnique im Gegensatz zu MaxScoreHardConstraintFixedSize durch die geringere Anzahl an Berechnungen deutlich schneller lösen lässt. Für das synthetische Beispiel konnte die optimale Lösung mit acht statt 21 Berechnungen in einem Viertel der Zeit bestimmt werden.

6 Experimente mit Realdaten und Untersuchungsergebnisse

In diesem Kapitel werden Untersuchungsergebnisse vorgestellt, die zum einen mit dem Objektzuordnungsverfahren für Polygone aus Kapitel 4 und zum anderen mit den in Kapitel 5 beschriebenen Schema-Matching-Ansätzen erzielt wurden. Die Verfahren werden auf drei Testgebiete mit unterschiedlichen Datensätzen angewendet. Abschnitt 6.1 stellt die Datensätze und Testgebiete vor und beschreibt notwendige Vorverarbeitungsschritte. Die statistischen Ergebnisse der Objektzuordnung werden in Abschnitt 6.2 gesondert für jedes Untersuchungsszenario präsentiert und analysiert. In Abschnitt 6.3 werden die Ergebnisse der Schema-Matching-Verfahren präsentiert, einander gegenübergestellt und Vor- und Nachteile der Verfahren diskutiert.

6.1 Datenquellen und Datenvorverarbeitung

Im Rahmen der vorliegenden Arbeit werden vier verschiedene Datenquellen verwendet. Die Auswahl der Testdaten war primär an die Bedingung geknüpft, dass Real-Welt-Objekte des gleichen räumlichen Gebiets durch Polygonobjekte beschrieben werden. Die geometrische Auflösung der Daten ist unterschiedlich und reicht von 1:1.000 bis 1:25.000. Die Datenquellen werden im nächsten Abschnitt vorgestellt, und es wird auf geometrische und semantische Besonderheiten hingewiesen. Anschließend werden die Testgebiete präsentiert.

6.1.1 Datenquellen

ALKIS

Das Amtliche Liegenschaftskatasterinformationssystem (ALKIS®) führt auf Basis des von der AdV[1] entwickelten AFIS-ALKIS-ATKIS-Referenzmodells (AAA-Modell) das Liegenschaftsbuch (ALB), den beschreibenden textlichen Teil, und die Liegenschaftskarte (ALK), den darstellenden graphischen Teil, in einem System zusammen. Es gewährleistet das Eigentum aus Artikel 14 Grundgesetz, da neben personenbezogenen Daten auch Flurstücksgrenzen, Gebäude und bestehende Nutzungsarten verbindlich beschrieben werden. Im Rahmen der vorliegenden Arbeit werden Informationen der Liegenschaftskarte verwendet. Die räumlichen Objekte werden vorwiegend mit terrestrischen Messverfahren erfasst, z.B. mittels Tachymetrie, und besitzen eine Genauigkeit im Zentimeter-Bereich. Die Objektdaten liegen im Referenzmaßstab von 1:1.000 überwiegend als Polygonobjekte vor. ALKIS-Objekte werden nach dem ALKIS-Objektartenkatalog klassifiziert, dessen Grundlage die GeoInfoDok 6.0.1 ist. Der Katalog gliedert sich in Objektbereiche (Ebene 1), die aus Objektartengruppen (Ebene 2) bestehen, die sich wiederum aus Objektarten (Ebene 3) zusammensetzen (AdV, 2008). In den verschiedenen Testgebieten werden unterschiedliche ALKIS-Objektartengruppen verwendet, die bei der Vorstellung der einzelnen Testgebiete in Abschnitt 6.1.2 genauer spezifiziert werden.

ATKIS

Das Amtliche Topographisch-Kartographische Informationssystem (ATKIS®) beschreibt auf Basis des AAA-Modells in einer geotopographischen Datenbasis die Topographie für ganz Deutschland und stellt sie in Form von digitalen Erdoberflächenmodellen, wie z.B. Landschaftsmodellen, Geländemodellen, Topographischen Karten und Orthophotos bereit. Im Rahmen dieser Arbeit werden Daten im Maßstab 1:10.000 verwendet, die dem Basis-Landschaftsmodell zuzuordnen sind. Aufgrund des kleineren Maßstabs werden bestimmte Objektgruppen, wie z.B. Straßen-, Eisenbahn- und Gewässernetze als Linien-Objekte repräsentiert. Für die vorliegende Untersuchung wird die Objektvielfalt jedoch auf Polygonobjekte begrenzt.

ATKIS-Objektinformationen werden vorwiegend aus der Interpretation von Luftbildern bzw. der Topographischen Karte, durch die Meldung von Veränderungsverursachern (Energieversorger, Deutsche Bahn, ÖPNV-Betreiber uvm.) oder durch Generalisierung gewonnen. Die Vorgabe der Lagegenauigkeit beträgt ± 3 m. Vielerorts werden linien- und punktförmige Objekte deutlich genauer erfasst. Landschaftsflächen werden in Form von verschiedenen Nutzungsarten dargestellt, wie z.B. Wohnbau-, Industrie- und Gewerbeflächen, Grünland und

[1] Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland – http://www.adv-online.de/

Wald. Für die Modellierung der tatsächlichen Nutzung werden hier, im Gegensatz zu ALKIS, die Grundstücksgrenzen nicht exakt berücksichtigt. Dadurch entstehen Unterschiede zwischen Objekten beider Datensätze.

OSM

Das OpenStreetMap-Projekt (OSM) wurde 2004 mit dem Ziel gegründet, allein aus Daten, die durch Freiwillige gesammelt werden, eine frei verfügbare digitale Karte der Welt herzustellen. In diesem Fall erzeugen Laien mit ihren mobilen Endgeräten (GPS-Empfänger oder Smartphones) Geodaten. Der OSM-Datensatz beinhaltet vielfältige Informationen. Eine Übersicht über den Attributkatalog des OSM-Projekts ist in OpenStreetMap (2019) zu finden. Der Attributkatalog stellt im Gegensatz zum AAA-Objektkatalog keine Vorschrift dar, sondern ist als Leitfaden zu verstehen. Neben sehr detaillierten Straßenverkehrsdaten – von Autobahnen über Wohngebietsstraßen bis hin zu Fußwegen – und verschiedenen Landnutzungsflächen, werden vor allem Gebäudeinformationen erfasst und annotiert. Detaillierte Gebäudegrundrisse existieren bereits für viele Städte. Aktuell werden die Geodaten nicht mehr allein durch Freiwillige im Feld erfasst und kartiert, sondern zunehmend von analogen Karten und aus Luftbildern digitalisiert oder sogar von lokalen Stadtverwaltungen kostenlos zur Verfügung gestellt.

Die Vollständigkeit und Genauigkeit der OSM-Daten ist sehr heterogen. Dies wird in zahlreichen Studien insbesondere durch Vergleiche mit amtlichen Daten festgestellt: Girres und Touya (2010); Haklay (2010); Zielstra und Zipf (2010); Mondzech und Sester (2011); Koukoletsos u. a. (2012); Neis u. a. (2012); Fan u. a. (2014). In nahezu allen Untersuchungen wird festgestellt, dass sowohl die Datendichte als auch der Detaillierungsgrad in großen Städten höher sind als in ländlichen Gebieten. Die Autoren begründen dies mit der höheren Anzahl an aktiven Projektmitgliedern in den Städten. Bauliche Veränderungen bei Gebäuden, wie z.B. Neubau, Abriss oder Anbau werden durch OSM-Mitglieder in der Nachbarschaft viel schneller wahrgenommen und erfasst. Demzufolge tragen sie zur Erhöhung der Aktualität des Karteninhaltes bei. Im Rahmen der vorliegenden Arbeit werden aus dem OSM-Datensatz ausschließlich Gebäudeobjekte verwendet.

GDF

Die Daten der Firma TomTom® liegen im Geographic Data Files-Format vor und werden im weiteren Verlauf der Arbeit kurz als GDF-Daten bezeichnet. Die Geodaten werden speziell für Zwecke der Fahrzeugnavigation erhoben und liegen für viele Gebiete in Westeuropa im Maßstab 1:25.000 vor. Aus diesem Grund weist das Straßennetz eine besondere Detaillierungsstufe auf. Neben Verkehrsdaten sind auch topographische Objekte enthalten, die in dieser Arbeit untersucht werden. Vermutlich werden sie nur als Hintergrundinformationen verwendet und sind somit weniger genau.

6.1.2 Testgebiete

Testgebiet A: ALKIS - OSM in Hannover

Das erste Testgebiet befindet sich in der Niedersächsischen Landeshauptstadt Hannover und überdeckt eine Fläche von 24 km^2 (4×6 km). Bei dieser Untersuchung werden ausschließlich, die in Abbildung 6.1 dargestellten Gebäudegrundrisse der ALKIS- und OSM-Daten verwendet.

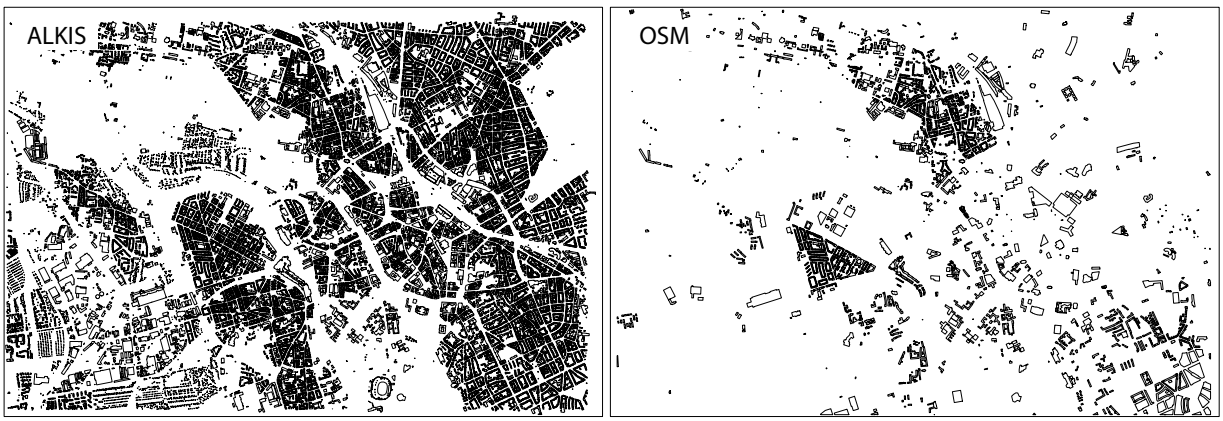


Abbildung 6.1: Testgebiet A: Gebäudeobjekte der Datensätze ALKIS (links) und OSM (rechts) in Hannover.

Gebäude wurden ausgewählt, weil sie aufgrund ihrer geringen Größe und der klaren Abgrenzung zu anderen Objekten einfach von Laien zu erfassen und zu digitalisieren sind. Abweichungen aufgrund unterschiedlicher Erfassungsmethoden werden als relativ klein eingeschätzt. Bei der Objektzuordnung werden keine großen Probleme erwartet.

Im Testgebiet stehen 23.663 ALKIS-Objekte 2.689 OSM-Objekten gegenüber. Im Zuordnungsprozess bringt diese Ungleichheit viele 1:0-Objektrelationen hervor. Bei einer Qualitätsuntersuchung des OSM-Datensatzes hinsichtlich der Vollständigkeit, nach Vorbild von Girres und Touya (2010), Zielstra und Zipf (2010) und Fan u. a. (2014) wirkt sich dies negativ aus. Dennoch können auch nicht zugeordnete Objekte hilfreich sein, da sie Hinweise auf veränderte Situationen (engl. Change Detection) liefern, z.B. auf neu gebaute bzw. abgerissene Gebäude.

Alle anderen Objektrelationen werden im Rahmen dieser Arbeit dazu genutzt, um semantische Korrespondenzen zwischen den Gebäudekategorien beider Datensätze abzuleiten. Mit den identifizierten Schemarelationen können Transformationsregeln definiert werden, die eine integrierte Nutzung aller Objektinformationen beider Datensätze oder eine Attributübertragung zwischen Objekten auch außerhalb des Testgebiets ermöglichen. Im Gegensatz zur Objektzuordnung wird die Zuordnung der Gebäudekategorien als schwieriger eingeschätzt, da sich die Anzahl der Objektklassen stark unterscheidet, was die Kombinationsmöglichkeiten der Objektklassen erhöht. Im Anhang geben A.1 und A.2 eine Übersicht der im Testgebiet vorhandenen 49 ALKIS- und 72 OSM-Objektklassenbezeichnungen und deren Objektanzahlen.

Während ALKIS-Gebäude mit Hilfe eines abgestimmten Objektartenkatalogs klassifiziert werden, der eine begrenzte Anzahl von Gebäudetypen detailliert beschreibt, gibt es für OSM-Daten kein verbindliches Regelwerk. Gebäudeobjekte können vom Erfasser mit dem Attribut *building* gekennzeichnet werden. Detaillierte Beschreibungen wie z.B. die Gebäudekategorie können wiederum mit Hilfe des Attributs *type* angegeben werden. Hierfür stehen einige Kategorien zur Auswahl, z.B. *type=Hotel, type=Church, type=Residential*. Allerdings steht es dem Erfasser frei, auch eigene Kategorien festzulegen. Die Möglichkeit, einen Freitext anzugeben, lässt die Anzahl der Kategorien stark ansteigen.

Jede ALKIS-Objektklasse besitzt neben einem Namen und einer detaillierten Beschreibung auch einen Klassencode, der im Zuordnungsprozess mit ausgewertet wird. Objekte semantisch ähnlicher Kategorien, z.B. A1121 (Allgemeinbildende Schule) und A1123 (Fachhochschule, Universität) weisen eine geringe Differenz der Klassencodes auf. Bei der Aggregation von Nachbarobjekten beeinflussen die Klassencodedifferenzen die Rangfolge der Objekte. Im Gegensatz dazu besitzen OSM-Objektklassen weder einen Klassencode noch einen direkten Hinweis auf die semantische Ähnlichkeit der Gebäudekategorien. Da ein Klassencode im Zuordnungsprozess benötigt wird, wurden die OSM-Objektklassen alphabetisch sortiert und durchnummeriert. Objekte semantisch ähnlicher Kategorien erhielten dadurch stark unterschiedliche Klassencodes, z.B. O81 (School) und O96 (University).

Testgebiet B: ALKIS - ATKIS in Hameln

Das zweite Testgebiet ist 12 km^2 (2×6 km) groß und befindet sich in Hameln, einer Stadt in Niedersachsen. Für die Untersuchung werden die in Abbildung 6.2 dargestellten Objekte der Datensätze ALKIS und ATKIS verwendet. Die Objekte beschreiben in unterschiedlichen Maßstäben die tatsächliche Nutzung, d.h. die Bodennutzung aufgrund der Bodenbedeckung, der bestehenden Gebäude oder baulichen Anlagen.

Die Besonderheit an diesem Testszenario ist, dass beide Datensätze zum AAA-Modell gehören und zum Teil identische Objektklassen verwenden. Folgende Fragestellung ist interessant: Werden die Objektklassen in den bisher getrennt modellierten und geführten Informationssystemen übereinstimmend verwendet? Dazu ist zu prüfen, ob die Objekte der verschiedenen Maßstäbe einander zugeordnet werden können und identische Objektklassen besitzen. Die identifizierten semantischen Korrespondenzen könnten in Zukunft genutzt werden, um Objektklassen der großmaßstäbigen ALKIS-Daten automatisch auf die kleinmaßstäbigen ATKIS-Daten zu übertragen. Die Definition von Transformationsregeln kann die derzeitige manuelle und ALKIS-unabhängige Klassifikation der ATKIS-Daten ablösen.

Im Anhang gibt B.1 eine Übersicht der im Testgebiet vorhanden 18 Objektklassen und deren Objektanzahlen. Es stehen 4.121 ALKIS-Objekte 834 ATKIS-Objekten gegenüber. In jedem Datensatz gibt es genau eine Objektklasse, die im jeweils anderen Datensatz nicht vertreten ist. Die Klasse Weg (42006) ist mit 910 Objekten nur in ALKIS vorhanden. In ATKIS gibt es diese Klasse laut Datenmodell nicht. Stattdessen werden Wege als sogenannte Überlagerungsobjekte linienförmig erfasst und gehören einer anderen Objektklasse an. Im Gegensatz dazu ist die Klasse Fläche zur Zeit unbestimmbar (43008) nur in ATKIS vorhanden und besitzt im Testgebiet 48 Objekte.

Abbildung 6.2: Testgebiet B: Objekte der tatsächlichen Nutzung der Datensätze ALKIS 1:1.000 (oben) und ATKIS 1:10.000 (unten) in Hameln.

Die Objektverteilung zeigt erwartungsgemäß, dass flächenhafte Objekte in ALKIS detaillierter beschrieben werden und fast jede Klasse eine deutlich größere Anzahl von Objekten besitzt. Im Testgebiet ist der Anteil der ATKIS-Objekte jedoch in drei Klassen höher. Während die Klassen Friedhof (41009) und Gehölz (43003) nur jeweils ein ATKIS-Objekt mehr besitzen, hat die Klasse Fläche gemischter Nutzung[2] (41006) gegenüber den 9 ALKIS-Objekten sogar 134 ATKIS-Objekte.

Abbildung 6.3 gibt einen detaillierten Einblick in beide Datensätze. Vier ALKIS-Objekte, gekennzeichnet durch gefüllte Polygone mit starken weißen Konturen, werden durch Aggregation einem ATKIS-Objekt, gekennzeichnet durch eine starke schwarze Kontur, geometrisch zugeordnet. Isoliert betrachtet, spiegelt die einseitig zusammengefasste 4:1-Objektrelation eine heterogene Schemarelation wider:

{Landwirtschaft, Landwirtschaft, Landwirtschaft, Industrie- u. Gewerbefläche} → Industrie- u. Gewerbefläche.

Obwohl die hellgrauen ALKIS-Objekte der Klasse Landwirtschaft[3] (43001) in dieser Relation zahlen- und flächenmäßig ein stärkeres Gewicht gegenüber dem dunkelgrauen ALKIS-Objekt der Klasse Industrie- u. Gewerbefläche[4] (41002) haben, besitzt das ATKIS-Objekt die Nutzung Industrie- u. Gewerbefläche. Der Unterschied zwischen den Nutzungsklassen entsteht vermutlich, weil beide Datensätze getrennt voneinander modelliert werden.

[2]Klassendefinition: Fläche gemischter Nutzung - bebaute Flächen einschließlich der mit ihr im Zusammenhang stehenden Freifläche (Hofraumfläche, Hausgarten), auf der keine Art der baulichen Nutzung vorherrscht. Solche Flächen sind insbesondere ländlich-dörflich geprägte Flächen mit land- und forstwirtschaftlichen Betrieben, Wohngebäuden u.a. sowie städtisch geprägte Kerngebiete mit Handelsbetrieben und zentralen Einrichtungen für die Wirtschaft und die Verwaltung (AdV, 2008).

[3]Klassendefinition: Landwirtschaft - ist eine Fläche für den Anbau von Feldfrüchten sowie eine Fläche, die beweidet und gemäht werden kann, einschließlich der mit besonderen Pflanzen angebauten Fläche. Die Brache, die für einen bestimmten Zeitraum (z. B. ein halbes oder ganzes Jahr) landwirtschaftlich unbebaut bleibt, ist als 'Landwirtschaft' bzw. 'Ackerland' zu erfassen (AdV, 2008).

[4]Klassendefinition: Industrie- u. Gewerbefläche - ist eine Fläche, die vorwiegend industriellen oder gewerblichen Zwecken dient (AdV, 2008).

Abbildung 6.3: Überlagerung der Beispieldaten ALKIS (gefüllte Polygone mit starken weißen Konturen) und ATKIS (nicht gefüllte Polygone mit schwarzen Konturen) im Testgebiet B. Die zugehörigen Objektklassen sind mit unterschiedlichen Schriftfarben gekennzeichnet: ALKIS (weiß) und ATKIS (schwarz).

Testgebiet C: ATKIS - GDF in Hannover-Wedemark

Das dritte Testgebiet befindet sich in der Region Hannover-Wedemark in Niedersachsen und ist 21 km^2 (4,2 × 5 km) groß. Für die Untersuchung werden die in Abbildung 6.4 dargestellten Objekte der Datensätze ATKIS und GDF verwendet. Beide Datensätze haben verschiedene Maßstäbe und beinhalten thematisch unterschiedliche Informationen. Folgende Fragestellung ist interesant: Besitzen beide Datensätze tatsächlich identische Objekte und können anhand derer semantische Korrespondenzen zwischen den Objektklassen aufgedeckt werden?

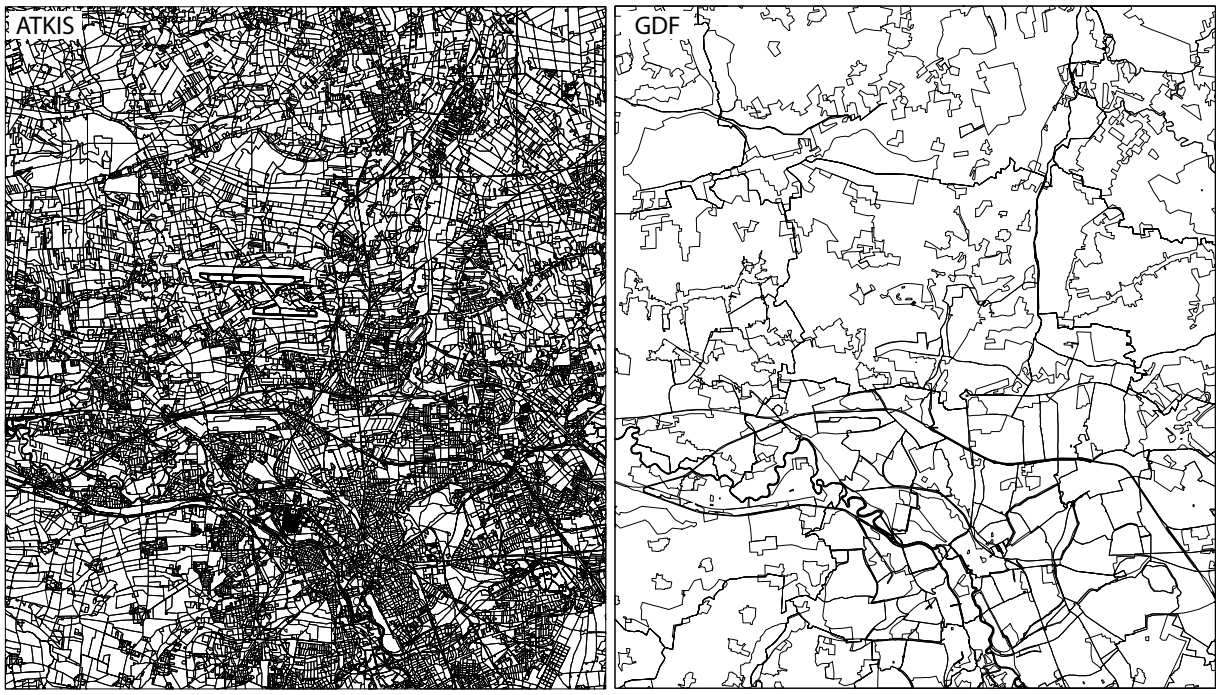


Abbildung 6.4: Testgebiet C: Objekte der Datensätze ATKIS 1:10.000 (links) und GDF 1:25.000 (rechts) in der Region Hannover-Wedemark.

Im Anhang geben C.1 und C.2 eine Übersicht der 63 ATKIS- und 19 GDF-Objektklassen und deren Objektanzahlen. Die Anzahl der im Ausschnitt dargestellten Objekte unterscheidet sich sehr stark zwischen den Datensätzen. Es stehen 21.925 ATKIS-Objekte gerade einmal 525 GDF-Objekten gegenüber. Auf den ersten Blick lassen sich wenige korrespondierende Objekte identifizieren.

Die Besonderheit an diesem Testszenario ist, dass innerhalb jedes einzelnen Datensatzes Objektüberlagerungen vorhanden sind. Das bedeutet, dass sich an einer geographischen Position gleich mehrere Objekte eines Datensatzes mit unterschiedlichen Objektklassen befinden können. Beispielsweise liegt ein Gewässerobjekt innerhalb

eines Verwaltungsbezirks und überlagert dieses Objekt auch geometrisch. Die Objektzuordnung verändert sich dahingehend, dass sich durch die Objektüberlagerungen innerhalb eines Datensatzes der Suchraum für potentielle Matching-Kandidaten vergrößert.

6.1.3 Datenvorverarbeitung

Für die Verwendung der unterschiedlichen Geodaten im Objektzuordnungsverfahren waren einige Vorverarbeitungsschritte notwendig. Alle Datensätze wurden in ein einheitliches Koordinatensystem transformiert, und die Auswahl wurde auf flächenhafte Objekte begrenzt. Für Testgebiet A wurden ausschließlich Gebäudeobjekte verwendet. Dazu wurden im OSM-Datensatz alle Objekte mit dem Attribut *building* selektiert. Von dieser Auswahl wurden die Objekte gelöscht, die keine semantische Zusatzinformation besitzen, um die Zuordnung zu den ALKIS-Gebäudekategorien zu ermöglichen. Es wurden nur Objekte verwendet, die entweder eine Angabe zur Gebäudekategorie, d.h. einen Eintrag beim Attribut *type*, wie z.B. *type=hotel, church, residential* oder einen Eintrag beim Attribut *name*, wie z.B. *name=Ernst-August-Carree* aufwiesen. Lediglich 50 % der im Testgebiet A erfassten OSM-Gebäude besitzen diese semantischen Zusatzinformationen. Für Testgebiet B wurde die Objektvielfalt auf Objekte der tatsächlichen Nutzung eingeschränkt. Hierzu zählen Objekte der Objektgruppen: Siedlung, Verkehr, Vegetation und Gewässer. Diese Auswahl verhindert, dass sich Objekte innerhalb der Datensätze überlagern. Für Testgebiet C war keine spezielle Objektauswahl notwendig.

Nach der Objektauswahl wurde speziell der OSM-Datensatz auf topologische Fehler hin untersucht. Es gab viele geringfügige Überlappungen bzw. Lücken zwischen benachbarten Gebäuden. Alle identifizierten Fehler wurden korrigiert, indem ihre Objektgeometrien verändert oder gelöscht wurden.

Objekte, die aus mehreren räumlich getrennten Objektteilen bestehen, sogenannte Multi-Part-Objekte, wurden in einzelne Objekte unterteilt. Dieser Spezialfall der Objektbildung wird im Rahmen der Arbeit nicht berücksichtigt. Abschließend erhielt jedes Objekt einen eindeutigen Identifikator und, falls nicht vorhanden, einen numerischen Klassencode, der die Zugehörigkeit zu einer bestimmten Objektklasse definiert.

6.2 Ergebnisse des Data-Matching

In diesem Abschnitt werden die Ergebnisse des Objektzuordnungsverfahrens für Polygone aus Kapitel 4 für alle drei Testszenarien vorgestellt. Das Verfahren identifiziert Objektkorrespondenzen, die anschließend für die Bestimmung von semantischen Ähnlichkeiten zwischen Objektklassen genutzt werden. Eine vollständige Objektzuordnung war nicht Ziel dieser Arbeit. Mit Hilfe von manuell erstellten Referenzdaten werden die Ergebnisse durch quantitative Maße bewertet.

6.2.1 Testgebiet A: ALKIS - OSM in Hannover

Tabelle 6.1 gibt einen Überblick über die Menge und Art der erzielten Objektrelationen R_o im Testgebiet A. Ohne Berücksichtigung von 1:0- bzw. 0:1-Relationen werden $|R_o| = 2.054$ Relationen ermittelt. Es werden insgesamt 3.108 (13 %) ALKIS-Gebäude 2.261 (84 %) OSM-Gebäuden zugeordnet. Die Objektrelationen repräsentieren Korrespondenzen zwischen 32 von 49 ALKIS- und 62 von 72 OSM-Gebäudeklassen. Objektklassen, deren Objekte in keiner Relation vertreten sind, sind in den Objektklassenübersichten A.1 und A.2 grau eingefärbt. Dazu zählen bei ALKIS u.a. die Klassen Stadion (A0917), Feuerwehr (A1172), Gebäude für Gewerbe und Industrie mit Wohnen (A2161) dazu und bei OSM Ruins (O79) und Supermarket (O87).

Tabelle 6.1: Ergebnisse des Data-Matching-Verfahrens für Testgebiet A: ALKIS - OSM in Hannover.

Kardinalität		R_o		Objektanzahl				Flächenanteil			
ALKIS:OSM		Anzahl	Prozent	ALKIS	OSM	Gesamt	Prozent	ALKIS in [ha]	Prozent	OSM in [ha]	Prozent
1:1		1.653	80,48	1.653	1.653	3.306	61,58	73,99	57,44	76,17	55,67
1:n	homogen	39	1,90	39	110	149	2,78	8,63	6,70	8,90	6,50
	heterogen	7	0,34	7	22	29	0,54	2,01	1,56	2,05	1,50
n:1	homogen	173	8,42	711	173	884	16,46	16,10	12,50	18,57	13,57
	heterogen	84	4,09	436	84	520	9,69	19,84	15,40	22,35	16,33
n:m	homogen	88	4,28	226	192	418	7,79	4,84	3,76	5,33	3,90
	heterogen	10	0,49	36	27	63	1,17	3,40	2,64	3,46	2,53
Gesamt		2.054	100	3.108	2.261	5.369	100	128,80	100	136,83	100

Gebäude sind durch ihre Außenwände klar definierte Objekte. Sie können von verschiedenen Personen einfach vermessen und digitalisiert werden. Erwartungsgemäß werden bei der Objektzuordnung mit einem Anteil von 80 % am häufigsten 1:1-Relationen identifiziert. Sie umfassen etwa 62 % aller zugeordneten Gebäude mit einem Flächenanteil von 56 %. Die jeweils größten Anteile haben die Gebäudekategorien Nichtöffentliches Gebäude (Wohngebäude) (A0931) mit 73 % und Nichtöffentliches Gebäude (Nebengebäude) (A0932) mit 10 % bei ALKIS und Apartments (O2) mit 57 % und Residential (O74) mit 6 % bei OSM. Allein durch 1:1-Relationen werden Objekte von 21 OSM- und 2 ALKIS-Gebäudeklassen vollständig zugeordnet. Sie beschreiben Gebäude mit einer besonderen Nutzung und besitzen im Durchschnitt nur sehr wenige Objekte, wie z.B. Parlament (A1111), Hallenbad (A2821), Arts centre (O6) oder Glasshouse (O40).

Die zweitstärkste Relation ist mit 12,5 % die einseitig zusammengefasste n:1-Relation. Einem OSM-Gebäude werden mehrere ALKIS-Gebäude zugeordnet. Zwei Drittel der Relationen sind homogen, d.h. es werden überwiegend Gebäude gleicher Kategorien zusammengefasst. Insgesamt umfassen diese Relationen 26,2 % der zugeordneten Objekte mit einem Flächenanteil von 30 %. Das bedeutet, dass OSM-Gebäude in n:1-Relationen im Durchschnitt größer sind als in 1:1-Relationen. Dominierende Objektklassen sind bei ALKIS wieder Nichtöffentliches Gebäude (Wohngebäude) (A0931) mit 71 % und Nichtöffentliches Gebäude (Nebengebäude) (A0932) mit 20 % und bei OSM Residential (O74) mit 55 %.

Abbildung 6.5 a) zeigt beispielhaft eine n:1-Relation. Geometrische Unterschiede sind offensichtlich und eventuell durch die Art der Aufnahme oder den Zweck, dem die Daten dienen, geschuldet. Bei einer Gebäudeaufnahme vor Ort wird sich ein Laie vermutlich an einer durchgehenden einheitlichen Fassade eines Baublocks orientieren. Die Trennung der einzelnen Gebäude auf Basis der grundstücksrechtlichen Informationen, wie sie in den amtlichen ALKIS-Daten enthalten sind, ist vor Ort nicht immer ersichtlich. Auffällig ist, dass für das OSM- Gebäude keine Details, wie z.B. abgeschrägte Blockecken oder geringe Vorsprünge erfasst wurden. Dies weist daraufhin, dass das Gebäude aus einem Luftbild digitalisiert wurde, in dem diese Informationen nicht erkennbar sind.

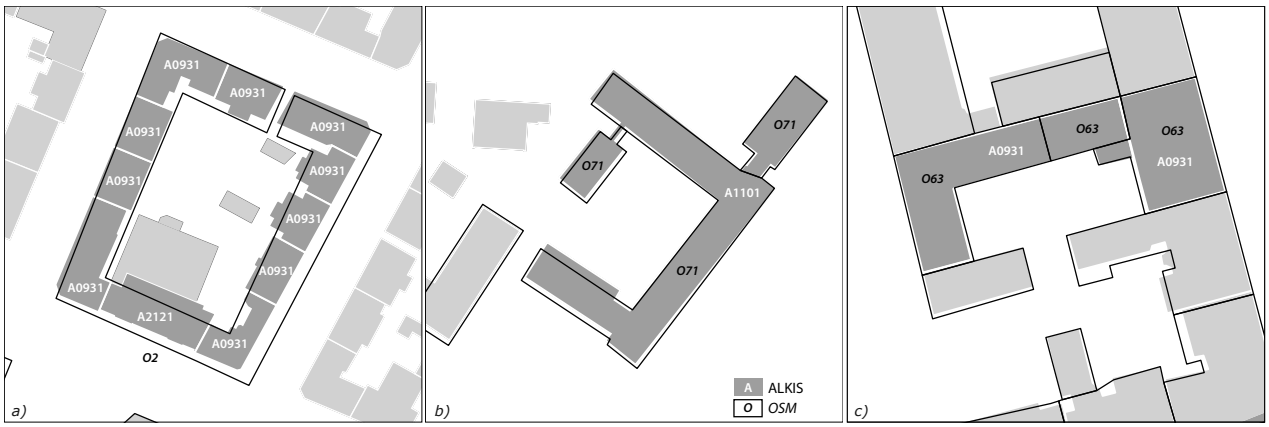


Abbildung 6.5: Beispiele für komplexe Objektrelationen im Testgebiet A: a) n:1-Relation zwischen 11 ALKIS-Objekten und einen OSM-Objekt, b) 1:n-Relation zwischen einem ALKIS- und drei OSM-Objekten und c) n:m-Relation zwischen zwei ALKIS- und drei OSM-Objekten.

Im Vergleich zu n:1-Relationen werden nur sehr wenige 1:n-Relationen identifiziert. Sie haben gemessen an allen Relationen nur einen Anteil von 2,3 % und repräsentieren 8 % aller zugeordneten Flächen. Von 46 Relationen sind sogar 85 % homogen. Am stärksten sind bei OSM die Objektklassen Apartments (O2) und Offices (O63) mit jeweils 20 % und bei ALKIS Nichtöffentliches Gebäude (Wohngebäude) (A0931) mit 46 % vertreten. Abbildung 6.5 b) zeigt beispielhaft eine homogene 1:n-Relation, bei dem einem ALKIS-Gebäude drei OSM-Gebäude zugeordnet werden. Es handelt sich hierbei um das Finanzamt Hannover-Mitte. Der Gebäudekomplex besitzt unterschiedliche Gebäudehöhen. Ein Laie könnte aufgrund der verschiedenen Etagenanzahlen drei Gebäudeteile erfasst haben.

Die beidseitig zusammengefassten n:m-Relationen umfassen nur 5 % der Objektrelationen, 6 % der Flächen und 9 % aller zugeordneten Objekte. Der Anteil an homogenen Relationen überwiegt deutlich mit 90 %. Abbildung 6.5 c) zeigt beispielhaft eine homogene 2:3-Relation. Auch bei diesen Relationen finden sich die dominierenden Gebäudekategorien der anderen Relationstypen wieder: Nichtöffentliches Gebäude (Wohngebäude) (A0931) mit 93 % bei ALKIS und Apartments (O2) mit 53 % bei OSM.

Bevor im nächsten Abschnitt die Gesamtähnlichkeitsmaße vorgestellt werden, die ausschlaggebend für die Objektzuordnung sind, wird die Ermittlung des Maßes an einem Beispiel wiederholt. Für die in Abbildung 6.5 a) dargestellte 11:1-Relation werden folgende Einzelwerte ermittelt: Beide Flächenparameter berechnen sich nach

Gleichung 4.2 zu $s_{i_A} = 0{,}88$ und $s_{i_B} = 0{,}80$. Die Werte verdeutlichen, dass der Anteil der ALKIS-Flächen innerhalb der OSM-Flächen größer ist. Der Ausrichtungsparameter wird nach Gleichung 4.3 zu $s_a = 0{,}92$ bestimmt und spiegelt die übereinstimmende Objektausrichtung wider. Durch Addition der Werte ergibt sich der geometrische Parameter von $s_g = 2{,}60$. Für die Objektzuordnung wird zusätzlich der Heterogenitätsparameter s_h berücksichtigt, der die Anzahl der beteiligten Objektklassen an der Gesamtzahl misst (Gleichung 4.4). Da eines der 11 ALKIS-Objekte einer anderen Objektklasse angehört, ergibt sich $s_h = 0{,}99$. Die endgültige Matching-Entscheidung wird auf Basis des Gesamtähnlichkeitsmaßes $s_t = s_g + s_h = 3{,}59$ aus Gleichung 4.5 getroffen.

Im Zuordnungsprozess wird zunächst eine Liste mit allen identifizierten Objektrelationen erstellt. Anschließend werden die Relationen entfernt, bei denen mindestens ein Flächenparameter kleiner als ein Schwellwert ist. Im Rahmen dieser Arbeit wird für alle Testgebiete der Schwellwert $s_{i_j} < 0{,}50$ gewählt, d.h. Objekte, die sich mit weniger als der Hälfte ihrer Flächen überlagern, werden als endgültige Matching-Kandidaten ausgeschlossen. Im Beispiel in Abbildung 6.6 b) werden beide identifizierte Objektrelationen verworfen, da jeweils ein Flächenwert unter dem festgelegten Schwellwert liegt. Aufgrund des Flächenschwellwerts bleiben 102 Relationen unberücksichtigt.

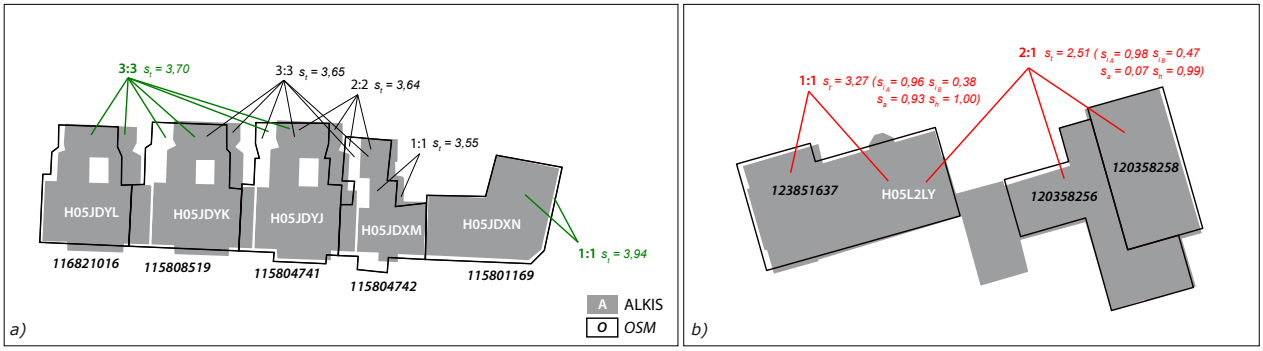


Abbildung 6.6: Beispiele für vom Verfahren identifizierte Objektrelationen in Testgebiet A. Während die rot markierten Relationen im Zuordnungsprozess aufgrund des Schwellwerts für die Flächenparameter und die schwarzen aufgrund der doppelten Objekteinträge wieder verworfen werden, sind die grün gekennzeichneten Relationen Bestandteil der endgültigen Ergebnisliste.

Die verbliebenen Objektrelationen werden noch auf doppelte Objekteinträge hin überprüft. Für das Beispiel in Abbildung 6.6 a) werden vom Zuordnungsverfahren fünf Objektrelationen identifiziert. Die beiden grün markierten Relationen verbleiben in der Ergebnisliste. Die schwarz markierten Relationen werden entfernt, da Relationsobjekte Teil anderer Relationen mit höheren Gesamtähnlichkeitsmaßen sind. Ein Experte hat hier eine andere Zuordnungsentscheidung getroffen. Er entscheidet sich sowohl für die grünen als auch für die schwarze 1:1-Relation.

