# Verifikation der Landnutzung durch Ähnlichkeitsanalyse der Landbedeckung gleicher Geodatenobjekte

Von der Fakultät für Elektrotechnik und Informatik

der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

**Doktor-Ingenieur**

genehmigte

**Dissertation**

von

**Dipl.–Math. Christian Becker**

geboren am 23. April 1978 in Hannover

**2013**

| | |
|---|---|
| Referent: | Prof. Dr.–Ing. Ostermann |
| Korreferent: | Prof. Dr.–Ing. Heipke |
| Tag der Promotion: | 16. August 2013 |

# Danksagung

Die vorliegende Dissertation entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Informationsverarbeitung (TNT) der Leibniz Universität Hannover.

Mein besonderer Dank gilt meinem Doktorvater Professor Dr.-Ing. Ostermann für die Betreuung der Arbeit, Professor Dr.-Ing. Heipke für die Übernahme des Korefferates und Prof. Dr.-Ing Rosenhahn für den Vorsitz. Ich danke Prof. Dr.-Ing. Liedtke, mich ans TNT geholt zu haben und für die vielen inspirierenden Diskussionen auch über den Beginn seines Ruhestandes hinaus.

Ich danke für die Möglichkeit am WiPKA-QS-Forschungsprojektes, gefördert durch das Bundesamt für Kartographie und Geodäsie, mitzuarbeiten. Die Arbeit erlaubte mir wertvolle Einblicke in die praktischen Fragestellungen der Prüfung von Geodaten. Sie inspirierten mich zu der ursprünglichen Idee für diese Arbeit.

Durch das Projekt wurde ich Teil des WiPKA-*Teams*, bestehend aus Mitarbeitern des TNT und des Institut für Photogrammetrie und Geoinformatik (IPI). Ich wünsche allen wissenschaftlichen Mitarbeitern solche Erfahrungen zu machen, wie ich sie mit Torsten Büschenfeld, Uwe Breitkopf, Markus Gehrke, Petra Helmholz, Sönke Müller, Martin Pahl, Karsten Vogt und Marcel Ziems machen konnte.

Die Gemeinschaft der Mitarbeiter am TNT sucht seines Gleichen. Die guten persönlichen Kontakte sind die Basis für einen intensiven interdisziplinären Austausch. Ich danke Torsten Büschenfeld für die tolle gemeinsame Zeit des fast täglichem Umgangs, fachlicher Diskussionen und erlebter Abenteuer. Martin Pahl danke ich für die Betreuung des WiPKA Projekts und dass ich ihm so oft über die Schulter gucken durfte. Sven Klomp, Stephan Preihs, Minh Nguyen, Tobias Elbrandt, Oliver Müller, Karsten Vogt, Yuri Vatis, Holger Meuel und Julia Schmidt (Liste nicht abschließend) danke ich für die vielen Diskussionen und den interessanten Austausch. Ich danke Matthias Schuh, Doris Jaspers-Göring, Silvia Scholl, Ursula Kemmner, Pia Bank und Hilke Brodersen, das TNT am Laufen zu halten und für die moralische und technische Unterstützung.

Trotz dieses tollen Umfelds war die Fertigstellung der Arbeit eine große Kraftanstrengung. Nicht zuletzt danke ich daher meiner Familie und meinen Freunden, die mich mich in dieser Zeit gleichzeitig bestärkt wie auch einen Gegenpol gebildet haben.

# Inhaltsverzeichnis

# Kurzfassung

Siedlungsflächen, Straßen und Flüsse der realen Welt werden als Geodatenobjekte in Geoinformationssystemen (GIS) erfasst. Geodaten sind, etwa aufbereitet als Karten, die Informationsgrundlage für viele wissenschaftliche und wirtschaftliche Fragestellungen. Eine regelmäßige Überprüfung von digitalen Geodaten ist daher entscheidend für deren Werterhaltung. Da Veränderungen der realen Welt meist vereinzelt und lokal begrenzt auftreten, ist ein wichtiger Schritt zur Automatisierung der Aktualisierung von Geodaten die Detektion dieser Bereiche.

Als Quelle für Informationen über den Zustand der realen Welt werden u. a. Satelliten- und Luftaufnahmen eingesetzt. Zwar existieren Verfahren, die aus den Fernerkundungsdaten die Landbedeckung für jeden Punkt der Oberfläche ermitteln können, Objekte werden aber meist über eine einheitliche Landnutzung definiert (die Objektklasse). Objekte solcher Objektklassen weisen üblicherweise Kombinationen von Landbedeckungen auf. Regelbasierte Verfahren zur Geodatenprüfung streben an, das Vorgehen menschlicher Experten durch die automatische Prüfung manuell formulierter Regeln nachzuahmen. Die komplexe und schwer zu überblickende manuelle Formulierung der Regelsätze geht aber zu Lasten der Qualität der Ergebnisse. Systeme ohne Regeln wurden bisher wenig untersucht und bieten methodisch starke Einschränkungen.

Das Ziel der vorliegenden Arbeit war die Entwicklung eines grundsätzlich neuen Ansatzes, um die Erkennung von fehlerhaften Geodatenobjekten ohne eine explizite Steuerung durch einen Bearbeiter zu ermöglichen. Das Verfahren betrachtet die Gesamtheit der potenziell veralteten Geodatenobjekte einer Szene und auf eine Landbedeckung voruntersuchte Fernerkundungsdaten. Die Anteile an Landbedeckungen werden für jedes Geodatenobjekt als Merkmal erfasst und die Geodatenobjekte einer Objektklasse so in einen gemeinsamen Merkmalsraum eingeordnet. Über die Mahalanobisdistanz wird für jedes Geodatenobjekt dessen Abstand zu einem durchschnittlichen Objekt ermittelt. Dieser als Abnormitätswert interpretierte Abstand sagt aus, inwieweit das Geodatenobjekt typisch für die Objektklasse ist. Indem nur die ungewöhnlichsten Objekte einer Objektklasse einer Nachprüfung unterzogen werden, wird der Prüfaufwand reduziert.

Das Verfahren wird in zwei umfangreichen Testszenen auf IKONOS Satellitendaten und deutschen ATKIS Geodaten getestet. Die Ergebnisse zeigen, dass das Verfahren trotz unterschiedlicher Testszenen und Schwankungen in der Landbedeckungsanalyse für unterschiedliche Objektklassen stabile Ergebnisse erzielt. Das neue Verfahren ist geeignet, den Prüfaufwand um 80 % zu reduzieren und dabei 80 % to 90 % der fehlerhaften Geodatenobjekte zu detektieren. Die Modellierung des Merkmalsraums erlaubt diese Ergebnisse selbst dann, wenn nur 50 % der Geodaten noch nicht veraltet sind.

**Stichworte** Verifikation, 2D-Vektordaten, GIS, Data Mining

# Abstract

Settlement areas, roads and rivers from the real world are digitally captured as polygons, lines or points and stored in geographical information systems (GIS) as *objects*. Content and structure of geographical data is usually defined by specifications that are traditionally implemented by human experts. Geo data is the basis for numerous applications. It provides data for maps as well as for analyzing economical, ecological and social patterns considering geographical relations. Thus, frequently updating the data to keep up with changes in the real world is crucial.

Changes in the real world often only affect a few local areas. Therefore, systems to locate areas of change are of high research interest. Satellite and airborne imagery is the main source of information about the state of the real world. Algorithms have been developed to analyze for land cover in the remote sensing data. Objects are more often describing areas of land use, though (specified as GIS feature). In this case, combinations of different land cover can be found in objects. Rule based systems allow to introduce expert knowledge to define which land cover combinations are to be expected in correct GIS objects. In practice, formulating the set of rules is complicated and updating the rules to new data and scenes is tedious. Systems with a higher level of automation are less researched and conceptually limited.

The system described in this thesis approaches the task in a fundamentally new way. Its aim is to determine the rate of conformance with the assigned GIS feature. Thus, partially incorrect objects may be detected as errors. Furthermore, GIS features are evaluated independently. The approach is implemented by analyzing land cover from remote sensing images and describing combinations of land cover in an object as attributes. Putting all attributes with a common GIS feature into an attribute space, a statistical model is used to determine the Mahalanobis distance of an object to an average object. The distance is interpreted as abnormality rating. Restricting further investigations to objects with high abnormality, the efficiency for updating geo data is increased.

Experiments on German ATKIS data and IKONOS satellite images for two extensive test scenes show that the new approach is able to successfully cope with various GIS features. Increasing the efficiency for analyzing geo data for out-dated areas by factor 5 while still detecting 80 to 90 % of incorrect objects compared to a manual check of the data is realistic. The approach is very robust. It performs well as long as the data is still at least 50 % correct.

**Keywords** verification, 2D vector data, GIS, data mining

# 1 Einleitung

## 1.1 Geodaten: Planungsgrundlage in einer komplexen und digitalen Welt

Kartendarstellungen sind seit jeher eine wichtige Grundlage zur Planung von Vorgehen und Abläufen. Schließlich erlauben Karten es, Gegebenheiten in einem räumlichen Kontext darzustellen. Durch digitale Kartendienste wie Google Maps[1], Microsoft Bing[2] oder OpenStreetMap[3] wird der Zugang zu solchen Daten vereinfacht. Die wirtschaftliche Bedeutung von Kartendaten steigt daher beständig an [15].

Karten sind dabei nur eine der Darstellungsmöglichkeiten von darstellungsunabhängigen *Geodaten.* Sie werden in Vektorformaten digital gespeichert und in Datenbanken, *Geoinformationssysteme* (GIS) genannt, verwaltet. Verschiedene GIS-Softwarelösungen unterstützen die Erzeugung, Visualisierung und Analyse von Geodaten [53].

Beispiele für die Nutzung von Geodaten als Planungsgrundlagen sind die Bestimmung des Grads der Flächenversiegelung [12, 14], die wiederum für den Hochwasserschutz [12] oder die Abwassergebührenberechnung [48] maßgeblich ist. Auch die Überwachung von Waldflächen, um die Rodung von Wäldern zu verfolgen [34], beruht auf Geodaten. Räumliche Analysen auf der Basis von Geodaten gehören zudem zu den grundlegenden Methoden von Biologen, Sozialforschern und weiteren Forschungsgebieten [51].

Allgemein spricht man auch bei einigen dieser Anwendungen von LULC-Anwendungen (engl. *Land Use/Land Cover* für *Oberflächennutzung/Oberflächenbedeckung*). Unter *Oberflächenbedeckung* versteht man dabei eine direkte Beschreibung der Erdoberfläche. So ist z.B. eine Unterscheidung in Flächen (genannt *Geodatenobjekte*) der *Objektklassen* wie Vegetation, Wasser und versiegelte Flächen denkbar. Dagegen beschreibt die *Oberflächennutzung* die Nutzung der Oberfläche durch den Menschen, also nun z.B. Objektklassen für Ackerland, Badesee oder Tagebau.

Es gibt Vorhaben, Geodaten für eine anwendungsübergreifende Nutzung von Geodaten vorzuhalten. In der Europäischen Union wurde hierfür das *Coordinated Information on the European Environment* Projekt (CORINE) aufgelegt [5]. In Deutschland dient das digitale Landschaftsmodell des *Amtlichen Topographisch-Kartographischen Informationssystem* (ATKIS) ähnlichen Zwecken [3, 46].

**Geodaten werden aufgrund von festgelegten Vorgaben erfasst** Um eine konsistente Erfassung der Geodaten sicher zu stellen, werden üblicherweise *Objektartenkataloge* festgelegt, in denen Standards für die Erfassung der Geodatenobjekte aufgeführt sind. Die Standardisierung umfasst dabei nicht nur die Eingrenzung auf bestimmte Objektklassen,

[1] http://www.maps.google.com
[2] http://www.bing.com/maps
[3] http://www.openstreetmap.org

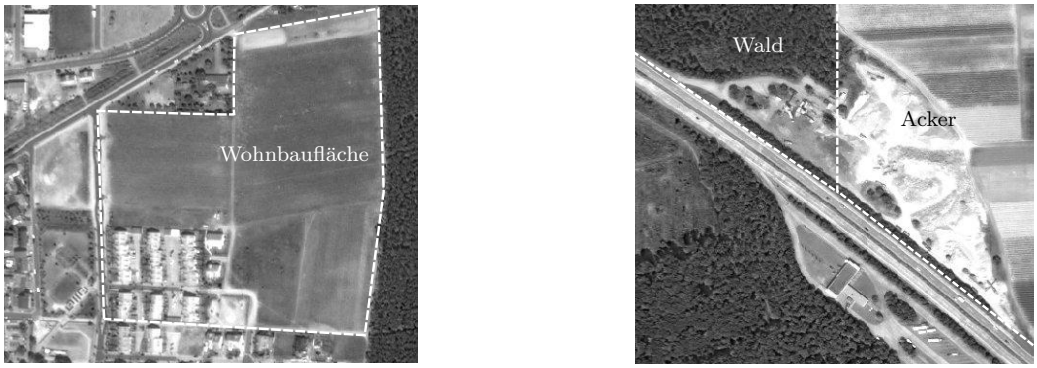



**Abbildung 1.1:** Geodaten-Fehler durch falsche Datengrundlage (links) und Bautätigkeit (rechts)

sondern auch Richtlinien, die die räumliche Genauigkeit der Erfassung festlegen. So kann beispielsweise eine *Mindesterfassungsgröße* definiert werden, um sicher zu stellen, dass nur Geodatenobjekte einer relevanten Größe festgehalten werden. Auch ist eine Beschränkung der Genauigkeit üblich, um den Erstellungsaufwand, bei gleichzeitig verlässlicher Genauigkeit, zu reduzieren. Beispiele aus dem Objektartenkatalog für die ATKIS-Geodaten für einen Maßstab 1:25 000 sind eine Mindesterfassungsfläche von 1 ha für Geodatenobjekte der Objektklasse *Grünland* und eine Positionsgenauigkeit von 3 m. Erfüllt eine Fläche nicht die Mindesterfassungsanforderungen, so wird sie einem ohnehin vorhandenen benachbarten Geodatenobjekt mit einer möglichst gut passenden Objektklasse zugewiesen.

Geodaten beschreiben im Optimalfall den Zustand einer Szene zum Zeitpunkt der Erfassung. Durch ungeeignete Quellen oder Bearbeitungsfehler kann es jedoch bereits bei der Erfassung zu Fehlern in den Geodaten kommen. Mit zeitlichem Abstand zum Erfassungsdatum veralten Teile der Geodaten zunehmend durch Veränderungen in der Szene (siehe Abbildung 1.1).

Die regelmäßige Prüfung von Geodaten ist also unumgänglich. Eine solche Prüfung ist allerdings sehr aufwändig. Laut [32] betrug die jährliche Aktualisierungsrate von Geodaten in Mitteleuropa gerade einmal 6-8% der Fläche. Einen maßgeblichen Teil des Aufwandes für die Fortführung von Geodaten stellt dabei die Lokalisierung der Abweichungen in der Szene dar [33].

**Fernerkundungsdaten sind Grundlage zur Erfassung und Fortführung von Geodaten**
Durch die Verfügbarkeit umfangreicher *Fernerkundungsdaten* von bildgebenden Satelliten und Flugzeugaufnahmen können Szenen effizient und in großem Umfang überwacht werden [27, 31]. Ein Überblick, wie tatsächliche Gegebenheiten, Fernerkundungsdaten und Geodaten zusammenhängen, wird in Abbildung 1.2 gezeigt: Die *Wirklichkeit* entspricht einem vollständigen Informationsgehalt der Gegebenheiten, die unabhängig von einer Darstellungsform oder Auffassung als abstraktes Konzept existiert. Um Erkenntnisse über die Gegebenheiten in der Wirklichkeit zu erhalten, muss die Szene durch Sensoren abgetastet werden. Diese können ein menschlicher Beobachter oder aber auch Sensoren auf Satelliten oder Flugzeugen sein, die etwa Fotos von der Erdoberfläche aufnehmen. Solche *Fernerkundungsdaten* liegen als Rasterbilder vor, bei denen ein engmaschiges Raster über die Oberfläche gelegt wird und die Sensormessungen für jedes Rasterelement (*Pixel*) dargestellt sind.

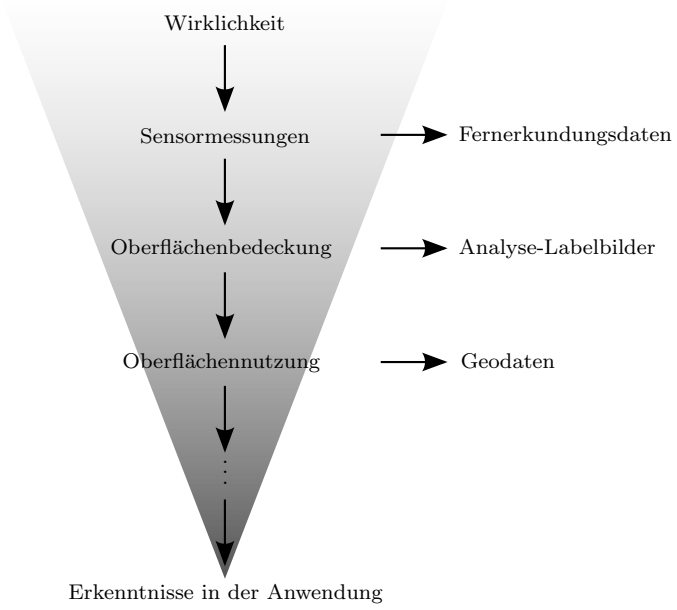



**Abbildung 1.2:** Abstraktionsebenen der Beschreibung einer Szene und ihre Datenrepräsentationen. (Darstellung entlehnt aus [17, Seite 3])

Die Durchsicht und Verarbeitung der Fernerkundungsdaten ist sehr arbeitsintensiv [20, 53]. Eine Automatisierung der Auswertung ist daher ein wichtiger Schritt zur effektiven Aktualisierung von Geodaten.