Häufigkeitsverteilung der Gesamtähnlichkeitsmaße

Abbildung 6.7 gibt einen Überblick über die Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t der endgültigen Zuordnungsliste. Die Relationen im Bereich $3{,}90 < s_t < 3{,}95$ machen mit 22,6 % den größten Anteil aus. Bei einem Maximalwert von 4,0 entspricht dies einer Ähnlichkeit von 97,5%. Für 19% der Objektrelationen werden sehr hohe Gesamtähnlichkeitsmaße von $s_t > 3{,}95$ bestimmt.

Mit abnehmenden Wert verringert sich die Anzahl der Relationen. Bis auf wenige Ausnahmen haben 1:1-Relationen immer den größten Anteil. Im Wertebereich $3{,}45 < s_t < 3{,}50$ übersteigt die Anzahl der 1:n-Relationen erstmalig die Anzahl der 1:1-Relationen. Drei Viertel aller Objektrelationen haben ein Gesamtähnlichkeitsmaß von $s_t > 3{,}6$, was einer Ähnlichkeit von 90 % entspricht. 96 % aller Relationen besitzen ein Gesamtähnlichkeitsmaß von $s_t > 3{,}45$ (86 %). Durch die Analyse der 1:1-Relationen wird deutlich, dass mit Abnahme des Gesamtähnlichkeitsmaßes ein Anstieg der Objektflächen zu verzeichnen ist.

Vergleich mit Referenzdaten

Anhand von manuell erstellten Referenzdaten werden die Zuordnungsergebnisse bewertet. Dafür ordnete ein Experte korrespondierende Objekte beider Datensätze geometrisch einander zu. Ziel der Arbeit ist es, zu prüfen, ob die zugeordneten Objekte identisch sind und nicht ob das Zuordnungsverfahren exakt die gleichen Relationen wie ein Experte identifiziert. Aus diesem Grund stützt sich die weitere Auswertung nicht auf Relationen, sondern auf die Objekte.

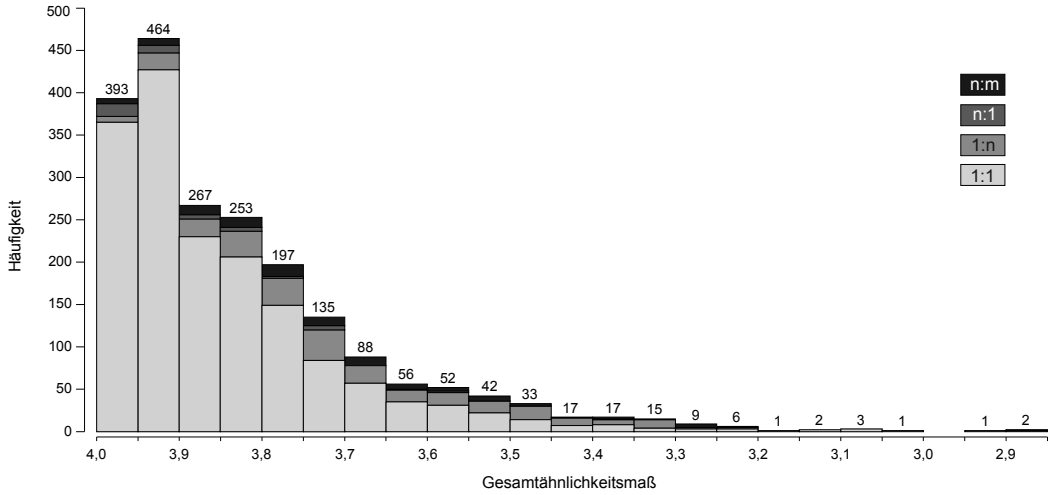


Abbildung 6.7: Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t für die im Zuordnungsprozess identifizierten endgültigen Relationen in Testgebiet A. Die Häufigkeiten sind für die unterschiedlichen Relationsarten dargestellt.

Die Qualität der Zuordnung lässt sich durch mehrere quantitative Maße beschreiben, die ausdrücken, wie genau das erzielte Ergebnis zu den Referenzdaten passt. Die verschiedenen Maße basieren auf disjunkten Teilmengen der Objekte, die zunächst vorgestellt werden. Die Teilmenge der *positiven Objekte* spiegelt die korrekte Zuordnung und die Teilmenge der *negativen Objekte* die fehlerhafte Zuordnung wider. Beide Teilmengen lassen sich weiter unterteilen. Die *richtig positive* Teilmenge (engl. True Positives (tp)) umfasst die korrekt erkannten Zuordnungen, während die *richtig negative* Teilmenge (engl. True Negatives (tn)) die korrekt nicht erkannten Zuordnungen beschreibt. Entsprechend fasst die *falsch positive* Teilmenge (engl. False Positives (fp)) die nicht korrekt erkannten Zuordnungen und die *falsch negative* Teilmenge (engl. False Negatives (fn)) die vom Algorithmus nicht erkannten Zuordnungen zusammen. Diese vier Teilmengen können in einer Konfusionsmatrix zusammengefasst werden. Tabelle 6.2 zeigt ganz links den Aufbau der allgemeinen Konfusionsmatrix und daneben die Matrizen für beide Datensätze getrennt, oben hinsichtlich der Objektanzahl und darunter bezogen auf die Flächen angegeben in Hektar.

Tabelle 6.2: Konfusionsmatrizen für Testgebiet A: ALKIS - OSM in Hannover. Links ist die Allgemeine Konfusionsmatrix, in der Mitte für ALKIS und rechts für OSM. Die obere Zeile bezieht sich auf die Objektanzahlen (#) und die untere Zeile auf die Flächen [ha].

		Referenz +	Referenz −	Gesamt
Match	+	tp	fp	$\sum M^+$
	−	fn	tn	$\sum M^-$
Gesamt		$\sum R^+$	$\sum R^-$	

ALKIS

#		Referenz +	Referenz −	Gesamt
Match	+	2.887	221	3.108
	−	197	20.358	20.555
Gesamt		3.084	20.579	23.663

Fläche [ha]		Referenz +	Referenz −	Gesamt
Match	+	122,05	6,75	128,80
	−	4,24	398,36	402,60
Gesamt		126,29	405,11	531,40

OSM

#		Referenz +	Referenz −	Gesamt
Match	+	2.201	60	2.261
	−	151	277	428
Gesamt		2.352	337	2.689

Fläche [ha]		Referenz +	Referenz −	Gesamt
Match	+	129,75	7,08	136,83
	−	5,84	33,20	39,04
Gesamt		135,59	40,28	175,87

Aus diesen Größen lassen sich wiederum Maße für die Güte der Zuordnung bestimmen. Die *Genauigkeit* (engl. Precision) entspricht der Exaktheit des Zuordnungsalgorithmus. Die korrekten Zuordnungen werden an allen durch den Algorithmus gefundenen Zuordnungen gemessen:

$$Genauigkeit = \frac{tp}{tp + fp} \ . \tag{6.1}$$

Im Gegensatz dazu ist die *Sensitivität* (engl. Recall) ein Vollständigkeitsmaß, das die korrekten Zuordnungen an der Summe aller richtigen Zuordnungen misst:

$$Sensitivität = \frac{tp}{tp + fn} \ . \tag{6.2}$$

Mit dem *F-Maß*, das die Genauigkeit und Sensitivität mit dem gewichteten harmonischen Mittel kombiniert, lässt sich die Güte der Zuordnung auch mit einem Maß angeben:

$$F\text{-}Maß = \frac{2 \cdot (Genauigkeit \cdot Sensitivität)}{Genauigkeit + Sensitivität} \ . \tag{6.3}$$

In Tabelle 6.9 werden die quantitativen Maße für alle Testgebiete zusammen aufgeführt. Der Genauigkeitswert ist für die zugeordneten OSM-Objekte mit 97,3 % am höchsten. Das bedeutet, weniger als 3 % der OSM-Objekte werden nicht korrekt zugeordnet. Die 60 fehlerhaft zugeordneten Objekte gehören 18 unterschiedlichen Objektklassen an, wobei den größten Anteil die Objektklassen Apartments (O2) mit 21 % und Residential (O74) mit 20 % ausmachen. Bezogen auf die Flächen nimmt der Genauigkeitswert leicht ab, was verdeutlicht, dass eher größere Objekte fehlerhaft zugeordnet werden.

Im Vergleich dazu werden bei ALKIS etwa doppelt so viele Objekte (7,1 %) falsch zugeordnet, während der Genauigkeitswert bezogen auf die Objektflächen gleich geblieben ist. Die 221 fehlerhaft zugeordneten Objekte verteilen sich auf 10 Objektklassen, wobei den größten Anteil die Objektklassen der Nichtöffentlichen Gebäude (A0931) mit 58 % und (A0932) mit 29 % ausmachen. Die Sensitivitätswerte beider Datensätze sind bezogen auf die Objektanzahlen gleich groß bei 93,6 %. Bei den Flächen ist ein leichter Anstieg der Werte zu verzeichnen.

6.2.2 Testgebiet B: ALKIS - ATKIS in Hameln

In gleicher Art und Weise wie für Testgebiet A werden in Tabelle 6.3 die Ergebnisse der Objektzuordnung für Testgebiet B präsentiert. Insgesamt werden $|R_o| = 522$ Relationen ermittelt, die 43 % der ALKIS-Objekte zu 72 % der ATKIS-Objekte zuordnen. In Abbildung 6.8 sind alle vom Verfahren zugeordneten Nutzungsflächen grau (tp) und rot (fp) gekennzeichnet. Die Objektrelationen repräsentieren Korrespondenzen zwischen allen ALKIS- und 16 von 17 ATKIS-Objektklassen. Die ATKIS-Objektklasse Schiffsverkehr (T42016) ist in keiner Objektrelation vertreten und ist in der Objektklassenübersicht B.1 grau eingefärbt.

Tabelle 6.3: Ergebnisse des Data-Matching-Verfahrens für Testgebiet B: ALKIS - ATKIS in Hameln.

Kardinalität ALKIS:ATKIS		R_o Anzahl	R_o Prozent	Objektanzahl ALKIS	Objektanzahl ATKIS	Objektanzahl Gesamt	Objektanzahl Prozent	Flächenanteil ALKIS in [ha]	Flächenanteil Prozent	Flächenanteil ATKIS in [ha]	Flächenanteil Prozent
1:1		181	34,67	181	181	362	15,02	272,59	31,70	299,02	30,46
1:n	homogen	20	3,83	20	55	75	3,11	88,16	10,25	92,98	9,47
	heterogen	2	0,38	2	4	6	0,25	1,53	0,18	1,97	0,20
n:1	homogen	80	15,33	358	80	438	18,17	127,11	14,78	145,36	14,81
	heterogen	212	40,61	1.116	212	1.328	55,10	299,30	34,80	357,52	36,42
n:m	homogen	7	1,34	19	16	35	1,45	22,88	2,66	28,18	2,87
	heterogen	20	3,83	115	51	166	6,89	48,41	5,63	56,73	5,78
Gesamt		522	100	1.811	599	2.410	100	859,98	100	981,76	100

In diesem Testszenario haben erwartungsgemäß die komplexen n:1-Relationen mit 56 % den größten Anteil. Dies ist in dem vorliegenden Maßstabsunterschied zwischen 1:1.000 der ALKIS-Objekte gegenüber 1:10.000 der ATKIS-Objekte begründet. Mehrere ALKIS-Objekte werden aggregiert und einem ATKIS-Objekt zugeordnet. Bei der Analyse der in Abbildung 6.8 rosa gefärbten Objekte, die vom Experten, aber nicht vom Verfahren identifiziert werden, werden einige n:1-Relationen nicht erkannt. Es gibt verschiedene Gründe. Hauptsächlich liegt es am zu klein gewählten Suchbereich, in dem nach potentiellen Matching-Kandidaten gesucht werden soll. Gegenüber Testgebiet A ist diesbezüglich keine Anpassung vorgenommen worden. Des Weiteren kann festgestellt werden, dass Objekte Teil sehr vieler Objektrelationen sind (vgl. Abbildung 6.9). Bei der Prüfung auf doppelte Objekteinträge und bezüglich des Flächenschwellwerts werden viele Relationen verworfen.

Alle identifizierten n:1-Relationen umfassen etwa 73 % der zugeordneten Objekte mit einem Flächenanteil von 50 %. Die Mehrheit der n:1-Relationen ist heterogen, d.h. es werden ALKIS-Objekte unterschiedlicher Objektklassen zusammengefasst. Bei 62 % der Relationen besitzt mindestens ein ALKIS-Objekt die gleiche Objektklasse wie das ATKIS-Objekt. Nur ein Viertel der n:1-Relationen sind homogen. Bei der Mehrheit (81 %) der

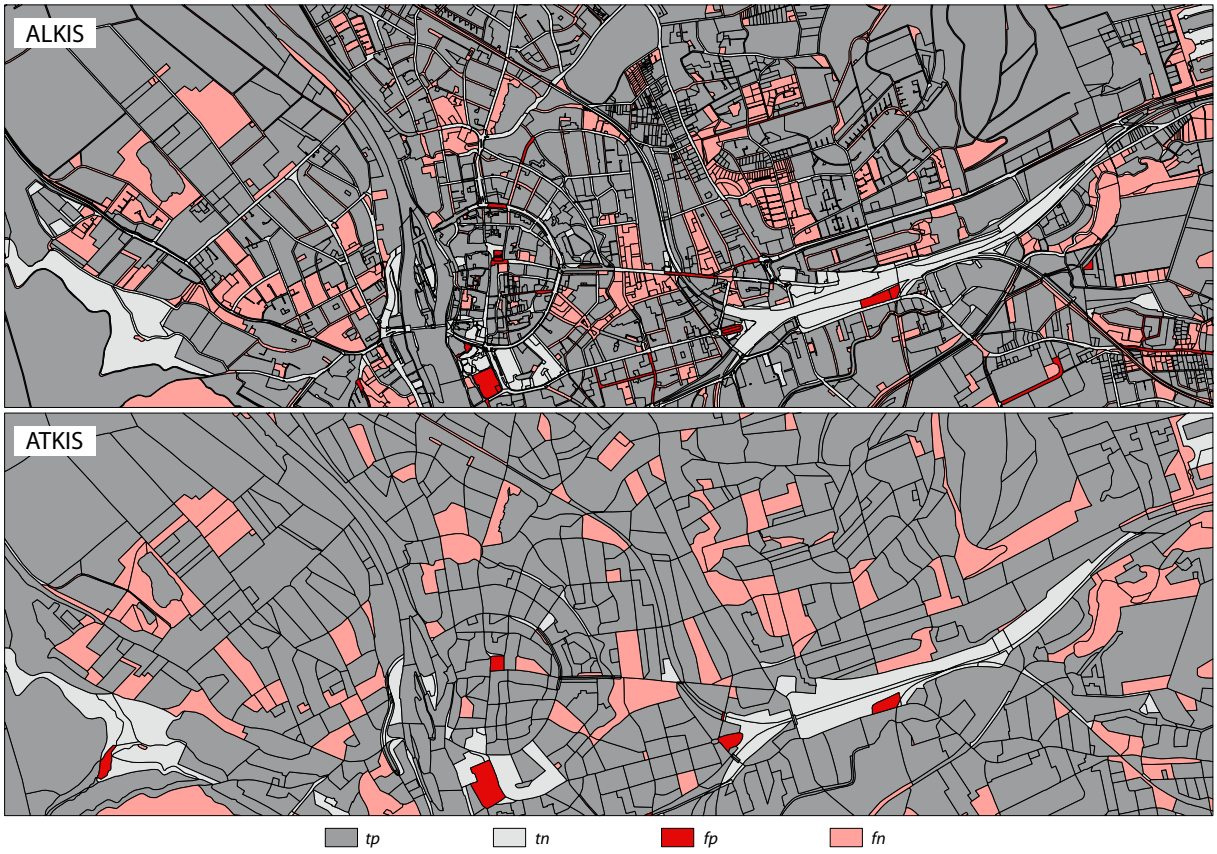


Abbildung 6.8: Testgebiet B: Ergebnis der Objektzuordnung: ALKIS 1:1.000 (oben) und ATKIS 1:10.000 (unten) - Teilmenge der korrekten (tp (grau), tn (hellgrau)) und der fehlerhaften Zuordnung (fp (rot), fn (rosa)).

Relationen stehen sich identische Objektklassen gegenüber. Den größten Anteil an den n:1-Relationen hat die Objektklasse Wohnbaufläche sowohl bei ALKIS (51 %) als auch bei ATKIS (40 %). Ein Viertel aller beteiligten ATKIS-Objekte gehört der Klasse Fläche gemischter Nutzung (T41006) an. Diesen Objekten werden ALKIS-Objekte von acht Klassen zugeordnet, überwiegend die der Klassen Wohnbaufläche (L41001) und Industrie- und Gewerbefläche (L41002).

Mit einem Anteil von 35 % stellen 1:1-Relationen die zweitstärkste Relationsart dar. Obwohl nur 15 % aller zugeordneten Objekte beteiligt sind, umfassen sie 30 % der Flächen. Durch die Analyse der Relationen wird aufgedeckt, dass die Objekte hauptsächlich zu flächenintensiven Objektklassen gehören, wie z.B. Wohnbaufläche (41001), Industrie und Gewerbe (41002), Landwirtschaft (43001) oder Stehendes Gewässer (44006). Die durchschnittliche Objektgröße beträgt ca. 1,5 Hektar.

Trotz des Maßstabsunterschiedes werden 1:n-Relationen identifiziert. Mit 4 % haben sie den geringsten Anteil am Zuordnungsergebnis. Ohne offensichtlichen Grund werden im kleinmaßstäbigen ATKIS-Datensatz Objekte unterteilt, die im großmaßstäbigen Datensatz ein Objekt repräsentieren. Verschiedene Datenstände könnten den Unterschied in der Modellierung erklären. In Abbildung 6.9 a) ist eine 1:n-Relation dargestellt. Der Experte hat vier ATKIS-Objekte einem ALKIS-Objekt zugeordnet. Das ATKIS-Objekt $p_{B_{20039}}$ gehört der Objektklasse Platz und alle anderen Objekte gehören der Klasse Fläche besonderer funktionaler Prägung an. Das Verfahren wählt aus vier identifizierten Objektrelationen letztendlich die 1:2-Relation mit dem höchsten Gesamtähnlichkeitsmaß aus. Dass die gewünschte 1:4-Relation nicht erkannt wird, ist vermutlich auf den zu kleinen Suchbereich zurückzuführen. Die 1:2-Relation ist homogen und besitzt in beiden Datensätzen die gleiche Objektklasse. Die Mehrheit der 22 1:n-Relationen sind homogen. In diesem Relationstyp sind Wald (43002) und Wohnbaufläche (41001) die dominierende Objektklassen. Die beteiligten ALKIS-Objekte sind mit einer Durchschnittsgröße von 4 Hektar deutlich größer als in den 1:1-Relationen.

Die n:m-Relationen haben mit 5 % ebenfalls einen geringen Einfluss am Zuordnungsergebnis. 75 % dieser Relationen sind heterogen. Die Relationsarten 1:n und n:m umfassen etwa 12 % aller zugeordneten Objekte und 18 % der zugeordneten Flächen.

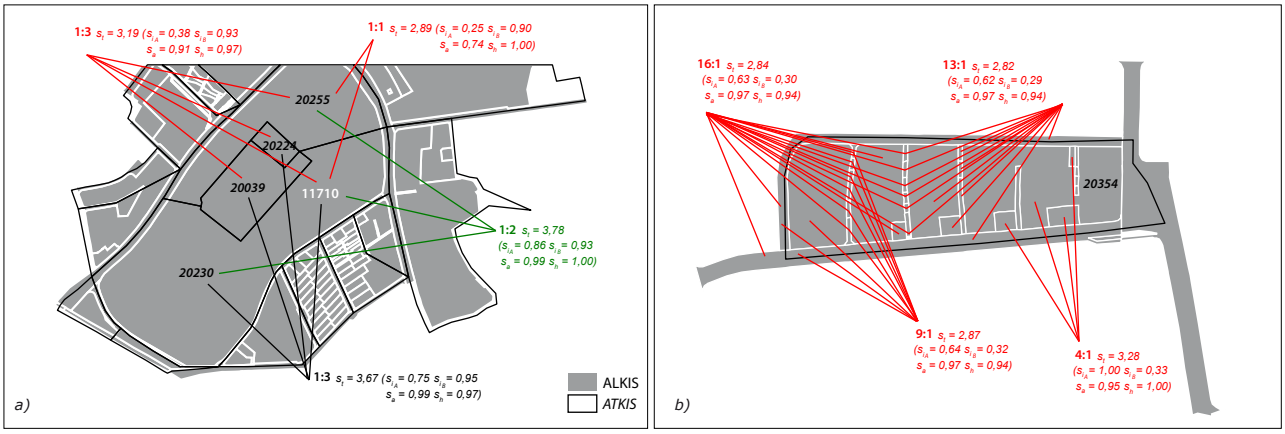


Abbildung 6.9: Beispiele für vom Verfahren identifizierte Objektrelationen in Testgebiet B. Während die rot markierten Relationen im Zuordnungsprozess aufgrund des Schwellwerts für die Flächenparameter wieder verworfen werden, ist die grün gekennzeichnete Relation Bestandteil der endgültigen Ergebnisliste. Bei a) wird die 1:2-Relation ausgewählt und in b) keine.

Häufigkeitsverteilung der Gesamtähnlichkeitsmaße

Abbildung 6.10 gibt den Überblick über die Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t der identifizierten Objektrelationen im Testgebiet B. Im Vergleich zu Testgebiet A gibt es nur wenige (1,5 %) und ausschließlich 1:1-Relationen mit einem hohen Gesamtähnlichkeitsmaß von $s_t > 3{,}95$. Beteiligt sind hierbei Objekte der Klassen Landwirtschaft (43001) und Wald (43002) mit einer Durchschnittsgröße von 9 Hektar. Mit abnehmenden s_t erhöht sich die Anzahl der Relationen und gleichzeitig verringert sich die durchschnittliche Objektgröße.

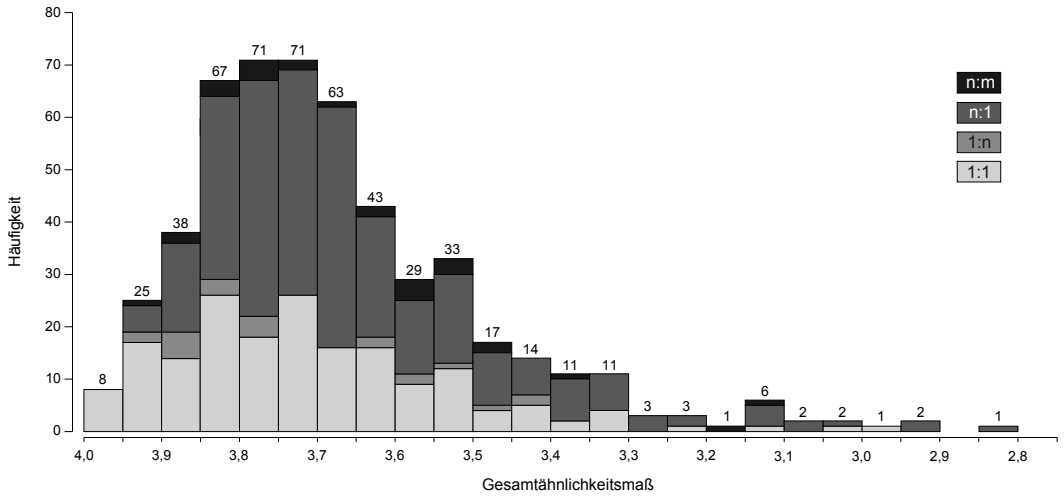


Abbildung 6.10: Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t für die im Zuordnungsprozess identifizierten endgültigen Relationen in Testgebiet B. Die Häufigkeiten sind für die unterschiedlichen Relationsarten dargestellt.

Die Hälfte aller Relationen (52 %) besitzt einen Gesamtähnlichkeitswert im Bereich $3{,}65 < s_t < 3{,}85$. Dies entspricht, bezogen auf den Maximalwert von $s_t = 4{,}0$, einer Ähnlichkeit von 91 bis 96 %. Für 86 % aller Relationen wird immer noch ein hoher Wert von $s_t > 3{,}50$ (87,5 % Ähnlichkeit) ermittelt. Der Anteil der n:1-Relationen ist, bis auf zwei Ausnahmen, in jeder Häufigkeitsklasse am größten.

Das Ergebnis zeigt, dass zwischen den zugeordneten Objekten Abweichungen bestehen. Aufgrund der unterschiedlichen Erfassungsmaßstäbe ist dies erwartungsgemäß. Bei 86 % aller Objektrelationen ist der Flächenparameter bezogen auf ALKIS größer als der auf ATKIS. Dies verdeutlicht, dass ATKIS-Objekte größer sind, als die ihr zugeordneten aggregierten ALKIS-Objekte. Beispielsweise werden ATKIS-Nutzungsflächen bis zur Straßenmitte hin erfasst, wohingegen bei ALKIS eigene flächenhafte Straßenobjekte modelliert werden (vgl. Abbildung 6.9).

Vergleich mit Referenzdaten

Der Experte erstellte auch für Testgebiet B Referenzdaten. Im Vergleich zu Testgebiet A war die Zuordnung schwieriger, da der Suchbereich weit über einen Baublock hinaus ging. Aufgrund der flächenhaften Abdeckung haben Nutzungsobjekte meist mehrere Nachbarn, die wiederum viele Nachbarn besitzen und somit alle zusammen Einfluss auf die Zuordnungsentscheidung haben können. Die Referenzdaten werden für dieses Testgebiet als unsicher eingeschätzt, da aufgrund des vorliegenden Maßstabsunterschiedes vom Experten größere Differenzen zwischen korrespondierenden Objekten erwartet und auch akzeptiert werden. Aus Tabelle 6.4 ergeben sich die in Tabelle 6.9 aufgeführten quantitativen Maße. Abbildung 6.8 visualisiert farblich alle Teilmengen tp, tn, fp und fn auf Objektebene.

Tabelle 6.4: Konfusionsmatrizen für Testgebiet B: ALKIS - ATKIS in Hameln. Links ist die Matrix für die ALKIS-Daten und rechts für ATKIS-Daten. Die obere Zeile bezieht sich auf die Objektanzahlen (#) und die untere Zeile auf die Flächen [ha].

ALKIS

#		Referenz +	Referenz −	Gesamt
Match	+	1.775	36	1.811
	−	1.660	650	2.310
Gesamt		3.435	686	4.121

Fläche [ha]		Referenz +	Referenz −	Gesamt
Match	+	851,17	8,81	859,98
	−	193,31	146,91	340,22
Gesamt		1.044,48	155,72	1.200,20

ATKIS

#		Referenz +	Referenz −	Gesamt
Match	+	591	8	599
	−	138	97	235
Gesamt		729	105	834

Fläche [ha]		Referenz +	Referenz −	Gesamt
Match	+	976,68	5,08	981,76
	−	149,10	69,15	218,24
Gesamt		1.125,78	74,23	1.200,00

In Abbildung 6.8 sind 1.775 ALKIS- und 591 ATKIS-Objekte grau markiert. Sie repräsentieren die richtig positive Teilmenge (tp), also Objekte, die vom Verfahren als Matching-Kandidaten identifiziert und vom Experten bestätigt werden. Alle hellgrauen Objekte werden weder vom Verfahren noch vom Experten zugeordnet und bilden somit die richtig negative Teilmenge (tn). Der Anteil der Matching-Kandidaten die zwar vom Verfahren erkannt werden, aber nicht korrekt sind (fp), sind rot markiert und mit 2 % bei ALKIS und 1 % bei ATKIS sehr gering. Aus den Werten tp und fp leiten sich für jeden Datensatz sehr hohe Genauigkeitswerte bezogen auf die Flächen ab, die sogar über den Genauigkeitswerten der Objektzahlen liegen. Das beste Ergebnis wird mit 99,5 % für ATKIS bestimmt.

Auffällig ist, dass der Experte in beiden Datensätzen deutlich mehr Objekte zuordnet als das entwickelte Verfahren. Die rosa gefärbten Objekte repräsentieren diese Teilmenge (fn). Abbildung 6.9 zeigt ausgewählte Beispiele, in denen das Zuordnungsverfahren versagt, weil die Aggregation zu früh abbricht. Dies kann zwei Gründe haben. Entweder bewirkt die Hinzunahme eines weiteren Nachbarobjektes keine Verbesserung des Gesamtähnlichkeitsmaßes oder der Suchbereich ist zu klein gewählt, so dass keine weiteren Objekte mehr aggregiert werden können. Im Vergleich zum entwickelten Verfahren ordnet der Experte im ALKIS Datensatz etwa doppelt so viele Objekte zu, was zu einem geringen Sensitivitätswert von 51,7 % führt. Bei ATKIS werden etwa ein Fünftel mehr Objekte durch den Experten zugeordnet. Viele dieser zusätzlich zugeordneten Objekte sind deutlich kleiner, was in relativ hohen Flächen-Sensitivitätswerten von 81,5 % (ALKIS) bzw. 86,8 % (ATKIS) resultiert.

Tabelle 6.5 gibt einen detaillierten Überblick über die quantitativen Maße pro Objektklasse und Datensatz. Für jede Objektklasse ist der höchste Prozentanteil grau hervorgehoben. Die Spalten fp und fn zeigen die Unterschiede zwischen dem Verfahren und der Referenzlösung. Ein großer Unterschied ist bei der ALKIS-Klasse Weg (42006) zu erkennen. Während der Experte für 73,3 % der Objekte Matching-Kandidaten im ATKIS-Datensatz identifiziert, obwohl es in ATKIS keine flächenhafte Entsprechung gibt, erkennt das Verfahren nur 8,4 %. Der Experte entscheidet in diesem Fall, Weg-Objekte, die z.B. zwischen zwei Nutzungsflächen liegen oder eine Nutzungsfläche umschließen, mit den Flächen zu aggregieren und dann zuzuordnen. Das vorgestellte Zuordnungsverfahren verfährt genauso, da es nicht in der Lage ist, räumlich getrennte Objekte einem anderen Objekt zuzuordnen. Durch die Hinzunahme von Wegen, die oft über Straßenblöcke hinaus gehen, verschlechtert sich häufig das Gesamtähnlichkeitsmaß, so dass letztendlich Zuordnungen ohne Weg-Objekte in die endgültige Zuordnungliste aufgenommen werden.

Ein ähnliches Verhalten ist für viele ALKIS-Straßenverkehrsobjekte (42001) ($fn = 43,71\,\%$) und schmale Fließgewässer (44001) ($fn = 42,11\,\%$) zu erkennen. Beide Objektklassen haben, anders als bei den Wegen,

Tabelle 6.5: Quantitive Teilmengen tp, tn, fp und fn pro Objektklasse und Datensatz für Testgebiet B: ALKIS - ATKIS in Hameln. Grau hinterlegte Werte spiegeln den jeweils höchsten Prozentanteil wider.

	ALKIS								ATKIS							
	tp		*tn*		*fp*		*fn*		*tp*		*tn*		*fp*		*fn*	
Code	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
41001	842	70,99	19	1,60	4	0,34	321	27,07	194	79,84	6	2,47	-	-	43	17,70
41002	290	72,68	38	9,52	2	0,50	69	17,29	40	74,07	6	11,11	1	1,85	7	12,96
41006	5	55,56	-	-	-	-	4	44,44	107	79,85	8	5,97	-	-	19	14,18
41007	66	66,67	8	8,08	3	3,03	22	22,22	42	91,30	1	2,17	2	4,35	1	2,17
41008	203	64,04	28	8,83	1	0,32	85	26,81	24	61,54	5	12,82	-	-	10	25,64
41009	1	20,00	1	20,00	-	-	3	60,00	5	83,33	-	-	-	-	1	16,67
42001	84	12,28	285	41,67	16	2,34	299	43,71	4	21,05	15	78,95	-	-	-	-
42006	76	8,35	160	17,58	7	0,77	667	73,30	-	-	-	-	-	-	-	-
42009	21	24,71	14	16,47	1	1,18	49	57,65	10	52,63	2	10,53	1	5,26	6	31,58
42010	19	20,43	47	50,54	-	-	27	29,03	4	22,22	14	77,78	-	-	-	-
42016	3	33,33	6	66,67	-	-	-	-	-	-	1	100,00	-	-	-	-
43001	101	63,92	2	1,27	-	-	55	34,81	84	80,00	2	1,90	-	-	19	18,10
43002	24	48,00	6	12,00	1	2,00	19	38,00	32	65,31	8	16,33	1	2,04	8	16,33
43003	5	35,71	2	14,29	-	-	7	50,00	6	40,00	2	13,33	-	-	7	46,67
43007	17	40,48	16	38,10	1	2,38	8	19,05	7	36,84	3	15,79	-	-	9	47,37
43008	-	-	-	-	-	-	-	-	15	31,25	23	47,92	3	6,25	7	14,58
44001	15	26,32	18	31,58	-	-	24	42,11	14	100,00	-	-	-	-	-	-
44006	3	75,00	-	-	-	-	1	25,00	3	60,00	1	20,00	-	-	1	20
$\sum$	1.775		650		36		1.660		591		97		8		138	

auch Objekte im ATKIS-Datensatz, jedoch ist das Verhältnis mit 2 % bei Straßenverkehr und 25 % bei Fließgewässer sehr gering. Hervorzuheben ist, dass alle ATKIS-Fließgewässer vom Verfahren erkannt und durch die Referenzlösung bestätigt werden.

6.2.3 Testgebiet C: ATKIS - GDF in Hannover-Wedemark

Bevor die Ergebnisse für Testgebiet C vorgestellt werden, wird erneut auf die Besonderheiten des Testszenarios hingewiesen. Im Vergleich zu den anderen Testgebieten beinhalten beide Datensätze thematisch unterschiedliche Informationen. Es ist nicht sicher, dass überhaupt identische Objekte vorhanden sind, die das Zuordnungsverfahren bestimmen kann. Des Weiteren ist der Suchraum, in dem potentielle Matching-Kandidaten gefunden werden müssen, vergrößert, da innerhalb jeden Datensatzes Objektüberlagerungen möglich sind. Das heißt, an einer geographischen Position können mehr als zwei Objekte mit unterschiedlichen Objektklassen vorhanden sein, die überprüft werden müssen.

Tabelle 6.6 gibt einen Überblick über die Menge und Art der erzielten Objektrelationen R_o im Testgebiet C. Insgesamt werden $|R_o| = 209$ Relationen ermittelt, in denen 1.638 (7,5 %) ATKIS-Objekte 219 (41,7 %) GDF-Objekten zugeordnet werden. Obwohl siebenmal mehr ATKIS-Objekte zugeordnet werden, überdecken die GDF-Objekte etwa 11 % mehr Fläche. Die Objektrelationen repräsentieren Korrespondenzen zwischen 29 von 63 ATKIS- und 15 von 19 GDF-Objektklassen. Objektklassen, deren Objekte in keiner Relation vertreten sind, sind in den Objektklassenübersichten C.1 und C.2 grau eingefärbt.

Tabelle 6.6: Ergebnisse des Data-Matching-Verfahrens für Testgebiet C: ATKIS - GDF in Hannover-Wedemark.

Kardinalität ATKIS:GDF		R_o Anzahl	R_o Prozent	Objektanzahl ATKIS	Objektanzahl GDF	Objektanzahl Gesamt	Objektanzahl Prozent	Flächenanteil ATKIS in [ha]	Flächenanteil Prozent	Flächenanteil GDF in [ha]	Flächenanteil Prozent
1:1		55	26,32	55	55	110	5,92	1.102,03	7,08	1.152,54	6,56
1:n	homogen	1	0,48	1	2	3	0,16	4,31	0,03	3,86	0,02
	heterogen	-	-	-	-	-	-	-	-	-	-
n:1	homogen	74	35,41	478	74	552	29,73	5.014,29	32,19	5.680,19	32,34
	heterogen	70	33,49	1.028	70	1.098	59,13	7.574,47	48,63	8.403,67	47,84
n:m	homogen	6	2,87	31	12	43	2,32	1.087,96	6,98	1.437,07	8,18
	heterogen	3	1,44	45	6	51	2,75	792,79	5,09	888,49	5,06
Gesamt		209	100	1.638	219	1.857	100	15.575,86	100	17.565,83	100

Mit einem Anteil von 69 % stellen die komplexen n:1 -Relationen die stärkste Relationsart dar, d.h. mehrere ATKIS -Objekte werden aggregiert und einem GDF -Objekt zugeordnet. Bei dem vorliegenden Maßstabsunterschied zwischen 1:10.000 von ATKIS gegenüber 1:25.000 von GDF war dies zu erwarten. Im Durchschnitt werden 10 ATKIS -Objekte einem GDF -Objekt zugeordnet. Das Verfahren findet eine Relation, in der 109 ATKIS -Objekte von sieben verschiedenen Objektklassen (Wald (# 52), Ackerland (# 22), Grünland (# 20), Fläche z.Z. unbestimmbar (# 6), Binnensee (# 6), Fläche gemischter Nutzung (# 1), Gehölz (# 1)) einem GDF-Objekt (Administrative area order 9) zugeordnet werden. Insgesamt sind 28 ATKIS -Objektklassen an allen n:1 -Relationen beteiligt. Die Objektklassen Heide (A4104), Moor, Moos (A4105) und Freizeitanlage (A2202) haben die höchsten Anteile. Alle 144 Relationen umfassen 89 % der zugeordneten Objekte und überdecken 80 % der zugeordneten Flächen. Es ist eine Gleichverteilung zwischen homogenen und heterogenen Relationen vorhanden.

Die zweitstärkste Relation ist mit 26 % die 1:1 -Relation. Diese 55 Relationen umfassen etwa 6 % aller zugeordneten Objekte mit einem Flächenanteil von 6,5 %. Insgesamt sind in ihnen 10 ATKIS - und 11 GDF -Objektklassen vertreten. Trotz des Maßstabsunterschiedes wird eine homogene 1:n -Relation identifiziert, indem ein Objekt der Klasse Wald, Forst (A4107) zwei Objekten der Klasse Land Cover: Forest (Woodland) (G7120) zuordnet wird. Das Gesamtähnlichkeitsmaß wird für diese Relation mit $s_t = 3,09$ ($s_{i_A} = 0,53$, $s_{i_B} = 0,59$, $s_a = 0,97$ und $s_h = 1,00$) bestimmt.

Im Testgebiet wird nur ein kleiner Anteil (4 %) an beidseitig zusammengefassten n:m -Relationen in die endgültige Ergebnisliste übernommen. Sie umfassen ca. 5 % aller zugeordneten Objekte mit einem Flächenanteil von 12 %. Insgesamt sind 6 ATKIS - und 3 GDF -Objektklassen an den 9 Relationen beteiligt. Durch das Verfahren werden auch Relationen mit sehr vielen Objekten identifiziert, z.B. 1617:3 mit $s_t = 2,93$ ($s_{i_A} = 0,76$, $s_{i_B} = 0,83$, $s_a = 0,72$ und $s_h = 0,62$). Nach Prüfung auf doppelte Objekteinträge werden solche Relationen wieder verworfen, da mindestens ein beteiligtes Objekt Teil einer anderen Relation mit einem höheren Gesamtähnlichkeitsmaß ist.

Häufigkeitsverteilung der Gesamtähnlichkeitsmaße

In gleicher Art und Weise wie für Testgebiet A und B gibt Abbildung 6.11 einen Überblick über die Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t der identifizierten Objektrelationen im Testgebiet C. Für 3 % der Objektrelationen werden sehr hohe Gesamtähnlichkeitsmaße von $s_t >$ 3,95 bestimmt. Die 7 Objektrelationen setzen sich aus drei 1:1 - und vier n:1 -Relationen mit wenigen Relationsteilnehmern zusammen. Insgesamt haben die n:1 -Relationen bis auf wenige Ausnahmen immer den größten Anteil. Die meisten Relationen, etwa 12 %, besitzen einen Gesamtähnlichkeitswert im Bereich von $3,75 < s_t < 3,80$. Dies entspricht bezogen auf den Maximalwert von $s_t = 4,0$ einer Ähnlichkeit von 94 % bis 95 %. Für 169 Relationen (80 %) wird ein hoher Wert von $s_t >$ 3,50 (87,5 % Ähnlichkeit) ermittelt.