Während ein Mensch bei der Durchsicht der Fernerkundungsdaten scheinbar unwillkürlich eine Vorstellung darüber entwickelt, was in den Daten zu sehen ist, müssen zur automatischen Erkennung zunächst Algorithmen angewendet werden. Welche Ergebnisse eine Analyse liefert, wird daher von der Wirklichkeit, den Eigenschaften der Sensoren und der Analysemethode bestimmt. Die Richtlinien zur Erfassung der Geodaten wie die Berücksichtigung von Mindesterfassungsgrößen führen dazu, dass Geodaten weiterhin von Analyseergebnissen abweichen können.

In dieser Arbeit stehen flächenhafte Geodatenobjekte im Fokus. Nicht betrachtet werden linienhafte Geodatenobjekte, wie sie für die Erfassung von Wegen, Straßen und Flüssen üblich sind. Die existierenden Ansätze für diese Objektarten prüfen etwa für durch Geodaten vorgegebene Straßenhypothesen die Existenz von maßgeblichen Strukturen [41, 52, 54]. Auch das Verfolgen von Netzstrukturen (z.B. Berücksichtigung von Straßenkreuzungen bei Straßen [16]) oder Effekte in der Umgebung der Geodatenobjekte (Erkennung von straßenbegleitenden Baumreihen [19]) werden genutzt.

Der in dieser Arbeit vorgestellte Ansatz markiert als Ergebnis Geodatenobjekte, die nach der Analyse wahrscheinlich fehlerhaft sind. Diese können dann wie gehabt von Experten mit Hilfe der existierenden Verarbeitungswerkzeuge und unter der Hinzunahme weiterer Daten genauer untersucht und ggf. korrigiert werden.

## 1.2 Stand der Technik

Das Prüfen von Geodaten kann auf verschiedenen Ebenen erfolgen. Unterschieden werden häufig Prüfungen auf

- das Vorhandensein der Geodatenobjekte
- die Genauigkeit der Lage der Geodatenobjekte
- die Korrektheit der zugewiesenen Objektklasse.

Verschiedene Herangehensweisen wurden in der Literatur bereits vorgestellt. Alle Ansätze verfolgen das Ziel, Änderungen in Geodaten räumlich aufzuzeigen, um als Hilfestellung für die eigentliche Aktualisierung zu dienen. Stets werden Fernerkundungsdaten genutzt, um Informationen über Änderungen zu erhalten. Jedoch gibt es bei der Art der Berücksichtigung der Fernerkundungsdaten konzeptionelle Unterschiede.

### 1.2.1 Pixelweise Objektklassenprüfung

Etliche Ansätze konzentrieren sich darauf, Objektklassen direkt aus den Fernerkundungsdaten zu erkennen. Jedem Pixel der Fernerkundungsdaten wird dazu eine Objektklasse zugewiesen. Schließlich ist ein pixelweise durchgeführter Abgleich mit bestehenden Geodaten möglich [42].

Genutzt werden pixelweise Klassifikationen der Fernerkundungsdaten, bei denen jedem Pixel – eine lokale Pixelumgebung berücksichtigend – eine Objektklasse zugewiesen wird [29]. Alternativ werden zunächst sich ähnelnde, benachbarte Pixel zu Segmenten zusammengefasst, um anschließend den Segmenten als ganzes Objektklassen zuzuweisen [43].

In einigen Fällen bestehen die Fernerkundungsdaten auch aus einer Zeitreihenaufnahme, bei der dieselbe Szene zu mehreren Zeitpunkten aufgenommen wird. Die Zeitreihe erlaubt es, charakteristische Szenenänderungen wie Vegetationsperioden auszuwerten [21]. Allerdings wird dann die Szene für diesen Zeitraum als inhaltlich unveränderlich angenommen, sodass die Ansätze dem generellen Vorgehen folgen.

Da die Objektklasse ohne Betrachtung der Geodatenobjekte für jeden Pixel individuell bestimmt wird, lassen sich auch kleine und lokale Änderungen gegenüber den Geodaten feststellen. Die Zuverlässigkeit der Objektklassenbestimmung ist somit maßgeblich für die Anwendbarkeit des Ansatzes.

Die in den zitierten Arbeiten behandelten Objektklassen haben gemein, dass sich die Objektklasse eines Pixels aus einer eng begrenzten lokalen räumlichen Nachbarschaft herleiten lässt. Sie entsprechen daher eher Oberflächenbedeckungen als Oberflächennutzungen. Wie das Beispiel in Abbildung 1.3 zeigt, besteht ein Geodatenobjekt der ATKIS-Objektklasse *Industrie- und Gewerbefläche* aber aus einer Vielzahl von Oberflächenbedeckungsflächen. Es ist die Vielfalt und räumliche Ausprägung, die die gezeigte Fläche zu einer Industrie- und Gewerbefläche machen. Die einzelnen Oberflächenbedeckungsflächen treten dazu meist in verschiedenen Oberflächennutzungsklassen auf, sodass das alleinige Vorhandensein nicht auf die Objektklasse schließen lässt. Erst bei Berücksichtigung dessen, welche Oberflächenbedeckungsflächen im räumlichen Kontext auftreten, ist eine solche Entscheidung möglich. Die Ergebnisse lokaler Verfahren sind also in der Lage, wichtige Hinweise auf das Vorhandensein von Oberflächennutzungsobjektklassen zu liefern, nicht aber eine abstrakte Oberflächennutzung.

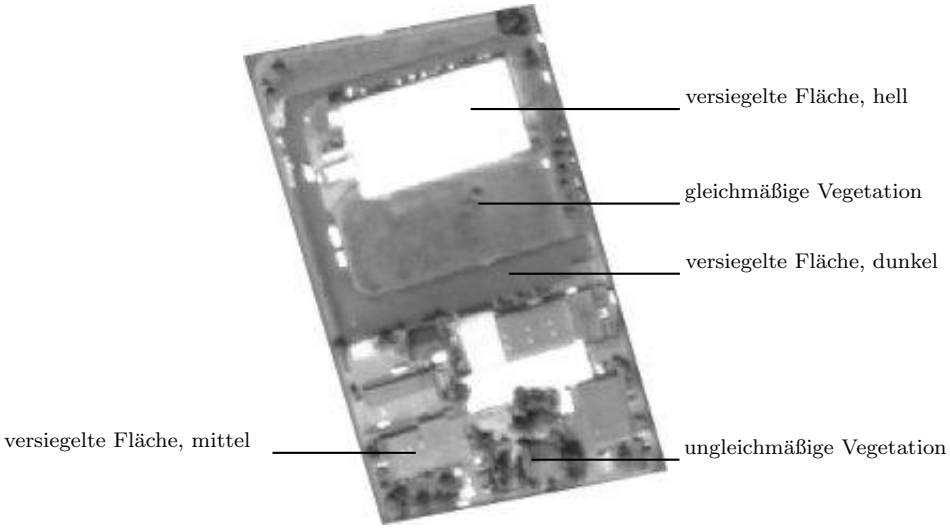



**Abbildung 1.3:** Die Oberflächennutzung *Gewerbegebiet* besteht aus einer Vielzahl von Oberflächenbedeckungsklassen.

### 1.2.2 Objektbezogene regelbasierte Überprüfung

Das an der Leibniz Universität Hannover von dem Institut für Informationsverarbeitung (TNT) und dem Institut für Photogrammetrie und GeoInformation (IPI) entwickelte System WiPKA nutzt die zu prüfenden Geodaten als Vorinformation. Flächen bekannter (und möglicherweise veralteter) Geodatenobjekte werden unabhängig voneinander in den Fernerkundungsdaten auf das Vorhandensein von Oberflächenbedeckung analysiert. Eine Besonderheit des Systems ist die Möglichkeit, nahezu beliebige Algorithmen zur Analyse der Fernerkundungsdaten integrieren zu können. Durch die hierarchische Analysesoftware GeoAIDA lassen sich Analysealgorithmen kombinieren und so auch komplexe Szenen auf die Oberflächenbedeckung hin untersuchen. Vorgestellt wurden Ansätze zur Gebäudeextraktion [38], Acker-/Grünlandunterscheidung [24], für Plantagen [49] und allgemeine Texturklassifikationsverfahren [4, 8].

Um Geodatenfehler zu ermitteln, werden in WiPKA für jedes Geodatenobjekt die Anteile der unterschiedlichen Oberflächenbedeckungsflächen im Verhältnis zur Gesamtgröße ermittelt und über manuell formulierte Regeln mit Sollwerten verglichen [26].

Die objektbezogene Betrachtung erlaubt es sowohl, lokale Erkenntnisse wie die Oberflächenbedeckung im Kontext des (möglicherweise veralteten) Geodatenobjektes zu bewerten als auch die Objektklasse bei der Bewertung zu berücksichtigen. Die manuelle Formulierung von Sollwerten und Regeln zur Bewertung der Geodaten erlaubt dabei große Freiheiten, stellt aber hohe Anforderungen an deren Erstellung: Von jeder Objektklasse muss bekannt sein, aus welchen Landbedeckungsflächen sich Geodatenobjekte dieser Klasse zusammensetzen. Dabei ist zu beachten, dass ein Mensch nur mit viel Erfahrung und technischem Sachverstand einschätzen kann, welche Ergebnisse ein Algorithmus für einen spezifischen Satz an Fernerkundungsdaten erzeugt. Die Zusammensetzung der Geodatenobjekte kann sich abhängig von der geographischen Lage und damit einhergehenden Unterschieden von Szene zu Szene unterscheiden, sodass die Konfiguration von erfahrenen Spezialisten kontinuierlich angepasst werden muss.

### 1.2.3 Objektbezogene Überprüfung durch Reklassifikation

Ein von Walter [50] entwickeltes System bestimmt für jedes Geodatenobjekt die Objektklasse neu. Entspricht die Neubestimmung nicht der ursprünglichen Objektklasse, werden die betroffenen Geodatenobjekte als fehlerhaft betrachtet.

Auch dieses System nutzt die zu überprüfenden Geodaten als Vorinformation: Zunächst dienen die möglicherweise veralteten, flächenhaften Geodatenobjekte der Einteilung der Szene in einzelne Flächen. Diese werden unabhängig voneinander in den Fernerkundungsdaten pixelweise auf das Vorhandensein von vordefinierten Oberflächenbedeckungen analysiert. Die Zusammensetzung eines Geodatenobjektes aus Oberflächenbedeckungen wird über den Anteil einer jeden möglichen Oberflächenbedeckung im Verhältnis zur Gesamtfläche des Geodatenobjekts beschrieben. Darüber hinaus werden für jedes Geodatenobjekt Mittelwert und Varianz in den einzelnen Kanälen der Satellitenbilder bestimmt und hieraus schließlich objektweise ein Merkmal formuliert.

Es wird nun das Merkmal jedes Geodatenobjektes betrachtet und dem Geodatenobjekt diejenige Objektklasse zugewiesen, zu der die Merkmale am ehesten „passen“ (*Reklassifikation*). Weicht die neue Objektklasse von der alten ab, so gilt das Geodatenobjekt als fehlerhaft. Die Parameter der Reklassifikationsentscheidung werden bei Walter auf Basis der zu prüfenden Geodaten automatisch bestimmt. Da die Geodaten teilweise veraltet sind, werden sie nicht direkt für einen Vergleich herangezogen, sondern es wird auf ihnen aufbauend ein stochastisches *Modell* gebildet, sodass für jedes Geodatenobjekt die plausibelste Objektklasse bestimmt werden kann.

Auch bei diesem Verfahren erlaubt die objektbezogene Betrachtung lokale Erkenntnisse wie die Oberflächenbedeckung und die Sensorwerte der Fernerkundungsdaten, im Kontext des Geodatenobjekt zu betrachten. Das automatische Training ersetzt die manuelle Regelformulierung und die Festsetzung von Sollwerten. Allerdings führt die Erkennung einer Fehlerfläche aufgrund der Reklassifikationsentscheidung zu starken Einschränkungen:

- Es muss stets eine geeignete Menge an Objektklassen geprüft werden, denn die Reklassifikation basiert auf einem Vergleich zwischen Objektklassen.
- Die Anzahl der Objektklassen darf nicht zu groß werden, da ansonsten die Objektklassen nicht mehr klar unterschieden werden können.
- Geodatenobjekte, die nur teilweise falsch sind, entsprechen weder der einen noch einer anderen Objektklasse. Daher können nur großflächige Änderungen erkannt werden.

Die Automatisierung ist ein großer Vorteil gegenüber der Fehlerauswertung des WiPKA-Systems, allerdings ist die Fehlererkennung zu unvollständig für die meisten Anwendungsszenarien.

### 1.2.4 Fazit

Pixelweise Analysen erlauben eine sehr hohe Genauigkeit bei der Bestimmung der Änderungsflächen. Allerdings lassen sich nur einfache Objektklassen aufgrund einer lokalen Nachbarschaft herleiten. Verfahren wie die Systeme WiPKA und das System von Walter werten Zusammenhänge daher geodatenobjektbasiert aus. Dabei wird es möglich, Messungen über die Fläche der Geodaten zu mitteln und in Bezug zueinander zu setzen. Dies

erhöht die Robustheit der Verfahren und erlaubt ein objektklassenspezifisches Vorgehen. Das System WiPKA ermöglicht durch dessen ausführliche manuelle Parametrisierbarkeit eine große theoretische Bandbreite bei der Fehlererkennung, ist andererseits aber nur schwer optimal einzustellen. Das System von Walter automatisiert die Fehlerbewertung, kann dabei aber nur grobe Fehler und diese auch nur in nicht zu komplexen Kontexten erkennen.

## 1.3 Ziele und Aufbau der Arbeit

Das Ziel dieser Arbeit ist die Entwicklung eines Verfahrens zur automatisierten Fehlerdetektion in Geodaten, welches die Vorteile einer objektweisen Bewertung ausnutzt. Es soll die Behandlung von Objektklassen ermöglichen, deren Objekte sich nicht ausschließlich aus einer charakteristischen Landbedeckung zusammensetzen. Darüber hinaus wird eine Vollautomatisierung angestrebt, ohne dabei die starken Einschränkungen des Walter-Verfahrens aufzuweisen.

Dazu soll die Gesamtheit aller Geodatenobjekte einer Objektklasse einer Ähnlichkeitsanalyse unterzogen werden, um ein statistisches Modell zu bestimmen. Dieses gibt wieder, wie sich solche Objekte aus Oberflächenbedeckungflächen zusammensetzen. Wenn – wie bei einer kontinuierlichen Aktualisierung gegeben – ein Großteil der Geodatenobjekte nach wie vor korrekt sind, gibt das Modell daher wieder, wie Geodatenobjekte jeder Objektklasse beschaffen sein dürfen.

Durch einen individuellen Vergleich der Geodatenobjekte mit diesem Normalitätsmodell wird für jedes Objekt dessen Stimmigkeit mit dem Modell bestimmt. Die Abweichung vom Modell wird als *Abnormität* gewertet. Diese kontinuierliche Bewertung kann als objektweise Fehlerbewertung genutzt werden. Da jede Objektklasse für sich betrachtet wird, entfallen die konzeptionellen Probleme des Reklassifikationsansatzes von Walter.

Zunächst werden in Kapitel 2 die Grundlagen für das Verständnis der Problematik der Qualitätsprüfung von Geodaten dargestellt. Im darauf folgenden Kapitel 3 wird allgemein in die Methodik der merkmalsbasierten Analyse eingeführt, auf die die in der Arbeit vorgestellten Entwicklungen aufbauen. In Kapitel 4 werden die bestehenden und bereits in der Einleitung erwähnten Lösungen – das WiPKA-System und das System von Walter – näher betrachtet. Das Konzept des neuen Ansatzes, die Bewertung von Geodatenobjekten über eine Abnormitätsbewertung, wird in Kapitel 5 entwickelt. Eine Leistungsabschätzung dieses neuen Ansatzes wird in Kapitel 6 vorgenommen, bevor die vorgestellten Konzepte in Kapitel 7 abschließend bewertet und zusammengefasst werden.

# 2 Grundlagen der Qualitätsprüfung von Geodaten

In diesem Kapitel wird ein Überblick über die für eine Auswertung notwendige Datengrundlage gegeben. Da der entwickelte Ansatz kaum spezifische Anforderungen an die Eingangsdaten stellt, beschränkt sich die Darstellung auf einen allgemeinen Überblick. Eine wichtige Annahme des Ansatzes aber ist, dass Informationen über die Szene als Draufsichten auf die zu analysierende Fläche zur Verfügung steht. Daher wird auf diesen Punkt verstärkt eingegangen.

## 2.1 Geodaten

Eine *Szene* ist ein zwei- oder dreidimensionaler Raum, der genutzt wird, um die Beschaffenheit einer (Erd-)Oberfläche festzuhalten [1]. Die Oberfläche wird als Ansammlung von *Geodatenobjekten* aufgefasst. Dies können Punkte, Linien oder Flächen sein. Diese Arbeit ist auf die Betrachtung von flächenhaften Geodatenobjekte beschränkt.

In zweidimensionalen Szenen spricht man von den $x$- und $y$-Koordinaten der Geodatenobjekte, welche die Position in einer Ebene beschreiben. Bei dreidimensionalen Szenen wird eine $z$-Dimension hinzugefügt, die üblicherweise als Höhe über Meeresniveau interpretiert wird. Im Rahmen dieser Arbeit wird die Betrachtung der Objekte auf die sogenannte *Kartengeometrie* eingeschränkt. Dabei werden alle Objekte ausschließlich im $x$-$y$-Unterraum betrachtet. Eine in dreidimensionalen Fällen vorhandene $z$-Komponente wird dabei senkrecht auf die $x$-$y$-Ebene projiziert. Geodatenobjekte können beliebige Eigenschaften, genannt *Attribute*, besitzen. Ein spezielles Attribut ist die *(Geo-)Objektklasse.* Dieses wird genutzt, um Geodatenobjekte in Gruppen, z.B. in die Objektklassen *Wohnbaufläche*, *Wald* zu unterteilen.

### 2.1.1 Objektartenkataloge

Geodaten dienen als Planungsgrundlage und zur Dokumentation für verschiedenste Anwendungen. Daher werden Geodaten stets unter Beachtung der Einsatzzwecke konzipiert und gesammelt.

Eine Menge von Geodaten mit einheitlichen *Erfassungskriterien* heißt *Objektartenkatalog.* Die Erfassungskriterien legen die zu berücksichtigenden Objektarten sowie die Herangehensweise bei der Erfassung fest. Die Objektklasseneinordnung kann durch *Attribute* erweitert werden. So kann etwa ein Geodatenobjekt einer Objektklasse *Plantage* über Attribute wie die spezifische Pflanze näher unterteilt werden.

Ein Beispiel ist der Objektartenkatalog ATKIS [3]. Der auf Deutschland beschränkte Datensatz ist als Datengrundlage für topographische Karten ausgelegt. Entsprechend finden

sich in dem Datensatz vor allem Informationen über Geländeformen und Details auf der Erdoberfläche.

Für diese Arbeit wird keine Unterscheidung zwischen Objektklassen und Attributen vorgenommen, sondern es wird ausschließlich von Objektklassen gesprochen. Dies stellt keine Einschränkung dar, solange sich die Kombination von Objektklasse und Attributen als Gruppierung von Geodatenobjekten auffassen lässt. Jede so bildbare Gruppe entspricht einer eigenständigen Objektklasse.

### 2.1.2 Generalisierung in Geodaten

Ein besonderer Aspekt der Erfassungskriterien für einen Objektartenkatalog betrifft die Festlegung von *Generalisierungskriterien*.

Diese Kriterien spezifizieren, welche Vereinfachungen für die Erfassung zugelassen bzw. gefordert sind. Sie sind notwendig, um die erfassten Informationen auf ein sinnvolles Maß einzuschränken und dabei eine verlässliche Qualität der Geodaten zu gewährleisten.

Beispiele für die Generalisierung in ATKIS sind Mindesterfassungsgrößen für flächenhafte Geodatenobjekte. Ist etwa eine Grünfläche kleiner als 1 ha wird sie nicht als eigenes Geodatenobjekt ausgewiesen, sondern ein benachbartes Geodatenobjekt wird auf die Grünfläche ausgeweitet.

Die Berücksichtigung der Regeln für die Generalisierung kann dazu führen, dass Geodaten und Wirklichkeit stark von einander abweichen, wenn man die Geodaten für andere als die geplanten Einsatzzwecke nutzen möchte.

Zudem können unabhängig voneinander durchgeführte Generalisierungen ohne weiteres zu unterschiedlichen Geodaten führen, da die Generalisierung einer Szene unter Abwägung unterschiedlichster und nicht immer eindeutiger Kriterien erfolgt [23].

### 2.1.3 Qualität von Geodaten

Die Qualität von Geodaten umfasst nach der Norm ISO 19113 verschiedene Ebenen von Qualität:

- *Vollständigkeit*: Das Vorhandensein, bzw. das Fehlen von Geodatenobjekten, Objektattributen und ihren Beziehungen.
- *Logische Konsistenz*: Die Einhaltung der Formate der Datenhaltung.
- *Positionsgenauigkeit*: Die Genauigkeit der Lage der Objekte.
- *Thematische Genauigkeit*: Korrektheit der Objektklassenzuordung und weiterer Attribute.

Qualität im Sinne der logischen Konsistenz lässt sich alleine die Geodaten betrachtend prüfen und wird daher nicht im Rahmen dieser Arbeit behandelt. Vollständigkeit, Positionsgenauigkeit und thematische Genauigkeit dagegen lassen sich weitgehend durch den Abgleich mit Fernerkundungsdaten prüfen und *sind* Gegenstand dieser Arbeit.

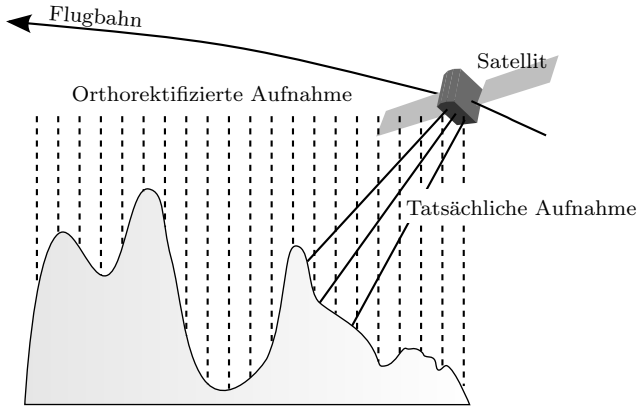



**Abbildung 2.1:** Orthorektifizierung von Satellitenaufnahmen.

## 2.2 Fernerkundungsdaten

Von Satelliten oder Flugzeugen gewonnene Daten, die Eigenschaften der Erdoberfläche wiedergeben, werden als *Fernerkundungsdaten* bezeichnet. Da kein direkter Kontakt zur Erdoberfläche besteht, können diese Daten nur aus Signalen hergeleitet werden. Bei aktiven Sensoren wie *Synthetic Aperture Radar* (SAR) [9] oder *Light Detection and Ranging* (*Lidar*) [7] werden Radar bzw. Laserstrahlen ausgesandt und die Rückstrahlung gemessen. Dagegen sind passive Sensoren wie Fotokameras auf externe Strahlenquellen, üblicherweise die Sonne, angewiesen [44].

Die gewonnenen Daten werden als Rasterbild dargestellt. Dabei wird die Oberfläche in ein gleichmäßiges Raster unterteilt und jedes Rasterelement (*Pixel*) enthält einen Wert, der die gesamte Fläche, die von dem Pixel abgedeckt wird, beschreibt[1] Der Abstand zwischen benachbarten Pixelzentren nennt sich die *geometrische Auflösung*. Im Fall, dass mehrere Werte für jedes Rasterelement erfasst werden, spricht man von *(Spektral-)Kanälen*. Pro Kanal wird ein Rasterbild genutzt. Als Mindestgröße für die Zwecke der Kartographierung wurde eine geometrische Mindestauflösung von 2,5 m ermittelt [31].

Für diese Arbeit wird vorausgesetzt, dass die Fernerkundungsdaten orthorektifiziert [35] vorliegen: Nach der Orthorektifizierung entsprechen die Daten einer direkten Draufsicht auf die Erdoberfläche, unabhängig von der Blickrichtung und den Abbildungseigenschaften des Sensors (siehe auch Abbildung 2.1). Zudem ist bekannt, welcher Teil der Erdoberfläche von den Daten beschrieben wird.

## 2.3 Vergleich der Datenrepräsentationen

Auch bei Übereinstimmung von Geokoordinatensystem und Ansicht unterscheidet sich die Flächenrepräsentation von Fernerkundungsdaten und Geodaten deutlich (Abbildung 2.2): Geodaten sind objektbasiert organisiert. Für jedes Geodatenobjekt wird die Form über ein Polygon beschrieben. Darüber hinaus wird für jedes Geodatenobjekt eine Attributliste geführt, die auch die Position umfasst. Eine Suche nach allen Geodatenobjekten einer spezifischen Objektklasse beispielsweise erfordert somit nur, die Liste aller Geodatenobjekte zu durchlaufen und ist daher unabhängig von der Größe der Szene. Allerdings muss

[1] In der Praxis handelt es sich nur um eine Näherung des tatsächliche Wertes.

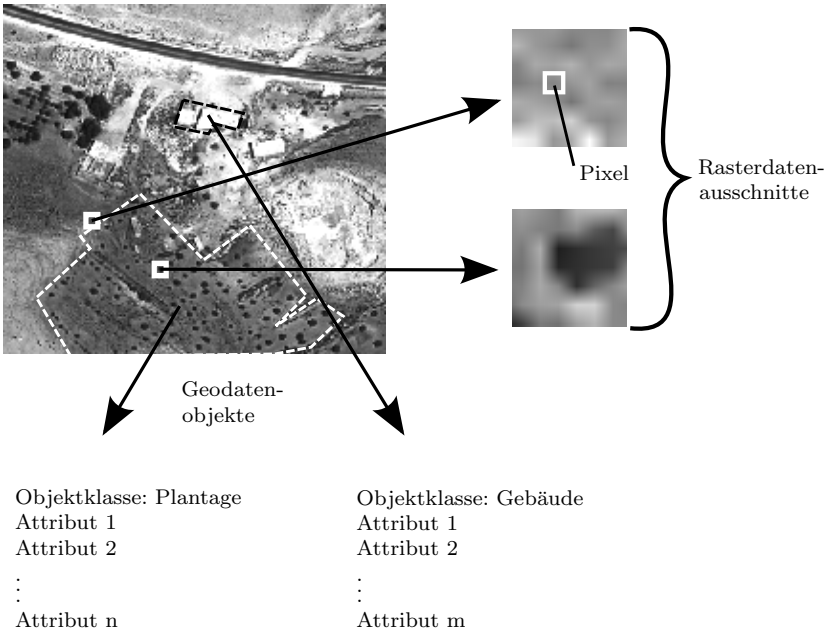



**Abbildung 2.2:** Geodatenobjekte: Die Form als Polygon (gestrichelt), weitere Werte als Attribute. Dazu Rasterdaten im Vergleich.

auch für lokale Abfragen jeweils die gesamte Objektliste durchsucht werden. Um etwa zu ermitteln, welche Geodatenobjekte sich in der Nähe eines Referenzpunktes befinden, muss die Liste aller Geodatenobjekte durchlaufen werden, um die Position zu prüfen.

Informationen können nur über die Verknüpfung mit Geodatenobjekten gespeichert werden. Wenn viele Informationen in der Szene gewonnen werden, müssen entsprechend viele Geodatenobjekte erzeugt werden. In diesem Fall steigt der Aufwand für ein Durchsuchen der Objektliste.

Fernerkundungsdaten, die in Rasterbilddarstellung vorliegen, erlauben dagegen direkt Informationen über einen beliebigen Pixel (der wiederum für eine Position in der Szene steht) zu hinterlegen. Die geometrische Genauigkeit ist nur von der räumlichen Auflösung und von der Qualität der Orthorektifizierung abhängig und nicht an Flächen (Geodatenobjekte) gebunden. Die Ermittlung von Informationen in der Nähe eines Referenzpunktes ist daher direkt möglich, unabhängig von der Szenengröße oder Menge an Geodatenobjekten. Die Suche nach Gebieten mit einer spezifischen Eigenschaft dagegen erfordert alle Pixel zu betrachten, ist also abhängig von der Größe der Szene und der geometrischen Auflösung.

In Rasterbilddarstellung ist die geometrische Auflösung eines Pixels für jedes Bild fest vorgegeben. Die Genauigkeit ist daher auf diese Größe beschränkt. Die Menge an Informationen ist für jeden Pixel immer gleich.

Da in Geodaten die Form eines Geodatenobjekts als Polygon gespeichert ist, kann über eine *Rasterung* der Daten eine Rasterbilddarstellung der Szene erzeugt werden. Die Genauigkeit ist durch die geometrische Auflösung beschränkt. Indem für jedes Geodatenobjekt eine Identifikationsnummer (ID) definiert wird, lassen sich die Objektflächen im Rasterbild darstellen. Für jeden betroffenen Pixel wird die ID als *Label* hinterlegt. In diesem

Fall spricht man von einem *Labelbild*. In Geodaten können zwei Geodatenobjekte auch eine gemeinsame Fläche belegen. Für eine Rasterung muss diese Mehrdeutigkeit entweder aufgelöst werden, oder das Labelbild muss mehrere Kanäle nutzen, um die alternativen Oberflächenbedeutungen abbilden zu können.

# 3 Grundlagen der merkmalsbasierten Analyse

Für die Anwendung des Abgleichs von Geodaten mit Fernerkundungsdaten ist an mehreren Stellen eine Analyse nötig. Einerseits gilt es für jeden Bildpunkt eine Landbedeckung festzustellen, andererseits sollen Geodatenobjekte analysiert werden. In diesem Kapitel werden die theoretischen Grundlagen der Analysetechniken beschrieben.

## 3.1 Merkmale und Merkmalsräume

Die Grundlage aller Analysen ist die Beschreibung der Bildpunkte bzw. Geodatenobjekte (allgemein *Entitäten*) über maschinenauswertbare Merkmale. Ein *Merkmal* $\mathbf{M} = (M_1, M_2, \ldots, M_n)$ beschreibt die berücksichtigten Eigenschaften einer Entität. Die unterschiedlichen $M_i$ werden als *Merkmalskomponenten* bezeichnet und beschreiben jeweils Teileigenschaften. Die Ausprägung eines Merkmals für eine spezifische Entität (der *Zustand* einer Entität) wird mit der Notation $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ dargestellt. $m_i$ entspricht dabei dem Wert der Merkmalskomponente $M_i$ mit $m_i \in \mathbb{R}$ für alle $i = 1, \ldots n$.

**Beispiel 1** *Wären die Entitäten Fahrzeuge, könnten deren zulässiges Gesamtgewicht und Höchstgeschwindigkeit durch entsprechende Merkmalskomponenten beschrieben werden (Abbildung 3.1).*

Ein *Merkmalsraum* $\mathcal{F}$ entspricht einem Vektorraum der Form $\mathbb{R}^n$, in dem dessen $n$ Dimensionen spezifische Eigenschaften darstellen. Entitäten mit entsprechenden Merkmalskomponenten können entsprechend ihres Zustands in einem gemeinsamen Merkmalsraum

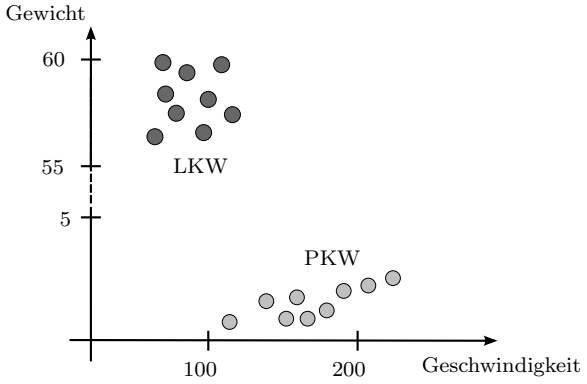



**Abbildung 3.1:** Fiktives Beispiel eines Merkmalsraums für Fahrzeuge. Merkmalskomponenten: Zulässiges Gesamtgewicht, Höchstgeschwindigkeit

positioniert werden. Durch die Positionierung in einen Merkmalsraum werden die Entitäten in einen gemeinsamen Bezugsrahmen gesetzt.

Die im weiteren Verlauf des Kapitels beschriebenen Verfahren basieren darauf, dass sich der Merkmalsraum wie ein euklidischer Vektorraum verhält. Für mathematische Gundlagen in entsprechenden Vektorräumen sei auf [30] (Seiten 148ff) verwiesen. Um die geforderten Eigenschaften sicherzustellen, kommen als Merkmalskomponenten nur Kenngrößen in Frage, die den Vorstellungen eines euklidischen Vektorraumes entsprechen. Insbesondere ist gefordert, dass mit dem Abstand zwischen Punkten, sowie Winkeln zwischen Geraden im Merkmalsraum eine Vorstellung verbunden ist.

Die Merkmale erlauben es, den Entitäten abhängig von ihrem Zustand eine Bedeutung zuzuweisen (*Klassifikation*, Abschnitt 3.3). Alternativ kann die Ähnlichkeit von Entitäten ausgewertet werden (*Abnormitätsanalyse*, Abschnitt 3.2).

### 3.1.1 Dimensionen des Merkmalsraum

Die Beschreibung von Entitäten über Merkmale wird um so detaillierter, je mehr Merkmalskomponenten zur Bildung des Merkmals genutzt werden. Mit jeder zusätzlichen Merkmalskomponente steigt per Definition die Dimension des Merkmalsraums.