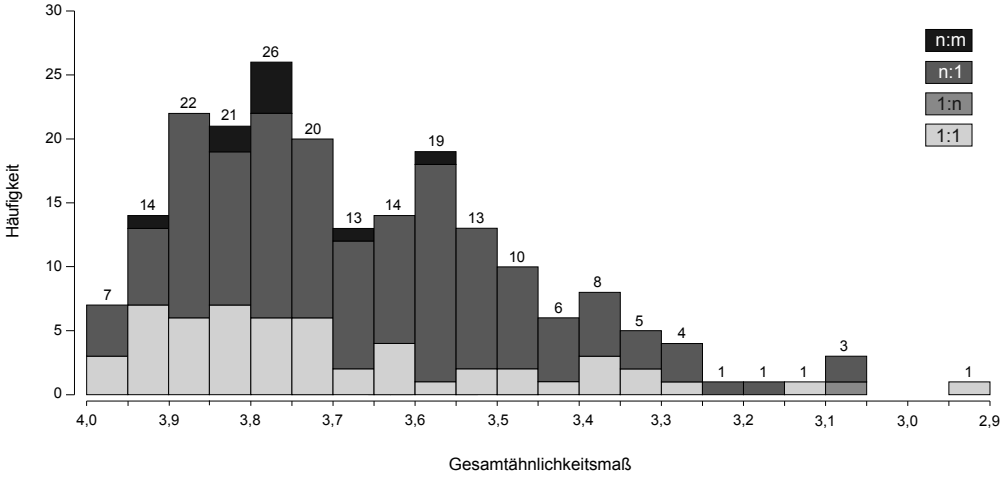


Abbildung 6.11: Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t für die im Zuordnungsprozess identifizierten endgültigen Relationen in Testgebiet C. Die Häufigkeiten sind für die unterschiedlichen Relationsarten dargestellt.

Die Analyse der Einzelparameter zeigt, dass zwischen den zugeordneten Objekten Größenunterschiede bestehen. Aufgrund der unterschiedlichen Erfassungsmaßstäbe ist dies plausibel. Bei 66% aller Objektrelationen ist der Flächenparameter bezogen auf ATKIS größer als der auf GDF. Dies verdeutlicht, dass GDF -Objekte größer sind als die ihr zugeordneten aggregierten ATKIS -Objekte.

Vergleich mit Referenzdaten

Durch einen Experten wurden auch für dieses Testgebiet Referenzdaten erstellt. Die Objektüberlagerungen innerhalb der Datensätze erschwerten jedoch die visuelle Zuordnung. Die Referenzdaten werden als relativ unsicher eingeschätzt. Aus Tabelle 6.7 ergeben sich die in Tabelle 6.9 aufgeführten quantitativen Maße. Auf die farbliche Darstellung der Teilmengen *tp*, *tn*, *fp* und *fn*, analog zu Testgebiet B wird an dieser Stelle bewusst verzichtet, da die Objektüberlagerungen in den Datensätzen und der geringe Anteil der identifizierten Objekte die Erkennbarkeit nicht gewährleisten.

Tabelle 6.7: Konfusionsmatrizen für Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Links ist die Matrix für die ATKIS-Daten und rechts für GDF-Daten. Die obere Zeile bezieht sich auf die Objektanzahlen (#) und untere Zeile auf die Flächen [ha].

ATKIS

#		Referenz +	Referenz −	Gesamt
Match	+	1.291	347	1.638
	−	1.998	18.289	20.287
Gesamt		3.289	18.636	21.925

Fläche [ha]		Referenz +	Referenz −	Gesamt
Match	+	9.804,14	5.771,72	15.575,86
	−	6.434,82	76.457,60	82.892,42
Gesamt		16.238,96	82.229,32	98.468,28

GDF

#		Referenz +	Referenz −	Gesamt
Match	+	190	29	219
	−	45	261	306
Gesamt		235	290	525

Fläche [ha]		Referenz +	Referenz −	Gesamt
Match	+	12.223,00	5.342,83	17.565,83
	−	4.262,80	142.444,93	146.707,73
Gesamt		16.485,80	147.787,76	164.273,56

In Tabelle 6.8 werden analog zu Testgebiet B die Teilmengen *tp*, *tn*, *fp* und *fn* pro Objektart und Datensatz angegeben. Es werden nur diejenigen Objektklassen aufgeführt, die vom Verfahren erkannt werden, d.h. 29 von 63 ATKIS-Klassen und 15 von 19 GDF-Klassen. Für jede Objektklasse ist der höchste Prozentanteil grau hervorgehoben. Für ATKIS werden, bis auf drei Ausnahmen, die höchsten Werte in Spalte *tn* erreicht. Sie kennzeichnen Objekte, für die es keine Entsprechungen im anderen Datensatz gibt, da sie weder vom Verfahren noch vom Experten identifiziert werden. Im Gegensatz dazu werden bei GDF die höchsten Prozentwerte für Spalte *tp* bestimmt. Damit werden die durch das Verfahren bestimmten Relationen bestätigt. Die Spalten *fp* und *fn* spiegeln die Unterschiede zwischen dem Verfahren und dem Experten wider. Im ATKIS-Datensatz stechen die Objektklassen Strom, Fluß, Bach (A5101) und Kanal (Schifffahrt) (A5102) besonders heraus. Hier werden mehr als 75 % ihrer Gruppenmitglieder vom Experten zugeordnet, aber nicht vom Verfahren. Abbildung 6.12 zeigt die Unterschiede in der Objektmodellierung, die bei der Objektzuordnung zu Problemen führt. In diesem Ausschnitt stehen 17 ATKIS-Objekte einem GDF-Objekt gegenüber. ATKIS-Objekte werden unterteilt, wenn z.B. Brücken oder Unterführungen diese Objekte kreuzen (siehe Vergrößerungen 6 und 7) oder ein Blattschnitt durch dieses Gebiet geht (siehe Vergrößerungen 1 und 6).

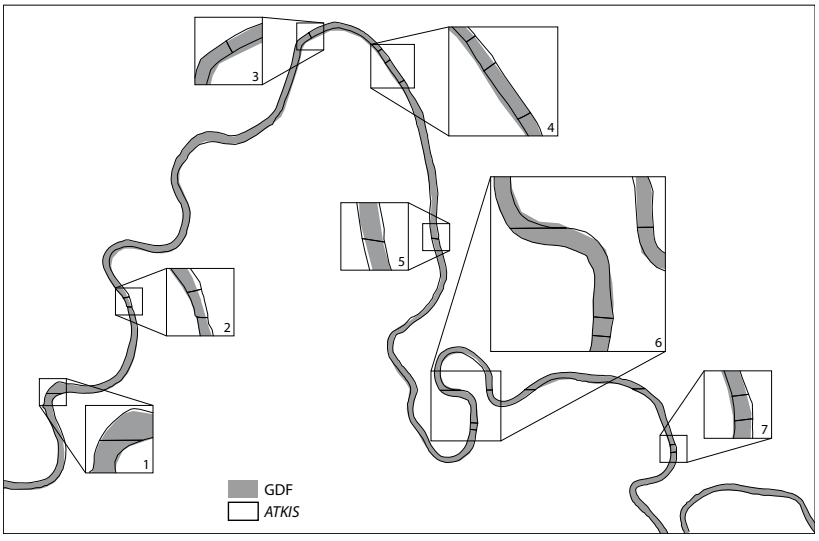


Abbildung 6.12: Unterschiede in der Objektmodellierung im Testgebiet C: ATKIS - GDF in Hannover-Wedemark.

Tabelle 6.8: Quantitative Teilmengen tp, tn, fp und fn pro Objektklasse und Datensatz für Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Grau hinterlegte Werte spiegeln den jeweils höchsten Prozentanteil wider.

	ATKIS							
	tp		*tn*		*fp*		*fn*	
Code	#	%	#	%	#	%	#	%
A2101	61	19,87	86	60,59	44	14,33	16	5,21
A2111	256	4,09	5.698	91,02	26	0,42	280	4,47
A2112	130	14,69	487	66,26	31	4,22	87	11,84
A2113	60	4,16	1.248	86,55	27	1,87	107	7,42
A2114	13	2,03	594	92,96	4	0,63	28	4,38
A2126	-	-	3	75,00	1	25,00	-	-
A2132	1	1,12	88	98,88	-	-	-	-
A2201	3	0,91	309	93,35	1	0,30	18	5,44
A2202	47	26,11	86	47,78	1	0,56	46	25,56
A2213	-	-	337	98,54	3	0,88	2	0,58
A2227	66	8,94	588	79,67	45	6,10	39	5,28
A2301	2	10,53	12	63,16	-	-	5	26,32
A3103	17	6,51	232	88,89	3	1,15	9	3,45
A3501	6	4,88	105	85,37	3	2,44	9	7,32
A3514	5	3,76	122	91,73	1	0,75	5	3,76
A4101	33	1,66	1.847	92,72	54	2,71	58	2,91
A4102	30	1,34	2.030	90,71	29	1,30	149	6,66
A4103	2	0,59	315	92,65	1	0,29	22	6,47
A4104	2	6,06	14	42,42	11	33,33	6	18,18
A4105	49	39,84	35	28,46	1	0,81	38	30,89
A4107	435	18,61	1.230	52,61	43	1,84	630	26,95
A4108	5	2,21	189	83,63	1	0,44	31	13,72
A4109	1	1,64	60	98,36	-	-	-	-
A4199	17	1,24	1.178	86,05	11	0,80	163	11,91
A5101	17	20,48	3	3,61	-	-	63	75,90
A5102	3	2,27	6	4,55	-	-	123	93,18
A5112	30	6,71	368	82,33	2	0,45	47	10,51
A7211	-	-	6	85,71	1	14,29	-	-
A7302	-	-	11	78,57	3	21,43	-	-
$\sum$	1.291				347			

	GDF							
	tp		*tn*		*fp*		*fn*	
Code	#	%	#	%	#	%	#	%
G1120	11	18,97	38	65,52	9	15,52	-	-
G3110	39	46,99	35	42,17	7	8,43	2	2,41
G3136	5	4,95	93	92,08	1	0,99	2	1,98
G4160	12	41,38	17	58,62	-	-	-	-
G4311	1	33,33	-	-	-	-	2	66,67
G4312	1	100,00	-	-	-	-	-	-
G4313	4	57,14	3	42,86	-	-	-	-
G4314	9	81,82	1	9,09	-	-	1	9,09
G4315	2	100,00	-	-	-	-	-	-
G7120	66	64,71	3	2,94	4	3,92	29	28,43
G7170	3	50,00	1	16,67	1	16,67	1	16,67
G7500	2	5,56	34	94,44	-	-	-	-
G9353	1	100,00	-	-	-	-	-	-
G9715	25	49,02	14	27,45	5	9,80	7	13,73
G9725	9	69,23	2	15,38	2	15,38	-	-
$\sum$	190				29			

Für den ATKIS-Datensatz wird in Bezug auf die Objekte ein Genauigkeitswert von 78,8 % erreicht. Das heißt, mehr als drei Viertel der vom Verfahren zugeordneten Objekte werden durch den Experten bestätigt. Der Sensitivitätswert halbiert sich allerdings auf 39,3 %. Der Experte kann doppelt so viele ATKIS-Objekte zuordnen wie das vorgestellte Verfahren. Bei den GDF-Objekten steigert sich der Genauigkeitswert auf 86,8 % und der Sensitivitätswert liegt mit 81 % nur knapp darunter.

6.2.4 Zusammenfassung der Data-Matching-Ergebnisse

Tabelle 6.9 stellt die quantitativen Maße Genauigkeit, Sensitivität und F-Maß der drei Testgebiete gegenüber. Die Genauigkeit entspricht der Exaktheit des Zuordnungsalgorithmus und die Sensitivität spiegelt die Vollständigkeit wider. Das F-Maß fasst beide Maße in einem Wert zusammen.

Für Testgebiet A, in dem zwei Gebäudedatensätze zugeordnet werden, wird eine sehr hohe Zuordnungsqualität erreicht. Alle Werte liegen über 92 %. Die vom Verfahren identifizierten Objektkorrespondenzen passen sehr gut zur Referenzlösung. Das definierte Gesamtähnlichkeitsmaß sowie die festgelegten Schwellwerte sind geeignet, um für dieses Testszenario zuverlässige Objektrelationen abzuleiten. Der Genauigkeitswert für die zugeordneten OSM-Objekte ist mit 97,3 % am höchsten. Bezogen auf die Flächen nimmt der Genauigkeitswert leicht ab. Daraus folgt, dass eher größere Objekte falsch zugeordnet werden. Der Sensitivitätswert bezogen auf die Flächen ist für die ALKIS-Objekte mit 96,6 % am höchsten. Das bedeutet, dass eher kleinere Objekte nicht vom Verfahren, aber vom Experten identifiziert werden.

Für Testgebiet B werden die Genauigkeitswerte nochmal gesteigert. Hier werden Objekte zweier Datensätze zugeordnet, die zum gleichen Datenmodell gehören, aber verschiedene Maßstäbe besitzen. Das Erstellen der Referenzlösung war schwierig, besonders in Hinblick auf den unterschiedlichen Umgang mit Weg-Objekten. Der Experte hat Wege überwiegend mit den benachbarten Nutzungsflächen zusammengefasst, während das

Tabelle 6.9: Zusammenfassung der quantitativen Maße Genauigkeit, Sensitivität und F-Maß zur Bewertung der Zuordnungsqualität von allen drei Testgebieten. Die Maße sind sowohl auf Objektanzahlen (#) als auch auf Flächen in % angegeben.

		Testgebiet A		Testgebiet B		Testgebiet C	
		ALKIS	OSM	ALKIS	ATKIS	ATKIS	GDF
Genauigkeit	# [%]	92,9	97,3	98,0	98,7	78,8	86,8
	Fläche [%]	94,8	94,8	99,0	99,5	62,9	69,6
Sensitivität	# [%]	93,6	93,6	51,7	81,1	39,3	80,9
	Fläche [%]	96,6	95,7	81,5	86,8	60,4	74,1
F-Maß	# [%]	93,2	95,4	67,7	89,0	52,4	83,7
	Fläche [%]	95,7	95,3	89,4	92,7	61,6	71,8

Verfahren Wege nur sehr selten aggregiert hat. Für ATKIS liegen die Genauigkeitswerte etwas über denen von ALKIS. Die höheren Genauigkeitswerte bezogen auf die Flächen besagen, dass Objekte, die vom Verfahren falsch identifiziert werden, klein sind. Die schlechten Vollständigkeitsmaße zeigen, dass viele vom Experten identifizierte Matching-Kandidaten nicht durch das Verfahren bestimmt werden. Hierzu zählen in erster Linie Straßen-, Weg- und Platz-Objekte. Der Unterschied zwischen den Datensätzen ist hinsichtlich der Objektanzahl sehr deutlich. Bei ALKIS werden nur 52 % der korrekten Zuordnungen durch das Verfahren gefunden, die aber flächenmäßig erheblich größer sind als die nicht gefundenen Objekte.

Im Vergleich zu den anderen Testgebieten werden in Testgebiet C die schlechtesten Zuordnungsqualitäten erzielt. Dies ist erwartungsgemäß, da hier Objekte von zwei thematisch verschiedenen Datensätzen zugeordnet werden, die unterschiedliche Maßstäbe besitzen. Sowohl für ein Verfahren als auch für einen Experten ist es eine Herausforderung, überhaupt ähnliche Objekte zu identifizieren. Der Genauigkeitswert für die zugeordneten GDF-Objekte ist mit 86,8 % am höchsten und liegt 8 % über dem der ATKIS-Objekte. Die falsch zugeordneten Objekte sind bei beiden Datensätzen flächenmäßig größer. Der Sensitivitätswert des großmaßstäbigen ATKIS-Datensatzes ist mit 39,3 % nochmals 12 % schlechter als bei Testgebiet B, was bedeutet, dass Objekte, die der Experte als Matching-Kandidaten identifiziert, nicht vom Verfahren bestimmt werden. Bei GDF ist der Sensitivitätswert mit 80,9 % doppelt so groß.

Die genaue Analyse der fehlerhaften Zuordnungsergebnisse verdeutlicht unterschiedliche Probleme. Einerseits werden zwar viele Objektkorrespondenzen vom Verfahren erkannt, aber nach der Prüfung auf doppelte Objekteinträge wieder verworfen. Das liegt daran, dass Objekte Teil verschiedener Relationen sind und für jedes Objekt die Relation mit dem höchsten Gesamtähnlichkeitsmaß ausgewählt wird (siehe Abb. 6.6 und 6.9). Die Auswahl geschieht ungeachtet dessen, ob Nachbarobjekte dadurch vielleicht nicht zugeordnet werden. Ein Experte entscheidet in diesen Situationen meist anders.

Andererseits werden durch den Experten Objekte zugeordnet, die große Flächenunterschiede aufweisen. Beispielsweise wird, wie in Abbildung 6.13, eine Kirche, deren Dach zerstört wurde, in einem Datensatz als gebäudeüberdeckende Fläche modelliert, während im anderen Datensatz nur Außenmauern erfasst werden. Für den Experten handelt es sich um das gleiche Objekt, da sowohl die Position als auch Details übereinstimmen. Die Relation wird nicht in die finale Zuordnungsliste aufgenommen, da ein Flächenparameter unterhalb des festgelegten Schwellwerts $s_{i_j} = 0{,}50$ liegt. Durch Reduzierung des Schwellwerts kann zwar dieser Spezialfall korrigiert werden, aber es können dadurch auch neue, ungewollte Fehlzuordnungen entstehen.

Des Weiteren wurde beobachtet, dass im OSM-Datensatz Gebäudekomplexe häufig nicht vollständig erfasst wurden und Lücken zwischen Gebäudeteilen vorhanden sind. Das Verfahren kann nur direkte Nachbarobjekte

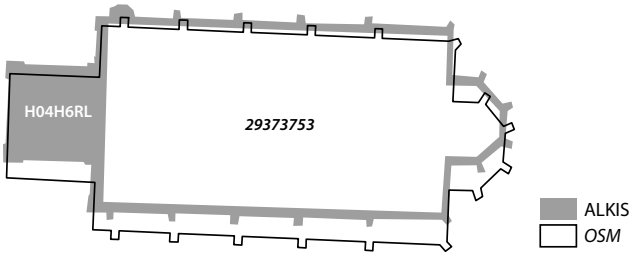


Abbildung 6.13: Beispiel für eine verworfene Objektrelation aufgrund des unterschrittenen Flächenparameters.

aggregieren. Räumlich entfernte Objekte werden vernachlässigt. Durch eine Veränderung der Nachbarschaftsdefinition könnten auch diese Objektrelationen gefunden werden.

Die Zuordnung von langgestreckten Objekten, die meist durch die Aggregation von Einzelobjekten erfolgt, ist für das vorgestellte Zuordnungsverfahren herausfordend. Die Bedingung, dass Nachbarobjekte nur aggregiert werden, wenn sich dadurch das Gesamtähnlichkeitsmaß verbessert, verhindert oftmals die korrekte Zuordnung. Besonders wenn kleinere Objekte aggregiert werden, hat der Ausrichtungsparameter einen negativen Einfluss.

Zusammenfassend lässt sich feststellen, dass das entwickelte Objektzuordnungsverfahren mit sehr einfachen geometrischen Parametern und festgelegten Schwellwerten zuverlässig Objektrelationen in durchaus unterschiedlichen Datensätzen identifizieren kann. Es ist fast unerheblich, ob Datensätze semantisch ähnliche Objektklassen, wie z.B. Gebäude (Testgebiet A) oder semantisch verschiedene Objektklassen (Testgebiet C) besitzen. Es wird beobachtet, dass ein größerer Maßstabsunterschied zwischen den Datensätzen einen geringen Anteil an korrekt erkannten Zuordnungen verursacht.

Für den anschließenden Schema-Matching-Prozess ist viel entscheidender, dass die Mehrheit der identifizierten Objektkorrespondenzen korrekt ist und nur sehr wenige falsche Zuordnungen bestimmt werden. Das Verfahren zeigt Schwächen bei sehr kleinen und sehr großen langgestreckten Objekten. Bei der Entwicklung des Zuordnungsverfahrens stand die Maximierung der Anzahl der zugeordneten Objekte nie im Vordergrund.

6.3 Ergebnisse des Schema-Matching

In diesem Abschnitt werden die Objektklassenzuordnungen für alle drei Testszenarien präsentiert, die auf den Ergebnissen der Objektzuordnung aus Abschnitt 6.2 basieren. Es werden alle identifizierten Objektrelationen berücksichtigt. Die Auswahl wurde nicht auf die positive Teilmenge beschränkt, die durch den Vergleich mit der Referenzzuordnung ermittelt wurde. Es werden Lösungen für die in Kapitel 5 eingeführten Schema-Matching-Verfahren bestimmt. Die Bewertung der Ergebnisse erfolgt mit Hilfe von manuell erstellten Referenzzuordnungen der Objektklassen.

Zunächst werden die Ergebnisse für Testgebiet B vorgestellt. ALKIS und ATKIS gehören zum gleichen Datenmodell und besitzen identische Objektklassen. Bei einheitlicher Klassifizierung der Objekte in beiden Datensätzen wird die Zuordnung identischer Objektklassen erwartet. Für dieses Testszenario hat das Objektzuordnungsverfahren die höchsten Genauigkeitswerte erzielt. Anschließend werden die Ergebnisse für Testgebiet A präsentiert, die auf Objektrelationen mit den höchsten Gesamtähnlichkeitsmaßen beruhen. Die Qualität der Objektrelationen, gemessen an der Höhe des Gesamtähnlichkeitsmaßes, findet jedoch keine Berücksichtigung. Zuletzt werden die Ergebnisse für Testgebiet C gezeigt.

6.3.1 Testgebiet B: ALKIS - ATKIS in Hameln

Die in Abschnitt 6.2.2 vorgestellten Objektrelationen repräsentieren Korrespondenzen zwischen 17 ALKIS- und 16 ATKIS-Objektklassen. Im Anhang gibt B.1 eine Übersicht über alle im Testgebiet B vertretenen Objektklassen und deren Objektanzahlen. Die semantische Übereinstimmung zwischen den Objektklassen ist sehr hoch. Lediglich die ALKIS-Objektklassen Weg (L42006) und Schiffsverkehr (L42016) und die ATKIS-Objektklasse Fläche zur Zeit unbestimmbar (T43008) haben im Untersuchungsgebiet keine eindeutige Entsprechung im jeweils anderen Datensatz.

Im weiteren Verlauf der Arbeit werden zur besseren Lesbarkeit alle Objektklassen mit den kürzeren Klassencodes bezeichnet. ALKIS-Objektklassen werden in diesem Testszenario mit einem vorangestellten L und ATKIS-Objektklassen mit T gekennzeichnet.

Häufigkeitsmatrizen

Für alle Schema-Matching-Verfahren werden als Eingabe Häufigkeitsmatrizen benötigt. Dazu wird hier, wie in Abschnitt 4.1.3 beschrieben, aus allen 522 Objektrelationen eine 17×16 Häufigkeitsmatrix H_R abgeleitet (Tabelle 6.10), in der ALKIS-Objektklassen Zeilen und ATKIS-Objektklassen Spalten repräsentieren. Es fließen auch die Relationen mit ein, die beim Vergleich mit der Referenzlösung als falsch klassifiziert wurden. Es werden Relationsanteile bezogen auf Flächeninhalte berechnet.

Tabelle 6.10: Häufigkeitsmatrix H_R für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). In jeder Zelle steht die Anzahl der Relationen bezogen auf den jeweiligen Flächenanteil, an dem beide Objektklassen beteiligt sind. Die hellgrau, grün und gelb hervorgehobenen Werte spiegeln Zuordnungen zwischen gleichen Objektklassen wider. Die orangefarbene Zelle stellt eine schwache Zuordnung dar.

H_R		ATKIS																
		T41001	T41002	T41006	T41007	T41008	T41009	T42001	T42009	T42010	T43001	T43002	T43003	T43007	T43008	T44001	T44006	$\sum$
ALKIS	L41001	165,47	2,04	49,29	1,70	0,94	0	0	0	0	2,06	0	0	0	0	0	0	221,49
	L41002	6,05	33,54	37,70	3,51	0,08	0	0	0,25	1	1	0	0	0,04	0,03	1	0	84,19
	L41006	0	0	1,59	0	0	0	0	0	0	0	0	0	0	0	0	0	1,59
	L41007	1,05	1	5,70	25,87	0,47	0	0	0,21	0	0	0	0	0	0,03	0	0	34,32
	L41008	5,01	0,24	1,79	2,53	17,08	0	0	2,11	0	23,85	1	0	1,29	0	0	0	54,91
	L41009	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
	L42001	4,99	0,88	3,82	1,35	0,09	0	1,56	0,61	0	0,26	0	2	0,14	3,03	0	0	18,74
	L42006	0,47	0,11	1,08	1,18	1,90	0	0	0,06	0	0,72	0,02	1	0	0	0	0	6,54
	L42009	0,02	0,09	0,03	0,15	0,88	0	0	6,26	0	0,21	0	0	1	0,13	0	0	8,75
	L42010	0	0	0	0	0	0	0	0	1,33	0,71	0	1	0	0,67	0	0	3,71
	L42016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	L43001	0	0,84	0	0	0,78	0	0	0	0	49,86	0	0	1,78	0	0	0	53,26
	L43002	0	0	0	0	0	0	0	0	0	0,35	16,98	0	0	0	0	0	17,33
	L43003	0	0	0	0	0	0	0	0	0	0,19	0,94	0	0	0	0	0	1,13
	L43007	0	0	0	0	0	0	0,44	0	0	0,64	0	1	0,74	0	0,01	0	2,82
	L44001	0	0	0	0	0	0	0	0	0	0,15	0,06	1	0	2	4,99	0	8,21
	L44006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
	$\sum$	183,06	38,75	101	36,27	22,21	1	2	9,49	2,33	80	19	6	5	5,89	7	3	522

Unter idealen Bedingungen, d.h. der Einhaltung einheitlicher Klassifikationsregeln, sollten in der Häufigkeitsmatrix nur auf der Hauptdiagonalen Relationsanteile größer 0 sein. In Tabelle 6.10 werden alle Relationswerte

hellgrau, grün bzw. gelb hervorgehoben, an denen gleiche Objektklassen beteiligt sind. Hier werden nicht immer Maximalwerte bestimmt. Beispielsweise wird für die gelbe Zuordnung L41006 → T41006 (Fläche gemischter Nutzung) nur ein Relationswert von 1,59 erzielt, wohingegen der Wert für L41001 (Wohnbaufläche) → T41006 (Fläche gemischter Nutzung) mit 49,29 deutlich höher ist. Das zeigt, dass die Zuweisung der Objektklassen in beiden Datensätzen unterschiedlich erfolgt.

Nur 91 von 272 Zellen (33 %) haben überhaupt Relationsanteile. Das Histogramm in Abbildung 6.14 gibt einen Überblick über die Verteilung und Höhe der Relationsanteile. Lediglich für 49 durch graue Rechtecke gekennzeichnete Objektklassenkombinationen (18 %) werden eine oder sogar mehrere Relationen zwischen ihren Objektinstanzen identifiziert. Für fünf Objektklassenkombinationen werden zwischen zwei und drei Relationen festgestellt. Den größten Anteil hat mit 165,47 (31,7 %) die Klassenkombination L41001 → T41001 (Wohnbaufläche). Beide Objektklassen haben im Testgebiet mit 28,8 % (ALKIS) bzw. 29,1 % (ATKIS) den jeweils größten Objektanteil.

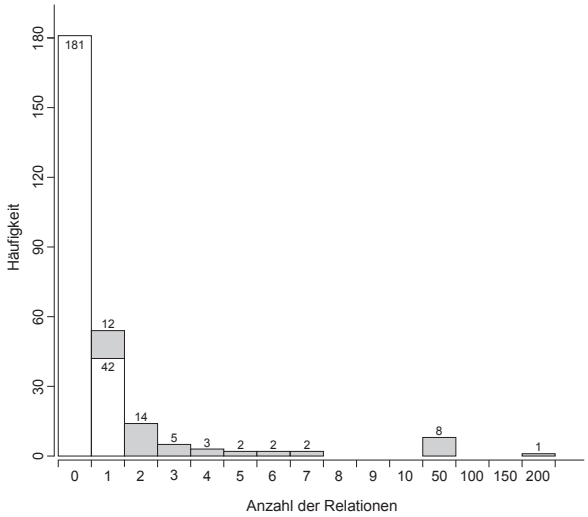


Abbildung 6.14: Histogramm der Relationsanteile der Häufigkeitsmatrix H_R für das Beispiel B: ALKIS - ATKIS in Hameln (17×16). *Die weißen Balken kennzeichnen Relationsanteile kleiner Eins.*

Die in der Häufigkeitsmatrix grün hervorgehobenen Objektklassenzuordnungen L41009 → T41009 (Friedhof) und L44006 → T44006 (Stehendes Gewässer) sind eindeutig. Während die erstgenannte Korrespondenz durch eine 1:5-Objektrelation bestätigt wird, umfasst die andere Zuordnung drei 1:1-Relationen. Beide Fälle spiegeln, trotz der geringen Zahlen, eindeutige Schemarelationen wider, da sich keine weiteren Objekte dieser Klassen unter den Objektrelationen befinden.

Im Gegensatz dazu repräsentiert die orangefarbene 1:1-Relation L44001 (Fließgewässer) → T43003 (Gehölz) eine schwache Korrespondenz. Lediglich 1 von 15 Fließgewässerobjekten mit einer Fläche von 1.794,92 m^2 gegenüber 223.501,95 m^2 werden einem von sechs Gehölzobjekten mit 2.089,47 m^2 von insgesamt 30.862,79 m^2 zugeordnet. Das Beispiel verdeutlicht, dass eine Auswertung hinsichtlich der Flächen eine andere Sicht auf die Zuordnung liefern kann. Aus diesem Grund werden zusätzlich Häufigkeitsmatrizen mit datensatzbezogenen, prozentualen Flächenanteilen bestimmt. Dafür werden die zugeordneten Objektflächen sowohl auf alle zugeordneten ALKIS- als auch ATKIS-Flächen bezogen:

$$h_{A_{(a_i,b_j)}} = \frac{F_{A_{(a_i,b_j)}}}{\sum_{j=1}^{m} F_{A_{a_i}}} \cdot 100 \quad \text{bzw.} \quad h_{B_{(a_i,b_j)}} = \frac{F_{B_{(a_i,b_j)}}}{\sum_{i=1}^{n} F_{B_{b_j}}} \cdot 100 \ , \tag{6.4}$$

wobei $F_{A_{(a_i,b_j)}}$ die zugeordneten ALKIS-Flächen und $F_{B_{(a_i,b_j)}}$ die zugeordneten ATKIS-Flächen kennzeichnen. Des Weiteren werden Häufigkeitswerte h_{AB} bezogen auf beide Datensätze abgeleitet:

$$h_{AB_{(a_i,b_j)}} = \frac{F_{A_{(a_i,b_j)}} + F_{B_{(a_i,b_j)}}}{\sum_{j=1}^{m} F_{A_{a_i}} + \sum_{i=1}^{n} F_{B_{b_j}}} \cdot 100 \ . \tag{6.5}$$

Tabelle 6.11 verdeutlicht den Unterschied anhand von drei Beispielen, die auch farblich in Tabelle 6.10 hervorgehoben sind. In der grünen Schemarelation werden Flächen der Objektklasse L41009 ausschließlich Flächen

von T41009 und umgekehrt zugeordnet. Diese eindeutige Korrespondenz ist durch den Wert 100 in allen Häufigkeitsmatrizen H_A, H_B und H_{AB} zu erkennen. Die orangefarbene Schemarelation ist keineswegs eindeutig. Während L44001-Objekte insgesamt sechs verschiedenen ATKIS-Objektklassen zugeordnet werden, werden T43003-Objekte fünf verschiedenen ALKIS-Objektklassen zugeordnet. Infolgedessen werden nur sehr kleine Werte bestimmt. Die gelbe Schemarelation setzt sich aus drei Objektrelationen zusammen, die aufgrund einer homogenen 1:1-Relation und zwei heterogenen komplexen Relationen insgesamt einen Relationsanteil von 1,59 erzielen. Während die Flächen von L41006 ausschließlich T41006 zugeordnet werden, gekennzeichnet durch $h_A = 100$, finden Objekte der Klasse T41006 in insgesamt acht ALKIS-Objektklassen Matching-Kandidaten. Da der Flächenanteil mit 10.221,59 m^2 gegenüber 1.281.805,57 m^2 sehr klein ist, ergeben sich nur sehr kleine h_B und h_{AB}-Werte. Für das Testgebiet B sind die vollständigen Häufigkeitsmatrizen H_A, H_B und H_{AB} im Anhang in Tabelle B.1 aufgeführt.

Tabelle 6.11: Beispielhafte Schemarelationen R_s mit Einzelhäufigkeiten h_p mit $p = \{R, A, B, AB\}$ der unterschiedlichen Häufigkeitsmatrizen.

Nr.	Schemarelationen R_s	h_R	h_A	h_B	h_{AB}
1	L41009 → T41009 (Friedhof)	1	100	100	100
2	L44001 → T43003 (Fließgewässer)→ (Gehölz)	1	0,80	6,77	1,53
3	L41006 → T41006 (Fläche gemischter Nutzung)	1,59	100	0,80	1,39

Referenzzuordnung der Objektklassen

Für die Bewertung der Ergebnisse wird eine Referenzzuordnung der Objektklassen durch den Vergleich der Objektklassennamen bestimmt. Identische Objektklassennamen werden einander zugeordnet. Die Objektklassen L42006, L42016 und T43008 sind im jeweils anderen Datensatz nicht vertreten. Sie werden unter Berücksichtigung der Relationsanteile und semantischer Ähnlichkeit mit anderen Klassen zusammengefasst.

Tabelle 6.12 stellt die Referenzzuordnung für Testgebiet B vor. Neben einzelnen Clusterhäufigkeiten h_p mit $p = \{R, A, B, AB\}$ werden für jede Matrix auch Gesamthäufigkeiten $H_{\text{Ref},p}$ aller Referenzcluster angegeben.

Tabelle 6.12: Referenzzuordnung für Testgebiet B: ALKIS - ATKIS mit Angabe der Einzelhäufigkeiten h_p mit $p = \{R, A, B, AB\}$ pro Cluster und den Gesamthäufigkeiten $H_{\text{Ref},p}$ aller Referenzcluster. Zeile $H_{\text{total},p}$ gibt die Gesamtmatrixinhalte an.

Nr.	Referenzschemarelationen R_s^*	h_R	h_A	h_B	h_{AB}
1	L41001 → T41001	165,47	74,45	90,57	82,58
2	L41002 → T41002	33,54	59,84	91,07	73,37
3	L41006 → T41006	1,59	100	0,80	1,39
4	L41007 → T41007	25,87	83,29	75,93	79,24
5	{L41008, L42006} → T41008	18,98	34,89	81,59	38,77
6	L41009 → T41009	1	100	100	100
7	L42001 → T42001	1,56	5,96	70,24	9,58
8	L42009 → T42009	6,26	75,02	77,30	76,19
9	L42010 → T42010	1,33	43,52	56,32	49,26
10	{L42016, L44001} → T44001	5,99	194,36	99,10	103,17
11	L43001 → T43001	49,86	98,90	81,62	89,16
12	L43002 → T43002	16,98	99,22	98,57	98,89
13	L43003 → T43003	0	0	0	0
14	L43007 → {T43007, T43008}	0,74	40,96	32,06	35,66
15	L44006 → T44006	3	100	100	100
	$H_{\text{Ref},p}$	332,16	1.110,41	1.055,19	937,24
	$H_{\text{total},p}$	522	1.700	1.600	1.259,68
	[%]	63,63	65,32	65,95	74,40

Insgesamt werden 15 Referenzschemarelationen bestimmt, die sich aus 12 einfachen (1:1) und drei komplexen (1:n bzw. n:1) Relationen zusammensetzen. Relation 13: L43003 → T43003 (Gehölz)[5] besitzt in allen Häufigkeitsmatrizen Nullwerte. In der Objektzuordnung war zwischen den 14 ALKIS - und 15 ATKIS -Objekten keine zuverlässige Zuordnung möglich. Dies zeigt erneut, dass sich die Klassifikationsregeln in beiden Datensätzen unterscheiden oder nicht einheitlich angewendet werden.

Im Testgebiet gibt es 910 ALKIS -Objekte der Klasse L42006 (Weg), denen kein ATKIS -Objekt derselben Klasse gegenübersteht. Für jedes vorgestellte Schema-Matching-Verfahren gilt jedoch die Bedingung, dass alle Objektklassen zugeordnet werden müssen. Basierend auf den Relationsanteilen werden die ALKIS -Klassen L42006 (Weg) und L41008 (Sport-, Freizeit- und Erholungsfläche) zusammengefasst und der ATKIS -Klasse T41008 (Sport-, Freizeit- und Erholungsfläche) zugeordnet. Gleichermaßen wird mit der Klasse Schiffsverkehr verfahren. Während im ALKIS -Datensatz neun Objekte der Klasse vorhanden sind, gibt es im ATKIS -Datensatz nur ein Schiffverkehrsobjekt, für das kein Matching-Kandidat gefunden wird. Daher wird die ATKIS -Klasse Schiffsverkehr in den Häufigkeitsmatrizen ausgeschlossen. Basierend auf den Relationsanteilen werden die Klassen L42016 (Schiffsverkehr) und L44001 (Fließgewässer) der Klasse T44001 (Fließgewässer) zugeordnet. Im Gegensatz dazu wird Schemarelation 14 allein auf Basis einer semantischen Analyse gebildet. Es werden T43007 (Unland, Vegetationslose Fläche) und T43008 (Fläche zur Zeit unbestimmbar) zusammengefasst und L43007 (Unland, Vegetationslose Fläche) zugeordnet.

Die Referenzzuordnung spiegelt im Durchschnitt nur 67,33 % der Objektrelationen wider. Die größte Übereinstimmung wird mit 74,40 % in H_{AB} erzielt, die die Flächenanteile bezogen auf beide Datensätze ausdrückt. Zur Referenzlösung gehören zwei eindeutige 1:1 -Schemarelationen: L41009 → T41009 (Friedhof) und L44006 → T44006 (Stehendes Gewässer) erkennbar am Maximalwert 100 in H_A, H_B und H_{AB}. Bei komplexen Schemarelationen müssen alle Einzelhäufigkeiten addiert werden, so dass der Maximalwert für eine 2:1 -Schemarelation 200 ist. Schemarelation 10 erzielt in der H_A -Matrix mit $h_{(\mathrm{L42016,T44001})} + h_{(\mathrm{L44001,T44001})} = 100 + 94{,}36 = 194{,}36$ fast den Maximalwert.

Im Rahmen der Untersuchungen werden Objektklassenzuordnungen basierend auf allen vier Häufigkeitsmatrizen mit dem einfachen Lösungsverfahren Max-Match, dem Heuristischen Verfahren sowie den Optimierungsverfahren (MaxScore, WeightedSumMaxScoreBalancedSize (WSMSBS), MaxScoreHardConstraintVariableSize (MSHCVS), MaxScoreHardConstraintFixedSize/MaxScoreHardConstraintFixedSizeUnique (MSHCFS-U), MaxScoreHardConstraintFixedSizeNonEmpty (MSHCFSNE)) bestimmt. Tabelle 6.13 fasst die Ergebnisse zusammen und hebt das beste Ergebnis pro Zeile grau hervor. Für Vergleichszwecke werden zusätzlich die Gesamtmatrixinhalte $H_{\mathrm{total},p}$ (Nr. 0a) sowie die Häufigkeiten der Referenzlösung $H_{\mathrm{Ref},p}$ (Nr. 0b) aufgelistet. In den folgenden drei Abschnitten werden die einzelnen Ergebnisse vorgestellt.

Einfaches Lösungsverfahren Max-Match

Mit dem in Abschnitt 5.2.1 beschriebenen Max-Match -Verfahren (Nr. 1) werden durch die Einführung einer Dummy-Spalte die notwendigen quadratischen Matrizen geformt und jeweils 17 Cluster gebildet. In der H_R -Matrix erzielen alle Cluster eine Gesamthäufigkeit von $H_{R(k=17)} = 334{,}29$, was 64,04 % des Gesamtinhalts entspricht. Das Ergebnis ist in Zeile $1 \cap 0a$ abzulesen. Der Vergleich über alle Matrizen zeigt, dass die Cluster in der H_{AB} -Matrix mit 73,39 % ($H_{AB(k=17)} = 924{,}52$) die größte Übereinstimmung haben. Allerdings sind zwei Nullcluster (NC) Bestandteil der Lösung. H_R besitzt drei Cluster der Häufigkeit 0 und H_A nur einen Nullcluster.