Bei nur wenigen Entitäten kann eine Erweiterung des Merkmalsraums dazu führen, dass die Entitäten sich zu weit im Merkmalsraum verstreuen. Dies ist auch anschaulich klar: Je mehr Details über eine Entität bekannt werden, desto unwahrscheinlicher wird es, dass sich zwei Entitäten in allen Belangen ähneln [45].

Daher ist es wichtig, das Merkmal auf die relevanten Merkmalskomponenten zu beschränken.

### 3.1.2 Singulärwertzerlegung

Die *Singulärwertzerlegung* (SVD) [45, Seiten 697-698] ist eine Hauptachsentransformation, bei der eine neue Orthogonalbasis für den Merkmalsraum berechnet wird.

Sei $A$ eine Matrix, die die Merkmalskomponenten aller Entitäten als Einträge besitzt. Jede Zeile stehe für eine Entität, die Spalten für die verschiedenen Merkmalskomponenten.

Dann lässt sich die $m \times n$-Matrix $A$ schreiben als

$$A = U\Sigma V^\top. \tag{3.1}$$

Dabei entspricht $U$ einer $m \times m$ und $V$ einer $m \times n$ Matrix. Matrix $U$ ist orthonormal, d.h. die Spaltenvektoren stehen paarweise orthogonal aufeinander und die Längen der Spaltenvektoren sind 1. Daraus folgt $U^{-1} = U^\top$. $\Sigma$ ist eine Diagonalmatrix mit nicht-negativen, aufsteigend sortierten Elementen, also $\sigma_{i,i} \geq \sigma_{i+1,i+1}$.

Die Relevanz für die Anwendung folgt aus folgenden Eigenschaften:

- Die Spaltenvektoren $v_1, v_2, \ldots, v_n$ von $V$ heißen Rechts-Singulärvektoren. Sie sind die Eigenvektoren der Matrix $A^\top A$. Sie erfassen die statistischen Eigenschaften der Merkmalskomponenten.
- Die Spaltenvektoren $u_1, u_2, \ldots u_n$ von $U$ heißen Links-Singulärvektoren. Sie sind die Eigenvektoren der Matrix $AA^\top$. Sie erfassen die statistischen Eigenschaften der Entitäten.

- Die Diagonalelemente $\sigma_1, \sigma_2, \ldots, \sigma_n$ der Matrix $\Sigma$ werden als Singulärwerte bezeichnet. Sie sind die Wurzel der Eigenwerte von $A^\top A$ bzw. $AA^\top$.

- $A$ lässt sich schreiben als

$$A = \sum_{i=1}^{\text{Rang(A)}} \sigma_i u_i {v_i}^\top \tag{3.2}$$

  Die Matrix lässt sich also als Linearkombination von Matrizen mit Rang 1 schreiben. Da die Singulärwerte monoton fallend sind, nimmt der Einfluss der Matrizen mit zunehmenden Index $i$ ab.

  Durch eine Beschränkung auf die Singulärvektoren mit den höchstwertigsten Singulärwerten lässt sich eine Reduzierung der Dimensionen des Merkmalsraums erreichen. Durch den Verzicht auf die niedrigstwertigsten Singulärvektoren wird ein minimaler Verlust an Informationen verbunden.

Die Dimensionsreduktion wird genutzt um den Merkmalsraum auf Kosten eines minimierten Informationsverlusts zu verdichten, oder aber um einen hochdimensionalen Raum zu visualisieren.

**Hauptkomponentenanalyse** Korrigiert man jede der Merkmalskomponenten vor der Singulärwertzerlegung um deren Mittelwert, so entspricht die Singulärwertzerlegung der *Hauptkomponentenanalyse* (PCA). $\sigma_i u_i {v_i}^\top$ aus Formel 3.2 entspricht dann der Abbildungsmatrix auf die $i$-te Hauptkomponente.

### 3.1.3 Abstände im Merkmalsraum

Eine grundlegende Fragestellung, der man sich stellen muss, wenn man die Positionen von Merkmalen bewerten will, ist die Interpretation von Abständen im Merkmalsraum, also der Differenz von Absolutpositionen.

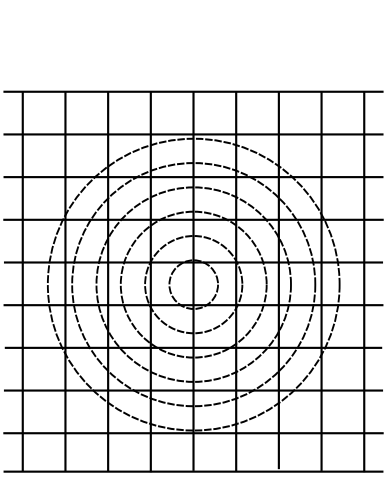
Als Standardabstand für zwei Punkte $a = (a_1, a_2, \ldots, a_n)$ und $b = (b_1, b_2, \ldots, b_n)$ gilt gewöhnlich der *euklidische Abstand*

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots (a_n - b_n)^2} \tag{3.3}$$

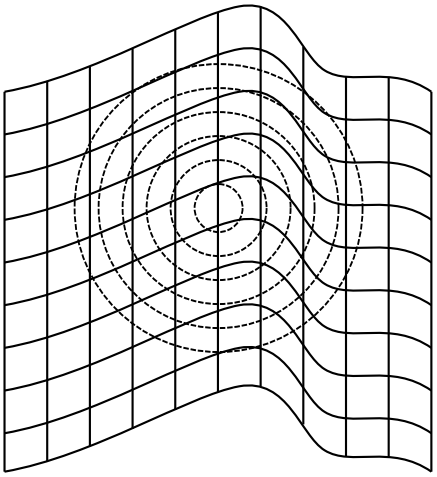
Der euklidische Abstand entspricht der üblichen Distanzmessung in der Alltagswelt. Überträgt man die Nutzung auf einen Merkmalsraum, impliziert die euklidische Distanz, dass allen Merkmalskomponenten gleich viel Gewicht beigemessen und die Abstandsberechnung als translationsinvariant angenommen wird: $d(a+t, b+t) = d(a, b)$. Die Abstände sind also nicht von der Position im Merkmalsraum abhängig.

Beide Annahmen sind häufig nicht gerechtfertigt. So können die Merkmalskomponenten unterschiedliche Wertebereiche darstellen. Werden etwa in einer Merkmalskomponente Werte in der Größenordnung 100-1000 abgebildet und in einer zweiten zwischen 0 und 1, beachtet der euklidische Abstand Unterschiede in der zweiten Komponente unzureichend. Eine Skalierung ist also sinnvoll.

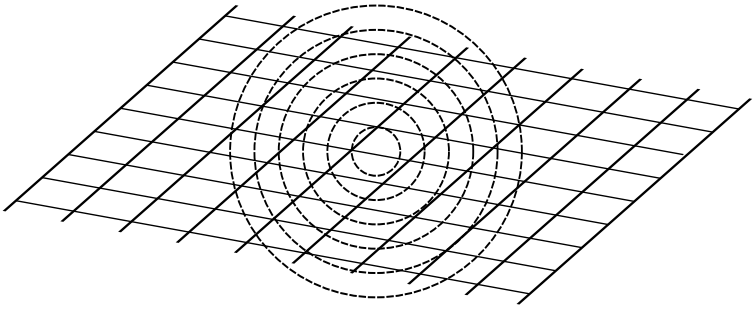
In Abbildung 3.2 sind die Möglichkeiten der Abstandsmessung exemplarisch dargestellt. Zur Betrachtung unterschiedlicher Abstandsmessungen wird die euklidische Distanz beibehalten. Stattdessen wird der Merkmalsraum durch eine Transformation verändert. In



(a) Keine Transformation



(b) Unabhängige Transformation der Komponenten



(c) Hauptachsentransformation

**Abbildung 3.2:** Darstellung unterschiedlicher Abstandsmessungen im Merkmalsraum. Dargestellt ist jeweils ein regelmäßiges Gitter, das durch eine Transformation in die dargestellte Form überführt wird. Im transformierten Raum kommt die euklidische Distanz zum Einsatz. Diese ist durch die Kreisstrukturen dargestellt. Punkte auf dem gleichen Ring haben den gleichen Abstand zum Mittelpunkt der Ringe.

Darstellung (a) ist ein unveränderter Merkmalsraum dargestellt. Um die Transformationen darzustellen ist jeweils ein regelmäßiges Gitter dargestellt. Das in (a) gezeigte unverzerrte Gitter ist in den Darstellungen (b) und (c) entsprechend verzerrt. In Abbildung (b) wird die Möglichkeit betrachtet, die Merkmalskomponenten unabhängig voneinander zu transformieren. Der Vorteil dieses Ansatzes ist, dass die Transformationen einfach realisiert werden können. So können lineare Streckungen der Achsen genauso vorgenommen werden, wie nicht-lineare Verzerrungen (in der Abbildung dargestellt). Einen Schritt weiter geht die in Abbildung (c) dargestellte *Hauptachsentransformation*. Durch die bereits aus dem vorigen Abschnitt bekannte Technik aus der linearen Algebra können paarweise lineare Abhängigkeiten zwischen Merkmalskomponenten zunächst kompensiert werden, bevor eine komponentenweise Transformation angewendet wird.

## 3.2 Merkmalsraumauswertung

Die Bewertung von Merkmalen auf Basis einer Stichprobe ist nicht die einzige Möglichkeit, Informationen über die beschriebenen Entitäten zu gewinnen. Eine weitere ist die *Merkmalsraumauswertung*. Hierbei steht die alleinige Betrachtung der Entitäten im Merkmalsraum im Vordergrund. In diesem Abschnitt werden einige grundlegende Verfahren zu Untersuchungen im Merkmalsraum beschrieben.

### 3.2.1 Normierung des Merkmalsraums

Zwei Entitäten gelten als ähnlich, wenn der Abstand ihrer Merkmale im Merkmalsraum gering ist. Die absoluten Abstände im Merkmalsraum sind wenig aussagekräftig. Ohne Bezugsgrößen ist nicht bekannt, inwieweit die Merkmale sich auch bei objektiv ähnlichen Entitäten unterscheiden können. Auch hier kann Abbildung 3.1 der Anschauung dienen. Die beiden Dimensionen des Merkmalsraums (Gesamtgewicht und Höchstgeschwindigkeit) werden in unterschiedlichen Einheiten beschrieben und die Zahlwerte haben unterschiedliche Größenordnungen. Eine Normierung der Merkmalskomponenten ist daher sinnvoll.

Darüber hinaus kann es dazu kommen, dass Dimensionen des Merkmalsraums in statistischer Abhängigkeit zueinander stehen. Dies ist der Fall, wenn sich zwei Merkmalskomponenten aufgrund einer gemeinsamen Ursache ändern. Im Beispiel aus Abbildung 3.1 scheint dies bei den PKW der Fall zu sein. So scheinen die Autos mit zunehmender Höchstgeschwindigkeit auch schwerer zu werden.

Um die Skalierung der Dimensionen und die paarweisen linearen Abhängigkeiten zwischen Dimensionen einer Merkmalsmenge zu messen, ist die Berechnung einer *Kovarianzmatrix* $\Sigma$ üblich.

**Kovarianzmatrix** Gegeben sei ein $n$-dimensionaler Merkmalsraum $\mathcal{F}$ mit den Merkmalskomponenten $M_1, M_2, \ldots, M_n$. Es bezeichne $\mathcal{T}$ die Menge der betrachteten Entitäten in diesem Merkmalsraum.

Die Kovarianzmatrix hat dann die Form:

$$\Sigma_{\mathcal{T}} := \begin{pmatrix} \operatorname{cov}(M_1, M_1), & \operatorname{cov}(M_1, M_2), & \dots, & \operatorname{cov}(M_1, M_n) \\ \operatorname{cov}(M_2, M_1), & \operatorname{cov}(M_2, M_2), & \dots, & \operatorname{cov}(M_2, M_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(M_n, M_1), & \operatorname{cov}(M_n, M_2), & \dots, & \operatorname{cov}(M_n, M_n) \end{pmatrix} \tag{3.4}$$

mit

$$\operatorname{cov}(M_i, M_j) := \frac{1}{|\mathcal{T}| - 1} \sum_{\mathbf{m} \in \mathcal{T}} (m_i - \mu_{\mathcal{T}})(m_j - \mu_{\mathcal{T}})^\top \tag{3.5}$$

mit $\nu^\top$ die Transponierte eines Vektors $\nu$ und $\mu_{\mathcal{T}}$ dem nach Formel 3.31 berechneten Mittelwert. Im Fall von $M_i = M_j$ ergibt sich die Varianz analog zu Formel 3.32.

Die Kovarianzmatrix zeigt sowohl die paarweise lineare Abhängigkeit (*Korrelation*) zwischen den Dimensionen, als auch die unterschiedliche Varianz der Dimensionen (auf der Hauptdiagonalen). Eine Kovarianzmatrix sei hier immer positiv definit [45, Seiten 702-703]. Im Fall dass die Matrix einen nicht vollen Rang aufweist, kann die Eigenschaft durch eine Merkmalsreduzierung, etwa mit einer SVD (Abschnitt 3.1.2), sicher gestellt werden.

### 3.2.2 Mahalanobis Transformation

Eine Kompensation der durch eine Kovarianzmatrix erfassten Verzerrung erfolgt über die *Mahalanobis Transformation*. Indem die Kovarianzmatrix als lineare Abbildung interpretiert wird, kann über die Inverse der Kovarianzmatrix ($\Sigma^{-1}$) diese Verzerrung kompensiert werden.

Allerdings kann die Kovarianzmatrix nicht direkt verwendet werden. Ähnlich wie im eindimensionalen Fall die Varianz $\sigma^2$ das Quadrat der Standardabweichung darstellt, handelt es sich auch bei der inversen Kovarianzmatrix um das Quadrat der eigentlich für die Transformation benötigten "Wurzel" $\Sigma^{-\frac{1}{2}}$.

Eine Matrix-Wurzel im allgemeinen Fall ist nicht definiert. Da aber eine Kovarianzmatrix, wie aus der Definition in Gleichung 3.4 zu erkennen, stets symmetrisch ist, kann sie über die Spektralzerlegung nach Jordan geschrieben werden als

$$\Sigma = \Gamma \Lambda \Gamma^{-1} \tag{3.6}$$

wobei $\Lambda = \operatorname{diag}(\lambda_i)$ eine Diagonalmatrix der Eigenwerte $\lambda_i$ von $\Sigma$ ist. $\Gamma$ ist eine Matrix deren Spalten die den Erwartungswerten zugehörigen normierten Eigenvektoren beinhaltet. Aus der Definition von $\Gamma$ folgt die Orthogonalität von $\Gamma$. Es gilt also insbesondere

$$\Gamma^{-1} = \Gamma^\top \tag{3.7}$$

Für $\Lambda$ gilt, dass für alle $r, s \in \mathbb{Z}$

$$\Lambda^{\frac{r}{s}} = \operatorname{diag}\left(\lambda_i^{\frac{r}{s}}\right) \tag{3.8}$$

Da Kovarianzmatrizen stets positiv definit und die Eigenwerte größer 0 sind, gilt schließlich

$$\Sigma^{\frac{r}{s}} = \Gamma \Lambda^{\frac{r}{s}} \Gamma^\top \tag{3.9}$$

Daraus folgt direkt

$$\Sigma^{\frac{1}{2}} = \Gamma\Lambda^{\frac{1}{2}}\Gamma^{\top} \quad \text{mit } \Lambda^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_i}) \tag{3.10}$$

und

$$\Sigma^{-\frac{1}{2}} = \Gamma\Lambda^{-\frac{1}{2}}\Gamma^{\top} \quad \text{mit } \Lambda^{-\frac{1}{2}} = \text{diag}(\frac{1}{\sqrt{\lambda_i}}) \tag{3.11}$$

Die Normierung des Merkmalsraums erfolgt, indem für alle Merkmale $\nu \in \mathcal{F}$ das entzerrte Merkmal $\nu'$ über die *Mahalanobis Transformation* wie folgt berechnet wird:

$$\nu' := \Sigma^{-\frac{1}{2}}\nu \tag{3.12}$$

**Beispiel 2** *Sei $\mathcal{F}$ ein zweidimensionaler Merkmalsraum mit den Merkmalskomponenten $x$ und $y$. Die $x$–Komponente nehme zufällig Werte an, deren Wahrscheinlichkeitsverteilung $X$ gaussverteilt mit Varianz 1 und Mittelwert 0 sei. Zudem existiere eine zweite Zufallsvariable $Y'$, die gaussverteilt mit Varianz 2 und Mittwert 0 sei. Die $y$–Komponente habe die Verteilung $\frac{1}{2}(X + Y')$. In Abbildung 3.3(a) ist der Merkmalsraum mit zufällig ermittelten 10000 Punkten dargestellt. Jeder Punkt entspricht dabei einem Merkmal. Eine Auswahl einiger Punkte ist zur besseren Übersicht hervorgehoben.*

*Die Kovarianzmatrix der Punkte lautet*

$$\Sigma = \begin{pmatrix} 0{,}99 & 0{,}48 \\ 0{,}48 & 1{,}22 \end{pmatrix}. \tag{3.13}$$

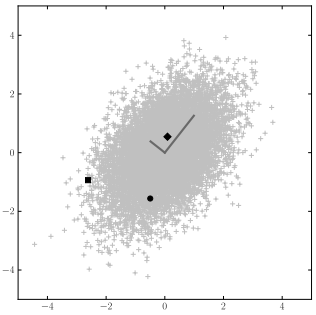
*Die ungeraden Werte resultieren aus der Schätzung einer endlichen Stichprobe der Verteilungen nach den Formeln 3.4 und 3.5. Die mit den Eigenwerten gewichteten normierten Eigenvektoren der Kovarianzmatrizen sind in den Abbildungen durch Pfeile dargestellt.*

*Die Entzerrung erfolgt nun über die Mahalanobis Transformation nach Gleichung 3.12. Die verschiedenen Schritte*

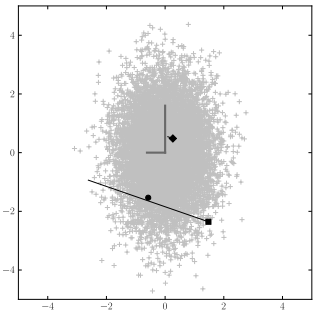
$$\begin{aligned} \nu' &= \Sigma^{-\frac{1}{2}}\nu && (3.14)\\ &= \Gamma\Lambda^{-\frac{1}{2}}\Gamma^{-1}\nu && (3.15)\\ &= \Gamma\Lambda^{-\frac{1}{2}}\nu_2 && (3.16)\\ &= \Gamma\nu_3 && (3.17) \end{aligned}$$

*sind in der Abbildung 3.3 in den Teilbildern (a)–(d) nachvollzogen. In (a) ist der anfängliche Merkmalsraum abgebildet. Die elliptische Form der Punktwolke ergibt sich aus der unterschiedlichen Varianz der beiden Merkmalskomponenten. Die* schiefe *Ausrichtung der Wolke folgt aus der Konstruktion der vertikalen y-Komponente, in die die x-Komponente additiv eingeht. Wie zu erwarten, zeigen die Eigenvektoren in die Richtung der Ausprägungen der Punktwolke. Das Ziel der Mahalanobis Transformation ist, die unterschiedliche Varianz der Dimensionen auszugleichen.*

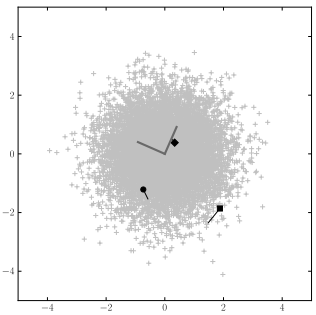
*In den Folgebildern werden die Auswirkungen der Teiloperationen (Gleichung 3.15) der Transformation betrachtet. Zunächst wird mit Hilfe der Inversen der Orthogonalmatrix der Merkmalsraum so gedreht und gespiegelt, dass die Wolke sich in Richtung der Achsen des Merkmalsraums ausrichtet (Abbildung (b)). Die Eigenvektoren drehen sich mit der*



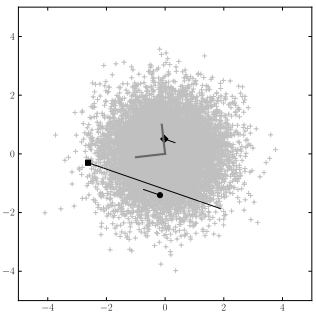
(a) Initialer Merkmalsraum



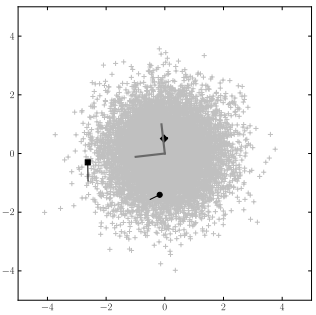
(b) Entkopplung der Dimensionen



(c) Normierung der Dimensionen



(d) Wiederherstellung des Zusammenhangs der Dimensionen, Endergebnis



(e) Absolute Verschiebung durch Transformation

**Abbildung 3.3:** Mahalanobis Transformation in einem zweidimensionalen Merkmalsraum mit zufällig erzeugten Punkten (graue Kreuze im Hintergrund). Die Eigenvektoren sind als graue Linien dargestellt, drei Punkte sind zur besseren Visualisierung durch Symbole hervorgehoben. Die Einzelschritte der Transformation sind in den Bildern (a) - (d) gezeigt. Die Positionsverschiebungen zum jeweilig vorhergehenden Schritt sind durch schwarze Linien angedeutet. Darstellung (e) zeigt die Auswirkung der gesamten Transformation auf die Ursprungsmenge.

*Punktwolke. Nun ist der Einfluss der x-Komponente auf die y-Komponente kompensiert. Somit können durch die Diagonalmatrix (welche als Einträge die multiplikativ Inversen der Längen der verschiedenen Eigenvektoren haben), die Dimensionen normiert werden (Abbildung (c), Gleichung 3.16). Die Punktwolke hat nach der Normierung keine Vorzugsrichtungen mehr und die Eigenvektoren zeigen in eine willkürliche Richtung. Im letzten Schritt (Gleichung 3.17) wird der ursprüngliche Zusammenhang wieder hergestellt, indem der Merkmalsraum zurückgedreht und –gespiegelt wird (Abbildung (d)).*

*Abbildung (e) zeigt die Auswirkungen der gesamten Transformation. Die Kovarianzmatrix der Punktmenge nach der Transformation lautet*

$$\Sigma = \begin{pmatrix} 1{,}0 & 0{,}0 \\ 0{,}0 & 1{,}0 \end{pmatrix}. \tag{3.18}$$

*Die Verteilung der Punkte ist aber auch nach der Transformation nicht gleichverteilt. Dies erkennt man an der mit dem Abstand zum Mittelpunkt abnehmenden Dichte der Punkte. Dies liegt daran, dass die Skalierung nur eine lineare Operation, die Verteilung der Punkte aber gaussverteilt ist.*

### 3.2.3 Mahalanobis Distanz

Im durch die Mahalanobis Transformation umgestalteten Merkmalsraum kann die gewöhnliche euklidische Distanz zur Abstandsmessung genutzt werden:

Für zwei Punkte $u, v$ und $w = u - v$ im nicht transfomierten Merkmalsraum mit Kovarianzmatrix $\Sigma$ folgt also

$$\begin{aligned}
d_{\text{Mah}}(u, v, \Sigma) &:= |u - v|_{\text{Mah}} = |w|_{\text{Mah}} && (3.19)\\
&:= \left(\Sigma^{-\frac{1}{2}} w\right)^{\top} \left(\Sigma^{-\frac{1}{2}} w\right) && (3.20)\\
&= w^{\top} {\Sigma^{-\frac{1}{2}}}^{\top} \left(\Sigma^{-\frac{1}{2}} w\right) && (3.21)\\
&= w^{\top} \left(\Gamma \Lambda^{-\frac{1}{2}} \Gamma^{\top}\right)^{\top} \left(\Gamma \Lambda^{-\frac{1}{2}} \Gamma^{\top} w\right) && (3.22)\\
&= w^{\top} \Gamma \left(\Gamma \Lambda^{-\frac{1}{2}}\right)^{\top} \left(\Gamma \Lambda^{-\frac{1}{2}} \Gamma^{\top} w\right) && (3.23)\\
&= w^{\top} \Gamma {\Lambda^{-\frac{1}{2}}}^{\top} \Gamma^{\top} \left(\Gamma \Lambda^{-\frac{1}{2}} \Gamma^{\top} w\right) && (3.24)\\
&= w^{\top} \Gamma {\Lambda^{-\frac{1}{2}}}^{\top} \left(\Gamma^{\top} \Gamma\right) \left(\Lambda^{-\frac{1}{2}} \Gamma^{\top} w\right) && (3.25)\\
&= w^{\top} \Gamma \left(\Lambda^{-\frac{1}{2}} \Lambda^{-\frac{1}{2}}\right) \Gamma^{\top} w && (3.26)\\
&= w^{\top} \left(\Gamma \Lambda^{-1} \Gamma^{\top}\right) w && (3.27)\\
&= w^{\top} \Sigma^{-1} w && (3.28)\\
&= (u - v)^{\top} \Sigma^{-1} (u - v) && (3.29)
\end{aligned}$$

Zur Abstandsbestimmung reicht also die Bestimmung der Kovarianzmatrix. Die Transformation des Merkmalsraums erfolgt bei der Distanzbestimmung implizit und eine Zerlegung der Matrix ist nicht nötig.

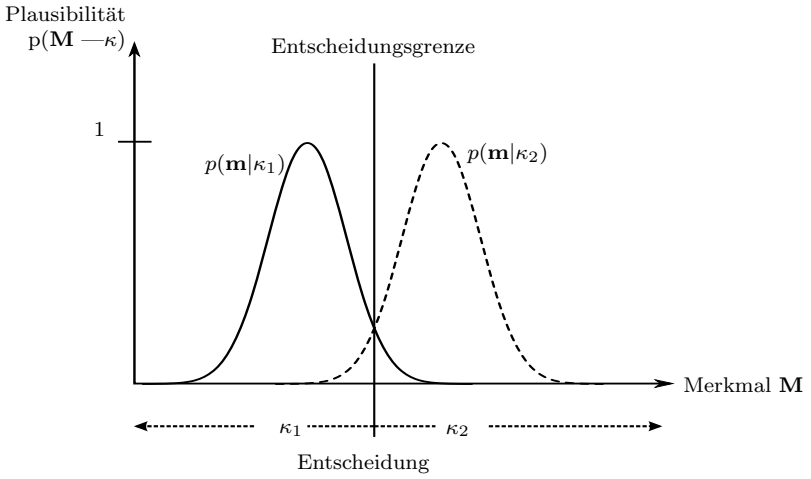



**Abbildung 3.4:** Darstellung der Maximum-Likelihood-Klassifikation. Es wird eine Entscheidung für Klasse $\kappa_1$ getroffen, falls $p(\mathbf{m}|\kappa_1) > p(\mathbf{m}|\kappa_2)$, ansonsten $\kappa_2$.

## 3.3 Merkmalsbasierte Klassifikationsverfahren

Um etwa für Pixel der Fernerkundungsdaten eine Landbedeckung zu bestimmen, werden die durch Pixel beschriebenen Merkmale durch Klassifikationsverfahren analysiert. Dazu wird der Algorithmus üblicherweise in einem Trainingsschritt über eine Stichprobe angelernt. Eine Stichprobe besteht aus repräsentativen Entitäten jeder in Erwägung gezogenen Klasse.

### 3.3.1 Maximum-Likelihood

Die Herangehensweise des *Maximum-Likelihood*-Verfahrens ist die Bestimmung der Plausibilität eines beobachteten Merkmals $\mathbf{m}$ für eine Entität einer angenommenen Klasse $\kappa$. Die Plausibilitätseinschätzung basiert auf einer angenommenen stochastischen Wahrscheinlichkeitsfunktion $p$, mit der $p(\mathbf{m}|\kappa_i)$ für jede Klasse $\kappa_i$ bestimmt wird. Eine Klasse $\kappa_k$ wird für eine Entität festgesetzt, falls

$$p(\mathbf{m}|\kappa_k) = \max\{p(\mathbf{m}|\kappa_1), p(\mathbf{m}|\kappa_2) \ldots p(\mathbf{m}|\kappa_n)\} \tag{3.30}$$

erfüllt ist.

Die Wahrscheinlichkeitsfunktion wird vorgegeben, offen sind zunächst die Parameter der Funktion. Daher ist ein Trainingsschritt nötig, bei dem für jede Klasse $\kappa$ eine Stichprobe $\mathcal{T}_\kappa$ bereitzustellen ist, aus der die Parameter der Wahrscheinlichkeitsfunktion geschätzt werden.

**Beispiel 3** *Bei der Annahme einer Gaussverteilung sind die Parameter $\mu_\kappa$ (Mittelwert) und $\sigma^2_\kappa$ (Varianz) für jede Klasse $\kappa$ zu schätzen. Übliche Schätzer sind:*

$$\hat{\mu}_\kappa = \frac{1}{|\mathcal{T}_\kappa|} \sum_{\boldsymbol{m} \in \mathcal{T}_\kappa} \boldsymbol{m} \tag{3.31}$$

*und*

$$\hat{\sigma}^2_\kappa = \frac{1}{|\mathcal{T}_\kappa| - 1} \sum_{\boldsymbol{m} \in \mathcal{T}_\kappa} (\hat{\mu}_\kappa - \boldsymbol{m})^2 \tag{3.32}$$

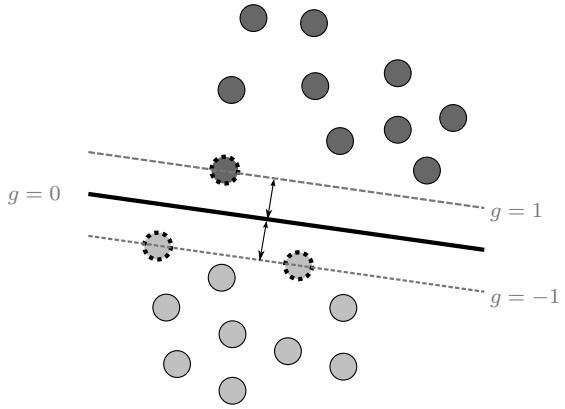



**Abbildung 3.5:** Zweidimensionaler Merkmalsraum, die Grenzlinie (fett) maximiert den Abstand zwischen Merkmalen der Klasse $\kappa_1$ (hell) und $\kappa_2$ (dunkel). Die Stützvektoren sind mit gepunktetem Rand dargestellt. Darstellung angelehnt an [11, Seite 262]

*Mit den vorliegenden (geschätzten) Parametern lässt sich die Plausibilität eines beliebigen Merkmals $\boldsymbol{m} \in \mathcal{F}$ wie folgt bestimmen:*

$$p(\boldsymbol{m}|\kappa) = p(\boldsymbol{m}|\hat{\mu}_\kappa, \hat{\sigma}^2_\kappa) = \frac{1}{\hat{\sigma}^2_\kappa} \exp(-\frac{(x - \hat{\mu}_\kappa)^2}{\hat{\sigma}^2_\kappa}). \tag{3.33}$$

*In Abbildung 3.4 ist das Prinzip für ein Merkmal mit nur einer Merkmalskomponente dargestellt. In diesem Fall formt der Merkmalsraum eine Gerade. In der Abbildung ist der Merkmalsraum auf der horizontalen Achse dargestellt. Die beiden Gausskurven geben die Wahrscheinlichkeit der Klassen 1 und 2 für ein Merkmal $\boldsymbol{m}$ wieder. Es wird auf Klasse 1 entschieden, falls $p_1(\boldsymbol{m}) > p_2(\boldsymbol{m})$. Ansonsten wird auf Klasse 2 entschieden.*

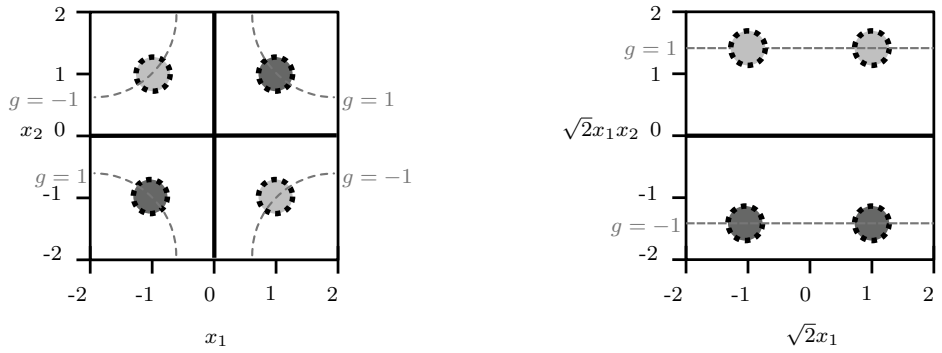
Als Voraussetzung des Verfahrens gilt, dass sich die Erwartungswerte hinreichend unterscheiden und die Varianz der Merkmale innerhalb der Klassen möglichst klein ist, damit es zu einer möglichst geringen Überschneidung der Wahrscheinlichkeitsdichten kommt. Ansonsten besteht die Gefahr von Fehlklassifikationen. Zudem beschränkt die Notwendigkeit der expliziten Formulierung einer Wahrscheinlichkeitsfunktion die Freiheit bei der Klassenmodellierung deutlich.

### 3.3.2 Support Vector Machine

Wie das Maximum-Likelihood-Verfahren ist auch das *Support Vector Machine* (SVM) Verfahren [47] ein Klassifikationsverfahren, bei dem jede Entität für sich klassifiziert wird.

Beim SVM-Verfahren wird der Merkmalsraum in Bereiche aufgeteilt. Der Bereich, in dem das Merkmal eines Pixels liegt, bestimmt die Klassenzugehörigkeit. Im klassischen Fall können zwei Klassen $\kappa_1$, $\kappa_2$ unterschieden werden.