Die Bewertung der Verfahren stützt sich in erster Linie auf die Schnittmenge der Clusterlösung mit der Referenzzuordnung. Die Ergebnisse werden in Zeile $1 \cap 0b$ präsentiert. Dazu werden die identischen Zellen zwischen beiden Lösungen betrachtet. Zwischen der H_R -Clusterlösung mit 17 Zellen und der Referenzzuordnung mit 18 Zellen ergibt sich eine Schnittmenge von 11 Zellen, die 292,56 Relationsanteile enthalten. Bezogen auf die Referenzzuordnung entspricht dies 88,38 %. Der größte Prozentwert mit 98,18 % wird für die H_{AB} -Matrix erzielt. Demzufolge wird sie als beste Lösung für das Max-Match-Verfahren ausgewählt und grau markiert.

Heuristisches Lösungsverfahren

Beim Heuristischen Verfahren (Nr. 2) wird, wie in Abschnitt 5.3 vorgestellt, der Min-Cut-Algorithmus rekursiv angewendet. Unter Maximierung der Clusterhäufigkeiten wird die Matrix in zwei Cluster unterteilt. Entstandene Cluster werden solange weiter geteilt, bis keine Teilung mehr möglich ist.

[5] Gehölz – Klassendefinition: ist eine Fläche, die mit einzelnen Bäumen, Baumgruppen, Büschen, Hecken und Sträuchern bestockt ist (AdV, 2008)

Tabelle 6.13: Ergebnisse der verschiedenen Schema-Matching-Verfahren für Testgebiet B: ALKIS - ATKIS in Hameln für H_R, H_A, H_B und H_{AB}. Zeile 0a gibt die Gesamtmatrixinhalte an, während Nr. 0b die Referenzzuordnung kennzeichnet. Zeile 1 präsentiert die Ergebnisse des einfachen Max-Match-Verfahrens, Zeile 2 des Heuristischen Verfahrens und Zeile 3 der Optimierungsverfahren mit dem größten $\emptyset H_{Ze}$. k steht für die Anzahl der Cluster, Spalte H kennzeichnet die Verfahrenslösung, NC repräsentiert die Anzahl der Nullcluster, während NZ die Anzahl der Nullzellen in den Clustern wiedergibt. Alle unter alleiniger Maximierung der Häufigkeiten entstandenen Lösungen sind zusätzlich mit $*$ gekennzeichnet. Unter der Clusteranzahl k steht der Schrankenwert $\{clsize_{var}\}$ bzw. $(clsize)$ in Klammern.

Nr.	Verfahren	H_R				H_A				H_B				H_{AB}			
		k	H	NC	NZ	k	H	NC	NZ	k	H	NC	NZ	k	H	NC	NZ
0a	H_{total}		522,00				1700,00				1600,00				1259,68		
0b	$H_{\text{Ref},p}$	15		1	2	15		1	2	15		1	2	15		1	2
	$0b \cap 0a$		332,16	[63,63%]			1110,41	[65,32%]			1055,19	[65,95%]			937,24	[74,40%]	
1	Max-Match	17		3	3	17		1	1	17		2	2	17		2	2
	$1 \cap 0a$		334,29	[64,04%]			1005,24	[59,13%]			1071,50	[66,97%]			924,52	[73,39%]	
	$1 \cap 0b$		292,56	[88,38%]			998,94	[89,96%]			1013,50	[96,05%]			920,20	[98,18%]	
2	Heur. Verfahren	16		2	2	$*$ 14		0	0	14		1	1	14		1	1
	$2 \cap 0a$		332,35	[63,67%]			1276,34	[75,08%]			1177,47	[73,59%]			1005,33	[79,81%]	
	$2 \cap 0b$		291,82	[87,86%]			1092,63	[98,40%]			1016,44	[96,33%]			926,28	[98,83%]	
3a	MaxScore $(*)$	15		3	3	14		0	0	16		2	2	16		1	1
	$3a \cap 0a$		397,57	[76,16%]			1276,34	[75,08%]			1078,33	[67,40%]			932,83	[74,05%]	
	$3a \cap 0b$		276,65	[83,29%]			1092,63	[98,40%]			1013,50	[96,05%]			928,51	[99,07%]	
3b	WSMSBS	$*$ 15		3	3	$*$ 14		0	0	$*$ 16		2	2	$*$ 16		1	1
	$3b \cap 0a$		397,57	[76,16%]			1276,34	[75,08%]			1078,33	[67,40%]			932,83	[74,05%]	
	$3b \cap 0b$		276,65	[83,29%]			1092,63	[98,40%]			1013,50	[96,05%]			928,51	[99,07%]	
3c	MSHCVS	15		2	2	$*$ 14		0	0	$*$ 16		2	2	$*$ 16		1	1
	$3c \cap 0a$	{1}	388,36	[74,40%]		{1}	1276,34	[75,08%]		{1}	1078,33	[67,40%]		{1}	932,83	[74,05%]	
	$3c \cap 0b$		311,19	[93,69%]			1092,63	[98,40%]			1013,50	[96,05%]			928,51	[99,07%]	
3d	MSHCFS-U	15		2	2	$*$ 14		0	0	$*$ 16		2	2	$*$ 16		1	1
	$3d \cap 0a$	(18)	388,36	[74,40%]		(19)	1276,34	[75,08%]		(17)	1078,33	[67,40%]		(17)	932,83	[74,05%]	
	$3d \cap 0b$		311,19	[93,69%]			1092,63	[98,40%]			1013,50	[96,05%]			928,51	[99,07%]	
3e	MSHCFSNE	15		0	1	$*$ 14		0	0	15		0	0	15		0	0
	$3e \cap 0a$	(18)	376,90	[72,20%]		(19)	1276,34	[75,08%]		(18)	1116,36	[69,77%]		(18)	971,04	[77,09%]	
	$3e \cap 0b$		323,67	[97,44%]			1092,63	[98,40%]			1017,24	[96,40%]			928,51	[99,07%]	

Die H_{AB}-Cluster werden sowohl bezogen auf die Gesamthäufigkeit (79,81 %) als auch auf die Referenzzuordnung (98,83 %) als beste Lösung grau markiert. Tabelle 6.14 zeigt beispielhaft die Entstehungsreihenfolge der Cluster für H_{AB} und gibt die Kantengewichte an, die in jedem Berechnungsschritt geschnitten werden. Als erstes entsteht das rote Cluster $Cl1$: L41001 → T41001 (Wohnbaufläche). Hier werden nur Kanten ohne Kantengewichte geschnitten. Zum Schluss werden das dunkelgraue $Cl13$: L41001 → T41001 (Wohnbaufläche) und das schwarze Cluster $Cl14$: L44006 → T44006 (Stehendes Gewässer) voneinander getrennt, indem zwei Kanten mit einer Summe von $h_{(\text{L42001,T41001})} + h_{(\text{L41001,T41006})} = 5,22 + 33,40 = 38,62$ geschnitten werden.

Für die verschiedenen Häufigkeitsmatrizen waren beim Heuristischen Verfahren unterschiedlich viele Berechnungsschritte notwendig, was sich auch in der Anzahl der Cluster widerspiegelt. Für H_R werden 16 Cluster gebildet, während in allen anderen Matrizen nur 14 Cluster entstehen. Die Anzahl der Nullcluster unterscheidet sich. Die H_A-Lösung hat keine Nullcluster, H_B und H_{AB} haben jeweils eins und H_R hat sogar zwei. In Tabelle 6.15 werden die Ergebnisse der vier verschiedenen Lösungen überlagert. Während hellgraue Zellen Zuordnungen in nur einer Haufigkeitsmatrix kennzeichnen, werden schwarze Zellen in allen Häufigkeitsmatrizen identifiziert. Insgesamt kommen 11 Zellen in allen vier Lösungen vor und repräsentieren somit zuverlässige Zuordnungen.

Optimierungsverfahren

Die in Abschnitt 5.4 vorgestellten Verfahren bestimmen garantiert optimale Lösungen, die auch komplexe Schemarelationen beinhalten können. Allerdings verursachen sie in Einzelfällen hohe Rechenzeiten.

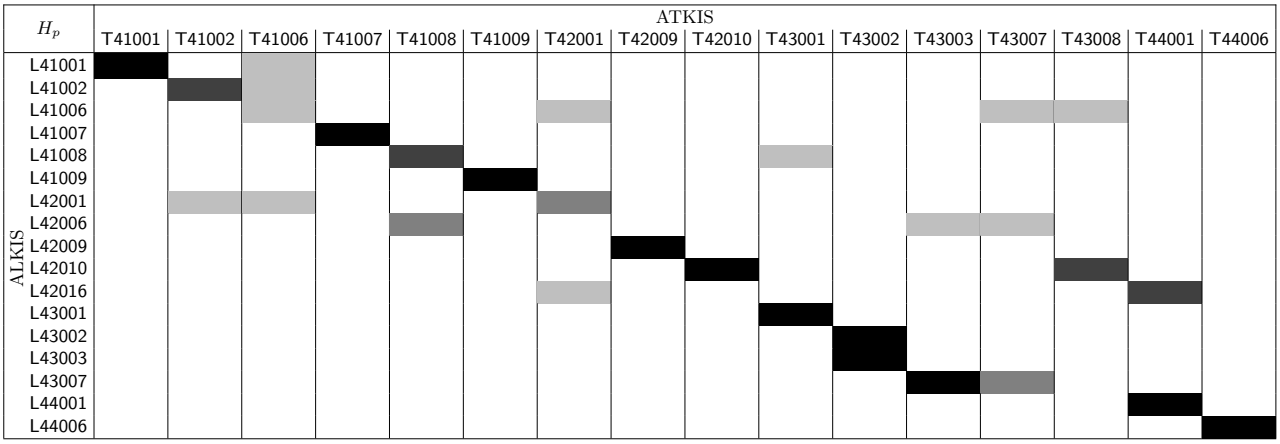
Als erstes werden die Ergebnisse des MaxScore-Verfahrens (Nr. 3a) vorgestellt, bei dem ausschließlich die Häufigkeiten der Cluster maximiert werden. Das Problem wird für jede mögliche Partition k gelöst. Für jede Häufigkeitsmatrix wird die Lösung mit der größten Durchschnittshäufigkeit pro Zelle $\emptyset H_{Ze}$ ausgewählt. Bezogen auf den Gesamtmatrixinhalt wird die $H_{R,(k=15)}$-Lösung mit 76,16 % als beste Lösung grau markiert. Drei der

Tabelle 6.14: Ergebnis des Heuristischen Verfahren für H_{AB} für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). Es werden insgesamt 14 Cluster mit 19 Zellen und einer Gesamthäufigkeit von $H_{AB(k=14)} = 1.005{,}33$ bestimmt.

H_{AB}	ATKIS T41001	T41002	T41006	T41007	T41008	T41009	T42001	T42009	T42010	T43001	T43002	T43003	T43007	T43008	T44001	T44006
ALKIS L41001	**82,58**	0,88	33,40	2,45	0,74	0	0	0	0	0,99	0	0	0	0	0	0
L41002	3,34	73,37	31,61	4,23	0,06	0	0	1,18	1,83	0,73	0	0	0,38	0,06	0,19	0
L41006	0	0	1,39	0	0	0	0	0	0	0	0	0	0	0	0	0
L41007	0,91	3,99	6,25	79,24	1,00	0	0	0,24	0	0	0	0	0	0,11	0	0
L41008	4,19	0,84	2,01	7,31	**30,03**	0	0	3,29	0	21,77	0,52	0	7,47	0	0	0
L41009	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0
L42001	5,22	2,73	**5,28**	4,07	0,97	0	9,58	2,06	0	0,85	0	4,57	5,71	8,49	0	0
L42006	0,50	0,30	1,53	2,44	**8,74**	0	0	3,74	0	1,13	0,12	4,32	0,26	0,09	0	0
L42009	0,04	0,25	0,02	0,98	4,29	0	0	**76,19**	0	0,25	0	0	14,57	1,13	0	0
L42010	0	0	0	0	0	0	0	0	**49,26**	0,54	0	9,32	0	**36,80**	0	0
L42016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8,31	0
L43001	0	0,46	0	0	0,46	0	0	0	0	89,16	0	0	1,20	0	0	0
L43002	0	0	0	0	0	0	0	0	0	0,46	98,89	0	0	0	0	0
L43003	0	0	0	0	0	0	0	0	0	0,24	1,42	0	0	0	0	0
L43007	0	0	0	0	0	0	7,96	0	0	1,52	0	**35,57**	**35,66**	0	0,48	0
L44001	0	0	0	0	0	0	0	0	0	0,16	0,08	1,53	0,10	6,26	94,86	0
L44006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Cluster	*Cl1*	*Cl2*	*Cl3*	*Cl4*	*Cl5*	*Cl6*	*Cl7*	*Cl8*	*Cl9*	*Cl10*	*Cl11*	*Cl12*	*Cl13*	*Cl14*
h_k	100	100	100,31	103,17	0	86,06	89,16	76,19	71,23	38,77	79,24	73,37	82,58	5,28
Geschnittene Kantengewichte	0	0	1,42	8,8	18,93	21,57	29,36	30,66	22,70	21,88	21,90	38,56	38,62	

Tabelle 6.15: Überlagerte Ergebnisse der Heuristischen Lösungen für alle Häufigkeitsmatrizen für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). Je dunkler die Zellen, desto mehr Lösungen beinhalten die Zuordnung.



15 Cluster sind allerdings Nullcluster. Gemessen an der Referenzzuordnung hat die $H_{AB,(k=16)}$-Lösung mit 99,07 % die größte Schnittmenge.

Anschließend werden die Ergebnisse des WeightedSumMaxScoreBalancedSize-Verfahrens (Nr. 3b) vorgestellt, bei dem die Optimierungsziele MaxScore - Bilden von Clustern mit maximalen Häufigkeiten und BalancedSize - Erzeugen von ausgewogenen Clustern hinsichtlich der Zellenanzahl, in einer kombinierten Zielfunktion (siehe Gl. (5.15)) mit dem Gewichtsfaktor $s = 0{,}5$ gleich gewichtet werden. Auffällig ist, dass in allen Häufigkeitsmatrizen diese Lösungen mit den jeweiligen MaxScore-Lösungen (Nr. 3a) übereinstimmen. Der Gewichtsfaktor $s = 0{,}5$ ist für dieses Beispiel zu dominant, verdeutlicht wird dies durch den Vergleich mit den normierten mittleren Gewichtsfaktoren: $s_R = 0{,}36$, $s_A = 0{,}15$, $s_B = 0{,}16$ und $s_{AB} = 0{,}19$. Die Matrixinhalte sind im Vergleich zur Matrixgröße zu gering und demzufolge unterstützt der Gewichtsfaktor $s = 0{,}5$ primär die Optimierung des ersten Ziels.

Die Bestimmung eines geeigneten Gewichtsfaktors, der für beide im Konflikt stehenden Optimierungsziele einen guten Kompromiss findet, ist sehr schwierig. Das Optimierungsmodell wurde dahingehend geändert, dass die Clustergröße hinsichtlich der Zellenanzahl als harte Bedingung eingeführt wurde. Im MaxScoreHardConstraint-VariableSize-Verfahren (Nr. 3c) wird eine variable Clustergröße, die die Differenz zwischen dem größten und dem kleinsten Cluster vorgibt, berücksichtigt, während im (MaxScoreHardConstraintFixedSize-Verfahren (Nr. 3d) eine feste Clustergesamtgröße beachtet wird. Für das Testgebiet liefern beide Verfahren für jede Matrix identische Ergebnisse. Die H_A-Cluster erzielen bezogen auf die Matrixgesamthäufigkeiten die beste Lösung, wohingegen die Schnittmenge zur Referenzzuordnung für die H_{AB}-Cluster am größten ist. Die einzelnen Lösungen stimmen sogar mit den MaxScore-Lösungen überein. Einzige Ausnahme bilden die H_R-Cluster, die einen Anstieg der

Schnittmenge zur Referenzzuordnung verzeichnen, obwohl die Schnittmenge zum Matrixgesamtinhalt gesunken ist.

Abschließend wird mit dem Verfahren MaxScoreHardConstraintFixedSizeNonEmpty (Nr. 3e) untersucht, welche Auswirkungen die Bedingung hat, keine Nullcluster in den Lösungen zuzulassen. Während die Ergebnisse für H_A identisch blieben, ist bei der H_R-Matrix die Schnittmenge mit der Referenzzuordnung gestiegen, aber die Schnittmenge zum Matrixgesamtinhalt gesunken. Für die H_B-Matrix haben sich alle Werte verbessert. Die beste Lösung wird für H_{AB} erreicht und entsprechend in der Tabelle grau hervorgehoben. Obwohl sich die Clustergesamthäufigkeit erhöht hat, ist keine Verbesserung hinsichtlich der Referenzzuordnung zu erkennen.

Zusammenfassung der Ergebnisse für Testgebiet B

Für die Häufigkeitsmatrix H_{AB} werden in Bezug auf die Schnittmengen zwischen den Verfahrenslösungen und der Referenzzuordnung (Nr. 0b) in jedem Fall die besten Ergebnisse erzielt (siehe Tab. 6.13). Die Optimierungsverfahren haben mit 99,07 % die größten Übereinstimmungen. Tabelle 6.16 gibt eine Übersicht über die Anzahl der identischen Cluster mit der Referenzzuordnung in allen H_{AB}-Verfahrenslösungen.

Tabelle 6.16: Testgebiet B: ALKIS - ATKIS in Hameln. Übersicht über die Anzahl der identischen Cluster zwischen den H_{AB}-Verfahrenslösungen und der Referenzzuordnung. Mit (x) *gekennzeichnete Relationen besitzen identische Clusterhäufigkeiten unter Vernachlässigung von Nullzellen.*

Nr.	Referenzschemarelationen R_s^*	Max-Match	Heur. Verfahren	MaxScore WSMSBS MSHCVS MSHCFS-U	MSHCFSNE
1	L41001 → T41001	x	x	x	x
2	L41002 → T41002	x	x	x	x
3	L41006 → T41006	x		x	x
4	L41007 → T41007	x	x	x	x
5	{L41008, L42006} → T41008		x		
6	L41009 → T41009	x	x	x	x
7	L42001 → T42001	x		x	x
8	L42009 → T42009	x	x	x	x
9	L42010 → T42010	x		x	
10	{L42016, L44001} → T44001		x	x	x
11	L43001 → T43001	x	x	x	x
12	L43002 → T43002	x		x	
13	L43003 → T43003				
14	L43007 → {T43007, T43008}	(x)		(x)	(x)
15	L44006 → T44006	x	x	x	x
	Σ	11 (12)	9	12 (13)	10 (11)

Die Optimierungsverfahren MaxScore, WSMSBS, MSHCVA und MSHCFS-U identifizieren 12 der 15 Referenzcluster. Mit dem einfachen Lösungsverfahren Max-Match werden 11 Cluster bestimmt. Das verdeutlicht, dass zwischen beiden Datensätzen überwiegend 1:1-Schemarelationen bestehen. Das MSHCFSNE-Verfahren bestimmt 10 der Referenzcluster exakt, während die geringsten Übereinstimmungen mit dem Heuristischen Verfahren erzielt werden.

Schemarelation 13 wird von keinem einzigen Verfahren gebildet. Dies ist erwartungsgemäß, da das Cluster lediglich eine Nullzelle umfasst und die Erzeugung von Nullclustern nicht zur Maximierung der Clustergesamthäufigkeiten beiträgt. Die Anzahl der identischen Cluster erhöht sich um die mit (x) gekennzeichneten Cluster, die hinsichtlich der Clusterhäufigkeit gleich sind, aber zusätzliche Nullzellen besitzen, die vernachlässigt werden.

Sieben Referenzcluster (Nr. 1, 2, 4, 6, 8, 11 und 15) werden von allen Verfahren identifiziert. Es sind ausschließlich 1:1-Schemarelationen mit großen Clusterhäufigkeiten (vgl. Tabelle 6.12). Relation 10: {L42016, L44001} (Schiffsverkehr, Fließgewässer) → T44001 (Fließgewässer) wird sowohl vom Heuristischen Verfahren als auch von allen Optimierungsverfahren identifiziert, da sie komplexe Relationen bestimmen können.

Obwohl die Optimierungsverfahren MaxScore, WSMSBS, MSHCVS und MSHCFS-U die beste Zuordnung in Bezug auf die Referenzzuordnung bestimmen, liefert der genaue Blick auf die Rechenzeiten bzw. die Vollständigkeit der durchgeführten Berechnungen Hinweise auf die tatsächliche Anwendbarkeit der einzelnen Verfahren in der Praxis. Tabelle 6.17 stellt die Rechenzeiten aller H_{AB}-Verfahrenslösungen vor und kennzeichnet diejenigen, die die größten Übereinstimmungen mit den Referenzschemarelationen R_s^* und die größte Schnittmenge mit der

Tabelle 6.17: Testgebiet B: ALKIS - ATKIS in Hameln. Übersicht über die benötigte Rechenzeit, die Anzahl der ausgeführten Rechenschritte pro H_{AB} -Verfahrenslösung, die besten Ergebnisse bezogen auf die Referenzschemarelationen R_s^*, Schnittmenge zur Referenzzuordnung ($\cap$ 0b) und zum Matrixinhalt ($\cap$ 0a).

Nr.	Verfahren	Rechenzeit	Rechenschritte	R_s^*	$\cap$ 0b	$\cap$ 0a
1	Max-Match	65.37 [sec]	1			
2	Heuristisches Verfahren	1.01 [sec]	13			×
3a	MaxScore	393.91 [sec]	15	×	×	
3b	WSMSBS	2152.88 [sec]	15	×	×	
3c	MSHCVS	>131 [h]	6 ($k = 16, ..., 13$)	×	×	
3d	MSHCFS-U	492.94 [sec]	59	×	×	
3e	MSHCFSNE	>17 [h]	19 ($k = 16, ..., 10$)		×	

Referenzzuordnung (Nr. 0b) und Matrixgesamthäufigkeit (Nr. 0a) haben. Es werden nur Zeiten berücksichtigt, die der IBM ILOG CPLEX Interactive Optimizer 12.5.1.0 für die Ermittlung der Lösungen braucht. Zeiten, die für die Erzeugung der Eingabedateien, für das Einlesen des Problems bzw. das Auslesen und Auswerten der Lösungen benötigt werden, bleiben unberücksichtigt.

Das Heuristische Verfahren liefert mit 1.01 Sekunden am schnellsten eine Lösung. Dafür waren 13 einzelne Min-Cut-Berechnungen notwendig. Das Ergebnis ist nicht garantiert optimal. Im Vergleich zu allen anderen Verfahren weist es zwar die geringste Übereinstimmung mit der Referenzzuordnung auf, aber mit 1.005,33 (79,81 %) die höchste Übereinstimmung mit dem Matrixgesamtinhalt. Der Vergleich zum Matrixgesamtinhalt ist sinnvoll, da Referenzzuordnungen von Expertenmeinungen beeinflusst sind und selten zur Verfügung stehen. Tabelle 6.32 zeigt ein Ranking der Verfahren hinsichtlich Rechenzeit, Referenzzuordnung und Matrixgesamtinhalt und stellt sie den Ergebnissen der anderen Testgebiete gegenüber.

Das Max-Match -Verfahren erzielt in etwas mehr als einer Minute (65.37 Sekunden) die zweitschnellste und gleichzeitig auch in Bezug auf die Referenzzuordnung die zweitbeste Lösung. Im Gegensatz dazu wird mit dem MaxScore -Verfahren in etwa sechseinhalb Minuten (393.91 Sekunden) eine optimale Lösung erzielt, die gleichzeitig die größte Übereinstimmung mit der Referenzzuordnung besitzt. Das Ergebnis zeigt, dass speziell für Testgebiet B die Beeinflussung der Clustergrößen, wie sie in den anderen Optimierungsverfahren vorgenommen wird, nicht notwendig ist. Das gleiche Ergebnis bestimmen auch die Verfahren WSMSBS, MSHCVS und MSHCFS-U. Das vereinfachte MSHCFS-U -Verfahren braucht etwas mehr als acht Minuten (492.94 Sekunden), um die optimale Lösung aus Lösungen von 59 $clsize$ -Werten zu bestimmen. WSMSBS führt 15 Rechenschritte in 36 Minuten durch. Der Gewichtsfaktor $s = 0{,}5$ bewirkt, dass die Maximierung der Häufigkeiten im Vordergrund steht.

Das MSHCVS -Verfahren bestimmt ebenfalls das beste Ergebnis bezogen auf die Referenzzuordnung. Bereits nach 3 Minuten Rechenzeit (223.27 Sekunden) und vier Berechnungsschritten wird aufgrund der stetigen Abnahme der Durchschnittshäufigkeit pro Zelle vorzeitig eine beste Lösung ausgewählt. Um diese Auswahl unter allen möglichen Lösungen zu bestätigen, wird die Berechnung für weitere Schrankenwerte $clsize_{var}$ und verschiedene Partitionen k fortgesetzt. Es wird u.a. eine Lösung für $clsize_{var} = 7$ und $k = 13$ bestimmt. Das heißt, es müssen 13 Cluster bestimmt werden, bei denen die Differenz zwischen dem größten und kleinsten Cluster maximal 7 Zellen betragen darf. Nach mehr als 59 Stunden Rechenzeit wird kein nachgewiesen optimales Ergebnis gefunden. Zwischen der optimalen Lösung und der unteren Schranke (engl. Lower Bound) besteht eine Lücke (engl. Gap) von 4 %. Da die Lücke recht klein ist, wird die Lösung als optimal angenommen. Basierend auf dieser Lösung wird der Schrankenwert auf $clsize_{var} = 2$ verringert. Die Vermutung, dass sich Rechenzeit und Lücke weiter vergrößern, hat sich bestätigt. Nach 72 Stunden wird eine Lösung bestimmt, die noch eine Lücke von 5,95 % aufweist. Die Durchschnittshäufigkeit pro Zelle verringert sich weiter. Es wird auf alle weiteren Berechnungen verzichtet und die vorzeitig ausgewählte Lösung als beste Verfahrenslösung übernommen.

Beim MSHCFSNE -Verfahren werden die Berechnungen ebenfalls vorzeitig abgebrochen. Ausgehend von der maximal möglichen Partition werden in 173.17 Sekunden 11 verschiedene $clsize$-Werte für $k = 16$ bis $k = 11$ getestet. Für $k = 10$ werden 8 Schrankenwerte geprüft. Die Berechnung des letzten Werts $clsize = 23$ erzielt nach mehr als 17 Stunden mit einer Lücke von 7,32 % keine nachgewiesene optimale Lösung. Bezogen auf die Schnittmenge zur Referenzzuordnung bestimmt das Verfahren wie alle anderen Optimierungsverfahren das beste Ergebnis. Beim Vergleich mit den Referenzschemarelationen fällt das Verfahren jedoch auf Rang 3 zurück, da es zwei Schemarelationen weniger bestimmt.

In Tabelle 6.18 werden die besten Lösungen hinsichtlich der Referenzzuordnung und dem Matrixinhalt gegenübergestellt. Während grau markierte Relationen Referenzcluster hervorheben, kennzeichnet □ identische Cluster zwischen den Lösungen und ∘ Nullcluster. Acht Schemarelationen, die gleichzeitig Referenzcluster dar-

Tabelle 6.18: Testgebiet B: ALKIS - ATKIS in Hameln. Gegenüberstellung der Lösungscluster in der H_{AB}-Matrix der Optimierungsverfahren MaxScore, WSMSBS, MSHCVS *und* MSHCFS-U*, als beste Lösung hinsichtlich der Referenzzuordnung, mit denen des Heuristischen Verfahrens, als beste Lösung hinsichtlich des Matrixgesamtinhalts. Die grau markierten Relationen kennzeichnen Referenzcluster, □ markiert identische Cluster zwischen den Lösungen und ○ Nullcluster.*

MaxScore, WSMSBS, MSHCVS, MSHCFS-U					Heuristisches Verfahren				
□	L41001	→	T41001		□	L41001	→	T41001	
□	L41002	→	T41002		□	L41002	→	T41002	
	L41006	→	T41006			L41006	→	T42001	○
□	L41007	→	T41007		□	L41007	→	T41007	
	L41008	→	T41008			{L41008, L42006}	→	T41008	
□	L41009	→	T41009		□	L41009	→	T41009	
	L42001	→	T42001			L42001	→	T41006	
	L42006	→	T43003						
□	L42009	→	T42009		□	L42009	→	T42009	
	L42010	→	T42010			L42010	→	{T42010, T43008}	
□	{L42016, L44001}	→	T44001		□	{L42016, L44001}	→	T44001	
□	L43001	→	T43001		□	L43001	→	T43001	
	L43002	→	T43002			{L43002, L43003}	→	T43002	
	L43003	→	T43008	○					
	L43007	→	T43007			L43007	→	{T43003, T43007}	
□	L44006	→	T44006		□	L44006	→	T44006	

stellen, stimmen in beiden Lösungen überein. Davon ist eine Schemarelation komplex. Die Zusammenfassung der ALKIS-Klassen {L42016 (Schiffsverkehr), L44001 (Fließgewässer) ist nachvollziehbar. Jedes Verfahren erzeugt auch ein Nullcluster. In der MaxScore-Lösung ist das Nullcluster L43003 (Gehölz) → T43008 (Fläche zur Zeit unbestimmbar) semantisch betrachtet plausibel. Beim Heuristischen Verfahren scheint die Zuordnung L41006 (Fläche gemischter Nutzung) → T42001 (Straßenverkehr) eher willkürlich. Auffällig ist, dass das Heuristische Verfahren fünf komplexe Schemarelationen bestimmt hat. Eine Analyse zeigt, dass Objektklassen entweder zusammengefasst werden, die semantisch verwandt sind, z.B. {L43002 (Wald), L43003 (Gehölz)} oder aber für die keine eindeutige Zuordnung möglich ist, z.B. {T43003 (Gehölz), T43007 (Unland, Vegetationslose Fläche)} oder {T42010 (Bahnverkehr), T43008 (Fläche zur Zeit unbestimmbar)}. Für semantisch unpräzise Klassen (T43007 und T43008) kann durch die eingesetzten Matching-Verfahren eine Konkretisierung ihrer Klassendefinition erreicht werden, weil die tatsächliche Klassifizierung der Objekte aufgedeckt werden kann.

Für Testgebiet B wird mit dem Heuristischen Verfahren das beste Ergebnis bezogen auf den Matrixgesamtinhalt und die Rechenzeit erzielt. Obwohl die Lösung nicht garantiert optimal ist, stimmt sie in 15 Zellen (78,9%) mit der besten Optimierungslösung MSHCFSNE, die gleichzeitig die zweitbeste Lösung bezogen auf den Matrixgesamtinhalt ist, überein. Bezogen auf die Referenzzuordnung steht das Heuristische Verfahren im Ranking nur auf dem vierten Platz. Da die Referenzzuordnung hauptsächlich durch den Namensvergleich der Objektklassen vollzogen wurde, konnte eine mögliche Falschklassifikation auf Objektebene nicht berücksichtigt werden. Das Ergebnis der Heuristischen Lösung zeigt, dass die Klassifikationsregeln in beiden Datensätzen überwiegend einheitlich angewendet werden. In sieben der neun einfachen Schemarelationen stehen sich immer identische Objektklassen gegenüber. In allen fünf komplexen 1:2- bzw. 2:1-Schemarelationen ist mindestens eine Objektklasse bei beiden Datensätzen identisch. Es gibt aber auch Hinweise auf abweichende Klassifikationsregeln. Beispielsweise werden ALKIS-Objekte der Klasse L43003 (Gehölz) nur Objekten der Klassen T43001 (Landwirtschaft) bzw. T43002 (Wald) zugeordnet und niemals T43003 (Gehölz).

6.3.2 Testgebiet A: ALKIS - OSM in Hannover

Die in Abschnitt 6.2.1 vorgestellten Objektrelationen repräsentieren Korrespondenzen zwischen 32 ALKIS- und 62 OSM-Objektklassen. Eine Übersicht der im Testgebiet A vertretenen Objektklassen mit Objektanzahlen werden im Anhang in A.1 und A.2 gegeben. Die Objektklassennamen beider Datensätze sind nicht identisch. Mit Hilfe eines Wörterbuchs können Korrespondenzen identifiziert werden, z.B. Krankenhaus → Hospital oder Polizei → Police. Im OSM-Datensatz sind einige Objektklassennamen sowohl in Einzahl als auch in Mehrzahl vertreten, z.B. Office/Offices oder Garage/Garages. Dies resultiert aus der Freiheit, die dem OSM-Datenerfasser bei der Annotation der Daten gegeben wird. Es wird erwartet, dass Klassen verschiedener Numeri im Rahmen des Schema-Matchings zusammengefasst werden. Im weiteren Verlauf der Arbeit werden ALKIS-Klassen mit einem vorangestellten A und OSM-Klassen mit O gekennzeichnet.

Häufigkeitsmatrizen

Analog zu Testgebiet B werden aus den 2.054 Objektrelationen die Häufigkeitsmatrizen H_R, H_A, H_B und H_{AB} abgeleitet. Aufgrund der langen Rechenzeiten in Testgebiet B werden die 32×62-Häufigkeitsmatrizen reduziert. In der H_R-Matrix werden dazu alle Objektklassen entfernt, deren Zeilen- bzw. Spaltensumme kleiner vier ist. Die Zahl wurde willkürlich gewählt. Die reduzierte H_R-Matrix umfasst 17 ALKIS- und 31 OSM-Objektklassen und beinhaltet 1.882 Objektrelationen, was 91,6 % der ursprünglichen Relationen entspricht. Der Anteil an Nullzellen ist mit 80,3 % besonders hoch.

Tabelle 6.19: Häufigkeitsmatrix H_R für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). *In jeder Zelle steht die Anzahl der Relationen bezogen auf den jeweiligen Flächenanteil, an denen beide Objektklassen beteiligt sind. Die grauen Zellen kennzeichnen die Referenzzuordnung.*

H_R		ALKIS																	
		A0931	A0932	A0933	A1101	A1121	A1122	A1123	A1124	A1134	A1141	A1151	A1171	A1401	A1701	A2121	A2361	A2601	Σ
OSM	O2	993,43	5,03	0,26	2	0	0	0	0	0	0	0	0	0	0	16,89	0	0	1.017,61
	O5	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	4
	O7	0	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	7
	O8	3	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	6
	O19	1	1	0	1	0	0	0	0	0	13	0	0	0	0	0	0	0	16
	O28	16,93	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17
	O37	3,12	62,18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65,31
	O38	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30
	O40	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	13
	O44	0,14	0	0	0	0	0	0	0	0	0,07	11,79	0	0	0	0	0	0	12
	O46	4,99	0,01	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	7
	O47	13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
	O48	9,39	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	16,39
	O49	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	O52	3	1,17	0	2,83	0	0	0	0	0	0	0	0	0	0	0	0	0	7
	O62	18,99	0	0,01	2	0	0	0	0	0	0	0	0	1	0	0	0	0	22
	O63	30,14	7,55	0	5	0	0	0	0	0	0	0	0	5	1	1	0	0	49,69
	O65	0	1,86	3	1	0	0	0	0	0	0	0	0	0	0	0	4,14	0	10
	O67	2,43	0	0	1	0	0	0	0	0	14,57	0	0	0	0	0	0	0	18
	O68	0	0	0,04	0	0	0	0	0	0	0	0	6,96	0	0	0	0	0	7
	O70	2	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	6
	O71	16,44	3,55	0	30,38	0	0	1	0	3,40	0	0	0	1	0	0	0	0	55,77
	O74	260,94	8,16	0	1	2	0	0	0	0	0	0	1	1,06	0	1,85	0	0	276
	O75	6	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	9
	O78	0	8	0	0,23	0	0	0,10	0	0	0	0	0	0	1	0	0	0	9,33
	O81	0	0,13	0	0	17,87	12	0	0	0	0	0	0	0	0	0	0	0	30
	O82	3	22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	31
	O92	2	1,27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,73	4
	O96	11,30	8,61	0,02	16	0	0	20,89	35,09	0	0	0	0	0	0	0	0	0	91,90
	O101	0	3,03	0	0	0	0	0	0	0	0	0	0	0	0,97	0	0	0	4
	O103	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23
	Σ	1.426,24	179,61	4,33	68,44	19,87	13	34,98	35,09	4,40	27,64	11,79	7,96	14,06	3,97	19,74	4,14	6,73	1.882

Tabelle 6.19 zeigt aus Platzgründen die reduzierte H_R-Häufigkeitsmatrix in transponierter Form. Alle anderen Matrizen sind im Anhang in Tabelle A.2 zu finden. Durch die Reduzierung werden bereits zwei eindeutige Schemarelationen entfernt: A1112 (Rathaus) → O94 (Townhall) und A2821 (Hallenbad) → O90 (Swimming pool). Dies bestätigt wieder, dass geringe Werte nicht grundsätzlich unbedeutend oder fehlerhaft sind, sondern gerade bedeutende Schemarelationen widerspiegeln können. Beide Datensätze besitzen Objektklassen, für die Relationen zu nur einer Objektklasse im jeweils anderen Datensatz identifiziert werden: A1124 (Forschungsinstitut), A1151 (Krankenhaus), A2361 (Parkhaus) und O38 (Garages), O40 (Glasshouse), O103 (Terrace). In keinem Fall lässt sich daraus eine eindeutige 1:1-Schemarelation ableiten.

Die Objektklassenkombination A0931 (Nichtöffentliches Geb. (Wohngebäude)) → O2 (Apartments) umfasst mit $|R_o| = 993{,}43$ mehr als die Hälfte aller Objektrelationen. Sie repräsentieren ungefähr ein Viertel der zugeordneten Flächen. Im Vergleich dazu deckt der zweitgrößte Relationsanteil $|R_o| = 260{,}94$ (13,9 %) bei A0931 → O74 (Residential) knapp ein Fünftel der Flächen ab. Diese hohen Werte repräsentieren bedeutsame Schemarelationen und begründen die Notwendigkeit von Wohngebäuden. Alle anderen Klassenkombinationen besitzen deutlich geringere Werte. Lediglich 23 der 104 Klassenkombinationen haben Relationsanteile größer zehn. Insgesamt besitzen die Objektklassen A0931 und A0932 Korrespondenzen zu 75 % aller OSM-Klassen. Dies weist auf die sehr allgemein definierten Klassenbeschreibungen hin, die für viele verschiedene Objekte gültig sind. Bestätigt wird dies durch die sehr ähnlichen Objektklassennamen A0931 (Nichtöffentliches Geb. (Wohngebäude)) und A0932 (Nichtöffentliches Geb. (Nebengebäude)).

Referenzzuordnung der Objektklassen

Die Referenzzuordnung der Objektklassen wird anhand der Objektklassennamen erstellt. In Tabelle 6.19 werden die Schemarelationen durch graue Zellen gekennzeichnet und in Tabelle 6.20 aufgelistet.

Tabelle 6.20: Referenzzuordnung für Testgebiet A: ALKIS - OSM mit Angabe der Einzelhäufigkeiten h_p und der Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ mit $p = \{R, A, B, AB\}$ aller Referenzcluster der vier verschiedenen Häufigkeitsmatrizen. Zeile H_{total} gibt die Gesamtmatrixinhalte an.

Nr.	Referenzschemarelationen R_s^*			h_R	h_A	h_B	h_{AB}
1	{A0931, A2121}	→	{O2, O28, O47, O74, O103}	1.326,04	172,05	493,07	131,58
2	A0932	→	{O40, O48}	5	0,31	20,90	0,59
3	{A0933, A2361}	→	{O37, O38, O65, O78}	7,14	138,34	50,69	72,00
4	A1101	→	{O70, O71}	33,38	57,02	120,33	63,08
5	{A1121, A1122}	→	{O52, O81}	29,87	195,41	99,87	129,58
6	{A1123, A1124}	→	O96	55,98	189,76	83,94	115,68
7	A1134	→	O5	0	0	0	0
8	A1141	→	{O19, O67}	27,57	97,76	181,79	126,58
9	A1151	→	O44	11,79	100	90,86	95,09
10	A1171	→	O68	6,96	94,79	99,29	96,88
11	A1401	→	{O7, O8, O46, O62, O63, O75, O82}	10	78,41	89,96	79,50
12	A1701	→	{O49, O101}	0,97	73,43	39,63	51,92
13	A2601	→	O92	0,74	11,06	27,32	17,58
			$H_{\mathrm{Ref},p}$	1.515,42	1.208,34	1.397,66	980,07
			H_{total}	1.882	1.700	3.100	1.331,29
			[%]	80,52	71,08	45,09	73,62

Insgesamt werden 13 Schemarelationen bestimmt, die sich aus vier einfachen (1:1), sechs einseitig (1:n bzw. n:1) und drei beidseitig zusammengefassten (n:m) Relationen zusammensetzen. Neben einzelnen Clusterhäufigkeiten h_p werden für jede Matrix auch Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ mit $p = \{R, A, B, AB\}$ aller Referenzcluster angegeben. Relation 7: A1134 (Museum) → O5 (Arts centre) besitzt in allen Häufigkeitsmatrizen Nullwerte. Das bedeutet, dass mit dem Objektzuordnungsverfahren keine zuverlässigen Korrespondenzen zwischen den neun ALKIS - und vier OSM -Objekten identifiziert werden konnten. Im Vergleich dazu werden für die folgenden fünf Schemarelationen hohe Häufigkeitswerte erzielt:

- {A0931 (Nichtöffentliches Geb. (Wohngebäude)), A2121 (Wohngeb. mit Handel und Dienstleistungen)} → {O2 (Apartments), O28 (Detached), O47 (House), O74 (Residential), O103 (Terrace)} (Nr. 1)
- {A1123 (Fachhochschule, Universität), A1124 (Forschungsinstitut)} → O96 (University) (Nr. 6)
- A1141 (Christliche Kirche) → {O19 (Church), O67 (Place of worship)} (Nr. 8)
- A1151 (Krankenhaus) → O44 (Hospital) (Nr. 9)
- A1171 (Polizei) → O68 (Police) (Nr. 10) .

Die Maximalwerte der Einzelhäufigkeiten sind abhängig von der Anzahl der an der Schemarelation beteiligten Objektklassen. Bei der ersten Schemarelation sind zwei ALKIS - und fünf OSM -Objektklassen beteiligt. Daraus leiten sich die Maximalwerte $h_A = 200$ und $h_B = 500$ ab. Obwohl in der Schemarelation drei Nullzellen vorhanden sind, werden folgende hohe Werte $h_A = 172{,}05$ und $h_B = 493{,}07$ bestimmt. Die gesamte Referenzzuordnung beinhaltet 16 Nullzellen, was einem Anteil von 37% entspricht. Die Übertragung der Referenzzuordnung auf alle vier Häufigkeitsmatrizen zeigt, dass mit 80,52 % die größte Übereinstimmung in der H_R -Matrix besteht. Im Vergleich zu Testgebiet B hat hier die Auswertung der einzelnen Flächen keinen Vorteil bewirkt. Die geringste Übereinstimmung wird für die H_B -Matrix erzielt, deren Häufigkeitswerte den prozentualen Flächenanteil der zugeordneten OSM -Flächen bezogen auf alle zugeordneten OSM -Flächen widerspiegeln.

Zusammenfassung der Ergebnisse für Testgebiet A

Tabelle 6.21 fasst die Ergebnisse der einzelnen Schema-Matching-Verfahren für Testgebiet A zusammen. Mit einer Ausnahme werden die größten Korrespondenzen zwischen den H_R -Verfahrenslösungen und dem Matrixgesamtinhalt (Nr. 0a) bzw. der Referenzzuordnung (Nr. 0b) erzielt und dementsprechend grau hervorgehoben. Das Heuristische Verfahren stellt mit 88,98 % (∩0a) bzw. 96,46 % (∩0b) die beste Lösung dar. Die Unterschiede zu den anderen Verfahren sind gering. Im Anhang wird in Tabelle A.4 das Ergebnis des Heuristischen Verfahrens für H_R präsentiert.

Tabelle 6.21: Ergebnisse der verschiedenen Schema-Matching-Verfahren für Testgebiet A: ALKIS - OSM in Hannover für H_R, H_A, H_B und H_{AB}. Zeile 0a gibt die Gesamtmatrixinhalte an, während Nr. 0b die Referenzzuordnung kennzeichnet. Zeile 1 präsentiert die Ergebnisse des einfachen Max-Match-Verfahrens, Zeile 2 des Heuristischen Verfahrens und Zeile 3 der Optimierungsverfahren mit dem größten $\varnothing H_{Ze}$. k steht für die Anzahl der Cluster, Spalte H kennzeichnet die Verfahrenslösung, NC repräsentiert die Anzahl der Nullcluster, während NZ die Anzahl der Nullzellen in den Clustern wiedergibt. Alle unter alleiniger Maximierung der Häufigkeiten entstandenen Lösungen sind zusätzlich mit $*$ gekennzeichnet. Unter der Clusteranzahl k steht der Schrankenwert $\{clsize_{var}\}$ bzw. $(clsize)$ in Klammern.

Nr.	Verfahren	H_R				H_A				H_B				H_{AB}			
		k	H	NC	NZ	k	H	NC	NZ	k	H	NC	NZ	k	H	NC	NZ
0a	H_{total}		1882,00				1700,00				3100,00				1331,29		
0b	$H_{\text{Ref},p}$	13		1	16	13		1	16	13		1	16	13		1	16
	$0b \cap 0a$		1515,42	[80,52%]			1208,34	[71,08%]			1397,66	[45,09%]			980,07	[73,62%]	
1	Max-Match	31		14	14	31		16	16	31		14	14	31		14	14
	$1 \cap 0a$		1204,27	[63,99%]			980,37	[57,67%]			936,75	[30,22%]			743,22	[55,83%]	
	$1 \cap 0b$		1121,07	[73,98%]			770,13	[63,73%]			697,48	[49,90%]			684,26	[69,82%]	
2	Heur. Verfahren	15		2	2	16		1	1	15		1	1	$*$ 16		2	2
	$2 \cap 0a$		1674,54	[88,98%]			1211,05	[71,24%]			2202,23	[71,04%]			1045,02	[78,50%]	
	$2 \cap 0b$		1461,83	[96,46%]			955,89	[79,11%]			1278,57	[91,48%]			893,82	[91,20%]	
3a	MaxScore $(*)$	17		3	3	15		1	2	17		1	1	16		2	2
	$3a \cap 0a$		1655,53	[87,97%]			1302,18	[76,60%]			2084,65	[67,25%]			1045,02	[78,50%]	
	$3a \cap 0b$		1442,83	[95,21%]			1045,65	[86,54%]			1111,69	[79,54%]			893,82	[91,20%]	
3b	WSMSBS	17		3	3	$*$ 15		1	2	$*$ 17		1	1	$*$ 16		2	2
	$3b \cap 0a$		1654,57	[87,92%]			1302,18	[76,60%]			2084,65	[67,25%]			1045,02	[78,50%]	
	$3b \cap 0b$		1443,83	[95,28%]			1045,65	[86,54%]			1111,69	[79,54%]			893,82	[91,20%]	
3c	MSHCVS	$*$ 17		3	3	14		0	3	16		1	1	16		2	2
	$3c \cap 0a$	{9}	1655,53	[87,97%]		{1}	1354,54	[79,68%]		{6}	2165,22	[69,85%]		{4}	1045,02	[78,50%]	
	$3c \cap 0b$		1442,83	[95,21%]			1044,36	[86,43%]			1208,17	[86,44%]			893,92	[91,21%]	
3d	MSHCFS-U	$*$ 17		3	3	13		0	1	16		1	1	15		2	2
	$3d \cap 0a$	(31)	1655,53	[87,97%]		(35)	1395,74	[82,10%]		(32)	2165,22	[69,85%]		(33)	1094,13	[82,19%]	
	$3d \cap 0b$		1442,83	[95,21%]			1110,65	[91,92%]			1208,17	[86,44%]			952,48	[97,18%]	
3e	MSHCFSNE	17		0	0	13		0	1	16		0	0	16		0	0
	$3e \cap 0a$	(31)	1610,44	[85,57%]		(35)	1395,74	[82,10%]		(32)	2151,32	[69,40%]		(32)	1019,96	[76,61%]	
	$3e \cap 0b$		1442,83	[95,21%]			1110,65	[91,92%]			1170,50	[83,75%]			882,90	[90,09%]	

Die Optimierungsverfahren MaxScore (Nr. 3a), MaxScoreHardConstraintVariableSize (Nr. 3c) und MaxScoreHardConstraintFixedSize (Nr. 3d) bestimmen für H_R identische Ergebnisse. Insgesamt erzeugen alle Optimierungsverfahren 17 Lösungscluster. Bis auf MaxScoreHardConstraintFixedSizeNonEmpty (Nr. 3e), das keine Nullcluster erlaubt, erzeugen die anderen Verfahren Lösungen mit jeweils drei Nullclustern. Im Vergleich dazu bestimmt das Heuristische Verfahren 15 Cluster mit zwei Nullclustern. Auch wenn in fast allen Verfahrenslösungen Nullcluster existieren, wird die Referenzschemarelation Nr. 7 in keinem Fall bestimmt. Für die Erzeugung dieses Clusters müssten insgesamt sechs Kanten mit Relationsanteilen von 8,40 geschnitten werden.

Tabelle 6.22 gibt eine detaillierte Übersicht über die Anzahl der mit der Referenzzuordnung übereinstimmenden Cluster in allen H_R-Verfahrenslösungen. Es gibt nur sehr wenige Korrespondenzen. Dies bestätigt, dass die Referenzzuordnung, die auf einer semantischen Analyse basiert, nicht zu den im Testgebiet vorliegenden und identifizierten Objektrelationen passt.

Das Heuristische Verfahren bestimmt fünf der 13 Referenzcluster. Es handelt sich um einfache und einseitig zusammengefasste Relationen. Die Anzahl erhöht sich um die beiden mit (×) gekennzeichneten Cluster, die hinsichtlich der Clusterhäufigkeit gleich sind, aber zusätzliche Nullzellen besitzen, die hier vernachlässigt werden. Beispielsweise besteht die komplexe Referenzschemarelation Nr. 3 aus acht Zellen, von denen sechs Nullzellen sind. Das Heuristische Verfahren fasst die beiden Zellen größer 0 als Cluster $Cl4$: {A0933 (Unterirdisches Gebäude), A2361 (Parkhaus)} → O65 (Parking) zusammen.

Das Max-Match-Verfahren bestimmt zwei der vier einfachen Referenzcluster. Die Optimierungsverfahren erzielen vier Referenzzuordnungen. Da das beste Ergebnis bezogen auf die Referenzzuordnung mit dem MSHCFS-U-Verfahren und der H_{AB}-Matrix erzielt wurde, werden zum Vergleich übereinstimmende Relationen mit aufgeführt und mit $\times_{AB}$ gekennzeichnet.

Tabelle 6.23 stellt die Rechenzeiten aller H_R-Verfahrenslösungen vor und kennzeichnet diejenigen, die die größten Übereinstimmungen mit den Referenzschemarelationen R_s^* und die größte Schnittmenge mit der Referenz-

Tabelle 6.22: Testgebiet A: ALKIS - OSM in Hannover. Übersicht über die Anzahl der identischen Cluster zwischen den H_R -Verfahrenslösungen und der Referenzzuordnung. Mit (x) *gekennzeichnete Relationen besitzen identische Clusterhäufigkeiten unter Vernachlässigung von Nullzellen. In der letzten Spalte kennzeichnet* x_{AB} *die identischen Cluster der* MSHCFS-U *-Lösung speziell für die H_{AB} -Matrix mit der Referenzzuordnung.*

Nr.	Referenzschemarelationen R_s^*	Max-Match	Heur. Verfahren	MaxScore WSMSBS MSHCVS MSHCFS-U MSHCFSNE
1	{A0931, A2121} → {O2, O28, O47, O74, O103}			
2	A0932 → {O40, O48}			
3	{A0933, A2361} → {O37, O38, O65, O78}		(x)	
4	A1101 → {O70, O71}		x	
5	{A1121, A1122} → {O52, O81}		(x)	(x_{AB})
6	{A1123, A1124} → O96			x_{AB}
7	A1134 → O5			
8	A1141 → {O19, O67}		x	x
9	A1151 → O44	x	x	x , x_{AB}
10	A1171 → O68	x	x	x , x_{AB}
11	A1401 → {O7, O8, O46, O62, O63, O75, O82}			
12	A1701 → {O49, O101}			
13	A2601 → O92		x	x
	Σ	2	5 (7)	4

zuordnung (Nr. 0b) und dem Matrixinhalt (Nr. 0a) haben. In Tabelle 6.32 werden die Verfahrensergebnisse den Ergebnissen der anderen Testgebiete gegenübergestellt.

Tabelle 6.23: Testgebiet A: ALKIS - OSM in Hannover. Übersicht über die benötigte Rechenzeit, die Anzahl der ausgeführten Rechenschritte pro H_R -Verfahrenlösung, die besten Ergebnisse bezogen auf die Referenzschemarelationen R_s^*, Schnittmenge zur Referenzzuordnung ($\cap$ 0*b*) *und zum Matrixinhalt* ($\cap$ 0*a*).

Nr.	Verfahren	Rechenzeit	Rechenschritte	R_s^*	$\cap$ 0b	$\cap$ 0a
1	Max-Match	743.44 [sec]	1			
2	Heuristisches Verfahren	1.82 [sec]	14	x	x	x
3a	MaxScore	428.52 [sec]	16			
3b	WSMSBS	1903.20 [sec]	16			
3c	MSHCVS	$>$ 170 [h]	24 $(k = 17, ..., 16)$			
3d	MSHCFS-U	43299.46 [sec]($\approx$ 12 [h])	168			
3e	MSHCFSNE	8184.73 [sec]	111 $(k = 17, ..., 12)$			

Das Heuristische Verfahren bestimmt die beste und in 1.82 Sekunden auch die schnellste Lösung. Allerdings ist das Ergebnis nicht garantiert optimal. Das MaxScore-Verfahren erzielt in etwas mehr als sieben Minuten die schnellste optimale Lösung, die gleichzeitig die zweitbeste Lösung hinsichtlich des Matrixgesamtinhalts ist. Die Optimierungsverfahren MSHCVS und MSHCFS-U bestimmen die gleiche Lösung wie MaxScore, benötigen aber sehr viel mehr Rechenzeit. Während MSHCFS-U alle Möglichkeiten in etwas mehr als 12 Stunden berücksichtigt, wird die Auswertung mit dem MSHCVS-Verfahren nach 24 Rechenschritten vorzeitig abgebrochen. Auch nach 170 Stunden konnte bei $k = 16$ und einer variablen Schranke von $clsize_{var} = 2$ keine optimale Lösung bestimmt werden. Im Gegensatz dazu ermittelt das WSMSBS-Verfahren in etwas mehr als einer halben Stunde die zweitbeste Lösung bezogen auf die Referenzzuordnung.

In Tabelle 6.24 wird die Lösung des Heuristischen Verfahrens der besten optimalen Lösung (MaxScore, MSHCVS, MSHCFS-U) gegenübergestellt. Beide Lösungen stimmen in acht Clustern überein, die mit □ gekennzeichnet sind. Davon sind sechs Relationen einfach und zwei komplex. Die Hälfte der Schemarelationen ist auch Teil der Referenzzuordnung. Beide Lösungen besitzen Nullcluster, die mit ∘ gekennzeichnet sind, dennoch gibt es keine Übereinstimmung. In der optimalen Lösung repräsentieren 14 der 17 Objektklassenzuordnungen 1:1-Schemarelationen. Bei den komplexen Relationen werden nur OSM-Objektklassen zusammengefasst. Im Vergleich dazu sind beim Heuristischen Verfahren neun der 15 Schemarelationen einfach und bei den sechs komplexen Relationen werden Objektklassen beider Datensätze zusammengefasst. Insgesamt stimmen 25 Zellen in beiden Lösungen überein und repräsentieren 98,45% der Heuristischen Lösung.

*Tabelle 6.24: Testgebiet A: ALKIS - OSM in Hannover. Gegenüberstellung der Lösungscluster in der H_R-Matrix des Heuristischen Verfahrens, als beste Lösung hinsichtlich der Referenzzuordnung und des Matrixgesamtinhalts, mit denen der besten Optimierungslösung (*MaxScore, MSHCVS, MSHCFS-U*), als zweitbeste Lösung hinsichtlich der Referenzzuordnung und des Matrixgesamtinhalts. Die grau markierten Relationen kennzeichnen Referenzcluster, □ markiert identische Cluster zwischen den Lösungen und ○ Nullcluster.*

Heuristisches Verfahren				
	A0931	→	{O2, O8, O28, O47, O48, O52, O62, O63, O74, O75, O103}	
□	A0932	→	{O7, O37, O38, O78, O82}	
	{A0933, A2361}	→	O65	
	A1101	→	{O70, O71}	
	{A1121, A1122}	→	O81	
□	A1123	→	O40	
□	A1124	→	O96	
	A1134	→	O49	○
□	A1141	→	{O19, O67}	
□	A1151	→	O44	
□	A1171	→	O68	
	A1401	→	O46	
□	A1701	→	O5	
	A2121	→	O101	○
□	A2601	→	O92	
MaxScore, MSHCVS, MSHCFS-U				
	A0931	→	{O2, O28, O46, O47, O48, O62, O63, O74, O75, O103}	
□	A0932	→	{O7, O37, O38, O78, O82}	
	A0933	→	O52	○
	A1101	→	O71	
	A1121	→	O81	
	A1122	→	O70	
□	A1123	→	O40	
□	A1124	→	O96	
	A1134	→	O101	○
□	A1141	→	{O19, O67}	
□	A1151	→	O44	
□	A1171	→	O68	
	A1401	→	O8	
□	A1701	→	O5	
	A2121	→	O49	○
	A2361	→	O65	
□	A2601	→	O92	

Die semantische Analyse zeigt, dass ähnliche Objektklassen erwartungsgemäß zusammengefasst werden, z.B. O37/38 (Garage(s)) oder O62/63 (Office(s)). Bei der Heuristischen Lösung werden zusätzlich die Objektklassen O70 (Public) und O71 (Public building) zusammengefasst. Vier Schemarelationen, die keine Referenzzuordnung darstellen, sind ebenso nachvollziehbar, so z.B. A1124 (Forschungsinstitut) → O96 (University) und A1123 (Universität) → O40 (Glasshouse). Laien, die OSM-Daten erfassen, prüfen nicht, ob Gewächshäuser oder Forschungsinstitute Teil einer universitären Einrichtung sind. Vielmehr wird der offensichtliche Zweck für die Annotation genutzt. Einige Zuordnungen repräsentieren zu Recht Nullcluster (○), da sie nicht nachvollziehbar sind, beispielsweise A0933 (Unterirdisches Gebäude) → O52 (Kindergarten) oder A2121 (Wohngebäude mit Handel und Dienstleistungen) → O49 (Industrial).

6.3.3 Testgebiet C: ATKIS - GDF in Hannover-Wedemark

Die in Abschnitt 6.3.3 vorgestellten Objektrelationen repräsentieren Korrespondenzen zwischen 29 ATKIS- und 15 GDF-Objektklassen. Im Anhang geben C.1 und C.2 eine Übersicht über alle im Testgebiet C vertretenen Objektklassen mit Objektanzahlen. Die Objektklassennamen beider Datensätze sind auch hier nicht identisch. Einige Korrespondenzen können mit einem Wörterbuch identifiziert werden, z.B. A2227 (Grünanlage) → G7170 (Land Use: Park, Garden) oder A4107 (Wald, Forst) → G7120 (Land Cover: Forest (Woodland)). Ob die identifizierten Objektrelationen diese Zuordnungen tatsächlich bestätigen, ist zu untersuchen. Im weiteren Verlauf der Arbeit werden ATKIS-Klassen mit einem vorangestellten A und GDF-Klassen mit G gekennzeichnet.

Häufigkeitsmatrizen

Analog zu den anderen Testgebieten werden aus allen Objektrelationen vier unterschiedliche Häufigkeitsmatrizen abgeleitet. Aufgrund der zu erwartenden langen Rechenzeiten bei den Optimierungsverfahren werden die 29×15-Matrizen mit 209 Objektrelationen reduziert. Dazu werden in der H_R-Matrix alle Objektklassen entfernt, deren Zeilen- bzw. Spaltensumme kleiner zwei ist. Während bei ATKIS dadurch mehr als 50 % der Objektklassen wegfallen, sind es bei GDF drei (20 %). Für die Schema-Matching-Verfahren stehen nun 13×12-Matrizen als Eingabe zur Verfügung. Tabelle 6.25 präsentiert die reduzierte H_R-Häufigkeitsmatrix. Sie umfasst 165 Objektrelationen, was 80 % der ursprünglichen Relationen entspricht. Der Anteil an Nullzellen ist mit 75% sehr hoch. Alle anderen Matrizen sind im Anhang in Tabelle C.1 zu finden.

Tabelle 6.25: Häufigkeitsmatrix H_R für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). In jeder Zelle steht die Anzahl der Relationen bezogen auf den jeweiligen Flächenanteil, an dem beide Objektklassen beteiligt sind. Die grauen Zellen kennzeichnen die Referenzzuordnung.

	H_R	GDF												
		G1120	G3110	G3136	G4160	G4313	G4314	G4315	G7120	G7170	G7500	G9715	G9725	$\sum$
ATKIS	A2101	10,59	25,61	2,99	0	0	0	0	1,35	0	0	2	0	42,54
	A2111	1,19	7,75	0,00	2	0	0	0	0	0	0	0,24	0	11,17
	A2112	1,23	0,91	0	0	0	0	0	0	0	0	19,36	0	21,50
	A2113	0,26	4,73	0,02	0	0	0	0	0	0	0	3,40	0	8,42
	A2114	0,15	1	0	4	0	0	0	0	0	0	0	0	5,15
	A2202	0	0	0	0	0	0	0	0	1	0	0	1,05	2,05
	A2227	0,21	0	0	1,12	0	0	0	0	2	0	0	0	3,33
	A3103	0	0	0	3,88	0	0	0	0	0	0	0	0	3,88
	A3514	0	0	0	0	0	0	0	0,00	0	2	0	0	2,00
	A4101	0	1,00	0,98	0	0	0	0	0	0	0	0	0	1,98
	A4105	0	0	0	0	0	0	0	0,35	0	0	0	1,56	1,91
	A4107	1,38	0	0	0	0	0	0	46,26	0	0	0	0,39	48,02
	A5112	0	0	0	0	2	9	2	0,04	0	0	0	0	13,04
	$\sum$	15	41	4	11	2	9	2	48	3	2	25	3	165

Die Objektklassenkombination A4107 (Wald, Forst) → G7120 (Land Cover: Forest (Woodland)) umfasst mit $|R_o| = 46{,}26$ mehr als ein Viertel aller Objektrelationen. Vier der 12 GDF-Objektklassen besitzen Relationen zu jeweils nur einer ATKIS-Objektklasse: G4313 (Water areas (DISPCLASS 3)), G4314 (Water areas (DISPCLASS 4)), G4315 (Water areas (DISPCLASS 5)) und G7500 (Area structures - Bridges and tunnels). Das Besondere ist, dass die ersten drei GDF-Klassen sogar denselben ATKIS-Zuordnungspartner A5112 (Binnensee, Stausee, Teich) haben. Es wird erwartet, dass diese GDF-Objektklassen im Rahmen des Schema-Matchings zusammengefasst werden. Im Vergleich dazu besitzt die ATKIS-Klasse A3103 (Platz) ausschließlich Relationen zur GDF-Klasse G4160 (Roads - Address area). Die Schemarelation ist nicht eindeutig, da es drei weitere ATKIS-Klassen gibt A2111 (Wohnbaufläche), A2114 (Fläche besonderer funktionaler Prägung), A2227 (Grünanlage), für die Relationsanteile zu G4160 bestimmt werden.

Referenzzuordnung der Objektklassen

Die Referenzzuordnung der Objektklassen wird auf Basis der Objektklassennamen und Relationsanteile bestimmt. Weil die ATKIS-Klasse A4101 (Ackerland) keine eindeutige Entsprechung im GDF-Datensatz hat, aber für den Vergleich zugeordnet werden muss, wird sie aufgrund der Relationsanteile mit den Klassen A2101 (Ortslage) und A2111 (Wohnbaufläche) zusammengefasst. In der reduzierten H_R-Matrix (Tabelle 6.25) werden die Schemarelationen durch graue Zellen gekennzeichnet und in Tabelle 6.26 aufgelistet. Für jede Referenzschemarelation R_s^* werden neben einzelnen Clusterhäufigkeiten h_p auch Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ mit $p = \{R, A, B, AB\}$ für jede Matrix angegeben. Insgesamt werden acht Referenzschemarelationen bestimmt, die sich aus drei einfachen (1:1), vier einseitig (1:n bzw. n:1) und einer beidseitig zusammengefassten (n:m) Relation zusammensetzen.

Der Relationsanteil der sechsten Relation A4105 (Moor, Moos) → G9725 (Land Cover: Moors and heathland) verdeutlicht erneut, dass der Ausschluss von Objektklassen mit kleinen Werten nicht immer sinnvoll ist. In diesem Beispiel führt der Wert $h_R = 1,56$ zu hohen Relationsanteilen in den anderen Häufigkeitsmatrizen und deutet somit auf eine zuverlässige Zuordnung hin. Im Vergleich dazu hat die siebente Relation A4107 (Wald, Forst) → G7120 (Land Cover: Forest (Woodland)) einen hohen Relationsanteil und erzielt auch hohe Werte in den anderen Häufigkeitsmatrizen. Die Übertragung der Referenzzuordnung auf alle vier Häufigkeitsmatrizen zeigt, dass mit 95,35 % die größte Übereinstimmung in der H_{AB}-Matrix besteht. Eine Auswertung hinsichtlich der Flächen ist somit sinnvoll.

Tabelle 6.26: Referenzzuordnung für Testgebiet C: ATKIS - GDF mit Angabe der Einzelhäufigkeiten h_p und der Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ mit $p = \{R, A, B, AB\}$ aller Referenzcluster der vier verschiedenen Häufigkeitsmatrizen. Zeile H_{total} gibt die Gesamtmatrixinhalte an.

Nr.	Referenzschemarelationen R_s^*			h_R	h_A	h_B	h_{AB}
1	{A2101, A2111, A4101}	→	{G1120, G3110, G3136}	50,11	293,96	289,02	199,19
2	{A2112, A2113}	→	G9715	22,76	115,47	82,81	87,28
3	{A2114, A3103}	→	G4160	7,88	124,64	78,55	93,88
4	{A2202, A2227}	→	G7170	3	136,66	100	111,89
5	A3514	→	G7500	2	51,88	100	65,13
6	A4105	→	G9725	1,56	98,91	94,16	96,32
7	A4107	→	G7120	46,26	96,19	90,78	93,29
8	A5112	→	{G4313, G4314, G4315}	13	99,99	300	138,89
			$H_{\mathrm{Ref},p}$	146,56	1.017,70	1.135,32	885,87
			H_{total}	165	1.300	1.200	929,09
			[%]	88,83	78,28	94,61	95,35

Zusammenfassung der Ergebnisse für Testgebiet C

Tabelle 6.27 fasst die Ergebnisse der einzelnen Schema-Matching-Verfahren aller Häufigkeitsmatrizen für Testgebiet C zusammen. Die größten Korrespondenzen werden zwischen den H_{AB}-Verfahrenslösungen und dem Matrixgesamtinhalt (Nr. 0a) bzw. der Referenzzuordnung (Nr. 0b) erzielt und dementsprechend grau hervorgehoben. Es gibt zwei Ausnahmen, einerseits beim Max-Match-Verfahren für die H_R-Matrix und andererseits beim MaxScoreHardConstraintFixedSizeNonEmpty-Verfahren für die H_B-Matrix.

Tabelle 6.27: Ergebnisse der verschiedenen Schema-Matching-Verfahren für Testgebiet C: ATKIS - GDF in Hannover-Wedemark für H_R, H_A, H_B und H_{AB}. Zeile 0a gibt die Gesamtmatrixinhalte an, während Nr. 0b die Referenzzuordnung kennzeichnet. Zeile 1 präsentiert die Ergebnisse des einfachen Max-Match-Verfahrens, Zeile 2 des Heuristischen Verfahrens und Zeile 3 der Optimierungsverfahren mit dem größten $\varnothing H_{Ze}$. k steht für die Anzahl der Cluster, Spalte H kennzeichnet die Verfahrenslösung, NC repräsentiert die Anzahl der Nullcluster, während NZ die Anzahl der Nullzellen in den Clustern wiedergibt. Alle unter alleiniger Maximierung der Häufigkeiten entstandenen Lösungen sind zusätzlich mit $*$ gekennzeichnet. Unter der Clusteranzahl k steht der Schrankenwert $\{clsize_{var}\}$ bzw. $(clsize)$ in Klammern.

Nr.	Verfahren	H_R				H_A				H_B				H_{AB}			
		k	H	NC	NZ	k	H	NC	NZ	k	H	NC	NZ	k	H	NC	NZ
0a	H_{total}		165,00				1300,00				1200,00				929,09		
0b	$H_{\mathrm{Ref},p}$	8		0	1	8		0	1	8		0	1	8		0	1
	$0b \cap 0a$		146,56	[88,83%]			1017,70	[78,28%]			1135,32	[94,61%]			885,87	[95,35%]	
1	Max-Match	12		2	2	12		2	2	12		2	2	12		2	2
	$1 \cap 0a$		111,96	[67,85%]			747,73	[57,52%]			714,92	[59,58%]			658,08	[70,83%]	
	$1 \cap 0b$		111,96	[76,39%]			675,61	[66,39%]			714,92	[62,97%]			658,08	[74,29%]	
2	Heur. Verfahren	10		0	0	10		0	0	$*$ 10		0	0	10		0	0
	$2 \cap 0a$		127,71	[77,40%]			937,76	[72,14%]			974,22	[81,18%]			802,58	[86,38%]	
	$2 \cap 0b$		126,40	[86,24%]			835,45	[82,09%]			974,22	[85,81%]			802,58	[90,60%]	
3a	MaxScore $(*)$	12		2	2	11		1	1	10		0	0	11		1	1
	$3a \cap 0a$		118,78	[71,98%]			900,05	[69,23%]			974,22	[81,18%]			778,17	[83,76%]	
	$3a \cap 0b$		118,52	[80,87%]			827,93	[81,35%]			974,22	[85,81%]			778,17	[87,84%]	
3b	WSMSBS	$*$ 12		2	2	$*$ 11		1	1	$*$ 10		0	0	$*$ 11		1	1
	$3b \cap 0a$		118,78	[71,98%]			900,05	[69,23%]			974,22	[81,18%]			778,17	[83,76%]	
	$3b \cap 0b$		118,52	[80,87%]			827,93	[81,35%]			974,22	[85,81%]			778,17	[87,84%]	
3c	MSHCVS	11		2	2	$*$ 11		1	1	9		0	0	10		1	1
	$3c \cap 0a$	{1}	128,64	[77,96%]		{1}	900,05	[69,23%]		{2}	1059,04	[88,25%]		{1}	838,04	[90,20%]	
	$3c \cap 0b$		128,64	[87,77%]			827,93	[81,35%]			1056,31	[93,04%]			838,04	[94,60%]	
3d	MSHCFS-U	11		2	2	$*$ 11		1	1	9		0	0	10		1	1
	$3d \cap 0a$	(14)	128,64	[77,96%]		(14)	900,05	[69,23%]		(16)	1059,04	[88,25%]		(15)	838,04	[90,20%]	
	$3d \cap 0b$		128,64	[87,77%]			827,93	[81,35%]			1056,31	[93,04%]			838,04	[94,60%]	
3e	MSHCFSNE	10		0	0	10		0	0	9		0	0	10		0	0
	$3e \cap 0a$	(15)	130,34	[79,00%]		(15)	946,81	[72,83%]		(16)	1059,04	[88,25%]		(15)	802,58	[86,38%]	
	$3e \cap 0b$		125,40	[85,56%]			874,70	[85,95%]			1056,31	[93,04%]			802,58	[90,60%]	

Die Optimierungsverfahren MSHCVS und MSHCFS-U erzielen die besten Lösungen. Sie liefern sogar für jede Häufigkeitsmatrix absolut identische Lösungen. Hinsichtlich der Rechenzeit gibt es jedoch gravierende Unter-

schiede. Tabelle 6.28 stellt die Anzahl der Rechenschritte und die benötigte Rechenzeit aller Verfahren gegenüber. Während MSHCVS für die Problemlösung 52 Rechenschritte und mehr als 36 Stunden benötigt, löst MSHCFS-U das Problem in nur knapp 16 Sekunden in 24 Rechenschritten. In Tabelle 6.32 wird das Ranking der Verfahren hinsichtlich Rechenzeit, Referenzzuordnung und Matrixgesamtinhalt den Ergebnissen der anderen Testgebiete gegenübergestellt.

Tabelle 6.28: Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Übersicht über die benötigte Rechenzeit, die Anzahl der ausgeführten Rechenschritte pro H_{AB} -Verfahrenlösung, die besten Ergebnisse bezogen auf die Referenzschemarelationen R_s^*, Schnittmenge zur Referenzzuordnung ($\cap 0b$) und zum Matrixinhalt ($\cap 0a$).

Nr.	Verfahren	Rechenzeit	Rechenschritte	R_s^*	$\cap$ 0b	$\cap$ 0a
1	Max-Match	4.45 [sec]	1			
2	Heuristisches Verfahren	0.36 [sec]	9	×		
3a	MaxScore	33.55 [sec]	11			
3b	WSMSBS	109.31 [sec]	15			
3c	MSHCVS	$>$ 129002.41 [sec] ($\approx$ 36 [h])	52 ($k = 12, ..., 2$)		×	×
3d	MSHCFS-U	16.25 [sec]	24		×	×
3e	MSHCFSNE	2588.09 [sec]	112 ($k = 10, ..., 2$)	×		

Die optimale Lösung für die H_{AB} -Matrix wird in Tabelle 6.29 präsentiert. Anders als erwartet, werden nur zwei der drei GDF -Objektklassen Water areas zusammengefasst. Stattdessen wird die Zuordnung A2111 (Wohnbaufläche) $\rightarrow$ G4315 (Water areas (Displayclass 4)) bestimmt, für die allerdings keine Relationsanteile identifiziert wurden. Aus semantischer Sicht ist diese Zuordnung falsch. Die Schemarelation A4101 (Ackerland) $\rightarrow$ G3136 (Postal district area) ist semantisch betrachtet auch wenig sinnvoll. Doch aufgrund ihrer Relationsanteile wird sie von anderen Objektklassen getrennt.

Tabelle 6.29: Optimale Lösung des MSHCVS - und MSHCFS-U -Verfahrens für H_{AB} für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). Es werden insgesamt 10 Cluster mit 15 Zellen und einer Gesamthäufigkeit von $H_{(k=10)} = 838{,}04$ bestimmt. Die Durchschnittshäufigkeit pro Zelle beträgt $\varnothing H_{Ze} = 55{,}87$.

H_{AB}		GDF											
		G1120	G3110	G3136	G4160	G4313	G4314	G4315	G7120	G7170	G7500	G9715	G9725
ATKIS	A2101	**64,67**	**56,17**	15,03	0	0	0	0	2,96	0	0	3,27	0
	A2111	7,30	7,75	0,05	0,62	0	0	0	0	0	0	1,11	0
	A2112	6,85	0,94	0	0	0	0	0	0	0	0	**75,82**	0
	A2113	1,76	2,33	1,10	0	0	0	0	0	0	0	**11,46**	0
	A2114	1,14	0,06	0	22,34	0	0	0	0	0	0	0	0
	A2202	0	0	0	0	0	0	0	0	41,16	0	0	9,04
	A2227	1,68	0	0	7,50	0	0	0	0	70,74	0	0	0
	A3103	0	0	0	71,54	0	0	0	0	0	0	0	0
	A3514	0	0	0	0	0	0	0	0,10	0	**65,13**	0	0
	A4101	0	4,74	43,47	0	0	0	0	0	0	0	0	0
	A4105	0	0	0	0	0	0	0	0,46	0	0	0	**96,32**
	A4107	1,40	0	0	0	0	0	0	**93,29**	0	0	0	0,89
	A5112	0	0	0	0	56,59	69,35	12,95	0,00	0	0	0	0

Das Heuristische Verfahren ist mit unter einer Sekunde für neun Rechenschritte das schnellste Verfahren und repräsentiert gleichzeitig die zweitbeste Lösung. Tabelle 6.30 zeigt die Zuordnung der Objektklassen. Die Lösung ist sogar bewiesen optimal, da sie mit der MSHCFSNE -Lösung identisch ist. Allerdings benötigt das Optimierungsverfahren für die Lösung des Problems cirka 42 Minuten Rechenzeit für 112 Berechnungen.

Beim Vergleich der besten Optimierungslösung (MSHCVS, MSHCFS-U) mit der Heuristischen Lösung wird deutlich, dass dreizehn Zellen übereinstimmen und diese 97,42% der Heuristischen Lösung repräsentieren. Für jede Häufigkeitsmatrix ermittelt das Heuristische Verfahren eine unterschiedliche Lösung mit jeweils zehn Cluster bestehend aus 15 Zellen. Keine der Lösungen beinhaltet Nullzellen oder Nullcluster. Zehn Zellen sind in allen vier Lösungen identisch und auch Teil der Referenzzuordnung. Tabelle 6.31 gibt eine detaillierte Übersicht über die Anzahl der mit den Referenzschemarelationen übereinstimmenden Cluster in allen H_{AB} -Verfahrenslösungen. Die größten Korrespondenzen erzielen das Heuristische und das MSHCFSNE -Optimierungsverfahren. Sie bestimmen sieben der acht Relationen exakt. Drei Referenzschemarelationen (Nr. 5, 6 und 7) werden von allen Verfahren bestimmt.

Das Heuristische Verfahren bestimmt als einziges Verfahren ein Lösungscluster, in dem alle drei GDF -Objektklassen Water areas zusammengefasst werden. Des Weiteren werden Korrespondenzen, die mit dem Wörterbuch

Tabelle 6.30: Ergebnis des Heuristischen und MSHCFSNE *-Verfahrens für H_{AB} für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). Es werden insgesamt 10 Cluster mit 15 Zellen und einer Gesamthäufigkeit von $H_{(k=10)} =$ 802,58 bestimmt. Die Durchschnittshäufigkeit pro Zelle beträgt $\emptyset H_{Ze} =$ 53,51.*

H_{AB}		GDF										
		G1120	G3110	G3136	G4160	G4313	G4314	G4315	G7120	G7170	G7500	G9715	G9725
ATKIS	A2101	**64,67**	56,17	15,03	0	0	0	0	2,96	0	0	3,27	0
	A2111	7,30	**7,75**	0,05	0,62	0	0	0	0	0	0	1,11	0
	A2112	6,85	0,94	0	0	0	0	0	0	0	0	75,82	0
	A2113	1,76	2,33	1,10	0	0	0	0	0	0	0	11,46	0
	A2114	1,14	0,06	0	22,34	0	0	0	0	0	0	0	0
	A2202	0	0	0	0	0	0	0	0	**41,16**	0	0	9,04
	A2227	1,68	0	0	7,50	0	0	0	0	**70,74**	0	0	0
	A3103	0	0	0	71,54	0	0	0	0	0	0	0	0
	A3514	0	0	0	0	0	0	0	0,10	0	65,13	0	0
	A4101	0	4,74	**43,47**	0	0	0	0	0	0	0	0	0
	A4105	0	0	0	0	0	0	0	0,46	0	0	0	96,32
	A4107	1,40	0	0	0	0	0	0	93,29	0	0	0	0,89
	A5112	0	0	0	0	**56,59**	**69,35**	**12,95**	0,00	0	0	0	0

Cluster	*Cl1*	*Cl2*	*Cl3*	*Cl4*	*Cl5*	*Cl6*	*Cl7*	*Cl8*	*Cl9*	*Cl10*
h_k	138,89	65,13	93,29	87,28	93,88	43,47	96,32	111,89	7,75	64,67

identifiziert wurden, teilweise: {A2227 (Grünanlage), A2202 (Freizeitanlage)} → G7170 (Land Use: Park, Garden) oder sogar ganz bestätigt: A4107 (Wald, Forst) → G7120 (Land Cover: Forest (Woodland)).