Die Einteilung des Merkmalsraums geschieht über die Positionierung einer Grenze, die die Bereiche gegeneinander abgrenzt. Die Grenze entspricht dabei immer einer Hyperebene im Merkmalsraum, also einem Teilraum des Merkmalsraums, der eine Dimension weniger hat als der Merkmalsraum selbst. Beispielsweise ist die Grenze in einem zweidimensionalen Merkmalsraum eine Gerade, man spricht daher auch von einer linearen Trennung.




(a) Merkmalsraum (b) Transformierter Merkmalsraum

**Abbildung 3.6:** Im anfänglichen Merkmalsraum (a) sind die Klassen $\kappa_1$ (hell), $\kappa_2$ (dunkel) nicht linear trennbar, im transfomierten Merkmalsraum (b) schon. Die Grenze (schwarz) im transformierten Merkmalsraum sind auch in (a) dargestellt, ebenso auch die Abstandslinien (grau). Die Darstellung ist angelehnt an [11, Seite 264].

Um trotz der einfachen Form der Grenzfunktion komplexe Verteilungen der Merkmale zu berücksichtigen, besitzt eine SVM die implizite Fähigkeit, den anfänglichen Merkmalsraum in einen höherdimensionierten Merkmalsraum zu transformieren. Das Ziel ist die Transformation in einen Merkmalsraum, in dem eine lineare Trennung möglich wird.

**Trennung im Merkmalsraum** Die Klassentrennung (oder Klassengrenze) im hoch transformierten Merkmalsraum wird mathematisch über die Definition einer *Diskriminanzfunktion* $g$ realisiert, die für ein Merkmal einen reellen Wert bestimmt. Negative Ergebniswerte entsprechen Klasse $\kappa_1$, positive entsprechend Klasse $\kappa_2$. Bei einer SVM wird die Grenze so positioniert, dass, eine Stichprobe analysierend, der Raum, der *zwischen* Merkmalen der beiden Klassen liegt, gleichmäßig verteilt wird. Davon verspricht man sich eine möglichst gute Verallgemeinerung des Stichprobentrainings. Dies wird erreicht, indem die Lage der Grenze nicht von allen Merkmalen der Stichprobe, sondern nur von einer Teilmenge, den *Stützvektoren*, abhängig gemacht wird (engl. *Support Vectors*).

Die Stützvektoren sind diejenigen Merkmale der Stichprobe, die im Merkmalsraum am nächsten zu Merkmalen der anderen Klasse liegen. Für diese Merkmale wird die Diskriminanzfunktion so gewählt, dass sie die Werte $g = -1$ (Klasse $\kappa_1$), bzw. $g = 1$ (Klasse $\kappa_2$) annimmt. Das Prinzip ist in Abbildung 3.5 exemplarisch dargestellt: Merkmale für Stichproben der beiden Klassen ($\kappa_1$: hell, $\kappa_2$: dunkel) liegen in einem zweidimensionalen Merkmalsraum. Als Stützvektoren werden die Punkte ausgewählt, die der anderen Klasse am Nächsten liegen. Die Entscheidungsgrenze, also an den Punkten, an den $g$ den Wert 0 annimmt, ist in schwarz dargestellt. In grau gestrichelt dargestellt, die Punkte des Merkmalsraums, an denen $g = 1$, bzw. $g = -1$ gilt. Bei der Klassifikation von unbekannten Merkmalen wird die Klasse $\kappa_1$ festgestellt, wenn $g < 0$, ansonsten $\kappa_2$.

**Transformation des Merkmalsraums** Die Nutzung einer Hyperebene des Merkmalsraums als Grenze erlaubt nur eine einfache Unterteilung des Merkmalsraums. Um auch komplexe Verteilungen der Klassen erkennen zu können, erlaubt es eine SVM, den Merkmalsraum geschickt in einen höher dimensionalen Merkmalsraum zu transformieren, sodass

dort eine lineare Trennung möglich wird.

**Beispiel 4** *Ein Beispiel (übernommen aus [11, Seiten 264–265]) ist in Abbildung 3.6 dargestellt: Im sogenannten* XOR*-Problem sind die Klassen wie in (a) dargestellt verteilt. In dem zweidimensionalen Merkmalsraum gibt es keine einzelne Gerade, die die beiden Klassen trennen kann. Überführt in einen Merkmalsraum mit der Basis* $1$, $\sqrt{2}x_1$, $\sqrt{2}x_2$, $\sqrt{2}x_1x_2$, $x_1^2$, $x_2^2$ *ist dies dagegen möglich. In 3.6 (b) ist eine zweidimensionale Projektion dieses Merkmalsraum zu sehen. Die vier Merkmale lassen sich jetzt durch die fett gedruckte Linie trennen, der Abstand der Merkmale (in diesem Fall gleichzeitig alles Stützvektoren) beträgt* $\sqrt{2}$.

**Die Unterscheidung von mehr als zwei Klassen** Um den SVM-Ansatz für die Unterscheidung zwischen mehr als zwei Klassen zu nutzen, werden mehrere SVM Durchläufe durchgeführt, die schließlich ausgewertet werden müssen, um eine Klassifikationsentscheidung zu erhalten. Dabei sind mehrere Verfahren gebräuchlich:

**One-against-all** Bei einem $n$-Klassen-Problem werden $n$ SVM.Durchläufe durchgeführt. Es wird jeweils eine Klasse $\kappa_i$ gegen eine Vereinigung aller übrigen Klassen getestet. Für jede Klassifikation werden alle Diskriminanzfunktion ausgewertet und die Ergebnisse verglichen. Die SVM mit dem größten Wert der Diskriminanzfunktion bestimmt das Ergebnis [6].

Dieser Ansatz hat den Vorteil, dass der Aufwand nur linear mit der Klassenzahl wächst.

**One-against-one** Für jede der $n$ Klassen werden $(n-1)$ verschiedene SVM trainiert. Jede SVM unterscheidet dabei jeweils zwischen zwei Klassen. Für jede Klassifikation werden also $\frac{n(n-1)}{2}$ SVM-Klassifikationen durchgeführt. Jede Klassenentscheidung wird als Stimme in einer Abstimmung interpretiert, um eine abschließende Klassifizierung zu erreichen [18].

Der Klassifikationsaufwand steigt quadratisch mit der Klassenzahl $n$.

**Baumauswertung** Die Herangehensweise bei der Baumauswertung entspricht dem vorgenannten One-against-one-Verfahren. Allerdings werden zu Beginn der Klassifikation nicht alle $\frac{n(n-1)}{2}$ SVM-Klassifikationen durchgeführt. Stattdessen bestimmt ein Baum (ein gerichteter azyklischer Graph mit eindeutigem Eintrittsknoten) die Reihenfolge der Berechnung. Zunächst berechnete SVM-Klassifikationen werden genutzt, um jeweils eine Klasse auszuschließen [40]. Somit reduziert sich der Klassifikationsaufwand darauf, $\log(\frac{n(n-1)}{2}) = O(n)$ SVM-Klassifikationen zu bestimmen. Beim Training müssen allerdings nach wie vor alle paarweisen SVMs berücksichtigt werden.

Vergleichsergebnisse [36] deuten an, dass *One-against-one* und die *Baumauswertung* dem *One-against-all* Ansatz überlegen sind.

### 3.3.3 Merkmale zur pixelweisen Oberflächenbedeckungsklassenbestimmung

Zur Klassifikation in Bildern, bei denen die Bildpunkte oder Pixel die Entitäten bilden, ist das Ergebnis ein Labelbild. Als Merkmalskomponenten kommen neben den Inten-

sitätswerten der Bilddaten weitere Kenngrößen in Frage, die die Zusammenhänge zwischen Pixeln berücksichtigen.[1]

Da für jeden Pixel jeweils ein Pixelbereich betrachtet wird, der den Pixel umschließt, werden diese Merkmale diesem zugerechnet. Diese Pixelnachbarschaft wird im folgenden mit $\mathcal{N}$ bezeichnet. Die Position eines Pixels wird über die Koordinaten $x$ und $y$ relativ zum Bildursprung angegeben und der Wert eines Pixels mit $p_{x,y}$. Alle Merkmale bestimmen für die Nachbarschaft $\mathcal{N}$ einen reellen Wert.

**Mittelwert**

$$\mu_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{p \in \mathcal{N}} p \tag{3.34}$$

Der Mittelwert dient einer gemittelten Beschreibung der Pixelnachbarschaft.

**Varianz**

$$\sigma^2_{\mathcal{N}} = \frac{1}{|\mathcal{N}| - 1} \sum_{p \in \mathcal{N}} (p - \mu_{\mathcal{N}})^2 \tag{3.35}$$

Während der Mittelwert eine gemittelte Beschreibung der Nachbarschaft ausdrückt, sagt die Varianz aus, wie stark die Pixel im Mittel vom Mittelwert abweichen. Eine geringe Varianz beschreibt daher eine homogene Nachbarschaft, eine hohe Varianz dagegen eine eher inhomogene. Die Art der Strukturierung geht allerdings verloren.

Weitere Merkmale wie *Local Binary Patterns* [39] und *Haralick Merkmale* [22] versuchen die Zusammenhänge zwischen Pixeln zu berücksichtigen. Da sie für die aktuelle Anwendung keine Vorteile gezeigt haben, wird auf sie hier nicht näher eingegangen.

[1] In der Literatur wird für die Kenngrößen häufig der Begriff *Merkmal* genutzt. In der Nomenklatur dieser Arbeit handelt es sich dagegen allgemeiner um Merkmalskomponenten.

# 4 Objektbasierte Verfahren zur Kontrolle von Geodatenobjekten

Im Unterschied zu einer direkten Analyse von Pixeln aus Fernerkundungsdaten, bzw. von geodatenunabhängigen Pixelgruppen (Abschnitt 1.2.1), werden bei einer objektbasierten Beschreibung Merkmale für einzelne Flächen (*Objekte*) erzeugt, die nicht direkt aus den Fernerkundungsdaten, sondern den Geodaten entlehnt sind.

Die zu prüfenden Geodaten werden in Bezug zu der Landbedeckung gesetzt, indem für jedes Geodatenobjekt dessen zugehörige Landbedeckung betrachtet wird. Die Landbedeckung wird so also zu Einheiten zusammengefasst. Somit wird nicht nur die Landbedeckung an sich, sondern auch das Zusammenspiel von Landbedeckungen innerhalb eines Geodatenobjekts beschreibbar.

Die Objektklasse der Geodatenobjekte erlaubt es, Geodatenobjekte zu gruppieren. Dies kann genutzt werden, um Gemeinsamkeiten etwa in der Landbedeckung zu analysieren und Unterschiede zwischen Objektklassen zu identifizieren.

In diesem Kapitel werden das WiPKA-System und das System von Walter näher betrachtet. Beide Ansätze haben gemein, dass sie eine *objektweise Beschreibung* einer Szene nutzen. Bei beiden Verfahren werden Geodatenobjekte entweder *akzeptiert* (korrekt) oder *verworfen* (inkorrekt).

## 4.1 WiPKA: Regelbasierte Überprüfung von Geodatenobjekten

Das am *Institut für Informationsverarbeitung (*TNT*)* und *Institut für Photogrammetrie und GeoInformation (*IPI) der *Leibniz Universität Hannover* entwickelte WiPKA-System ist ein umfangreiches System zur Überprüfung von Geodaten.

Zur Analyse der Fernerkundungsdaten kommen verschiedene Verfahren zum Einsatz. Beispiele sind Algorithmen zur Erkennung von Straßen und Flüssen, Gebäuden, sowie die Ermittlung der Oberflächenbedeckung aufgrund von Texturen. Die Koordinierung der Analyseaufrufe und die Fusion der Ergebnisse erfolgt durch das am TNT entwickelte hierarchische Analysetool GeoAIDA. Eine umfangreiche Übersicht über die Verfahren findet sich in [25].

Bislang unveröffentlicht sind Details zur Kontrolle der Geodaten an sich. Die Kontrolle von flächenhaften Geodatenobjekten wird daher im Folgenden näher beschrieben.

### 4.1.1 Überprüfung von Geodaten

Pixelweise Klassifikationsverfahren wie eine SVM (Abschnitt 3.3.2) erzeugen zunächst ein Labelbild der Szene, in dem mehrere Oberflächenbedeckungsklassen unterschieden werden. Die unterscheidbaren Oberflächenbedeckungsklassen hängen von den Eingangsdaten

und den Szenen ab. Empirische Untersuchungen haben ergeben, dass sich aus optischen Satellitenbildern 4–5 Oberflächenbedeckungsklassen verlässlich unterscheiden lassen.

Die Überprüfung erfolgt durch manuell formulierte und parametrisierte Regeln, die den Aufbau der Geodatenobjekte durch Oberflächenbedeckungsflächen überprüfen. Es werden folgende Regeln eingesetzt:

1. Für jede Objektklasse werden die ermittelten Oberflächenbedeckungsklassen unterteilt in *positive* und *negative* Oberflächenbedeckungen. Positive Oberflächenbedeckungen sprechen für das Vorhandensein einer Objektklasse, negative sprechen dagegen. Für die Regel wird das Verhältnis der Flächenanteile von positiven zu negativen Oberflächenbedeckungsflächen berechnet und mit einem Schwellwert abgeglichen. Übersteigt der negative Anteil den Schwellwert wird das Geodatenobjekt verworfen.

   Da Oberflächenbedeckungsklassen nur binär in positiv und negativ unterschieden werden, wird meist nur eine Oberflächenbedeckungsklasse als positiv gewertet. Die weiteren Klassen werden entsprechend als negativ eingeschätzt.

2. Durch die oben stehende Regel werden Geodatenobjekte als Ganzes überprüft. Die Nutzung eines Flächenverhältnisses als Entscheidungskriterium führt dazu, dass die Absolutgrößen von Flächen nicht betrachtet werden. Für den Fall, dass bestimmte Oberflächenbedeckungsflächen besondere negative Relevanz haben, werden daher zusammenhängende Flächen dieser Oberflächenbedeckungsflächen einzeln betrachtet und auf Größe und Kompaktheit untersucht. Werden hierfür vorher bestimmte Schwellwerte überschritten, so führt dies zum Markieren des Geodatenobjekts als potentieller Fehler.

Im Fall von vergleichsweise kleinen Geodatenobjekten (wie bei ATKIS üblich) spielt die zweite Regel keine große Rolle, da jede große negativ eingeschätzte Landbedeckung zugleich das Verhältnis stark belastet, sodass bereits die erste Regel zu einer Erkennung des Geodatenobjekts als fehlerhaft führen würde. Bei großen Geodatenobjekten allerdings kehrt sich die Bedeutung der Regeln um.

### 4.1.2 Bewertung des Verfahrens

Zur Analyse und Demonstration der Methodik wird das Verfahren anhand von Testdaten evaluiert, die auch später in Kapitel 6 für die Analyse des neu entwickelten Verfahrens genutzt werden. Genutzt wurden die Geodatenobjekte der Objektklassen *Wohnbaufläche* und *Wald* der Szene Weiterstadt. Die Landbedeckungsanalyse wurde durch eine SVM-Analyse bestimmt und ermittelte die Landbedeckungsklassen *Bäume*, *Häusergruppen*, *Gras/Acker*, sowie *Industriehallen*.

Die Regeln unterscheiden die möglichen Landbedeckungsklassen in positiv und negativ. In Tabelle 4.1 ist die Aufteilung der Objektklassen aufgeführt. Abbildung 4.1 stellt dar, welcher Anteil von positiver Fläche zur Gesamtfläche in den Geodatenobjekten der Szene zu beobachten sind (gestrichelt: Wald, gepunktet: Wohnbaufläche).

Bei der Betrachtung der Kurven fällt auf, dass sich nur etwa 60% der Wohnbauflächenobjekte und 40% der Waldobjekte vollständig aus positiv eingeschätzter Oberflächenbedeckung zusammensetzen (die Werte sind auf die Ordinate durch Linien abgetragen). Dies

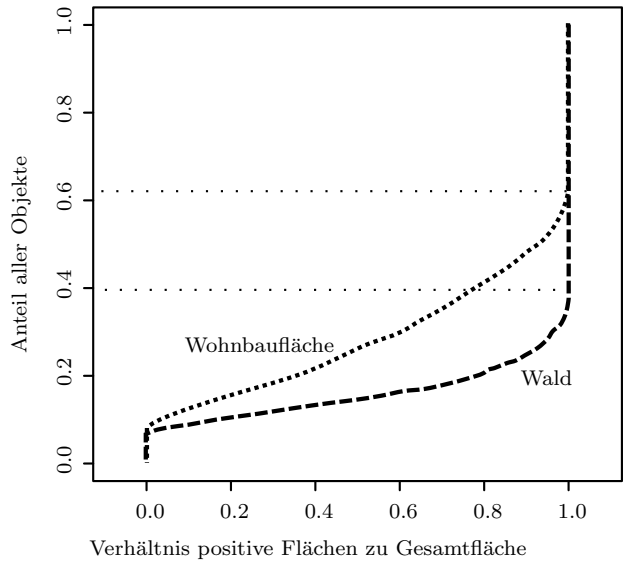



**Abbildung 4.1:** Anteil verworfener Geodatenobjekte für die möglichen Schwellwerte für die ATKIS Objektklassen Wohnbaufläche (gepunktet) und Wald (gestrichelt).