Die besten Optimierungsverfahren MSHCVS und MSHCFS-U bestimmen sechs Referenzschemarelationen exakt. Die MaxScore- und WSMSBS-Lösung sind identisch und bilden fünf der acht Referenzcluster. Das verdeutlicht, dass der Gewichtsfaktor $s = 0{,}5$ beim WSMSBS-Verfahren zu dominant ist, so dass das Ziel MaxScore gegenüber BalancedSize vorrangig optimiert wird. Das Max-Match-Verfahren erzielt mit 70,83% ($\cap$0a) bzw. 74,29% ($\cap$0b) die schlechteste Lösung. Es werden lediglich drei Referenzschemarelationen erkannt. Die geringe Schnittmenge bezogen auf die Referenzzuordnung ist erwartungsgemäß, da fünf der acht Relationen komplex sind und vom Verfahren nicht bestimmt werden können.

Die n:m Referenzschemarelation Nr. 1 wird von keinem Verfahren identifiziert. Ein Grund dafür ist, dass das Bilden von großen Clustern sich nachteilig auf das Auswahlkriterium Durchschnittshäufigkeit pro Zelle $\emptyset H_{Ze}$ auswirkt. Wird in der MSHCFSNE-Lösung in Tabelle 6.30 anstatt der Cluster $Cl6$, $Cl9$ und $Cl10$ die Referenzschemarelation Nr. 1 gebildet, vergrößert sich zwar die Gesamthäufigkeit der Cluster auf $H_{(k=8)} = 885{,}86$, aber gleichzeitig steigt auch die Anzahl der Clusterzellen von 15 auf 21. Weil sich die Durchschnittshäufigkeit pro Zelle von 53,51 auf 42,18 verringert, wird diese Lösung nicht als beste Lösung ausgewählt.

Tabelle 6.31: Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Übersicht über die Anzahl der identischen Cluster zwischen den H_{AB} -Verfahrenslösungen und der Referenzzuordnung.

Nr.	Referenzschemarelationen R_s^*			Max-Match	Heur. V. MSHCFSNE	MSHCVS MSHCFS-U	WSMSBS MaxScore
1	{A2101, A2111, A4101}	→	{G1120, G3110, G3136}				
2	{A2112, A2113}	→	G9715		×	×	
3	{A2114, A3103}	→	G4160		×	×	×
4	{A2202, A2227}	→	G7170		×	×	×
5	A3514	→	G7500	×	×	×	×
6	A4105	→	G9725	×	×	×	×
7	A4107	→	G7120	×	×	×	×
8	A5112	→	{G4313, G4314, G4315}		×		
			Σ	3	7	6	5

6.3.4 Zusammenfassung aller Schema-Matching-Ergebnisse

Tabelle 6.32 stellt die Ergebnisse aller Schema-Matching-Verfahren der drei Testgebiete gegenüber. Für jedes Gebiet wird ein Ranking der Verfahren in den Kategorien Rechenzeit, Schnittmenge zum Matrixgesamtinhalt ($\cap$0a) und zur Referenzzuordnung ($\cap$0b) erstellt. Die Werte in Klammern repräsentieren das Ranking ohne Berücksichtigung der Referenzzuordnung, da bei Datensätzen mit vielen Objektklassen bzw. mit unterschiedlichen Objektanzahlen, die durch den Experten ermittelte Referenzzuordnung als unsicher eingestuft wird. Das beste Verfahren erhält in der entsprechenden Kategorie eine 1. Mit zunehmender Verschlechterung steigt die Zahl.

Tabelle 6.32: Ranking der Verfahrensergebnisse aller Testgebiete hinsichtlich Rechenzeit, Schnittmenge zum Matrixgesamtinhalt ($\cap$ 0a) und zur Referenzzuordnung ($\cap$ 0b). Die Werte in Klammern repräsentieren das Ranking ohne Berücksichtigung der Referenzzuordnung.

Nr.	Verfahren	Testgebiet A (H_R)			Testgebiet B (H_{AB})			Testgebiet C (H_{AB})			Σ	Ranking
		Zeit	$\cap 0a$	$\cap 0b$	Zeit	$\cap 0a$	$\cap 0b$	Zeit	$\cap 0a$	$\cap 0b$		
1	Max-Match	3	5	4	2	4	3	2	4	4	31 (20)	4 (4)
2	Heur. Verfahren	1	1	1	1	1	2	1	2	2	12 (7)	1 (1)
3a	MaxScore	2	2	3	3	3	1	4	3	3	24 (17)	2 (2)
3b	WSMSBS	4	3	2	5	3	1	5	3	3	29 (23)	3 (6)
3c	MSHCVS	7	2	3	7	3	1	7	1	1	32 (27)	5 (7)
3d	MSHCFS-U	6	2	3	4	3	1	3	1	1	24 (19)	2 (3)
3e	MSHCFSNE	5	4	3	6	2	1	6	2	2	31 (21)	4 (5)

Das Heuristische Verfahren erreicht in sechs von neun Fällen den ersten Platz und belegt mit 12 Punkten auch den ersten Platz im Gesamtranking. Die Optimierungsverfahren MaxScore und MSHCFS-U belegen mit 24 Punkten den zweiten Platz. Ohne Berücksichtigung der Referenzzuordnung schneidet MaxScore mit zwei Punkten weniger sogar besser als MSHCFS-U ab. Die Rechenzeit des MSHCFS-U-Verfahren für Testgebiet A beeinflusst das Ranking negativ. Während das MaxScore-Verfahren eine optimale Lösung in 7 Minuten für die 31×17-Häufigkeitsmatrix bestimmt, benötigt das MSHCFS-U-Verfahren ca. 12 Stunden. Mit zunehmender Matrixgröße steigt die Rechenzeit an. Lange Rechenzeiten schränken den Einsatz von Verfahren in der Praxis stark ein. Der Vergleich der Heuristischen Lösungen mit den besten Optimierungslösungen zeigt, dass in allen Testgebieten die optimalen Lösungen mehr als 91% der Heuristischen Lösungen bestätigen. Demnach sind die Heuristischen und optimalen Lösungen vergleichbar und die Anwendung des Verfahren wird empfohlen.

Auf dem dritten Platz steht das WSMSBS-Verfahren, bei dem beide Optimierungsziele MaxScore und BalancedSize mit $s = 0{,}5$ gleichgewichtet werden. Ohne Berücksichtigung der Referenzzuordnung fällt das Verfahren auf den sechsten Platz zurück. Die Rechenzeiten beeinflussen auch hier das Ranking negativ. Für die Anwendung des Verfahrens in der Praxis kommt erschwerend hinzu, dass die Wahl des Gewichtsfaktors, der einen Kompromiss beider Ziele bewirken soll, schwierig ist.

Das einfache Max-Match-Lösungsverfahren, das nur einfache Schemarelationen bestimmen kann, belegt sowohl mit als auch ohne Berücksichtigung der Referenzlösung, den vierten Rang. Nur, wenn semantische Unterschiede zwischen den Datensätzen durch 1:1-Relationen abgebildet werden können, liefert das Verfahren zuverlässige Ergebnisse.

Am schlechtesten schneidet das MSHCVS-Verfahren ab. Obwohl die Ergebnisse in Hinblick auf den Matrixgesamtinhalt mit zu den besten gehören, verhindern die sehr langen Rechenzeiten die Nutzung des Verfahrens. Die variable Clustergröße, die nur die Differenz zwischen dem größten und dem kleinsten Cluster vorgibt, verursacht eine gesteigerte Anzahl an zu berücksichtigen Lösungen.

Zusammenfassend lässt sich feststellen, dass sich sowohl das Heuristische Verfahren als auch das Optimierungsverfahren MaxScoreHardConstraintFixedSizeUnique für die Zuordnung von Objektklassen auf Basis von Objektrelationen gut eignen. Obwohl die Lösungen des Heuristischen Verfahrens nicht optimal sind und die Schemarelationen nur einfach oder einseitig komplex sein können, kann es für große Häufigkeitsmatrizen schnell sehr gute Ergebnisse bestimmen. Werden in den zu untersuchenden Datensätzen allerdings beidseitig komplexe Korrespondenzen vermutet, ist das MSHCFS-U-Verfahren zu wählen. Es findet zwar garantiert optimale Lösungen, aber dies kann deutlich höhere Rechenzeiten benötigen.

Auf Basis der Untersuchungsergebnisse werden, für Datensätze mit vergleichbaren Maßstäben, Häufigkeitsmatrizen mit Relationsanteilen H_R als Eingabe empfohlen. Im Gegensatz dazu eignen sich bei zunehmenden Maßstabsunterschieden Häufigkeitsmatrizen mit prozentualen Flächenanteilen beider Datensätze H_{AB}. Es sei an dieser Stelle noch einmal darauf hingewiesen, dass mit dem vorliegenden instanzbasierten Verfahren nur Korrespondenzen zwischen Objektklassen gefunden werden, für die es im Testgebiet auch Objekte gibt. Für alle anderen Objektklassen fehlen diese Informationen. Somit können die Regeln nie vollständig sein.

7 Zusammenfassung und Ausblick

Datenintegration, Datenaustausch, Verschmelzung von Daten und Datenaktualisierung sind im wissenschaftlichen Umfeld nach wie vor aktuelle Themen. Geoinformationen sind heutzutage für viele Entscheidungen in Politik, Wirtschaft und Verwaltung unverzichtbar. Die Fülle an raumbezogenen Daten verschiedener Disziplinen, die zunehmend auch über Webdienste zur Verfügung stehen, müssen dafür immer häufiger kombiniert werden. Im Datenintegrationsprozess verursacht gerade die Kombination verschiedener Datensätze erhebliche Probleme. Dies ist vor allem auf strukturelle, geometrische und semantische Unterschiede der Daten zurückzuführen.

In der vorliegenden Arbeit wurde ein zweistufiges Zuordnungsverfahren entwickelt, das einen Beitrag zur geometrischen und semantischen Datenintegration leistet. Zuerst werden Objektkorrespondenzen zwischen Polygonobjekten zweier sich überlagernder geographischer Datensätze bestimmt. Anschließend werden aus den Objektrelationen automatisch semantische Korrespondenzen zwischen den Objektklassen beider Datensätze auf Schemaebene abgeleitet. Beide Ansätze sind auch unabhänig voneinander einsetzbar.

Zu Beginn der Arbeit werden Grundlagen vorgestellt, die sowohl für die Zuordnung auf Objekt- und Objektklassenebene benötigt werden. Dazu zählen verschiedene Ähnlichkeitsmaße, unterschiedliche Relationstypen für Objekt- und Schemaebene, Graphentheorie und die Ganzzahlige lineare Optimierung. Anschließend liefert der aktuelle Stand der Forschung einen Überblick über verschiedene Objektzuordnungverfahren speziell für Vektordaten und stellt ausgewählte Schema-Matching-Verfahren im geographischen Kontext vor. Die Herausforderungen bei der Zuordnung von Objekten und Objektklassen wurden aufgezeigt.

Das in dieser Arbeit entwickelte Verfahren führt im ersten Schritt, der Objektzuordnung, eine geometrische Analyse der zugrunde liegenden Objektinstanzen durch. Für die Bewertung der identifizierten Objektkorrespondenzen wurde ein Gesamtähnlichkeitsmaß entwickelt, das sich aus einem geometrischen und einem semantischen Parameter zusammensetzt. Der geometrische Parameter wertet Überlagerungsflächen und Ausrichtungen der Objekte zueinander aus. In Hinblick auf das sich anschließende Schema-Matching werden eindeutige gegenüber unspezifischen Relationen gestärkt, die semantisch einfach nachzuvollziehen sind. Der semantische Parameter spiegelt die Anzahl der beteiligten Objektklassen pro Objektrelation wider. Das Objektzuordnungsverfahren kann Datensätze mit verschiedenen Maßstäben berücksichtigen. Die Datensätze müssen nicht als Partition vorliegen, es sind auch Objektüberlagerungen innerhalb der Datensätze zugelassen. Dadurch vergrößert sich der Suchraum für potentielle Matching-Kandidaten. Das entwickelte Verfahren kann sowohl einfache (1:1) als auch komplexe (1:n/n:1/n:m) Objektrelationen bestimmen. Komplexe Relationen werden durch das Aggregieren von Nachbarobjekten in einem bestimmten Umkreis gebildet. Dafür werden Nachbarobjekte solange zusammengefasst, wie sich das Gesamtähnlichkeitsmaß verbessert oder Objekte im Suchradius enthalten sind.

Das Objektzuordnungsverfahren wurde in drei verschiedenen Testgebieten mit vier unterschiedlichen Datensätzen getestet. Die Datensätze unterscheiden sich geometrisch und thematisch. Die geometrische Auflösung der Daten reicht von 1:1.000 bis 1:25.000. Neben semantisch vergleichbaren Objekten, wie z.B. Gebäuden, standen sich auch semantisch verschiedenartige Objekte gegenüber. Die Zuordnung von semantisch ähnlichen Objekten grenzt die Objektauswahl ein und verringert den Suchraum für potentielle Matching-Kandidaten. Allerdings stehen diese semantischen Informationen nicht immer vorab zur Verfügung.

Die Tests haben gezeigt, dass das entwickelte Verfahren, im Vergleich zu manuell erstellten Referenzlösungen, für alle Testszenarien hohe Zuordnungsgenauigkeiten erzielt. Bei Datensätzen mit ähnlichem Maßstab liegt der Anteil an falschen bzw. nicht identifizierten richtigen Relationen bei nur 5% . Bei Datensätzen mit unterschiedlichen Maßstäben verschlechtern sich die Vollständigkeitsmaße. Das bedeutet, dass vom Verfahren weniger Matching-Kandidaten als vom Experten bestimmt wurden. Überwiegend sind dies sehr kleine oder sehr große Objekte.

Die Untersuchungen zeigen, dass die im Objektzuordnungsverfahren verwendeten sehr einfachen Ähnlichkeitsmaße und festgelegten Schwellwerte für die Identifikation von korrekten Objektrelationen in verschiedenen Testszenarien geeignet sind. Die Maximierung der Objektzuordnung stand dabei nie im Vordergrund. Das Verfahren wurde entwickelt, um zuverlässige Eingangsdaten für das sich anschließende Schema-Matching zu bestimmen. Dennoch kann die Anzahl der Objektrelationen weiter erhöht werden, indem z.B. die Nachbarschaftsdefinition gelockert wird, die für die Aggregation von Nachbarobjekten ausschlaggebend ist. Aktuell können nur direkt angrenzende Objekte berücksichtigt werden. Aufgrund von geometrischen Ungenauigkeiten müssen in Zukunft

auch räumlich entfernte Objekte beachtet werden. Zusätzlich sollte das in Abschnitt 4.2 vorgestellte Objektzuordnungsverfahren für unterschiedliche Geometriedimensionen mit dem Verfahren für Polygone kombiniert werden, um zukünftig auch Objekte unterschiedlicher Geometriedimensionen einander zuzuordnen. Besonders bei Maßstabsdifferenzen ist die Zuordnung von Objekten mit unterschiedlichen Geometriedimensionen sehr wahrscheinlich.

Identifizierte Objektkorrespondenzen sollten zukünftig als sogenannte Links in Datenbanken abgespeichert werden, um einen vereinfachten Zugriff zu ermöglichen. Links können beispielsweise im RDF-Format (Resource Description Framework) beschrieben werden und über standardisierte Abfragesprachen wie SPARQL oder OWL (Web Ontology Language) genutzt werden. Gerade für Aktualisierungsprozesse sind diese Links vorteilhaft, da Veränderungen an einem Objekt gleich auf korrespondierende Objekte übertragen werden können. Denkbar sind Geometrieanpassungen oder Ergänzungen der Attributinformationen. Je mehr Verlinkungen erzeugt werden, desto mehr Daten können angereichert und gemeinsam ausgewertet werden. Für Datensätze, die zum gleichen Datenmodell gehören, aber unterschiedliche Maßstäbe besitzen, können Links sogar für automatisierte Generalisierungsprozesse genutzt werden. Das Erzeugen von Links für geographische Daten stellt eine eigene umfassende Thematik dar. Für einen ausführlicheren Überblick über das Thema Linked Data sei an dieser Stelle auf Kuhn u. a. (2014) und Poblet u. a. (2019) verwiesen.

Der zweite Schritt des entwickelten Verfahrens entspricht der Zuordnung der Objektklassen auf Schemaebene. Eingangsdaten sind die Ergebnisse der Objektzuordnung. Es wurden alle identifizierten Objektrelationen verwendet, ungeachtet der Prüfung eines Experten. Für die Zukunft sollten entweder nur korrekte Relationen als Eingangsdaten verwendet oder hinsichtlich ihrer Qualität, z.B. gemessen am Gesamtähnlichkeitsmaß, berücksichtigt werden. Aufgrund verschiedener Sichtweisen können die Objektrelationen entweder als Häufigkeitsmatrix oder als bipartiter Graph interpretiert werden. Insgesamt wurden für jedes Testgebiet vier unterschiedliche Häufigkeitsmatrizen abgeleitet. Neben Relationsanteilen wurden auch datensatzbezogene prozentuale Flächenanteile berechnet, die korrespondierende Flächen widerspiegeln und dadurch eine andere Sicht auf die Korrespondenzen zeigen.

Für die Zuordnung der Objektklassen wurden existierende Graphalgorithmen verwendet. Die Objektklassen beider Datensätze repräsentieren zwei Knotenteilmengen, zwischen denen mit Relationsanteilen gewichtete Kanten verlaufen. Zu Beginn wurde der Ansatz des Maximalen Matchings (Max-Match) gewählt, der quadratische Häufigkeitsmatrizen voraussetzt und nur 1:1-Relationen zwischen Objektklassen bestimmen kann. Um diese Beschränkungen zu überwinden, wurde der Ansatz des Minimalen-2-Schnitts (Min-Cut) angewendet, was die Häufigkeitsmatrix in nur zwei Teile unterteilt. Die Zuordnung der Objektklassen entspricht vielmehr der Unterteilung eines Graphen in k Teile. Dies stellt allerdings ein $\mathcal{NP}$-vollständiges Problem dar, für das kein Algorithmus mit polynomieller Laufzeit existiert.

Aus diesem Grund wurde ein Heuristisches Verfahren entwickelt, das das einfache Min-Cut-Verfahren rekursiv auf die Häufigkeitsmatrix anwendet. Als Ergebnis entstehen in der Zuordnungsmatrix rechteckige Cluster, die sich nicht schneiden und nur pro Zeile und Spalte eine Zuordnung zulassen. Näherungsverfahren können keine optimalen Ergebnisse garantieren, dafür aber effizient sein. Für dieses Näherungsverfahren gibt es die Einschränkung, dass keine beidseitigen komplexen n:m-Schemarelationen bestimmt werden können. Für die Bewertung des Verfahrens wurden neben der Erstellung von manuellen Referenzzuordnungen auch Optimierungsverfahren entwickelt, die exakte Lösungen bestimmen.

Dazu wurde das $\mathcal{NP}$-schwere Problem in ein mathematisches Modell mit ganzzahligen Variablen überführt und mit dem existierenden Optimierer (IBM ILOG CPLEX Interactive Optimizer 12.5.1.0) gelöst. Das primäre Optimierungsziel ist die Maximierung der Häufigkeiten innerhalb der Cluster (MaxScore). Dadurch wird garantiert, dass die Zuordnung auf Objektklassen durch die identifizierten Objektrelationen bestätigt wird. Das Erzeugen von ausgewogenen Clustern hinsichtlich der Zellenanzahl pro Cluster (BalancedSize) wurde als zweites Optimierungsziel eingefügt. Beide Ziele wurden kombiniert, einerseits gleichgewichtet (WeightedSumMaxScoreBalancedSize) und anderseits, indem das zweite Optimierungsziel, die Clustervariabilität, die die Differenz zwischen dem größten und kleinsten Cluster vorgibt, als harte Bedingung eingeführt wurde (MaxScoreHardConstraintVariableSize). Die Lösung mit der größten Durchschnittshäufigkeit pro Zelle wird als beste Lösung ausgewählt. Aufgrund der langen Rechenzeiten wurde das Problem dahingehend vereinfacht, dass die Clustergesamtgröße als feste harte Bedingung ersetzt wurde (MaxScoreHardConstraintFixedSize). Die Festlegung von Clustergesamtgrößen erhöht die Anzahl der Rechenschritte. Abschließend wurde das Modell nochmals erweitert, in dem keine Nullcluster zugelassen werden (MaxScoreHardConstraintFixedSizeNonEmpty).

Die Tests haben gezeigt, dass für Datensätze mit vergleichbaren Maßstäben Häufigkeitsmatrizen mit Relationsanteilen am besten geeignet sind. Bei größeren Maßstabsunterschieden bringt die Auswertung der zugeordneten

Flächen Vorteile. Alle Verfahren wurden in den Kategorien Rechenzeit, Schnittmenge zum Matrixgesamtinhalt und zur Referenzzuordnung verglichen und bewertet.

Für das Heuristische Verfahren wurden in allen Testgebieten sehr gute Ergebnisse mit den im Vergleich kürzesten Rechenzeiten erzielt. Die Optimierungverfahren MaxScore und MaxScoreHardConstraintFixedSize belegten im Gesamtranking Platz 2. Hinsichtlich der Rechenzeiten gibt es große Unterschiede. Bereits für eine 31×17-Matrix benötigt das MaxScoreHardConstraintFixedSize-Optimierungverfahren ca. 12 Stunden. Das Verfahren MaxScoreHardConstraintVariableSize schneidet aufgrund der sehr langen Rechenzeiten am schlechtesten ab. Das verdeutlicht, dass lange Rechenzeiten den Einsatz von Verfahren in der Praxis verhindern und somit gute und effiziente Näherungsverfahren notwendig sind.

Der Vergleich der Heuristischen Lösungen mit den besten Optimierungslösungen zeigt, dass in allen Testgebieten die optimalen Lösungen mehr als 91% der Heuristischen Lösungen bestätigen. Demnach sind die heuristischen und optimalen Lösungen vergleichbar und die Anwendung des Näherungsverfahrens wird empfohlen. Ein weiterer Vorteil des Verfahrens ist, dass keine manuelle Unterstützung notwendig ist. Demnach können mit dem Heuristischen Verfahren in Zukunft auch große Schemas (Large-Scale-Schema-Matching) bewältigt werden.

Die identifizierten Korrespondenzen auf Schemaebene spiegeln die tatsächlich vorliegenden semantischen Relationen in den Daten wider. Der hier vorgestellte instanzbasierte Ansatz ist dadurch resistent gegenüber Fehlern, die durch die manuelle Annotation der Objektklassenzugehörigkeit verursacht worden sind. In Zukunft können aus den semantischen Relationen Transformationsregeln abgeleitet werden, die für einen Datenaustausch, für alle Arten der Datenintegration oder für die Entwicklung neuer Schemas genutzt werden können, um Daten aus verschiedenen Disziplinen zusammenzuführen.

Für den Sonderfall, dass wie in Testgebiet B zwischen ALKIS und ATKIS Objektklassenkorrespondenzen zwischen Datensätzen vom gleichen Datenmodell bestimmt werden, können diese Ergebnisse auch für einen Generalisierungsprozess genutzt werden. In einer eigenen früheren Arbeit wurde ein Verfahren entwickelt, das aus der Überlagerung von Objektinstanzen zweier Datensätze Übergangsmatrizen (engl. Transition Matrices) in einem iterativen Lernprozess ableitet und als Kontrollparameter für ein optimiertes Aggregationsverfahren verwendet (Kieler u. a., 2009a). Das heißt, in den Übergangsmatrizen sind semantische Distanzen angegeben, die Informationen über mögliche Aggregationspartner widerspiegeln. Aus den Ergebnissen des entwickelten Zuordnungsverfahrens könnten ebenfalls Übergangsmatrizen abgeleitet werden. Alle Lösungscluster repräsentieren semantische Übereinstimmungen und würden dementsprechend in der Übergangsmatrix Kosten gleich 0 erhalten. Im Gegensatz dazu erhalten alle Zellen mit Häufigkeiten gleich 0 hohe Kosten, da hier keine Objektklassenänderung gewünscht wird. Für alle anderen Zellen könnten normierte Kostenwerte bestimmt werden. Die tatsächliche Anwendbarkeit des Verfahrens ist zu untersuchen.

A Testgebiet A: ALKIS - OSM in Hannover

A.1 Semantik der Datensätze

ALKIS					
Code	Bezeichnung	Anz.	Code	Bezeichnung	Anz.
A0916	Zuschauertribüne, überdacht	1	A1172	Feuerwehr	5
A0917	Stadion	1	A1401	Gebäude für Handel und DL[2]	201
A0922	Vorratsbehälter, Speicherbauwerk	9	A1701	Gebäude für Gewerbe und Industrie	222
A0929	Schöpfwerkgebäude	1	A1731	Tankstelle	5
A0931	Nichtöffentliches Geb.[1] (Wohngebäude)	15.735	A1771	Bergwerk in Betrieb	1
A0932	Nichtöffentliches Geb. (Nebengebäude)	6.147	A2111	Wohngebäude mit Öffentlich	10
A0933	Unterirdisches Gebäude	270	A2121	Wohngeb. mit Handel und DL	109
A1101	Öffentliches Gebäude	278	A2131	Wohngeb. mit Gewerbe und Industrie	8
A1111	Parlament	1	A2151	Geb. für Handel u. DL m. Wohnen	8
A1112	Rathaus	2	A2161	Geb. für Gewerbe u. Industrie m. Wohnen	10
A1115	Gericht	4	A2328	Gebäude für Schienenverkehr	5
A1117	Kreisverwaltung	1	A2348	Gebäude für Schifffahrt	2
A1121	Allgemeinbildende Schule	61	A2361	Parkhaus	28
A1122	Berufs-, Fach-, Volkshochschule	16	A2362	Parkdeck	3
A1123	Fachhochschule, Universität	70	A2501	Betriebsgebäude für Versorgung	205
A1124	Forschungsinstitut	46	A2541	Sendeturm, Fernmeldeturm	2
A1132	Theater, Oper	4	A2601	Betriebsgebäude für Entsorgung	28
A1134	Museum	9	A2728	Gebäude für landwirtschaftlichen Betrieb	1
A1136	Veranstaltungsgebäude	2	A2741	Gewächshaus, Treibhaus	2
A1138	Gebäude für kulturelle Zwecke	1	A2801	Gebäude für Erholung	5
A1141	Christliche Kirche	45	A2811	Sporthalle	4
A1143	Kapelle	11	A2821	Hallenbad	1
A1145	Gotteshaus einer anderen Religionsgem.	1	A2831	Tribüne	27
A1151	Krankenhaus	36	A3911	Schornstein	1
A1171	Polizei	17			

Abkürzungen: [1] Geb: Gebäude, [2] DL: Dienstleistungen 23.663

Abbildung A.1: Übersicht über die ALKIS-Objektklassen und deren Objektanzahlen für das Testgebiet A: ALKIS - OSM in Hannover. ALKIS besitzt 49 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.1) enthalten.

OSM								
Code	type	Anz.	Code	type	Anz.	Code	type	Anz.
O2	Apartments	1.199	O42	Greenhouse	1	O74	Residential	428
O4	Aquarium	1	O43	Hangar	4	O75	Restaurant	11
O5	Arts centre	4	O44	Hospital	14	O76	Retail	3
O6	Artwork	2	O46	Hotel	8	O78	Roof	58
O7	Attraction	8	O47	House	21	O79	Ruins	2
O8	Bank	6	O48	Hut	34	O80	Sauna	1
O13	Bunker	3	O49	Industrial	13	O81	School	37
O14	Cafe	2	O50	Information	1	O82	Service	44
O18	Chapel	1	O52	Kindergarten	9	O84	Social facility	2
O19	Church	20	O53	Library	4	O86	Storage tank	1
O20	Collapsed	1	O54	Manufacture	2	O87	Supermarket	2
O22	Commercial	2	O57	Monument	4	O88	Supermarket; Apa	1
O24	Construction	5	O58	Museum	1	O89	Surveillance	1
O26	Crematorium	1	O59	Nightclub	2	O90	Swimming pool	1
O28	Detached	18	O60	No	2	O91	Theatre	3
O29	Disused	1	O62	Office	28	O92	Toilets	5
O31	Entrance	1	O63	Offices	88	O93	Tower	2
O33	Fast food	2	O65	Parking	13	O94	Townhall	1
O35	Food court	2	O66	Pharmacy	1	O96	University	124
O36	Fuel	1	O67	Place of worship	20	O98	Viewpoint	1
O37	Garage	85	O68	Police	12	O100	Water tower	1
O38	Garages	37	O69	Pub	3	O101	Works	4
O39	Gasometer	1	O70	Public	6	O103	Terrace	29
O40	Glasshouse	13	O71	Public building	77	O9999	Unknown type	143
								2.689

Abbildung A.2: Übersicht über die OSM-Objektklassen und deren Objektanzahlen für das Testgebiet A: ALKIS - OSM in Hannover. OSM besitzt 72 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.1) enthalten.

A.2 Häufigkeitsmatrizen

Tabelle A.1: Häufigkeitsmatrix H_A für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17)*. In jeder Zelle steht der prozentuale Flächenanteil der zugeordneten ALKIS-Flächen bezogen auf alle zugeordneten ALKIS-Flächen.*

H_A		ALKIS																
		A0931	A0932	A0933	A1101	A1121	A1122	A1123	A1124	A1134	A1141	A1151	A1171	A1401	A1701	A2121	A2361	A2601
OSM	O2	46,07	5,82	49,66	0,99	0	0	0	0	0	0	0	0	0	0	75,18	0	0
	O5	0,16	0,09	0	0	0	0	0	0	0	0	0	0	1,52	9,58	0	0	0
	O7	0	4,53	0	0	0	0	0	0	6,37	0	0	0	0	0	0	0	0
	O8	1,18	0	0	8,55	0	0	0	0	0	0	0	0	6,62	0	0	0	0
	O19	0,03	0,40	0	0,60	0	0	0	0	0	44,51	0	0	0	0	0	0	0
	O28	0,54	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O37	0,08	9,78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O38	0	5,60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O40	0	0	0	0	0	0	5,01	0	0	0	0	0	0	0	0	0	0
	O44	0,22	0	0	0	0	0	0	0	0	2,24	100	0	0	0	0	0	0
	O46	1,80	0,03	0	0	0	0	0	0	0	0	0	0	11,01	0	0	0	0
	O47	5,38	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O48	0,11	0,31	0	0	0	0	0	0	0	0	0	0	0,18	0	0	0	3,65
	O49	0,11	17,76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O52	0,21	0,56	0	1,72	0	0	0	0	0	0	0	0	0	0	0	0	0
	O62	3,17	0	1,65	5,65	0	0	0	0	0	0	0	0	41,56	0	0	0	0
	O63	3,67	4,63	0	7,02	0	0	0	0	0	0	0	0	15,57	12,12	13,14	0	0
	O65	0	15,92	38,34	1,96	0	0	0	0	0	0	0	0	0	0	0	100	0
	O67	0,17	0	0	0,24	0	0	0	0	0	53,25	0	0	0	0	0	0	0
	O68	0	0	3,68	0	0	0	0	0	0	0	0	94,79	0	0	0	0	0
	O70	0,16	0	0	1,29	0	0,99	0	0	0	0	0	0	0	0	0	0	0
	O71	2,48	3,69	0	55,74	0	0	3,08	0	93,63	0	0	0	17,24	0	0	0	0
	O74	32,91	4,63	0	0,61	3,60	0	0	0	0	0	0	5,21	2,65	0	11,68	0	0
	O75	0,32	0,66	0	4,72	0	0	0	0	0	0	0	0	3,65	0	0	0	0
	O78	0	3,49	0	0,30	0	0	2,15	0	0	0	0	0	0	4,87	0	0	0
	O81	0	0,17	0	0	96,40	99,01	0	0	0	0	0	0	0	0	0	0	0
	O82	0,07	11,60	2,09	0	0	0	0	0	0	0	0	0	0	0	0	0	85,29
	O92	0,03	0,14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11,06
	O96	0,83	4,11	4,57	10,64	0	0	89,76	100	0	0	0	0	0	0	0	0	0
	O101	0	6,00	0	0	0	0	0	0	0	0	0	0	0	73,43	0	0	0
	O103	0,30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Σ	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Tabelle A.2: Häufigkeitsmatrix H_B für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). *In jeder Zelle steht der prozentuale Flächenanteil der zugeordneten OSM-Flächen bezogen auf alle zugeordneten OSM-Flächen.*

H_B		ALKIS																	
		A0931	A0932	A0933	A1101	A1121	A1122	A1123	A1124	A1134	A1141	A1151	A1171	A1401	A1701	A2121	A2361	A2601	Σ
OSM	O2	94,52	1,52	0,91	0,42	0	0	0	0	0	0	0	0	0	0	2,62	0	0	100
	O5	51,07	3,52	0	0	0	0	0	0	0	0	0	0	23,65	21,77	0	0	0	100
	O7	0	80,96	0	0	0	0	0	0	19,04	0	0	0	0	0	0	0	0	100
	O8	35,24	0	0	56,58	0	0	0	0	0	0	0	0	8,19	0	0	0	0	100
	O19	1,30	2,08	0	6,28	0	0	0	0	0	90,35	0	0	0	0	0	0	0	100
	O28	99,56	0,44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	O37	6,83	93,17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	O38	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	O40	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	100
	O44	6,04	0	0	0	0	0	0	0	0	3,09	90,86	0	0	0	0	0	0	100
	O46	77,02	0,14	0	0	0	0	0	0	0	0	0	0	22,85	0	0	0	0	100
	O47	99,90	0,10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	O48	70,26	20,90	0	0	0	0	0	0	0	0	0	0	6,24	0	0	0	2,60	100
	O49	4,28	95,12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	O52	30,06	8,79	0	61,15	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	O62	45,16	0	0,12	17,04	0	0	0	0	0	0	0	0	37,67	0	0	0	0	100
	O63	48,39	7,84	0	28,57	0	0	0	0	0	0	0	0	10,56	1,23	3,41	0	0	100
	O65	0	40,06	4,32	9,24	0	0	0	0	0	0	0	0	0	0	0	46,37	0	100
	O67	6,52	0	0	2,03	0	0	0	0	0	91,44	0	0	0	0	0	0	0	100
	O68	0	0	0,71	0	0	0	0	0	0	0	0	99,29	0	0	0	0	0	100
	O70	29,83	0	0	54,25	0	15,92	0	0	0	0	0	0	0	0	0	0	0	100
	O71	13,12	2,29	0	66,08	0	0	2,06	0	12,12	0	0	0	4,34	0	0	0	0	100
	O74	95,96	1,67	0	0,32	0,88	0	0	0	0	0	0	0,31	0,35	0	0,51	0	0	100
	O75	21,15	4,85	0	63,32	0	0	0	0	0	0	0	0	10,69	0	0	0	0	100
	O78	0	55,40	0	8,64	0	0	31,18	0	0	0	0	0	0	4,78	0	0	0	100
	O81	0	0,13	0	0	57,63	42,24	0	0	0	0	0	0	0	0	0	0	0	100
	O82	4,78	87,96	0,73	0	0	0	0	0	0	0	0	0	0	0	0	0	6,53	100
	O92	46,76	25,92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27,32	100
	O96	3,54	2,06	0,10	10,35	0	0	42,76	41,18	0	0	0	0	0	0	0	0	0	100
	O101	0	60,37	0	0	0	0	0	0	0	0	0	0	0	39,63	0	0	0	100
	O103	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Tabelle A.3: Häufigkeitsmatrix H_{AB} für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). *Jeder Zelle beinhaltet den prozentualen Flächenanteil der zugeordneten Flächen bezogen auf beide Datensätze.*

H_{AB}		ALKIS																
		A0931	A0932	A0933	A1101	A1121	A1122	A1123	A1124	A1134	A1141	A1151	A1171	A1401	A1701	A2121	A2361	A2601
OSM	O2	62,42	2,33	1,43	0,59	0	0	0	0	0	0	0	0	0	0	4,95	0	0
	O5	0,31	0,17	0	0	0	0	0	0	0	0	0	0	2,78	13,33	0	0	0
	O7	0	8,89	0	0	0	0	0	0	9,74	0	0	0	0	0	0	0	0
	O8	2,52	0	0	16,23	0	0	0	0	0	0	0	0	7,33	0	0	0	0
	O19	0,06	0,69	0	1,18	0	0	0	0	0	59,59	0	0	0	0	0	0	0
	O28	1,09	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O37	0,17	17,76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O38	0	10,88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O40	0	0	0	0	0	0	9,55	0	0	0	0	0	0	0	0	0	0
	O44	0,45	0	0	0	0	0	0	0	0	2,63	95,09	0	0	0	0	0	0
	O46	3,60	0,05	0	0	0	0	0	0	0	0	0	0	14,94	0	0	0	0
	O47	10,64	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O48	0,22	0,59	0	0	0	0	0	0	0	0	0	0	0,37	0	0	0	3,03
	O49	0,21	31,26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	O52	0,41	1,01	0	3,56	0	0	0	0	0	0	0	0	0	0	0	0	0
	O62	5,95	0	0,23	8,48	0	0	0	0	0	0	0	0	39,27	0	0	0	0
	O63	6,68	5,85	0	12,44	0	0	0	0	0	0	0	0	12,59	2,16	5,25	0	0
	O65	0	22,88	7,80	3,28	0	0	0	0	0	0	0	0	0	0	0	64,20	0
	O67	0,35	0	0	0,45	0	0	0	0	0	66,99	0	0	0	0	0	0	0
	O68	0	0	1,26	0	0	0	0	0	0	0	0	96,88	0	0	0	0	0
	O70	0,32	0	0	2,56	0	2,06	0	0	0	0	0	0	0	0	0	0	0
	O71	4,14	2,83	0	60,52	0	0	2,44	0	19,93	0	0	0	7,05	0	0	0	0
	O74	50,92	2,35	0	0,42	1,40	0	0	0	0	0	0	0,63	0,60	0	0,97	0	0
	O75	0,64	1,16	0	8,78	0	0	0	0	0	0	0	0	5,37	0	0	0	0
	O78	0	6,59	0	0,58	0	0	4,03	0	0	0	0	0	0	4,82	0	0	0
	O81	0	0,15	0	0	71,28	58,30	0	0	0	0	0	0	0	0	0	0	0
	O82	0,14	20,48	1,09	0	0	0	0	0	0	0	0	0	0	0	0	0	11,75
	O92	0,06	0,30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17,58
	O96	1,35	2,75	0,20	10,49	0	0	57,57	58,11	0	0	0	0	0	0	0	0	0
	O101	0	10,92	0	0	0	0	0	0	0	0	0	0	0	51,92	0	0	0
	O103	0,58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A.3 Ergebnis des Heuristischen Verfahrens für H_R

Tabelle A.4: Ergebnis des Heuristischen Verfahrens für H_R für das Testgebiet A: ALKIS-OSM in Hannover (31×17). Es werden insgesamt 15 Cluster mit 33 Zellen und einer Gesamthäufigkeit von $H_{R(k=15)} = 1.674,54$ bestimmt.

H_R	ALKIS A0931	A0932	A0933	A1101	A1121	A1122	A1123	A1124	A1134	A1141	A1151	A1171	A1401	A1701	A2121	A2361	A2601
O2	993,43	5,03	0,26	2	0	0	0	0	0	0	0	0	0	0	16,89	0	0
O5	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
O7	0	6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
O8	3	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0
O19	1	1	0	1	0	0	0	0	0	13	0	0	0	0	0	0	0
O28	16,93	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O37	3,12	62,18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O38	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O40	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0
O44	0,14	0	0	0	0	0	0	0	0	0,07	11,79	0	0	0	0	0	0
O46	4,99	0,01	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
O47	13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O48	9,39	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
O49	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O52	3	1,17	0	2,83	0	0	0	0	0	0	0	0	0	0	0	0	0
O62	18,99	0	0,01	2	0	0	0	0	0	0	0	0	1	0	0	0	0
O63	30,14	7,55	0	5	0	0	0	0	0	0	0	0	5	1	1	0	0
O65	0	1,86	3	1	0	0	0	0	0	0	0	0	0	0	0	4,14	0
O67	2,43	0	0	1	0	0	0	0	0	14,57	0	0	0	0	0	0	0
O68	0	0	0,04	0	0	0	0	0	0	0	0	6,96	0	0	0	0	0
O70	2	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0
O71	16,44	3,55	0	30,38	0	0	1	0	3,40	0	0	0	1	0	0	0	0
O74	260,94	8,16	0	1	2	0	0	0	0	0	0	1	1,06	0	1,85	0	0
O75	6	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
O78	0	8	0	0,23	0	0	0,10	0	0	0	0	0	0	1	0	0	0
O81	0	0,13	0	0	17,87	12	0	0	0	0	0	0	0	0	0	0	0
O82	3	22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5
O92	2	1,27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,73
O96	11,30	8,61	0,02	16	0	0	20,89	35,09	0	0	0	0	0	0	0	0	0
O101	0	3,03	0	0	0	0	0	0	0	0	0	0	0	0,97	0	0	0
O103	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Cluster	*Cl1*	*Cl2*	*Cl3*	*Cl4*	*Cl5*	*Cl6*	*Cl7*	*Cl8*	*Cl9*	*Cl10*	*Cl11*	*Cl12*	*Cl13*	*Cl14*	*Cl15*
h_k	11,79	6,96	29,87	7,14	1,00	27,57	0	0,73	2,00	13,00	0	35,09	33,38	128,18	1.377,82
Geschnittene Kantengewichte	0,21	1,04	3,13	4,15	5,97	6,43	7,40	9,27	16,06	21,98	22,77	35,91	38,05	35,09	

B Testgebiet B: ALKIS - ATKIS in Hameln

B.1 Semantik der Datensätze

Bezeichnung	ALKIS Code	ALKIS Anz.	AKTIS Code	AKTIS Anz.
Wohnbaufläche	L41001	1.186	T41001	243
Industrie und Gewerbefläche	L41002	399	T41002	54
Fläche gemischter Nutzung	L41006	9	T41006	134
Fläche besonderer funktionaler Prägung	L41007	99	T41007	46
Sport-, Freizeit- und Erholungsfläche	L41008	317	T41008	39
Friedhof	L41009	5	T41009	6
Straßenverkehr	L42001	684	T42001	19
Weg	L42006	910		
Platz	L42009	85	T42009	19
Bahnverkehr	L42010	93	T42010	18
Schiffsverkehr	L42016	9	T42016	1
Landwirtschaft	L43001	158	T43001	105
Wald	L43002	50	T43002	49
Gehölz	L43003	14	T43003	15
Unland, Vegetationslose Fläche	L43007	42	T43007	19
Fläche zur Zeit unbestimmbar			T43008	48
Fließgewässer	L44001	57	T44001	14
Stehendes Gewässer	L44006	4	T44006	5
		4.121		834

Abbildung B.1: Übersicht über die Objektklassen und deren Objektanzahl für das Testgebiet B: ALKIS - ATKIS in Hameln. ALKIS und ATKIS besitzen jeweils 17 Objektklassen. Das Objekt der grau eingefärbten Objektklasse ist in keiner Relationen der Objektzuordnung (Tabelle 6.3) enthalten.

B.2 Häufigkeitsmatrizen

Tabelle B.1: Häufigkeitsmatrizen H_A (oben), H_B (mitte) und H_{AB} (unten) für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). In H_A steht in jeder Zelle der prozentuale Flächenanteil der zugeordneten ALKIS-Flächen bezogen auf alle zugeordneten ALKIS-Flächen, während in H_B der prozentuale Flächenanteil der zugeordneten ATKIS-Flächen bezogen auf alle zugeordneten ATKIS-Flächen steht. H_{AB} beinhaltet den prozentualen Flächenanteil der zugeordneten Flächen bezogen auf beide Datensätze.

H_A	ATKIS																
ALKIS	T41001	T41002	T41006	T41007	T41008	T41009	T42001	T42009	T42010	T43001	T43002	T43003	T43007	T43008	T44001	T44006	$\sum$
L41001	74,45	0,49	22,273	1,55	0,30	0	0	0	0	0,94	0	0	0	0	0	0	100
L41002	5,01	59,84	29,323	3,09	0,03	0	0	0,62	0,65	1,11	0	0	0,19	0,03	0,11	0	100
L41006	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	100
L41007	2,21	4,86	8,946	83,30	0,52	0	0	0,12	0	0	0	0	0	0,05	0	0	100
L41008	9,82	0,95	2,914	6,91	17,78	0	0	1,92	0	55,16	0,60	0	3,95	0	0	0	100
L41009	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	100
L42001	35,57	7,81	20,586	9,35	0,93	0	5,96	1,23	0	6,79	0	2,71	4,28	4,79	0	0	100
L42006	10,62	2,42	17,008	15,28	17,11	0	0	5,85	0	27,35	1,14	2,81	0,36	0,06	0	0	100
L42009	0,30	1,04	0,119	3,10	5,36	0	0	75,02	0	2,76	0	0	11,54	0,77	0	0	100
L42010	0	0	0	0	0	0	0	0	43,52	16,25	0	6,49	0	33,74	0	0	100
L42016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	100
L43001	0	0,32	0	0	0,22	0	0	0	0	98,90	0	0	0,56	0	0	0	100
L43002	0	0	0	0	0	0	0	0	0	0,78	99,22	0	0	0	0	0	100
L43003	0	0	0	0	0	0	0	0	0	31,66	68,34	0	0	0	0	0	100
L43007	0	0	0	0	0	0	4,78	0	0	26,83	0	26,47	40,96	0	0,96	0	100
L44001	0	0	0	0	0	0	0	0	0	1,09	0,19	0,80	0,07	3,49	94,36	0	100
L44006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100

H_B	ATKIS															
ALKIS	T41001	T41002	T41006	T41007	T41008	T41009	T42001	T42009	T42010	T43001	T43002	T43003	T43007	T43008	T44001	T44006
L41001	90,57	2,06	55,07	5,77	7,53	0	0	0	0	1,03	0	0	0	0	0	0
L41002	2,64	91,06	33,53	6,06	0,31	0	0	5,91	43,68	0,58	0	0	2,14	1,25	0,57	0
L41006	0	0	0,80	0	0	0	0	0	0	0	0	0	0	0	0	0
L41007	0,63	3,42	5,10	75,93	2,62	0	0	0,80	0	0	0	0	0	1,19	0	0
L41008	2,74	0,76	1,57	7,66	75,71	0	0	9,90	0	13,98	0,48	0	26,75	0	0	0
L41009	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
L42001	3,01	1,59	3,08	2,61	1,00	0	70,24	3,28	0	0,43	0	15,65	8,09	29,38	0	0
L42006	0,28	0,15	0,85	1,37	5,88	0	0	2,81	0	0,58	0,06	7,05	0,21	0,15	0	0
L42009	0,03	0,14	0,01	0,59	3,50	0	0	77,30	0	0,14	0	0	17,88	2,21	0	0
L42010	0	0	0	0	0	0	0	0	56,32	0,32	0	12,74	0	40,29	0	0
L42016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3,74	0
L43001	0	0,81	0	0	3,44	0	0	0	0	81,62	0	0	12,72	0	0	0
L43002	0	0	0	0	0	0	0	0	0	0,33	98,57	0	0	0	0	0
L43003	0	0	0	0	0	0	0	0	0	0,13	0,83	0	0	0	0	0
L43007	0	0	0	0	0	0	29,76	0	0	0,79	0	57,79	32,06	0	0,32	0
L44001	0	0	0	0	0	0	0	0	0	0,08	0,05	6,77	0,16	25,24	95,36	0
L44006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
$\sum$	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

H_{AB}	ATKIS															
ALKIS	T41001	T41002	T41006	T41007	T41008	T41009	T42001	T42009	T42010	T43001	T43002	T43003	T43007	T43008	T44001	T44006
L41001	82,58	0,88	33,40	2,45	0,74	0	0	0	0	0,99	0	0	0	0	0	0
L41002	3,34	73,37	31,61	4,23	0,06	0	0	1,18	1,83	0,73	0	0	0,38	0,06	0,19	0
L41006	0	0	1,39	0	0	0	0	0	0	0	0	0	0	0	0	0
L41007	0,91	3,99	6,25	79,24	1,00	0	0	0,24	0	0	0	0	0	0,11	0	0
L41008	4,19	0,84	2,01	7,31	30,03	0	0	3,29	0	21,77	0,52	0	7,47	0	0	0
L41009	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
L42001	5,22	2,73	5,28	4,07	0,97	0	9,58	2,06	0	0,85	0	4,57	5,71	8,49	0	0
L42006	0,50	0,30	1,53	2,44	8,74	0	0	3,74	0	1,13	0,12	4,32	0,26	0,09	0	0
L42009	0,04	0,25	0,02	0,98	4,29	0	0	76,19	0	0,25	0	0	14,57	1,13	0	0
L42010	0	0	0	0	0	0	0	0	49,26	0,54	0	9,32	0	36,80	0	0
L42016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8,31	0
L43001	0	0,46	0	0	0,46	0	0	0	0	89,16	0	0	1,20	0	0	0
L43002	0	0	0	0	0	0	0	0	0	0,46	98,89	0	0	0	0	0
L43003	0	0	0	0	0	0	0	0	0	0,24	1,42	0	0	0	0	0
L43007	0	0	0	0	0	0	7,96	0	0	1,52	0	35,57	35,66	0	0,48	0
L44001	0	0	0	0	0	0	0	0	0	0,16	0,08	1,53	0,10	6,26	94,86	0
L44006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

C Testgebiet C: ATKIS - GDF in Hannover-Wedemark

C.1 Semantik der Datensätze

ATKIS					
Code	Bezeichnung	Anz.	Code	Bezeichnung	Anz.
A2101	Ortslage	307	A2345	Schwimmbecken	30
A2111	Wohnbaufläche	6.260	A3103	Platz	261
A2112	Industrie- und Gewerbefläche	735	A3301	Flughafen	61
A2113	Fläche gemischter Nutzung	1.442	A3302	Flugplatz, Landeplatz	2
A2114	Fläche besonderer funktionaler Prägung	639	A3303	Rollbahn	12
A2121	Bergbaubetrieb	24	A3304	Vorfeld	3
A2122	Deponie	9	A3401	Hafen	22
A2124	Werft	2	A3402	Hafenbecken	22
A2126	Kraftwerk	4	A3501	Bahnhofsanlage	123
A2127	Umspannstation	6	A3502	Raststätte	5
A2128	Förderanlage	39	A3513	Tunnel	3
A2129	Kläranlage, Klärwerk	19	A3514	Brücke, Überführung, Unterführung	133
A2131	Ausstellungsgelände, Messegelände	1	A4101	Ackerland	1.992
A2132	Gärtnerei	89	A4102	Grünland	2.238
A2134	Wasserwerk	1	A4103	Gartenland	340
A2201	Sportanlage	331	A4104	Heide	33
A2202	Freizeitanlage	180	A4105	Moor, Moos	123
A2211	Freilichttheater	1	A4107	Wald, Forst	2.338
A2213	Friedhof	342	A4108	Gehölz	226
A2221	Stadion	7	A4109	Sonderkultur	61
A2222	Sportplatz	444	A4111	Nasser Boden	42
A2223	Schießanlage	31	A4120	Vegetationslose Fläche	3
A2224	Schwimmbad, Freibad	29	A4199	Fläche, z.Z. unbestimmbar	1.369
A2225	Zoo	2	A5101	Strom, Fluß, Bach	83
A2226	Freizeitpark, Safaripark, Wildgehege	25	A5102	Kanal (Schifffahrt)	132
A2227	Grünanlage	738	A5112	Binnensee, Stausee, Teich	447
A2228	Campingplatz	20	A5303	Schleuse	10
A2230	Golfplatz	19	A5304	Schleusenkammer	4
A2301	Tagebau, Grube, Steinbruch	19	A7211	Insel	7
A2302	Halde, Aufschüttung	4	A7302	Naturschutzgebiet	14
A2314	Absetz- und Erdfaulbecken, Schlammteich	3	A7311	Wasser- und Heilquellenschutzgebiet	10
A2316	Turm	4			
					21.925

Abbildung C.1: Übersicht über die ATKIS-Objektklassen und deren Objektanzahl für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark. ATKIS besitzt 63 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.6) enthalten.

GDF		
Code	Bezeichnung	Anz.
G1119	Administrative area order 8	12
G1120	Administrative area order 9	58
G3110	Built-up area	83
G3136	Postal district area	101
G4160	Roads - Address area	29
G4311	Water areas (DISPCLASS 1)	3
G4312	Water areas (DISPCLASS 2)	1
G4313	Water areas (DISPCLASS 3)	7
G4314	Water areas (DISPCLASS 4)	11
G4315	Water areas (DISPCLASS 5)	2
G7110	Land Use: Building	7
G7120	Land Cover: Forest (Woodland)	102
G7170	Land Use: Park, Garden	6
G7500	Area structures - Bridges and tunnels	36
G9353	Land Use: Company ground	1
G9710	Land Cover: Beach, Dune and Plain sand	1
G9715	Land Use: Industrial area	51
G9720	Land Use: Industrial harbour area	1
G9725	Land Cover: Moors and heathland	13
		525

Abbildung C.2: Übersicht über die GDF-Objektklassen und deren Objektanzahl für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark. GDF besitzt 19 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.6) enthalten.

C.2 Häufigkeitsmatrizen

Tabelle C.1: Häufigkeitsmatrizen H_A (oben), H_B (mitte) und H_{AB} (unten) für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13 × 12). In H_A steht in jeder Zelle der prozentuale Flächenanteil der zugeordneten ATKIS-Flächen bezogen auf alle zugeordneten ATKIS-Flächen, während in H_B der prozentuale Flächenanteil der zugeordneten GDF-Flächen bezogen auf alle zugeordneten GDF-Flächen steht. H_{AB} beinhaltet den prozentualen Flächenanteil der zugeordneten Flächen bezogen auf beide Datensätze.

H_A		GDF												
		G1120	G3110	G3136	G4160	G4313	G4314	G4315	G7120	G7170	G7500	G9715	G9725	∑
ATKIS	A2101	50,05	39,07	7,31	0	0	0	0	1,80	0	0	1,77	0	100
	A2111	48,91	48,52	0,10	0,36	0	0	0	0	0	0	2,11	0	100
	A2112	21,63	2,54	0	0	0	0	0	0	0	0	75,83	0	100
	A2113	25,47	30,21	4,69	0	0	0	0	0	0	0	39,64	0	100
	A2114	72,11	3,24	0	24,64	0	0	0	0	0	0	0	0	100
	A2202	0	0	0	0	0	0	0	0	62,71	0	0	37,29	100
	A2227	21,54	0	0	4,51	0	0	0	0	73,95	0	0	0	100
	A3103	0	0	0	100	0	0	0	0	0	0	0	0	100
	A3514	0	0	0	0	0	0	0	48,12	0	51,88	0	0	100
	A4101	0	22,97	77,03	0	0	0	0	0	0	0	0	0	100
	A4105	0	0	0	0	0	0	0	1,09	0	0	0	98,91	100
	A4107	3,26	0	0	0	0	0	0	96,19	0	0	0	0,54	100
	A5112	0	0	0	0	39,55	53,31	7,13	0,01	0	0	0	0	100

H_B		GDF											
		G1120	G3110	G3136	G4160	G4313	G4314	G4315	G7120	G7170	G7500	G9715	G9725
ATKIS	A2101	86,87	90,61	66,77	0	0	0	0	8,88	0	0	16,42	0
	A2111	4,78	4,48	0,04	2,73	0	0	0	0	0	0	0,77	0
	A2112	4,46	0,60	0	0	0	0	0	0	0	0	75,81	0
	A2113	1,12	1,32	0,67	0	0	0	0	0	0	0	7,00	0
	A2114	0,74	0,04	0	20,62	0	0	0	0	0	0	0	0
	A2202	0	0	0	0	0	0	0	0	31,68	0	0	3,98
	A2227	1,12	0	0	18,72	0	0	0	0	68,32	0	0	0
	A3103	0	0	0	57,93	0	0	0	0	0	0	0	0
	A3514	0	0	0	0	0	0	0	0,05	0	100	0	0
	A4101	0	2,95	32,52	0	0	0	0	0	0	0	0	0
	A4105	0	0	0	0	0	0	0	0,30	0	0	0	94,16
	A4107	0,92	0	0	0	0	0	0	90,78	0	0	0	1,86
	A5112	0	0	0	0	100	100	100	0,00	0	0	0	0
	∑	100	100	100	100	100	100	100	100	100	100	100	100

H_{AB}		GDF											
		G1120	G3110	G3136	G4160	G4313	G4314	G4315	G7120	G7170	G7500	G9715	G9725
ATKIS	A2101	64,67	56,17	15,03	0	0	0	0	2,96	0	0	3,27	0
	A2111	7,30	7,75	0,05	0,62	0	0	0	0	0	0	1,11	0
	A2112	6,85	0,94	0	0	0	0	0	0	0	0	75,82	0
	A2113	1,76	2,33	1,10	0	0	0	0	0	0	0	11,46	0
	A2114	1,14	0,06	0	22,34	0	0	0	0	0	0	0	0
	A2202	0	0	0	0	0	0	0	0	41,16	0	0	9,04
	A2227	1,68	0	0	7,50	0	0	0	0	70,74	0	0	0
	A3103	0	0	0	71,54	0	0	0	0	0	0	0	0
	A3514	0	0	0	0	0	0	0	0,10	0	65,13	0	0
	A4101	0	4,74	43,47	0	0	0	0	0	0	0	0	0
	A4105	0	0	0	0	0	0	0	0,46	0	0	0	96,32
	A4107	1,40	0	0	0	0	0	0	93,29	0	0	0	0,89
	A5112	0	0	0	0	56,59	69,35	12,95	0,00	0	0	0	0

Abbildungsverzeichnis

1.1 Schematischer Verfahrensablauf für die automatische Zuordnung von Objektklassen zweier Datensätze A und B auf Basis von geometrischen Objektzuordnungen. 11

2.1 Zwei Darstellungen eines Gebäudeobjekts. 14
2.2 Verschiedene Relationsarten: Eine Ähnlichkeitsbeziehung besteht zwischen den Objekten Platz mit Grünfläche aus Datensatz A (schwarz) und dem Square-Objekt aus Datensatz B (grau), während eine Nachbarschaftsbeziehung zwischen den Objekten innerhalb eines Datensatzes vorliegt, z.B. City Park und Square bzw. Park und Platz. 14
2.3 Einfaches, hierarchisches Schema zur Beschreibung von Gewässerdaten. Die Objektklassen Fluss, Kanal, See und Teich besitzen im Gegensatz zu den rot gekennzeichneten, abstrakten Objektklassen eigene Instanzen mit identischen Attributen, wie z.B: Id und Name, aber auch unterschiedliche Attribute, um sich von anderen Klassen abzugrenzen. 15
2.4 Probleme bei der Objektzuordnung: a) geometrische Unterschiede hinsichtlich Position, Objektform und -größe, aber auch bzgl. der Attributinformationen, b) unterschiedliche Maßstäbe können zu komplexen Objektrelationen (1:n/n:1 bzw. n:m) führen und c) unterschiedliche Geometriedimensionen (Polygon- und Linienobjekte) 17
2.5 Objektzuordnung nach Van Wijngaarden u. a. (1997). Durch die geometrische Überlagerung der schwarzen und grauen Objekte werden Schnittflächen bestimmt und Überlagerungsverhältnisse s_{i_j} mit $j = A, B$ abgeleitet. Basierend auf den Verhältnissen werden einfache (L4) und komplexe Objektrelationen (L2 und L3) bestimmt, die letztendlich die Objektzuordnungen $p_{A_1} \rightarrow p_{B_1}$, $p_{A_2} \rightarrow \{p_{B_2}, p_{B_3}\}$, $\{p_{A_3}, p_{A_4}\} \rightarrow p_{B_4}$ und $\{p_{A_5}, p_{A_6}\} \rightarrow \{p_{B_6}, p_{B_7}, p_{B_8}\}$ widerspiegeln. 19
2.6 Die Objektzuordnung nach dem Vorbild von van Wijngaarden u. a. (1997) erzielt bei Datensätzen, in denen mehrere Objekte das gleiche Gebiet überlagern, keine zuverlässige Lösung. 20
2.7 Knoten-Matching nach Mustière und Devogele (2008). Die schwarzen Knoten v_T des detaillierten Datensatzes werden als vollständige (a, b), unvollständige (c) und unmögliche (d) Matching-Kandidaten zu den grauen Knoten v_C annotiert. Darauf aufbauend werden die Relationen zwischen den Knoten als sicher (a, b) bzw. unsicher (c) klassifiziert. Zwischen Knoten beider Datensätze wird in a) eine 1:1 -Relation und in b) eine 1:n -Relation identifiziert. Die gezeigten Beispiele wurden aus Mustière und Devogele (2008) entnommen und abgewandelt. 21
2.8 Buffer Growing nach Walter (1997) für die Identifizierung von komplexen n:m -Relationen zwischen Linienobjekten der Datensätze A (schwarz) und B (grau). 22
2.9 Schema-Matching zwischen Schema A und Schema B. Die gestrichelte Verbindungslinie zwischen den grau hervorgehobenen Schemaelementen Bank kennzeichnet eine 1:1 -Schemarelation. 23
2.10 Klassifikation von Schema-Matching-Ansätzen (aus Rahm und Bernstein (2001)). 24

3.1 Hausdorff-Distanz $d_{Haus}(L_A, L_B)$ zwischen a) Linien etwa gleicher Länge und b) Linien unterschiedlicher Länge. Die dünn gestrichelten Linien repräsentieren die minimalen Abstände in den Stützpunkten zwischen beiden Linien, wohingegen die dick gestrichelten Linien die Maximalwerte widerspiegeln. Daraus leitet sich in a) $d_{Haus}(L_A, L_B) = \vec{d}_{Haus}(L_A, L_B)$ und in b) $d_{Haus}(L_A, L_B) = \vec{d}_{Haus}(L_B, L_A)$ ab. 30
3.2 a) Distanzmaße für Polygone: Distanz zwischen den Objektschwerpunkten $d_S(p_A, p_B)$ im Vergleich zur Distanz zwischen den Polygonrändern $d_R(p_A, p_B)$. b) Symmetrische Differenz $p_A \Delta p_B = (p_A \backslash p_B) \cup (p_B \backslash p_A)$. 30
3.3 Vereinfachte Bestimmung der Ausrichtung von langgestreckten und gekrümmten Linienobjekten. Der Richtungswinkel α beschreibt die Orientierung der geradlinigen, gestrichelten Verbindung zwischen Anfangs- und Endpunkt einer Linie und τ beschreibt die Orientierung des mit grau gekennzeichneten, minimal umschließenden Rechtecks. 31
3.4 Bestimmung der Ausrichtung und Länge von Polygonobjekten. a) Die Objektausrichtung kann durch den Richtungswinkel τ zur Längsseite des MERs beschrieben werden. Die Objektlänge wird durch die Ausdehnung der Längsseite l_{MER} angenähert. b) Für mäandernde Objekte ist die Länge des Hauptskeletts l_G zuverlässiger. 31

3.5 Geometrische Formmerkmale für Polygonobjekte: a) Kompaktheitsmaß bezogen auf einen Kreis C_C bzw. auf ein Quadrat C_S, b) Langgestrecktheit El und c) Rechtwinkligkeit Re. 32
3.6 Der Knotengrad $deg(l_S)$ gibt als topologisches Ähnlichkeitsmaß die Anzahl der inzidenten Kanten wieder. 32
3.7 Die Rechts-Links-Relation für die Zuordnung von Linien unter Berücksichtigung der semantischen Informationen der angrenzenden Flächenobjekte. 33
3.8 Konzeptionelles Nachbarschaftmodell der topologischen Relationen für einfache flächenhafte Objekte nach Bruns und Egenhofer (1996). 33
3.9 Einfache Objektrelationen zwischen Objekten zweier Datensätze A und B: a) geometrisch reine 1:1 -Relationen zwischen Objekten gleicher Geometriedimension, b) geometrisch gemischte 1:1-Relation zwischen Objekten unterschiedlicher Geometriedimensionen, c) 1:0 -Relation aufgrund zu großer Mindestabstände. 35
3.10 Komplexe Objektrelationen zwischen Objekten zweier Datensätze A und B: a) geometrisch reine 1:n -Relation, b) geometrisch gemischte 1:n -Relation, c) geometrisch reine n:m-Relation 36
3.11 Verschiedene Graphtypen: a) ungerichteter und ungewichteter Graph, b) gerichteter Graph und c) gerichteter und gewichteter Graph. 37
3.12 Vollständig bipartiter Graph mit Gewichten w_{ij} mit $i = 1, ..., n$ und $j = 1, ..., m$. 37
3.13 Für einen bipartiten Graph werden verschiedene Matchingtypen dargestellt: a) Nicht erweiterbares Matching mit $M = 3$, b) und c) maximale und perfekte Matchings mit $M = 4$. Die starken Kanten kennzeichnen die Zuordnung. 38
3.14 Bestimmung eines maximalen Matchings mittels augmentierender Pfade: Ausgangspunkt ist das in a) präsentierte, nicht erweiterbare Matching $M = 3$, anschließend wird in b) ein verbessernder, alternierender Pfad $(a_1, b_1, a_3, b_3, a_4, b_2, a_2, b_4)$ bestimmt und in c) daraus ein verbessertes, maximales und perfektes Matching $M = 4$ abgeleitet. 38
3.15 Beispiel für ein gewichtetes maximales Matching in einem bipartiten Graphen mit den Mengen $A = \{a_1, a_2, a_3, a_4, a_5\}$ und $B = \{b_1, b_2, b_3, b_4\}$. Die starken Kanten kennzeichnen das Ergebnis des maximalen Matchings mit $\sum = 70$. Die gestrichelten Kanten markieren die Kanten zum Dummy-Element mit dem Gewicht 0. 39
3.16 Minimaler-2-Schnitt in a) ungewichteten und b) gewichteten Graphen. In a) gibt es drei verschiedene minimale Schnitte mit der gleichen Kantenanzahl $|C| = 2$: 1. min-Cut: $V_1 = \{v_6\}$ und $V_2 = \{v_1, v_2, v_3, v_4, v_5\}$, 2. min-Cut: $V_1 = \{v_1\}$ und $V_2 = \{v_2, v_3, v_4, v_5, v_6\}$ und 3. min-Cut: $V_1 = \{v_3\}$ und $V_2 = \{v_1, v_2, v_4, v_5, v_6\}$. In b) sind zwei minimale Schnitte mit $w(C) = 6$ bei $|C| = 2$ und $|C| = 3$ möglich, wobei der 1. min-Cut mit dem ersten Schnitt aus a) identisch ist und der 2. min-Cut andere Kanten schneidet: $V_1 = \{v_5, v_6\}$ und $V_2 = \{v_1, v_2, v_3, v_4\}$ 40
3.17 Minimaler Schnitt für den gewichteten Graphen aus Abb. 3.15. Der Graph wird in die Partitionen $P_A = \{\{a_1, a_2, a_4, a_5\}, \{a_3\}\}$ und $P_B = \{\{b_1, b_3, b_4\}, \{b_2\}\}$ geteilt. Der Schnitt ist durch eine starke Linie und die geschnittenen Kanten mit Strichpunktlinien gekennzeichnet. Der minimale Schnitt ermöglicht folgende 1:1-Zuordnung zwischen den Partitionen: $\{a_1, a_2, a_4, a_5\} \rightarrow \{b_1, b_3, b_4\}$ und $a_3 \rightarrow b_2$. Die Summe der geschnittenen Kanten beträgt $w(C) = 7$. 40
3.18 Graphische Lösung eines LP und eines IP. Das zweidimensionale, konvexe Polytop (graue Fläche) stellt den Lösungsraum dar und wird durch die Restriktionsgleichungen g_1, g_2 und die Koordinatenachsen x_1, x_2 begrenzt. Die rot markierten Eckpunkte sind mögliche Lösungspunkte des LPs und die schwarzen Punkte innerhalb des Polytops die des IPs. Der vergrößerte rote Punkt kennzeichnet das LP -Optimum und der schwarze das IP -Optimum. Die rot gestrichelten Linien sind Linien gleicher Kosten und erhöhen sich mit Abstand vom Koordinatenursprung. 42

4.1 Schematische Darstellung des Zuordnungsprozesses von Polygonobjekten. Die als Ergebnis identifizierten finalen Objektrelationen werden in einer Häufigkeitsmatrix zusammengefasst und dienen als Eingabe für das Schema-Matching. 45
4.2 Bei der Überlagerung von verschiedenen Polygonobjekten p_A und p_B werden unterschiedliche Werte für die Flächenparameter s_{i_j} und den Ausrichtungsparameter s_a bestimmt. 46
4.3 Bildung einer komplexen Objektrelation zwischen dem schwarz umrandeten GDF-Objekt von Datensatz A und den grauen ATKIS-Objekten aus Datensatz B durch die Zusammenfassung von Nachbarobjekten unter Berücksichtigung ihrer Objektklassen. Zur Bewahrung der Übersichtlichkeit, werden die sich überlagernden Objekte beider Datensätze räumlich getrennt dargestellt. Erläuterungen zu den vorliegenden Objektklassen sind im Anhang in der Übersichten C.1 und C.2 zu finden. 47

4.4 Die Einführung des Heterogenitätsparameter s_h kann die in a) dargestellte einfache Objektrelation mit geringerem geometrischen Parameter gegenüber der in b) dargestellten komplexen Objektrelation für die endgültige Relationsauswahl bevorzugen. 48
4.5 Auswertung der komplexen Objektrelationen für die Objektklassenzuordnung. Die Objektrelationen in a) und c) sind Vertreter von homogenen Schemarelationen und werden wie 1:1-Relationen berücksichtigt. Im Gegensatz dazu repräsentieren b) und d) heterogene Schemarelationen, erkennbar an den unterschiedlichen Grautönen und den Heterogenitätswerten $s_h < 1$. Die Bestimmung von s_h basiert in diesem Beispiel auf $|A| = 2$ und $|B| = 3$. 49
4.6 Schematische Darstellung des Zuordnungsprozesses von Polygon- und Linienobjekten. 50
4.7 Die unterschiedliche Skelettbildung bei a) zwei benachbarten Polygonen p_1 und p_2 und b) bei einem aggregierten Polygon. In c) wird die Verlängerung des linienhaften Objektes l bis hin zur Skelettlinie des Polygons l_G gezeigt, die notwendig ist, um die topologische Verbindung zu erhalten. 50
4.8 In a) wird die Bestimmung des besten Matching-Kandidaten für die graue Kante e_B des kleinmaßstäbigen Maßstabs gezeigt. Beide schwarzen Kanten e_{A_1} und e_{A_2} des großmaßstäbigen Datensatzes sind Matching-Kandidaten, da jede Kante das Pufferpolygon von e_B schneidet. Innerhalb der grauen Schnittpolygone werden die fett markierten Kantenteile auf die Winkeldifferenz hin überprüft. Aufgrund der geringeren Winkeldifferenz besitzt e_{A_2} eine höhere Match-Qualität als e_{A_1}. In b) und c) werden auf Basis der zugeordneten Knoten, gekennzeichnet durch Ellipsen, Beispiele für die Zuordnung der Kanten präsentiert. Während in b) beide Endpunkte von e_B einen Matching-Kandidaten besitzen, hat in c) entweder nur ein Endpunkt oder gar kein Endpunkt einen Matching-Kandidaten. 51

5.1 Min-Cut-Lösung für das synthetische Beispiel. Darstellung der Partitionierung im a) bipartiten Graph und b) in der Häufigkeitsmatrix: $P_A = \{\{\text{Wa, Ac, Si, Fl, St}\}, \text{Ba}\}$ und $P_B = \{\{\text{Bw, Nw, Lw, Ac, Ge, Ve}\}, \text{Sc}\}$. 55
5.2 Rekursive Unterteilung der Häufigkeitsmatrix des synthetischen Beispiels mit dem Min-Cut-Algorithmus. In jedem Verfahrensschritt entstehen zwei Cluster. Das hellgraue Cluster wird weiter unterteilt. Insgesamt sind vier Berechnungen des minimalen Schnitts notwendig. Es entstehen fünf Cluster mit einer Häufigkeit von $H_{(k=5)} = 592$. 56
5.3 MaxScore-Lösung: Clusterbildung unter Maximierung der Häufigkeiten $H_{(k)}$ für $k = 2, 3, 4, 5, 6$. Zellen, die zu einem Cluster gehören, haben den gleichen farbigen Hintergrund. 58
5.4 BalancedSize-Lösung: Clusterbildung unter Berücksichtigung ausgewogener Clustergrößen bezüglich der Zellenanzahl $H_{a(k)}$ für $k = 2, 3, 4, 5, 6$. Die präsentierte Lösung stellt eine von vielen optimalen Lösungen dar. Zellen, die zu einem Cluster gehören, haben den gleichen farbigen Hintergrund. 59
5.5 BalancedScore-Lösung: Clusterbildung unter Berücksichtigung ausgewogener Clustergrößen bezüglich der Häufigkeiten $H_{b(k)}$ für $k = 2, 3, 4, 5, 6$. Zellen, die zu einem Cluster gehören, haben den gleichen farbigen Hintergrund. 60
5.6 Geometrische Darstellung der Kombination von zwei Optimierungszielen: a) mittels gewichteter Summe und einem Gewichtsfaktor $s = 0{,}5$, b) durch Maximierung von H bei gleichzeitiger Beschränkung der Clustervariabilität als harte Restriktion, welches entweder die Clustergrößendifferenz Δ_{clsize} bzw. die Clusterhäufigkeitsdifferenz $\Delta_{clscore}$ entspricht. Die roten gestrichelten Linien sind Linien gleicher Qualität. 61
5.7 Clusterbildung bei gleichgewichteter Kombination von zwei Optimierungszielen mit dem Gewichtsfaktor $s = 0{,}5$: WeightedSumMaxScoreBalancedSize (oben) und WeightedSumMaxScoreBalancedScore (unten). 62

6.1 Testgebiet A: Gebäudeobjekte der Datensätze ALKIS (links) und OSM (rechts) in Hannover. . . 68
6.2 Testgebiet B: Objekte der tatsächlichen Nutzung der Datensätze ALKIS 1:1.000 (oben) und ATKIS 1:10.000 (unten) in Hameln. 70
6.3 Überlagerung der Beispieldaten ALKIS (gefüllte Polygone mit starken weißen Konturen) und ATKIS (nicht gefüllte Polygone mit schwarzen Konturen) im Testgebiet B. Die zugehörigen Objektklassen sind mit unterschiedlichen Schriftfarben gekennzeichnet: ALKIS (weiß) und ATKIS (schwarz). 71
6.4 Testgebiet C: Objekte der Datensätze ATKIS 1:10.000 (links) und GDF 1:25.000 (rechts) in der Region Hannover-Wedemark. 71
6.5 Beispiele für komplexe Objektrelationen im Testgebiet A: a) n:1-Relation zwischen 11 ALKIS-Objekten und einen OSM-Objekt, b) 1:n-Relation zwischen einem ALKIS- und drei OSM-Objekten und c) n:m-Relation zwischen zwei ALKIS- und drei OSM-Objekten. 73

6.6 Beispiele für vom Verfahren identifizierte Objektrelationen in Testgebiet A. Während die rot markierten Relationen im Zuordnungsprozess aufgrund des Schwellwerts für die Flächenparameter und die schwarzen aufgrund der doppelten Objekteinträge wieder verworfen werden, sind die grün gekennzeichneten Relationen Bestandteil der endgültigen Ergebnisliste. 74
6.7 Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t für die im Zuordnungsprozess identifizierten endgültigen Relationen in Testgebiet A. Die Häufigkeiten sind für die unterschiedlichen Relationsarten dargestellt. 75
6.8 Testgebiet B: Ergebnis der Objektzuordnung: ALKIS 1:1.000 (oben) und ATKIS 1:10.000 (unten) - Teilmenge der korrekten (tp (grau), tn (hellgrau)) und der fehlerhaften Zuordnung (fp (rot), fn (rosa)). 77
6.9 Beispiele für vom Verfahren identifizierte Objektrelationen in Testgebiet B. Während die rot markierten Relationen im Zuordnungsprozess aufgrund des Schwellwerts für die Flächenparameter wieder verworfen werden, ist die grün gekennzeichnete Relation Bestandteil der endgültigen Ergebnisliste. Bei a) wird die 1:2-Relation ausgewählt und in b) keine. 78
6.10 Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t für die im Zuordnungsprozess identifizierten endgültigen Relationen in Testgebiet B. Die Häufigkeiten sind für die unterschiedlichen Relationsarten dargestellt. 78
6.11 Häufigkeitsverteilung der Gesamtähnlichkeitsmaße s_t für die im Zuordnungsprozess identifizierten endgültigen Relationen in Testgebiet C. Die Häufigkeiten sind für die unterschiedlichen Relationsarten dargestellt. 81
6.12 Unterschiede in der Objektmodellierung im Testgebiet C: ATKIS - GDF in Hannover-Wedemark. 82
6.13 Beispiel für eine verworfene Objektrelation aufgrund des unterschrittenen Flächenparameters. . . 84
6.14 Histogramm der Relationsanteile der Häufigkeitsmatrix H_R für das Beispiel B: ALKIS - ATKIS in Hameln (17×16). Die weißen Balken kennzeichnen Relationsanteile kleiner Eins. 87

A.1 Übersicht über die ALKIS-Objektklassen und deren Objektanzahlen für das Testgebiet A: ALKIS - OSM in Hannover. ALKIS besitzt 49 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.1) enthalten. 109
A.2 Übersicht über die OSM-Objektklassen und deren Objektanzahlen für das Testgebiet A: ALKIS - OSM in Hannover. OSM besitzt 72 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.1) enthalten. 110

B.1 Übersicht über die Objektklassen und deren Objektanzahl für das Testgebiet B: ALKIS - ATKIS in Hameln. ALKIS und ATKIS besitzen jeweils 17 Objektklassen. Das Objekt der grau eingefärbten Objektklasse ist in keiner Relationen der Objektzuordnung (Tabelle 6.3) enthalten. . . . 115

C.1 Übersicht über die ATKIS-Objektklassen und deren Objektanzahl für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark. ATKIS besitzt 63 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.6) enthalten. 117
C.2 Übersicht über die GDF-Objektklassen und deren Objektanzahl für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark. GDF besitzt 19 Objektklassen. Objekte der grau eingefärbten Objektklassen sind nicht in den Relationen der Objektzuordnung (Tabelle 6.6) enthalten. 118

Tabellenverzeichnis

2.1 Auszug aus den Ergebnissen der Objektzuordnung für die GDF-Objektklasse RoadElement für das zweite Testgebiet aus der Arbeit von Volz (2006). . . . 28
2.2 Korrelationswerte für die GDF-Objektklasse RoadElement aus den in Tabelle 2.1 angegebenen Zuordnungsergebnissen. . . . 28

4.1 Allgemeine Häufigkeitsmatrix H als Ergebnis des Data-Matchings zwischen Menge A (Objektklassen in Datensatz A) und Menge B (Objektklassen in Datensatz B). Jede Zelle der Häufigkeitsmatrix repräsentiert eine Klassenkombination und beinhaltet eine Trefferzahl $h_{ij} = H(a_i, b_j)$ mit $i = 1, ..., n$ und $j = 1, ..., m$. . . . 48
4.2 Häufigkeitsmatrix H für die in Abbildung 4.5 präsentierten Objektrelationen. Neben den einzelnen Relationsanteilen, die mit dem Index des Beispiels gekennzeichnet sind, werden in fett die Gesamtrelationsanteile und in [ ] die prozentualen Anteile angegeben. . . . 49

5.1 Synthetisches Beispiel: Die Gesamthäufigkeit der 6×7 Matrix beträgt $H_{\text{total}} = 600$. . . . 54
5.2 Max-Match-Lösung für das synthetische Beispiel in Matrixdarstellung. Die grau unterlegten Werte kennzeichnen die Zuordnung mit $H_{(k=7)} = 372$. . . . 55
5.3 Vergleich der Clusterbildung unter verschiedenen Bedingungen: MaxScore; BalancedSize; BalancedScore. $clsize_k$ gibt die Anzahl der Zellen pro Cluster, $clscore_k$ die einzelnen Clusterhäufigkeiten und $\emptyset H_{Ze}$ die Durchschnittshäufigkeit pro Zelle wieder. . . . 59
5.4 Auflistung der einzelnen Zielfunktionsterme für die gleichgewichtete Kombination zweier Optimierungsziele nach Gleichung 5.16 mit $s = 0{,}5$ für WeightedSumMaxScoreBalancedSize und WeightedSumMaxScoreBalancedScore. $clsize$ repräsentiert die Anzahl aller Zellen in den Clustern und $\emptyset H_{Ze} = H/clsize$ gibt die Durchschnittshäufigkeit pro Zelle an. . . . 61
5.5 Ergebnisse für MaxScoreHardConstraintVariableSize (oben) und MaxScoreHardConstraintVariableScore (unten) für das synthetische Beispiel. Die Clustervariabilität $clsize_{var}$ bzw. $clscore_{var}$ wurde als harte Bedingung eingefügt. Die mit $*$ gekennzeichneten Zielfunktionswerte repräsentieren die MaxScore-Lösung. Die grau hervorgehobenen Lösungen repräsentieren die optimalen Lösungen mit der höchsten Durchschnittshäufigkeit pro Zelle $\emptyset H_{Ze}$. . . . 63
5.6 Schranken für $clsize_o$ und $clsize_u$ für das synthetische Beispiel. . . . 64
5.7 Ergebnisse für MaxScoreHardConstraintFixedSize für das synthetische Beispiel. Die mit $*$ gekennzeichneten Zielfunktionswerte repräsentieren die MaxScore-Lösung und $clsize$ ist die maximale Anzahl der Zellen, die zur Clusterbildung erlaubt sind. Die grau hervorgehobene Lösung repräsentiert die optimale Lösung mit der höchsten Durchschnittshäufigkeit pro Zelle $\emptyset H_{Ze}$. . . . 65