**Tabelle 4.1:** Bedeutung der Landbedeckungsklassen für die Objektklassen Wohnbaufläche und Wald (P: positiv, N: negativ)

| Landbedeckung | Status für Wohnbaufläche | Status für Wald |
|---|---|---|
| Bäume | N | P |
| Häusergruppen | P | N |
| Gras/Acker | N | N |
| Industriehallen | N | N |

unterstreicht, dass sich Objektklassen nicht grundsätzlich anhand einer einzelnen Oberflächenbedeckung erkennen lassen. Die Geodatenobjekte werden als fehlerhaft markiert, wenn der Anteil einen vorher festzusetzenden Wert unterschreitet.

Um die Regeln zu entwickeln, muss der Bearbeiter daher versuchen, gedanklich alle möglichen Situationen vorwegzunehmen. Im Falle der Geodatenüberprüfung bedeutet dies, sich alle möglichen Ausprägungen von beispielsweise Wohnbauflächenobjekten vorzustellen und die entscheidenden Kriterien durch Regeln zu beschreiben. Experten für diese Aufgabe sind Bearbeiter, die es gewohnt sind, manuell Geodaten durch die Betrachtung von Fernerkundungsdaten zu überprüfen.

Allerdings unterscheidet sich das Planen von Regeln grundlegend von dem Vorgehen, das ein Bearbeiter bei einem direkten Abgleich von Geodaten durch Fernerkundungsdaten zeigt. Ein Planung von Regeln erfordert, sich alle möglichen Ausprägungen für etwa Wohnbauflächenobjekten vorzustellen und die entscheidenden Kriterien durch Regeln zu beschreiben. Bei einer traditionellen rein manuellen Prüfung der Daten dagegen wird jeder existierende Fall tatsächlich betrachtet und jeweils entschieden, ob die Situation zulässig ist. Dabei werden die Kriterien gegebenenfalls unweigerlich und unbewusst angepasst[1].

Abbildung 4.1 verdeutlicht die Problematik, einen optimalen Schwellwert für die Kriterien manuell zu bestimmen. Während für die Waldobjekte die Kurve einen flachen Verlauf hat, ist die Wohnbauflächenkurve deutlich steiler. Dadurch führen Unterschiede beim gewählten Schwellwert bei Wald zu nur geringen Änderungen bei der Anzahl der als fehlerhaft markierten Geodatenobjekte. Bei der Objektklasse Wohnbaufläche hat die Wahl des Schwellwertes größere Konsequenzen. Qualitative Ergebnisse des WiPKA-Systems werden in Abschnitt 6 gezeigt.

## 4.2 Walter: Automatische Reklassifikation von Geodatenobjekten

Ein von Volker Walter entwickeltes System [50] überprüft die Richtigkeit der einem Geodatenobjekt zugewiesenen Objektklasse dadurch, dass für jedes Objekt die Objektklasse durch eine überwachte Klassifikation neu bestimmt wird. Entspricht die Neubestimmung nicht der ursprünglichen Objektklasse, werden die betroffenen Geodatenobjekte als fehlerhaft betrachtet.

Als Klassifikationsverfahren kommt ein *Maximum Likelihood* Verfahren (Abschnitt 3.3.1) zum Einsatz. Die Geodatenobjekte fungieren als Entitäten, als Anlernstichprobe wird die Gesamtheit alle Geodatenobjekte herangezogen. Hierdurch entfällt eine manuelle Vorauswahl einer Stichprobe, allerdings ist die Stichprobe nicht fehlerfrei.

Als Merkmale für die Objektklassenklassifikation werden genutzt:

- Mittelwert und Varianz der Intensitätswerte der Eingangsbilder für jeweils das gesamte Geodatenobjekt. Die Definitionen aus Abschnitt 3.3.3 nutzend, entspricht die Nachbarschaft $\mathcal{N}$ der gesamten Fläche des behandelten Geodatenobjekts.

[1] Zu Untersuchungen bezüglich der Zuverlässigkeit menschlicher Entscheidungsfindung bei Bewertungssituationen und die mangelhafte Selbsteinschätzung siehe [28]

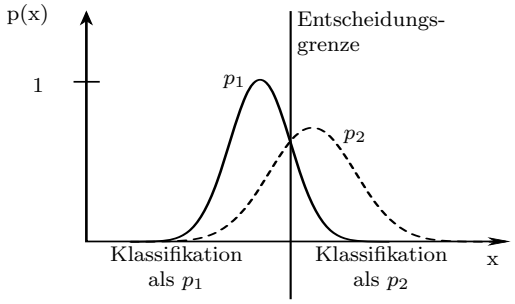



**Abbildung 4.2:** Maximum Likelihood Klassifikation, bei der sich die Merkmale der Klassen nicht sicher unterscheiden lassen.

- Durch eine weitere nicht näher beschriebene Maximum Likelihood Klassifikation wird jedem Pixel des Eingangsbildes $I$ eine Oberflächenbedeckungsklasse zugewiesen. Die Landbedeckung eines Geodatenobjekts wird über das in Abschnitt 5.3.1 beschriebene Landbedeckungshistogrammmerkmal dargestellt.

### 4.2.1 Bewertung des Verfahrens

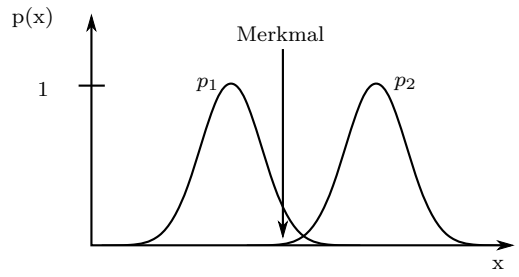
Durch das automatische Training der Objektklassenklassifikation entfällt eine manuelle Konfiguration des Systems. Durch diese Vollautomatisierung des Verfahrens sind die Anforderungen an den Benutzer niedrig. Das Vorgehen erlaubt allerdings nur eine Behandlung von Objekten, die vollständig einer der dem System bekannten Objektklassen entsprechen. Für Fälle in denen dem System die richtige Objektklasse eines Objekts nicht bekannt ist, oder auch in denen sich ein Objekt stückweise aus mehreren Objektklassen zusammen setzt, ist die Formulierung als Reklassifikation nicht geeignet.

Doch selbst wenn die Fragestellung auf Geodatenobjekte beschränkt wird, die sich vollständig einer der Objektarten zurechnen lassen, ist die Überprüfung Einschränkungen unterworfen:

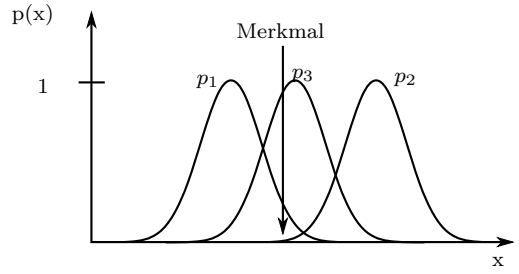
Es wird vorausgesetzt, dass sich der Merkmalsraum so unterteilen lässt, dass sich Objekte der gleichen Objektklasse genau einem Bereich zurechnen lassen und sich die Bereiche verschiedener Objektklassen nicht überlappen. Andernfalls ist eine Situation gegeben, wie sie in Abbildung 4.2 beispielhaft dargestellt ist. Hier werden alle Objekte der Klasse der jeweils plausibelsten Klasse zugeschlagen. Folglich lässt das Merkmal nicht sicher auf die Klasse schließen.

Um dies zu verhindern, müssen die Merkmale eine Unterscheidung der Objektklasse zunächst einmal prinzipiell erlauben. Das Problem wird bei einer Maximum Likelihood Klassifikation dadurch verschärft, dass sich die zu erwartenden Merkmale durch eine berechenbare Wahrscheinlichkeitsfunktion beschreiben lassen können müssen.

Die auf den ersten Blick triviale Forderung nach geeigneten Merkmalen für die Unterscheidung von Objektklassen ist eine wesentliche Fragestellung. Die Reklassifikationsaufgabe fordert nicht nur, dass die Merkmale die Eigenschaften einer Klasse gut wiedergeben, sondern auch, dass sich die Merkmale von den Merkmalen jeder anderen Objektklasse deutlich unterscheiden. Je größer die Anzahl an Objektklassen, desto problematischer gestaltet sich eine Unterscheidung. Doch auch eine zu geringe Anzahl an Objektklassen ist




(a) Objekt klassifiziert als $p_1$




(b) Objekt klassifiziert als $p_3$

**Abbildung 4.3:** Ob ein inkorrektes $p_1$ Objekt mit dargestelltem Merkmal als fehlerhaft erkannt wird, hängt von den anderen Objektklassen ab. In (a) wird das Objekt als $p_1$ zugehörig eingeschätzt. Wird aber dem Merkmalsraum eine dritte Klasse $p_3$ hinzugefügt (b), so führt dies auch zu einer veränderten Einschätzung des Objekts (nun $p_3$).

problematisch. Angenommen, ein Objekt habe eine falsche Objektklasse, dann müßte diese Objektklasse dem System bekannt sein, um einen Fehler überhaupt aufzudecken. Vergleiche hierzu Darstellung 4.3. Angenommen es gäbe ein Objekt der Objektklasse $p_1$ mit einem Merkmal wie in der Abbildung dargestellt. Das Objekt sei fehlerhaft, das heißt die Objektklasse wäre falsch. Dann würde in Fall (a) dennoch die Klasse $p_1$ festgestellt. Erst wenn eine weitere Objektklasse dem Merkmalsraum hinzugefügt wird, kann auf eine abweichende Objektklasse erkannt werden (b).

## Zusammenfassung

Beide Verfahren vereint, dass die Überprüfung von Geodaten nicht auf einen direkten pixelweisen Abgleich von Analysedaten und Geodaten reduziert wird. Stattdessen berücksichtigen beide den Aufbau von Geodatenobjekten durch Oberflächenbedeckungsobjekte. Das Analysesystem WiPKA (Abschnitt 4.1) erlaubt eine sehr große Flexibilität bei der Konfiguration der Fehlererkennung, stellt hierfür aber große Anforderungen an die Benutzer. Die mögliche Komplexität der Regeln ist daher praktisch beschränkt. Das System von Walter (Abschnitt 4.2) dagegen beherrscht eine vollautomatische Fehlererkennung. Die Methodik schränkt die Erkennung allerdings auf die thematische Genauigkeit ein. Darüber hinaus ist die Erweiterbarkeit auf eine große Anzahl an Objektklassen nicht möglich, da die Problemkomplexität mit der Anzahl der Klassen steigt.

# 5 Neuer Ansatz: Prüfung von Geodaten über eine Gütefunktion

In diesem Kapitel wird der neu entwickelte Ansatz beschrieben. Die Aufgabenstellung ist weiterhin die Prüfung von Geodatenobjekten um fehlerhafte Geodatenobjekte zu erkennen. Die angestrebten Merkmale des neuen Verfahrens sind

- Bewertung von Geodatenobjekten durch einen kontinuierlichen Gütewert. Im Vergleich zu einer binären *korrekt/falsch* Entscheidung der Geodatenobjekte ermöglicht die kontinuierliche Bewertung auch einen Umgang mit Geodatenobjekten, die sich für das System nicht eindeutig bewerten lassen.

- Behandlung von Objektklassen, die sich nicht eindeutig als Landbedeckung identifizieren lassen. Die Geodatenobjekte dieser Objektklassen können aus verschiedenen Landbedeckungen bestehen.

- Berücksichtigung von Geodatenobjekten, die sich in ihrer aktuellen Landbedeckung nicht vollständig einer bekannten Objektklasse zurechnen lassen.

- Selbstständige Anpassung an aktuell vorliegende Szene, keine explizite Konfiguration.

## 5.1 Herangehensweise des neuen Verfahrens

Die neue Herangehensweise besteht darin, für jedes Geodatenobjekt dessen *Normalität* zu schätzen. Die vom Verfahren behandelte Fragestellung ist also nicht etwa wie bei Walter (Abschnitt 4.2), ob ein gegebenes Wohnbauflächenobjekt nicht vielleicht ein Waldobjekt sein sollte, sondern ob es, verglichen nur mit anderen Wohnbauflächenobjekten, ein gewöhnliches Objekt ist. Dies fußt auf der Vorstellung, dass die in Erfassungskriterien hinterlegten Spezifikationen für Geodatenobjekte für die Erstellung der in der Szene befindlichen Geodatenobjekte bereits berücksichtigt wurden. Sind also etwa Beschränkungen in der Genauigkeit der Abgrenzung zwischen Wohnbauflächenobjekten und einer anderen Objektart vorgesehen, so wird sich dies bei den Wohnbauflächen in der vorliegenden Szene beobachten lassen. Somit ist eine Abweichung nicht mehr ungewöhnlich und wird vom System entsprechend weniger beachtet.
Das System lässt sich in eine Abfolge mehrerer Arbeitsschritte unterteilen (Abbildung 5.1). Die Abnormitätsanalyse wird immer nur für die Geodatenobjekte einer spezifischen Objektklasse vorgenommen. Verschiedene Objektklassen werden unabhängig voneinander behandelt. Die Trennung der Objektklassen erfolgt unter der Annahme, dass die Abnormität der Objekte einer Objektklasse unabhängig von den Objekten der weiteren Objektklassen ist.

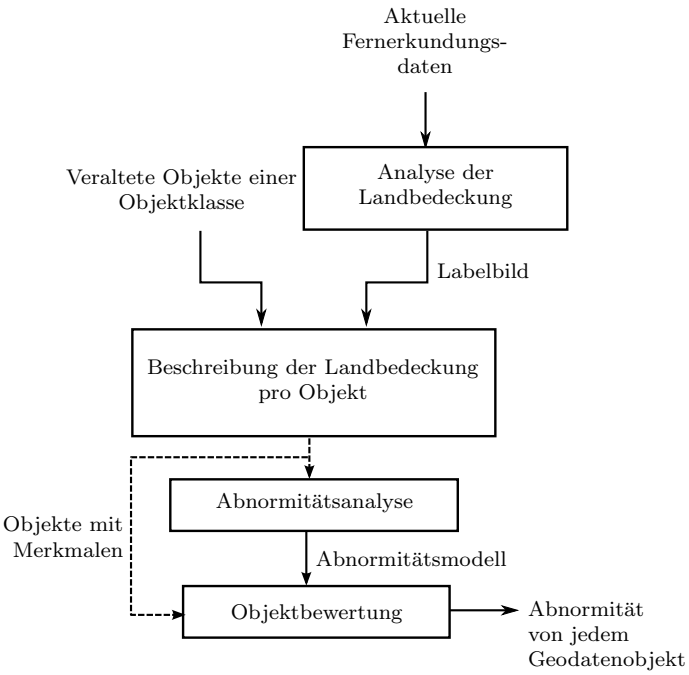



**Abbildung 5.1:** Arbeitsschritte der Abnormitätsbewertung von Geodatenobjekten

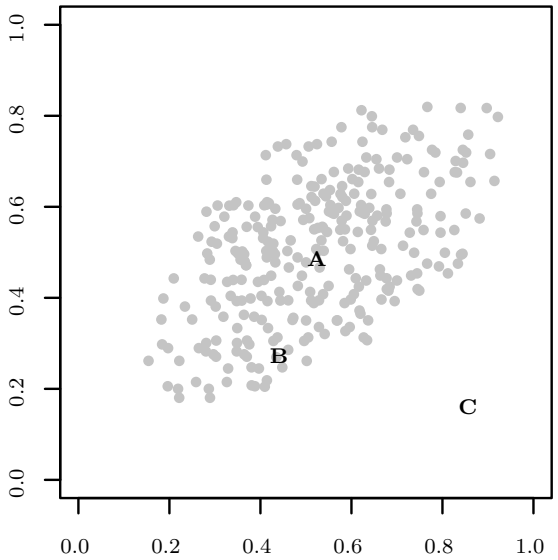



**Abbildung 5.2:** Merkmalsraum mit Objekten. Relative Position bestimmt Normalität: $A$ normaler als $B$, $B$ normaler als $C$.

Die Landbedeckung wird analysiert und für jedes Geodatenobjekt durch Merkmale beschrieben. Alle Geodatenobjekte einer Objektklasse werden in den durch die Merkmale aufgespannten Merkmalsraum eingeordnet.

Geodatenobjekte, deren Merkmale sich im Merkmalsraum räumlich nahe sind, weisen ähnliche Merkmale auf. Bereiche im Merkmalsraum, in der sich Geodatenobjekte konzentrieren, deuten daher auf übliche Merkmale für die gemeinsame Objektklasse hin. Da fehlerhafte Geodatenobjekte anders beschaffen sind als korrekte Objekte, nehmen sie im Merkmalsraum entsprechend Positionen ein, in denen sich kaum Geodatenobjekte befinden. Während das in Abbildung 5.2 durch $A$ markierte Objekt nahe des Häufungspunkts der Gesamtmenge positioniert ist, liegt Objekt $B$ am Rand der Menge. Es scheint daher weniger gewöhnlich zu sein. Objekt $C$ schließlich liegt weit entfernt von allen anderen Objekten, was auf ein sehr ungewöhnliches Merkmal hinweist. Das Ziel der Analyse ist eine kontinuierliche Bewertung der Geodatenobjekte über die Beschreibung mit einem Gütewert, der *Abnormität*.