6.1 Ergebnisse des Data-Matching-Verfahrens für Testgebiet A: ALKIS - OSM in Hannover. . . . 72
6.2 Konfusionsmatrizen für Testgebiet A: ALKIS - OSM in Hannover. Links ist die Allgemeine Konfusionsmatrix, in der Mitte für ALKIS und rechts für OSM. Die obere Zeile bezieht sich auf die Objektanzahlen (#) und die untere Zeile auf die Flächen $[ha]$. . . . 75
6.3 Ergebnisse des Data-Matching-Verfahrens für Testgebiet B: ALKIS - ATKIS in Hameln. . . . 76
6.4 Konfusionsmatrizen für Testgebiet B: ALKIS - ATKIS in Hameln. Links ist die Matrix für die ALKIS-Daten und rechts für ATKIS-Daten. Die obere Zeile bezieht sich auf die Objektanzahlen (#) und die untere Zeile auf die Flächen $[ha]$. . . . 79
6.5 Quantitive Teilmengen tp, tn, fp und fn pro Objektklasse und Datensatz für Testgebiet B: ALKIS - ATKIS in Hameln. Grau hinterlegte Werte spiegeln den jeweils höchsten Prozentanteil wider. . . . 80
6.6 Ergebnisse des Data-Matching-Verfahrens für Testgebiet C: ATKIS - GDF in Hannover-Wedemark. . . . 80
6.7 Konfusionsmatrizen für Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Links ist die Matrix für die ATKIS-Daten und rechts für GDF-Daten. Die obere Zeile bezieht sich auf die Objektanzahlen (#) und untere Zeile auf die Flächen [ha]. . . . 82
6.8 Quantitative Teilmengen tp, tn, fp und fn pro Objektklasse und Datensatz für Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Grau hinterlegte Werte spiegeln den jeweils höchsten Prozentanteil wider. . . . 83

6.9 Zusammenfassung der quantitativen Maße Genauigkeit, Sensitivität und F-Maß zur Bewertung der Zuordnungsqualität von allen drei Testgebieten. Die Maße sind sowohl auf Objektanzahlen (#) als auch auf Flächen in % angegeben. 84

6.10 Häufigkeitsmatrix H_R für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). In jeder Zelle steht die Anzahl der Relationen bezogen auf den jeweiligen Flächenanteil, an dem beide Objektklassen beteiligt sind. Die hellgrau, grün und gelb hervorgehobenen Werte spiegeln Zuordnungen zwischen gleichen Objektklassen wider. Die orangefarbene Zelle stellt eine schwache Zuordnung dar. 86

6.11 Beispielhafte Schemarelationen R_s mit Einzelhäufigkeiten h_p mit $p = \{R, A, B, AB\}$ der unterschiedlichen Häufigkeitsmatrizen. 88

6.12 Referenzzuordnung für Testgebiet B: ALKIS - ATKIS mit Angabe der Einzelhäufigkeiten h_p mit $p = \{R, A, B, AB\}$ pro Cluster und den Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ aller Referenzcluster. Zeile $H_{\mathrm{total},p}$ gibt die Gesamtmatrixinhalte an. 88

6.13 Ergebnisse der verschiedenen Schema-Matching-Verfahren für Testgebiet B: ALKIS - ATKIS in Hameln für H_R, H_A, H_B und H_{AB}. Zeile 0a gibt die Gesamtmatrixinhalte an, während Nr. 0b die Referenzzuordnung kennzeichnet. Zeile 1 präsentiert die Ergebnisse des einfachen Max-Match-Verfahrens, Zeile 2 des Heuristischen Verfahrens und Zeile 3 der Optimierungsverfahren mit dem größten $\varnothing H_{Ze}$. k steht für die Anzahl der Cluster, Spalte H kennzeichnet die Verfahrenslösung, NC repräsentiert die Anzahl der Nullcluster, während NZ die Anzahl der Nullzellen in den Clustern wiedergibt. Alle unter alleiniger Maximierung der Häufigkeiten entstandenen Lösungen sind zusätzlich mit $*$ gekennzeichnet. Unter der Clusteranzahl k steht der Schrankenwert $\{clsize_{var}\}$ bzw. $(clsize)$ in Klammern. 90

6.14 Ergebnis des Heuristischen Verfahren für H_{AB} für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). Es werden insgesamt 14 Cluster mit 19 Zellen und einer Gesamthäufigkeit von $H_{AB(k=14)} = 1.005{,}33$ bestimmt. 91

6.15 Überlagerte Ergebnisse der Heuristischen Lösungen für alle Häufigkeitsmatrizen für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). Je dunkler die Zellen, desto mehr Lösungen beinhalten die Zuordnung. 91

6.16 Testgebiet B: ALKIS - ATKIS in Hameln. Übersicht über die Anzahl der identischen Cluster zwischen den H_{AB}-Verfahrenslösungen und der Referenzzuordnung. Mit (x) gekennzeichnete Relationen besitzen identische Clusterhäufigkeiten unter Vernachlässigung von Nullzellen. 92

6.17 Testgebiet B: ALKIS - ATKIS in Hameln. Übersicht über die benötigte Rechenzeit, die Anzahl der ausgeführten Rechenschritte pro H_{AB}-Verfahrenslösung, die besten Ergebnisse bezogen auf die Referenzschemarelationen R_s^*, Schnittmenge zur Referenzzuordnung ($\cap 0b$) und zum Matrixinhalt ($\cap\, 0a$). 93

6.18 Testgebiet B: ALKIS - ATKIS in Hameln. Gegenüberstellung der Lösungscluster in der H_{AB}-Matrix der Optimierungsverfahren MaxScore, WSMSBS, MSHCVS und MSHCFS-U, als beste Lösung hinsichtlich der Referenzzuordnung, mit denen des Heuristischen Verfahrens, als beste Lösung hinsichtlich des Matrixgesamtinhalts. Die grau markierten Relationen kennzeichnen Referenzcluster, □ markiert identische Cluster zwischen den Lösungen und ∘ Nullcluster. 94

6.19 Häufigkeitsmatrix H_R für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). In jeder Zelle steht die Anzahl der Relationen bezogen auf den jeweiligen Flächenanteil, an denen beide Objektklassen beteiligt sind. Die grauen Zellen kennzeichnen die Referenzzuordnung. 95

6.20 Referenzzuordnung für Testgebiet A: ALKIS - OSM mit Angabe der Einzelhäufigkeiten h_p und der Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ mit $p = \{R, A, B, AB\}$ aller Referenzcluster der vier verschiedenen Häufigkeitsmatrizen. Zeile H_{total} gibt die Gesamtmatrixinhalte an. 96

6.21 Ergebnisse der verschiedenen Schema-Matching-Verfahren für Testgebiet A: ALKIS - OSM in Hannover für H_R, H_A, H_B und H_{AB}. Zeile 0a gibt die Gesamtmatrixinhalte an, während Nr. 0b die Referenzzuordnung kennzeichnet. Zeile 1 präsentiert die Ergebnisse des einfachen Max-Match-Verfahrens, Zeile 2 des Heuristischen Verfahrens und Zeile 3 der Optimierungsverfahren mit dem größten $\varnothing H_{Ze}$. k steht für die Anzahl der Cluster, Spalte H kennzeichnet die Verfahrenslösung, NC repräsentiert die Anzahl der Nullcluster, während NZ die Anzahl der Nullzellen in den Clustern wiedergibt. Alle unter alleiniger Maximierung der Häufigkeiten entstandenen Lösungen sind zusätzlich mit $*$ gekennzeichnet. Unter der Clusteranzahl k steht der Schrankenwert $\{clsize_{var}\}$ bzw. $(clsize)$ in Klammern. 97

6.22 Testgebiet A: ALKIS - OSM in Hannover. Übersicht über die Anzahl der identischen Cluster zwischen den H_R-Verfahrenslösungen und der Referenzzuordnung. Mit (x) gekennzeichnete Relationen besitzen identische Clusterhäufigkeiten unter Vernachlässigung von Nullzellen. In der letzten Spalte kennzeichnet x_{AB} die identischen Cluster der MSHCFS-U -Lösung speziell für die H_{AB}-Matrix mit der Referenzzuordnung. 98

6.23 Testgebiet A: ALKIS - OSM in Hannover. Übersicht über die benötigte Rechenzeit, die Anzahl der ausgeführten Rechenschritte pro H_R-Verfahrenlösung, die besten Ergebnisse bezogen auf die Referenzschemarelationen R_s^*, Schnittmenge zur Referenzzuordnung ($\cap 0b$) und zum Matrixinhalt ($\cap\, 0a$). 98

6.24 Testgebiet A: ALKIS - OSM in Hannover. Gegenüberstellung der Lösungscluster in der H_R-Matrix des Heuristischen Verfahrens, als beste Lösung hinsichtlich der Referenzzuordnung und des Matrixgesamtinhalts, mit denen der besten Optimierungslösung (MaxScore, MSHCVS, MSHCFS-U), als zweitbeste Lösung hinsichtlich der Referenzzuordnung und des Matrixgesamtinhalts. Die grau markierten Relationen kennzeichnen Referenzcluster, □ markiert identische Cluster zwischen den Lösungen und ∘ Nullcluster. 99

6.25 Häufigkeitsmatrix H_R für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). In jeder Zelle steht die Anzahl der Relationen bezogen auf den jeweiligen Flächenanteil, an dem beide Objektklassen beteiligt sind. Die grauen Zellen kennzeichnen die Referenzzuordnung. . . . 100

6.26 Referenzzuordnung für Testgebiet C: ATKIS - GDF mit Angabe der Einzelhäufigkeiten h_p und der Gesamthäufigkeiten $H_{\mathrm{Ref},p}$ mit $p = \{R, A, B, AB\}$ aller Referenzcluster der vier verschiedenen Häufigkeitsmatrizen. Zeile H_{total} gibt die Gesamtmatrixinhalte an. 101

6.27 Ergebnisse der verschiedenen Schema-Matching-Verfahren für Testgebiet C: ATKIS - GDF in Hannover-Wedemark für H_R, H_A, H_B und H_{AB}. Zeile 0a gibt die Gesamtmatrixinhalte an, während Nr. 0b die Referenzzuordnung kennzeichnet. Zeile 1 präsentiert die Ergebnisse des einfachen Max-Match-Verfahrens, Zeile 2 des Heuristischen Verfahrens und Zeile 3 der Optimierungsverfahren mit dem größten $\varnothing H_{Ze}$. k steht für die Anzahl der Cluster, Spalte H kennzeichnet die Verfahrenslösung, NC repräsentiert die Anzahl der Nullcluster, während NZ die Anzahl der Nullzellen in den Clustern wiedergibt. Alle unter alleiniger Maximierung der Häufigkeiten entstandenen Lösungen sind zusätzlich mit $*$ gekennzeichnet. Unter der Clusteranzahl k steht der Schrankenwert $\{clsize_{var}\}$ bzw. $(clsize)$ in Klammern. 101

6.28 Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Übersicht über die benötigte Rechenzeit, die Anzahl der ausgeführten Rechenschritte pro H_{AB}-Verfahrenlösung, die besten Ergebnisse bezogen auf die Referenzschemarelationen R_s^*, Schnittmenge zur Referenzzuordnung ($\cap\, 0b$) und zum Matrixinhalt ($\cap\, 0a$). 102

6.29 Optimale Lösung des MSHCVS- und MSHCFS-U-Verfahrens für H_{AB} für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). Es werden insgesamt 10 Cluster mit 15 Zellen und einer Gesamthäufigkeit von $H_{(k=10)} = 838{,}04$ bestimmt. Die Durchschnittshäufigkeit pro Zelle beträgt $\varnothing H_{Ze} = 55{,}87$. 102

6.30 Ergebnis des Heuristischen und MSHCFSNE-Verfahrens für H_{AB} für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). Es werden insgesamt 10 Cluster mit 15 Zellen und einer Gesamthäufigkeit von $H_{(k=10)} = 802{,}58$ bestimmt. Die Durchschnittshäufigkeit pro Zelle beträgt $\varnothing H_{Ze} = 53{,}51$. 103

6.31 Testgebiet C: ATKIS - GDF in Hannover-Wedemark. Übersicht über die Anzahl der identischen Cluster zwischen den H_{AB}-Verfahrenslösungen und der Referenzzuordnung. 103

6.32 Ranking der Verfahrensergebnisse aller Testgebiete hinsichtlich Rechenzeit, Schnittmenge zum Matrixgesamtinhalt ($\cap$0a) und zur Referenzzuordnung ($\cap$0b). Die Werte in Klammern repräsentieren das Ranking ohne Berücksichtigung der Referenzzuordnung. 104

A.1 Häufigkeitsmatrix H_A für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). In jeder Zelle steht der prozentuale Flächenanteil der zugeordneten ALKIS-Flächen bezogen auf alle zugeordneten ALKIS-Flächen. 111

A.2 Häufigkeitsmatrix H_B für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). In jeder Zelle steht der prozentuale Flächenanteil der zugeordneten OSM-Flächen bezogen auf alle zugeordneten OSM-Flächen. 112

A.3 Häufigkeitsmatrix H_{AB} für das Testgebiet A: ALKIS - OSM in Hannover in transponierter Form (31×17). Jeder Zelle beinhaltet den prozentualen Flächenanteil der zugeordneten Flächen bezogen auf beide Datensätze. 112

A.4 Ergebnis des Heuristischen Verfahrens für H_R für das Testgebiet A: ALKIS-OSM in Hannover (31×17). Es werden insgesamt 15 Cluster mit 33 Zellen und einer Gesamthäufigkeit von $H_{R(k=15)} = 1.674,54$ bestimmt. 113

B.1 Häufigkeitsmatrizen H_A (oben), H_B (mitte) und H_{AB} (unten) für das Testgebiet B: ALKIS - ATKIS in Hameln (17×16). In H_A steht in jeder Zelle der prozentuale Flächenanteil der zugeordneten ALKIS-Flächen bezogen auf alle zugeordneten ALKIS-Flächen, während in H_B der prozentuale Flächenanteil der zugeordneten ATKIS-Flächen bezogen auf alle zugeordneten ATKIS-Flächen steht. H_{AB} beinhaltet den prozentualen Flächenanteil der zugeordneten Flächen bezogen auf beide Datensätze. 116

C.1 Häufigkeitsmatrizen H_A (oben), H_B (mitte) und H_{AB} (unten) für das Testgebiet C: ATKIS - GDF in Hannover-Wedemark (13×12). In H_A steht in jeder Zelle der prozentuale Flächenanteil der zugeordneten ATKIS-Flächen bezogen auf alle zugeordneten ATKIS-Flächen, während in H_B der prozentuale Flächenanteil der zugeordneten GDF-Flächen bezogen auf alle zugeordneten GDF-Flächen steht. H_{AB} beinhaltet den prozentualen Flächenanteil der zugeordneten Flächen bezogen auf beide Datensätze. 119

Literaturverzeichnis

Abadi, D. J., Marcus, A., Madden, S., Hollenbach, K. J., 2007. Scalable Semantic Web Data Management Using Vertical Partitioning. In: *Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007*. S. 411–422.

Abdolmajidi, E., Mansourian, A., Will, J., Harrie, L., 2015. Matching authority and VGI road networks using an extended node-based matching algorithm. *Geo-spatial Information Science* 18 (2-3), S. 65–80.

AdV, 2008. Dokumentation zur Modellierung der Geoinformationen des amtlichen Vermessungswesens (GeoInfoDok 6.0). Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland. `http://www.adv-online.de/AAA-Modell/Dokumente-der-GeoInfoDok/GeoInfoDok-6.0/binarywriterservlet?imgUid=42b23fd2-1153-911a-3b21-718a438ad1b2&uBasVariant=11111111-1111-1111-1111-111111111111` (Letzter Zugriff: 27.04.2017).

Alt, H., Godau, M., 1995. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications* 5 (01-02), S. 75–91.

Amin, M., Khan, W., Hussain, S., Bui, D.-M., Banos, O., Kang, B., Lee, S., 2016. Evaluating Large-Scale Biomedical Ontology Matching Over Parallel Platforms. *IETE Technical Review* 33 (4), S. 415–427.

Arnold, P., Rahm, E., 2014. Enriching Ontology Mappings with Semantic Relations. *Data & Knowledge Engineering* 93 (C), S. 1–18.

Aumueller, D., Do, H.-H., Massmann, S., Rahm, E., 2005. Schema and Ontology Matching with COMA++. In: *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. SIGMOD '05. ACM, New York, NY, USA, S. 906–908.

Bellahsene, Z., Bonifati, A., Rahm, E. (Hrsg.), 2011. Schema Matching and Mapping. Springer-Verlag, Berlin, Deutschland.

Belussi, A., Catania, B., Podestà, P., 2005. Towards topological consistency and similarity of multiresolution geographical maps. In: *GIS'05: Proceedings of the 13th annual ACM international workshop on Geographic information systems*. ACM Press, New York, NY, USA, S. 220–229.

Bernstein, P. A., Madhavan, J., Rahm, E., 2011. Generic Schema Matching, Ten Years Later. *Proceedings of the Very Large Data Bases Endowment* 4 (11), S. 695–701.

Bishr, Y., 1997. Semantic Aspects of Interoperable GIS. Dissertation, Wageningen Agricultural University and International Institute for Aerospace Survey and and Earth Science (ITC), Enschede, Niederlande.

Bruns, H. T., Egenhofer, M. J., 1996. Similarity of Spatial Scenes. In: Kraak, J.-M., Molenaar, M. (Hrsg.), *Seventh International Symposium on Spatial Data Handling, Delft, The Netherlands*. Taylor & Francis, London, S. 173–184.

Chen, C., Shahabi, C., Kolahdouzan, M., Knobloc, C., 2006. Automatically and Efficiently Matching Road Networks with Spatial Attributes in Unknown Geometry Systems. In: *3rd Workshop on Spatio-Temporal Database Management (STDBM'06)*.

Dahlhaus, E., Johnson, D. S., Papadimitriou, C. H., Seymour, P. D., Yannakakis, M., 1992. The Complexity of Multiway Cuts (Extended Abstract). In: *Proceedings of the Twenty-fourth Annual ACM Symposium on Theory of Computing*. STOC '92. ACM, New York, NY, USA, S. 241–251.

Dantzig, G. B., 1963. Linear Programming and Extensions. Princeton University Press, Princeton, NJ, USA.

Diestel, R., 2000. Graphentheorie, 2. Auflage. Springer-Verlag Berlin Heidelberg.

Diez, Y., Lopez, M. A., Sellarès, J., 2008. Noisy Road Network Matching. In: Cova, T. J., Miller, H. J., Beard, K., Frank, A. U., Goodchild, M. F. (Hrsg.), *Geographic Information Science*. Bd. 5266 von Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg, S. 38–54.

Do, H.-H., Rahm, E., 2007. Matching Large Schemas: Approaches and Evaluation. *Information Systems* 32 (6), S. 857–885.

Duchateau, F., Bellahsene, Z., Coletta, R., 2008. A Flexible Approach for Planning Schema Matching Algorithms. In: Meersman, R., Tari, Z. (Hrsg.), *On the Move to Meaningful Internet Systems: OTM 2008*. Bd. 5331 von Lecture Notes in Computer Science. Springer-Verlag Berlin Heidelberg, S. 249–264.

Duckham, M., Worboys, M., 2005. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science* 19, S. 537–557.

Dunkars, M., 2004. Multiple Representation Databases for Topographic Information. Dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden.

Egenhofer, M., Herring, J., 1991. Categorizing Binary Topological Relationships Between Regions, Lines, and Points in Geographic Databases. Tech. rep., Department of Surveying Engineering, University of Maine, Orono, ME.

Egenhofer, M. J., 1989. A Formal Definition of Binary Topological Relationships. In: *3rd International Conference, FODO 1989 on Foundations of Data Organization and Algorithms*. Springer-Verlag New York, Inc., New York, NY, USA, S. 457–472.

Ehrgott, M., 2005. Multicriteria Optimization. Springer-Verlag, Berlin, Heidelberg.

Ehrig, M., Staab, S., 2004. QOM - Quick Ontology Mapping. In: *In Proceedings 3rd International Semantic Web Conference (ISWC04)*. Bd. 3298 von Lecture Notes in Computer Science. Springer Berlin Heidelberg, S. 683–697.

Ehrig, M., Staab, S., Sure, Y., 2005. Bootstrapping Ontology Alignment Methods with APFEL. In: Gil, Y., Motta, E., Benjamins, V., Musen, M. (Hrsg.), *The Semantic Web - ISWC 2005*. Bd. 3729 von Lecture Notes in Computer Science. Springer Berlin Heidelberg, S. 186–200.

Europäische Union, 2007. Richtlinie 2007/2/EG des Europäischen Parlaments und des Rates vom 14. März 2007 zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE). Amtsblatt der Europäischen Union, L 108, 25. April 2007.

Euzenat, J., Shvaiko, P., 2007. Ontology Matching. Springer-Verlag, Heidelberg (DE).

Euzenat, J., Shvaiko, P., 2013. Ontology Matching, 2. Auflage. Springer-Verlag, Heidelberg (DE).

F. Cruz, I., Palandri Antonelli, F., Stroe, C., 2009. AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. *Proceedings of the VLDB Endowment* 2, S. 1586–1589.

Falleri, J.-R., Huchard, M., Lafourcade, M., Nebut, C., 2008. Metamodel Matching for Automatic Model Transformation Generation. In: Czarnecki, K., Ober, I., Bruel, J.-M., Uhl, A., VÃPlter, M. (Hrsg.), *Model Driven Engineering Languages and Systems*. Bd. 5301 von Lecture Notes in Computer Science. Springer Berlin Heidelberg, S. 326–340.

Fan, H., Yang, B., Zipf, A., Rousell, A., 2016. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. *International Journal of Geographical Information Science* 30 (4), S. 748–764.

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality Assessment for Building Footprints Data on OpenStreetMap. *International Journal of Geographical Information Science* 28 (4), S. 700–719.

Ford, L. R., Fulkerson, D. R., 1958. Maximal Flow through a Network. *Canadian Journal of Mathematics* 8, S. 399–404.

Fréchet, M., 1906. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Mathematico di Palermo* 22, S. 1–74.

Girres, J.-F., Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS* 14, S. 435–459.

Goldschmidt, O., Hochbaum, D., 1988. Polynomial algorithm for the k-cut problem. In: *Foundations of Computer Science, 1988., 29th Annual Symposium on*. S. 444–451.

Goldschmidt, O., Hochbaum, D. S., 1994. A Polynomial Algorithm for the k-Cut Problem for Fixed k. *Mathematics of Operations Research* 19 (1), S. 24–37.

Grätzer, G., 1978. General lattice theory. Bd. 75 von Pure and Applied Mathematics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York.

Gross, A., Hartung, M., Kirsten, T., Rahm, E., 2010. On Matching Large Life Science Ontologies in Parallel. In: *Proceedings of the data integration in the life sciences conference*. Bd. 6254. Springer LNCS, S. 35–49.

Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design* 37 (4), S. 682–703.

Hamming, R., 1950. Error Detecting and Error Correcting Codes. *The Bell System Technical Journal* 29 (2), S. 147–160.

Hao, J., Orlin, J., 1994. A Faster Algorithm for Finding the Minimum Cut in a Directed Graph. *Journal of Algorithms* 17 (3), S. 424–446.

Harrie, L., Hellström, A.-K., 1999. A prototype system for propagating updates between cartographic data sets. *The Cartographic Journal* 36 (2), S. 133–140.

Haunert, J.-H., Sester, M., 2008. Area Collapse and Road Centerlines based on Straight Skeletons. *GeoInformatica* 12 (2), S. 169–191.

Haunert, J.-H., Wolff, A., 2016. Handbuch der Geodäsie. Springer Spektrum, Berlin, Heidelberg, Kap. Räumliche Analyse durch kombinatorische Optimierung, S. 1–39.

Hess, G. N., Iochpe, C., Castano, S., 2007. An Algorithm and Implementation for GeoOntologies Alignment. In: Davis, Clodoveu Augusto, J., Monteiro, A. M. V. (Hrsg.), *Advances in Geoinformatics*. Springer Berlin Heidelberg, S. 151–164.

Hopcroft, J. E., Karp, R. M., 1973. An n5/2 Algorithm for Maximum Matchings in Bipartite Graphs. *SIAM J. Comput.* 2 (4), S. 225–231.

Hu, W., Qu, Y., 2006. Block Matching for Ontologies. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L. (Hrsg.), *The Semantic Web - ISWC 2006*. Bd. 4273 von Lecture Notes in Computer Science. Springer Berlin Heidelberg, S. 300–313.

Hu, W., Qu, Y., 2008. Falcon-AO: A practical ontology matching system. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 6(3), S. 237–239.

Hu, W., Qu, Y., Cheng, G., 2008. Matching Large Ontologies: A Divide-and-conquer Approach. *Data Knowl. Eng.* 67 (1), S. 140–160.

Kalfoglou, Y., Schorlemmer, M., 2003. Ontology Mapping: The State of the Art. *The Knowledge Engineering Review* 18 (1), S. 1–31.

Kamidoi, Y., Yoshida, N., Nagamochi, H., 2007. A Deterministic Algorithm for Finding All Minimum k-Way Cuts. *SIAM Journal on Computing* 36 (5), S. 1329–1341.

Karger, D. R., 2000. Minimum Cuts in Near-linear Time. *J. ACM* 47 (1), S. 46–76.

Karger, D. R., Stein, C., 1993. An $O(n^2)$ algorithm for minimum cuts. In: *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*. STOC '93. ACM, New York, NY, USA, S. 757–765.

Karger, D. R., Stein, C., 1996. A New Approach to the Minimum Cut Problem. *J. ACM* 43 (4), S. 601–640.

Kieler, B., Haunert, J.-H., Sester, M., 2009a. Deriving scale-transition matrices from map samples for simulated annealing-based aggregation. *Annals of GIS* 15 (2), S. 107–116.

Kieler, B., Huang, W., Haunert, J.-H., Jiang, J., 2009b. Matching River Datasets of Different Scales. In: Sester, M., Bernard, L., Paelke, V. (Hrsg.), *Advances in GIScience*. Lecture Notes in Geoinformation and Cartography. Springer, S. 135–154.

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

King, V., Rao, S., Tarjan, R., 1994. A Faster Deterministic Maximum Flow Algorithm. *Journal of Algorithms* 17 (3), S. 447–474.

Kokla, M., Kavouras, M., 2005. Semantic Information in Geo-Ontologies: Extraction, Comparison, and Reconciliation. *Journal on Data Semantics III* 3534, S. 125–142.

Koukoletsos, T., Haklay, M. M., Ellul, C., 2012. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS* 16 (4), S. 477–498.

Kuhn, H. W., 1955. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly* 2, S. 83–97.

Kuhn, W., Kauppinen, T., Janowicz, K., 2014. Linked Data - A Paradigm Shift for Geographic Information Science. In: Duckham, M., Pebesma, E., Stewart, K., Frank, A. U. (Hrsg.), *Geographic Information Science*. Springer International Publishing, Cham, S. 173–186.

Lee, Y., Sayyadian, M., Doan, A., Rosenthal, A. S., 2007. eTuner: tuning schema matching software using synthetic scenarios. *VLDB Journal* 16 (1), S. 97–122.

Li, L., Goodchild, M., 2012. Automatically and accurately matching objects in geospatial datasets. In: Shi, W., Goodchild, M., Lees, B., Leung, Y. (Hrsg.), *Advances in Geo-Spatial Information Science*. Bd. 38 von ISPRS Book Series. CRC Press, Taylor & Francis Group, London, UK, S. 71–80.

Li, L., Goodchild, M. F., 2010. Automatically and accurately matching objects in geospatial datasets. In: *Proceedings of theory, data handling and modelling in geospatial information science*. Hong Kong.

Li, L., Goodchild, M. F., 2011. An optimisation model for linear feature matching in geographical data conflation. *Internal Journal of Image and Data Fusion* 2 (4), S. 309–328.

Luan, X., 2012. A structure-based approach for matching road junctions with different coordinate systems. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* I-4, S. 41–46.

Lüscher, P., Burghardt, D., Weibel, R., 2007. Matching road data of scales with an order of magnitude difference. In: *XXIII International Cartographic Conference*. International Cartographic Association, S. online.

Mascret, A., Devogele, T., Berre, I., Hénaff, A., 2006. Coastline matching process based on the discrete Fréchet distance. In: Riedl, A., Kainz, W., Elmes, G. (Hrsg.), *Progress in Spatial Data Handling*. Springer Berlin Heidelberg, S. 383–400.

Miller, G. A., 1995. WordNet: A Lexical Database for English. *Magazine Communications of the ACM* 38 (11), S. 39–41.

Mondzech, J., Sester, M., 2011. Quality Analysis of OpenStreetMap Data Based on Application Needs. *Cartographica: The International Journal for Geographic Information and Geovisualization* 46 (2), S. 115–125.

Munkres, J., 1957. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics* 5 (1), S. 32–38.

Mustière, S., Devogele, T., 2008. Matching Networks with Different Levels of Detail. *GeoInformatica* 12 (4), S. 435–453.

Nagamochi, H., Ibaraki, T., 1992. Computing Edge-connectivity in Multigraphs and Capacitated Graphs. *SIAM Journal on Discrete Mathematics* 5 (1), S. 54–66.

Nagamochi, H., Katayama, S., Ibaraki, T., 1999. A Faster Algorithm for Computing Minimum 5-Way and 6-Way Cuts in Graphs. In: Asano, T., Imai, H., Lee, D., Nakano, S.-i., Tokuyama, T. (Hrsg.), *Computing and Combinatorics*. Bd. 1627 von Lecture Notes in Computer Science. Springer Berlin Heidelberg, S. 164–173.

Neis, P., Zielstra, D., Zipf, A., 2012. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011. *Future Internet* 4, S. 1–21.

Nocedal, J., Wright, S., 2006. Numerical Optimization. Springer Series in Operations Research and Financial Engineering. Second Edition, Berlin.

Noy, N. F., Musen, M. A., 2003. The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping. *International Journal of Human-Computer Studies* 59 (6), S. 983–1024.

OpenStreetMap, 2019. DE:Map Features - OpenStreetMap Wiki. `http://wiki.openstreetmap.org/wiki/DE:Map_Features` (Letzter Zugriff: 14.05.2019).

Ottmann, T., Widmayer, P. (Hrsg.), 2002. Algorithmen und Datenstrukturen, 4. Auflage. Spektrum Akademischer Verlag.

Papadimitriou, C. H., Steiglitz, K., 1982. Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall, Inc.

Paulheim, H., 2008. On Applying Matching Tools to Large-scale Ontologies. In: *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008) Collocated with the 7th International Semantic Web Conference (ISWC-2008)*. Bd. 431 von CEUR Workshop Proceedings. CEUR-WS.org.

Penninga, F., Verbree, E., Quak, W., van Oosterom, P., 2005. Construction of the Planar Partition Postal Code Map Based on Cadastral Registration. *GeoInformatica* 9 (2), S. 181–204.

Peukert, E., Berthold, H., Rahm, E., 2010. Rewrite Techniques for Performance Optimization of Schema Matching Processes. In: *Proceedings of the 13th International Conference on Extending Database Technology*. EDBT '10. ACM, New York, NY, USA, S. 453–464.

Poblet, M., Casanovas, P., Rodríguez-Doncel, V., 2019. Introduction to Linked Data. Springer International Publishing, Cham, S. 1–25.

Rahm, E., Bernstein, P. A., 2001. A survey of approaches to automatic schema matching. *VLDB Journal* 10 (4), S. 334–350.

Rahm, E., Peukert, E., 2018a. Holistic Schema Matching. Springer International Publishing, S. 1–5.

Rahm, E., Peukert, E., 2018b. Large-Scale Schema Matching. Springer International Publishing, S. 1–6.

Saleem, K., Bellahsene, Z., Hunt, E., 2008. PORSCHE: Performance ORiented SCHEma mediation. *Information Systems* 33 (7-8), S. 637–657.

Schulze, M., Thiemann, F., Sester, M., 2014. Using Semantic Distance To Support Geometric Harmonization Of Cadastral And Topographical Data. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-2, S. 15–22.

Shvaiko, P., Euzenat, J., 2005. A survey of schema-based matching approaches. *Journal on Data Semantics* 4, S. 146–171.

Smiljanic, M., van Keulen, M., Jonker, W., 2006. Using Element Clustering to Increase the Efficiency of XML Schema Matching. In: *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. S. 45.

Stoer, M., Wagner, F., 1997. A simple min-cut algorithm. *J. ACM* 44 (4), S. 585–591.

Thorup, M., 2008. Minimum k-way cuts via deterministic greedy tree packing. In: *Proceedings of the 40th annual ACM symposium on Theory of computing*. STOC '08. ACM, New York, NY, USA, S. 159–166.

Tversky, A., 1977. Features of similarity. *Psychological Review* 84, S. 327–352.

Uitermark, H., Vogels, A., Oosterom, P. v., 1999. Semantic and Geometric Aspects of Integrating Road Networks. In: *Proceedings of the Second International Conference on Interoperating Geographic Information Systems*. INTEROP '99. London, UK, UK, S. 177–188.

van Wijngaarden, F., van Putten, J., van Oosterom, P., Uitermark, H., 1997. Map integration - update propagation in a multi-source environment. In: *Proceedings of the 5th ACM international workshop on Advances in geographic information systems*. GIS '97. Las Vegas, Nevada, United States, S. 71–76.

Vivid Solutions, 2005. RoadMatcher User Guide - RoadMatcher Version 1.4. `http://www.vividsolutions.com/products.asp?catg=spaapp&code=roadmatcher` (Letzter Zugriff: 10.07.2014).

Volz, S., 2006. Modellierung und Nutzung von Relationen zwischen Mehrfachrepräsentationen in Geo-Informationssystemen. Dissertation, Universität Stuttgart, Holzgartenstr. 16, 70174 Stuttgart.

Walter, V., 1997. Zuordnung von raumbezogenen Daten - am Beispiel von ATKIS und GDF. Dissertation, Deutsche Geodätische Kommission, Reihe C, Nr. 480, München.

Walter, V., Fritsch, D., 1999. Matchting spatial data sets: a statistical approach. *Internal Journal Geographical Information Science* 13, S. 445–473.

Werder, S., Kieler, B., Sester, M., 2010. Semi-automatic Interpretation of Buildings and Settlement Areas in User-generated Spatial Data. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, New York, NY, USA, S. 330–339.

Yang, B., Zhang, Y., Luan, X., 2013. A probabilistic relaxation approach for matching road networks. *International Journal of Geographical Information Science* 27 (2), S. 319–338.

Yuan, S., Tao, D. C., 1999. Development of Conflation Components. In: *In Proceedings of Geoinformatics, Ann Arbor*. S. 1–13.

Zhang, M., 2009. Methods and Implementations of Road-Network Matching. Dissertation, Technische Universität München.

Zhang, M., Meng, L., 2007. An iterative road-matching approach for the integration of postal data. *Computers, Environment and Urban Systems* 31 (5), S. 597–615.

Zhang, X., Ai, T., Stoter, J., Zhao, X., 2014. Data matching of building polygons at multiple map scales improved by contextual information and relaxation. *ISPRS Journal of Photogrammetry and Remote Sensing* 92, S. 147– 163.

Zielstra, D., Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of 13th AGILE International Conference on Geographic Information Science*. Guimaraes, Portugal.

Danksagung

Mein größter Dank gilt Frau Prof. Dr.-Ing. habil. Monika Sester. Sie ermöglichte mir die Mitarbeit in dem von der Deutschen Forschungsgemeinschaft geförderten Forschungsprojekt „Automatic semantic transformation between Geo-Ontologies“ des deutsch-chinesischen Bündelprojekts „Interoperation of 3D Urban Geoinformation“. Ihre Ideen und fachlichen Anregungen haben meine Arbeit stets vorangebracht. Sie förderte mich persönlich und fachlich, indem sie mir die Teilnahme an Auslandsaufenthalten und Konferenzen ermöglichte sowie Aufgaben in der Lehre und im Institutsalltag übertrug. Selbst nach Ablauf des Forschungsprojekts hat sie mir mit ihrem unerschütterlichen Optimismus immer das Gefühl gegeben, an mich zu glauben und mich unterstützt, die Arbeit auch nach langer Zeit abzuschließen.

Vielen Dank den Korreferenten Herrn Prof. Dr.-Ing. habil. Jan-Henrik Haunert und Herrn Prof. Dr.-Ing. habil. Christian Heipke für die freundliche Übernahme des Korreferats. Mein besonderer Dank gilt Herrn Haunert, der mir in vielen fachlichen Diskussionen geholfen hat, eine Lösung für meine Problemstellung mit Hilfe der Linearen Programmierung zu finden. Des Weiteren danke ich ihm für die sehr ausführlichen Korrekturen der ersten Entwürfe dieser Arbeit.

Mein Dank gilt auch allen Kolleginnen und Kollegen des Instituts, die mich stets fachlich, persönlich und moralisch unterstützt haben. Daniel Eggert danke ich für die sehr geduldige Unterstützung bei der Java-Programmierung. Ich danke Nora Meyer-Spradow und Christoph Dold für das tolle Arbeitsklima im gemeinsamen Büro, die stets selbstverständliche gegenseitige Hilfe und die gemeinsamen freundschaftlichen Aktivitäten. Ein großes Dankeschön gilt zwei ganz besonderen Menschen – Sabine Hofmann und Karen-Insa Wolf. Während meiner Zeit an der Universität sind sie zu wahren Freundinnen geworden, da sie mich in jeglicher Hinsicht motiviert und in meinem Können bestärkt haben. Ihr seid einfach wunderbar und echte Vorbilder!

Von Herzen danke ich meiner ganzen Familie, die während der unzähligen Schreibwochenenden die Betreuung meines Sohns übernommen hat! Ganz besonders danke ich meinem Freund Fabian Dahms und meiner Schwester Anett Kieler. Sie haben mich immer liebevoll unterstützt, umsorgt, aufgebaut und immer wieder angetrieben. Mein größtes Glück ist aber mein Sohn Enno Matthis.

Das Zitat von Marie Curie fasst die Situation des Schreibens dieser Arbeit sehr gut zusammen.

> „Man merkt nie, was schon getan wurde, man sieht immer nur, was noch zu tun ist.“

Lebenslauf

Persönliche Daten

Birgit Kieler
geboren am 24. Januar 1979 in Berlin

Schulische Ausbildung

1985 – 1990	Georg-Krausz-Oberschule, Berlin
1990 – 1991	2. Polytechnische Oberschule „Erich Baron“, Berlin
1991 – 1998	Friedrich-List-Gymnasium, Berlin
11. Juni 1998	Allgemeine Hochschulreife

Berufliche Ausbildung

09/1998 – 02/2001	Ausbildung zur Vermessungstechnikerin, Bezirksamt Charlottenburg-Wilmersdorf, Berlin

Studium

10/2001 – 06/2006	Diplomstudiengang Geodäsie und Geoinformatik, Leibniz Universität Hannover
06. Juni 2006	Abschluss als Diplom-Ingenieurin

Berufserfahrung

02/2001 – 08/2001	Vermessungstechnikerin, Bezirksamt Charlottenburg-Wilmersdorf, Berlin
06/2006 – 08/2013	Wissenschaftliche Mitarbeiterin, Institut für Kartographie und Geoinformatik, Leibniz Universität Hannover
02/2014 – 05/2014	Wissenschaftliche Mitarbeiterin, Institut für Kartographie und Geoinformatik, Leibniz Universität Hannover
seit 01/2016	Angestellte, Senatsverwaltung für Stadtentwicklung und Wohnen, Berlin Abteilung Geoinformation, Referat Geodateninfrastruktur