Das Ziel ist es also, durch Methoden der Merkmalsraummodellierung (Abschnitt 3.2), die charakteristischen Eigenschaften eines *normalen* Objekts zu ermitteln und im Umkehrschluss die Abnormität eines beliebigen Objektes zu bestimmen.

## 5.2 Fernerkundungsdaten als Referenz für Geodaten

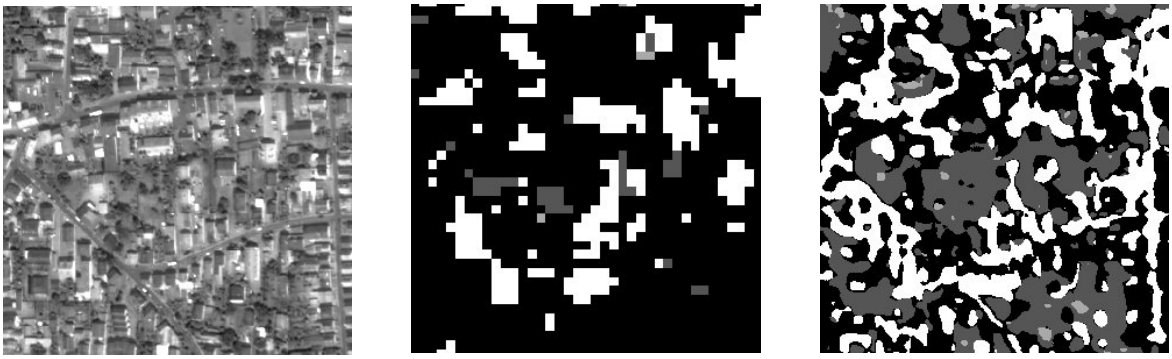
Wie auch bei den existierenden Verfahren werden die den Geodaten gegenüber zu stellenden Informationen aus Fernerkundungsdaten gewonnen. Fernerkundungsdaten in ihrer ursprünglichen Form enthalten Informationen nur als Messwerte von Sensoren. Für Satellitenbilder beispielsweise steht jeder Pixel für die Spektralinformation an der von ihm abgebildeten Stelle der Erdoberfläche.

Daher ist es sinnvoll, die Daten einer ersten Verarbeitung zu unterziehen, um in den Sensormessungen inhärent enthaltene Informationen auswertbar zu machen. Für (flächenhafte) Geodatenobjekte ist dies vor allem die Landbedeckung. Die für diesen Ansatz zu ermittelnde Landbedeckung ist nicht zwingend gleichbedeutend mit den Objektklassen der Geodatenobjekte. Die Anforderungen sind:

- Geodatenobjekte können aus mehreren Flächen unterschiedlicher Landbedeckungen bestehen.
- Die gleiche Landbedeckung darf in Objekten unterschiedlicher Objektklasse auftreten.
- Die Entscheidung für Landbedeckungen sollte sich an der Eignung der Daten und den Eigenschaften der Verfahren orientieren, die für die Landbedeckungsanalyse zur Verfügung stehen.

Es sind prinzipiell alle Verfahren geeignet, die jedem Pixel in den Fernerkundungsdaten eine Landbedeckung zuweisen. Das neue System hat den Anspruch, sich selbstständig den zur Verfügung stehenden Daten anzupassen.

Die vergleichsweise geringen Anforderungen an die zu ermittelnden Landbedeckungen sind ein Vorteil beim Umgang mit den Analysemethoden. Die Selbstanpassung des Systems erlaubt mehr Freiheiten bei einem manuellen Training. Wie in Abbildung 5.3 dargestellt,



Wohnbauflächen | Landbedeckung 1 | Landbedeckung 2

**Abbildung 5.3:** Das Ergebnis von zwei Analysen einer SVM mit unterschiedlicher Lernstichprobe. Die Landbedeckung ist mit Grauwerten markiert. Von hell nach dunkel: Gras, Bäume, Hallen, Häuser.

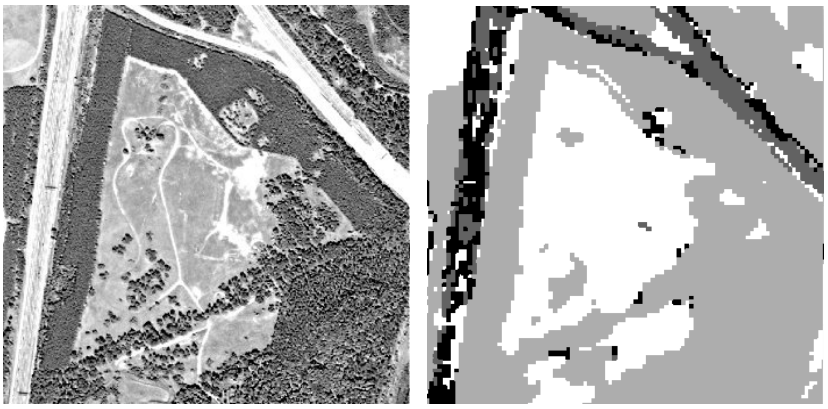
hängt die ermittelbare Landbedeckung stark von der verwendeten Lernstichprobe ab. Zwar ist die Erkennungsleistung weiter von der Eignung der Landbedeckungsanalyse abhängig, doch passt sich das System an eine gegebene Analyse an, um im Rahmen der bestehenden Möglichkeite gute Resultate zu ermöglichen.

Da die erwartete Landbedeckung für ein Geodatenobjekt nicht wie bei WiPKA (Abschnitt 4.1) über manuelle Parameter- oder Regelsätze spezifiziert wird, ist nicht relevant, ob die ermittelte Landbedeckung menschlichen Einschätzungen entspricht. Bei empirischen Untersuchungen ließ sich beispielsweise beobachten, dass an Übergängen zwischen verschiedenen Landbedeckungen häufig eine Häuser-Landbedeckung analysiert wurde. Eine manuelle Überprüfung in den Fernerkundungsdaten fand aber keine Häuser. Dem Augenschein nach kommt es regelmäßig zu einer Falscheinschätzung von Schatten durch die genutzte Landbedeckungsanalyse. Durch die Selbstanpassung des neuen Systems an die Szene sollen solche regelmäßigen Effekte kompensiert werden.

Im Rahmen dieser Arbeit wurde eine SVM (Abschnitt 3.3.2) zur Ermittlung der Landbedeckung eingesetzt. Als Pixelmerkmale kamen Mittelwert und Varianz (Abschnitt 3.3.3) zum Einsatz. Wie bereits gesehen, ist das Verfahren üblich für diese Aufgabe. Es existieren zwar etliche Alternativen für Verfahren und Merkmale, deren Nutzung die absolute Qualität der Ergebnissen beeinflussen könnte. Um die Effektivität der vorgestellten Methoden zu bewerten, reicht eine SVM mit den einfachen Merkmalen jedoch aus.

Die Qualität der Ergebnisse der Landbedeckungsanalyse ist insofern entscheidend, als dass die Fernerkundungsdaten im weiteren Verlauf nicht weiter betrachtet werden. In dem folgenden Abschnitt dienen die Labelbilder der durchgeführten Analyse als Darstellung der Wirklichkeit. Entscheidend ist daher, dass die Analyse möglichst ohne Berücksichtigung der Geodaten erfolgt, um eine Beeinflussung durch die unsicheren Daten zu vermeiden.

Ein Beispiel für eine Bestimmung der Landbedeckung mit dem SVM-Verfahren ist in Abbildung 5.4 dargestellt. Im Satellitenbild (a) ist für Menschen eine von Wald umgebene zentrale Wiesenfläche erkennbar. Durch den Wald verlaufen am Bildrand mehrspurige Straßen. In (b) ist das Ergebnis der Landbedeckungsanalyse abgebildet. Dem Verfahren wurden Stichproben von Gras, Bäumen, (Industrie-)hallen, und Häusern gezeigt. Folglich



(a) Satellitenbild (Ausschnitt) (b) Labelbild der Landbedeckung

**Abbildung 5.4:** Beispiel für ein Satellitenbild und das Ergebnis einer SVM-Landbedeckungsanalyse. Die Landbedeckung ist mit Grauwerten markiert. Von hell nach dunkel: Gras, Bäume, Hallen, Häuser.

konnten die mehrspurigen Straßen nicht als solche erkannt werden. Stattdessen wurden die Pixel den am ehesten passenden Hallen und Häusern zugeordnet.

## 5.3 Objektbeschreibung durch Merkmale

Durch die Landbedeckungsanalyse steht für jeden Pixel statt des Spektralwerts nun die aussagefähigere Landbedeckung zur Verfügung. Das Ergebnis wird durch ein Labelbild repräsentiert. Der Zusammenhang zwischen Landbedeckung und den individuellen Geodatenobjekten wird über Objektmerkmale beschrieben. Die Gestaltung der Merkmale bestimmt, welche Aspekte bei der Normalitätsbetrachtung berücksichtigt werden. Für die Betrachtung in einem Merkmalsraum müssen alle Geodatenobjekte die gleichen Merkmalskomponenten aufweisen. Das Merkmal muss daher so formuliert werden, dass es alle möglichen Geodatenobjekte stimmig wiedergibt.

Das Objektmerkmal ist so zu wählen, dass sich die Merkmale korrekter Geodatenobjekte möglichst ähneln. Zwar ist die Normalitätsanalyse die Aufgabe eines späteren Schrittes (Abschnitt 5.4), doch muss durch die Gestaltung des Merkmals eine Grundlage gelegt werden.

Um die Ähnlichkeit von Merkmalen korrekter Geodatenobjekte sicher zu stellen, muss das Merkmal die Gestaltung eines Geodatenobjekts ausreichend verallgemeinern. Bei der Merkmalsentwicklung und Auswahl muss also eine Abwägung zwischen der Verallgemeinerung des Erscheinungsbildes von Geodatenobjekten einerseits und der Unterscheidbarkeit von korrekten und falschen Geodatenobjekten andererseits getroffen werden.

### 5.3.1 Landbedeckungsanteile in einem Geodatenobjekt

Um die Häufigkeit des Auftretens der unterschiedlichen Landbedeckungen innerhalb eines Geodatenobjekts $o$ zu beschreiben, ist das *Landbedeckungshistogramm* ein geeignetes Merkmal:

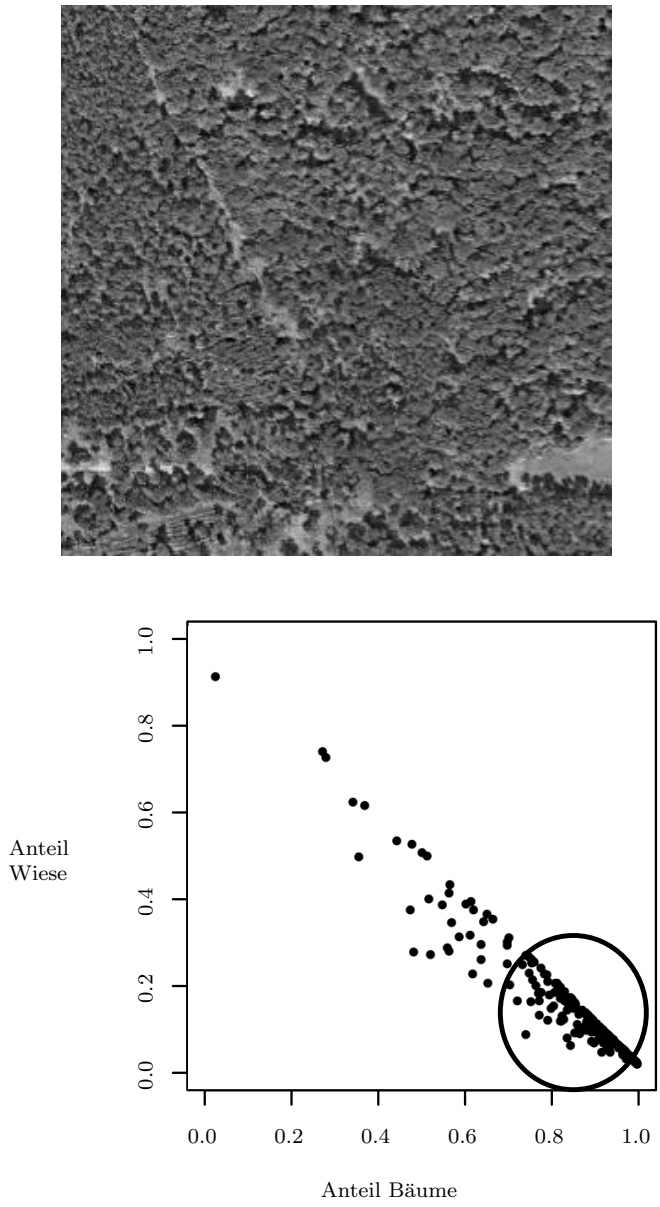



**Abbildung 5.5:** In ATKIS-Geodaten stellen Bäume die dominierende Landbedeckung von Wald (obere Abbildung). Das Landbedeckungshistogramm kann das Vorhandensein gut abbilden. Untere Abbildung: Auszug des Landbedeckungshistogrammmerkmalsraums mit 255 Waldobjekten für die Komponenten Wald- und Wiesenanteile.

Für die Menge der $n$ Landbedeckungsklassen $l_1, l_2, \ldots l_n$ wird für jede Landbedeckungsklasse $l_i$ die Merkmalskomponente $M_{l_i}$ aufgebaut. Der Wert wird bestimmt durch

$$m_{l_i}(o) := \frac{1}{|o|} \sum_{\text{Pixel } p \in o} \delta_{l_i}(p) \tag{5.1}$$

mit

$$\delta_{l_i}(p) := \begin{cases} 1 & \text{, falls } p = l_i \\ 0 & \text{, sonst.} \end{cases}$$

Aus der Definition folgt

$$\sum_{i=1}^{n} m_{l_i}(o) = 1 \tag{5.2}$$

Jede Merkmalskomponente beschreibt den Anteil der jeweiligen Landbedeckung am Geodatenobjekt. Die räumliche Verteilung der Landbedeckung innerhalb des Objekts hat keine Auswirkungen auf das Merkmal. Es ist daher beispielsweise gut für die Beschreibung von Wäldern in ATKIS-Geodaten geeignet, in denen zwar sowohl Baumbestand und Lichtungen auftreten dürfen, aber keine Gebäude:

Abbildung 5.5 zeigt einen Merkmalsraum von 255 Waldobjekten. Jedes Geodatenobjekt wurde über das Landbedeckungshistogramm mit vier Landbedeckungen beschrieben. Für die Darstellung wurden die beiden Merkmalskomponenten Baum- und Wiesenbedeckung herausgegriffen. Die Geodatenobjekte konzentrieren sich bemerkenswert auf die Hauptdiagonale der Darstellung. Die Anteile von Bäumen und Wiesen ergeben zusammen genommen also häufig die Gesamtfläche. Der Großteil der Objekte konzentriert sich zudem auf den mit einem Kreis markierten Bereich von über 70% Baumanteil. Das Landbedeckungsmerkmal verfügt also über eine gute Verallgemeinerung. Dass es auch Merkmale weit außerhalb des markierten Bereichs gibt ist nicht überraschend. Schließlich handelt es sich um teilweise fehlerhafte Daten. So ist das einzeln liegende Geodatenobjekt mit einem Baumanteil nahe null Prozent und einem Wiesenanteil von über 90 Prozent ein falsches Waldobjekt.

Allgemein ausgedrückt verspricht das Merkmal *Landbedeckungshistogramm* gute Ergebnisse für Objektarten, bei denen das Auftreten bestimmter Landbedeckungen maßgeblich für die Korrektheit der Geodatenobjekte ist.

Dieses Merkmal hat eine lange Tradition als robustes Merkmal zur Beschreibung von Flächen. So wird es bei dem Geographieanalyseprogramm FRAGSTATS (engl. *Spatial Pattern Analysis Program for Quantifying Landscape Structure*) [37] genauso berücksichtigt wie bei den existieren Geodatenprüfansätzen von Walter und WiPKA. Daher wird es auch in dieser Arbeit als Ausgangspunkt für die Analyse genutzt.

## 5.4 Abnormitätsanalyse und Objektbewertung

Die Aufgabe der Abnormitätsanalyse ist die Bewertung der Geodatenobjekte mit einem Gütemaß. Die Ansprüche an das Gütemaß sind

- Möglichst viele korrekte Geodatenobjekte müssen eine niedrige Bewertung erhalten.

- Möglichst viele falsche Geodatenobjekte müssen eine hohe Bewertung erhalten.

Die Zuweisung des Gütemaßes basiert auf der Analyse des durch die Merkmale aufgespannten Merkmalsraums. Durch eine Bewertung der Position des Merkmals eines Geodatenobjekts im Vergleich zu der Masse an korrekten Geodatenobjekten wird die Abnormität des Geodatenobjekts bestimmt. Die anschauliche Definition von Ähnlichkeit von Merkmalen gilt es zu formalisieren. Die Abnormitätsbewertung eines Geodatenobjekts wird zurückgeführt auf den Abstand von dessen Merkmal zu einem festgelegten Referenzpunkt. Dies ist eine übliche Herangehensweise im Feld der Anomaliedetektion [45, Seiten 651 ff]. Bei der Auswahl und Anpassung der Methoden sind verschiedene Überlegungen zu berücksichtigen.

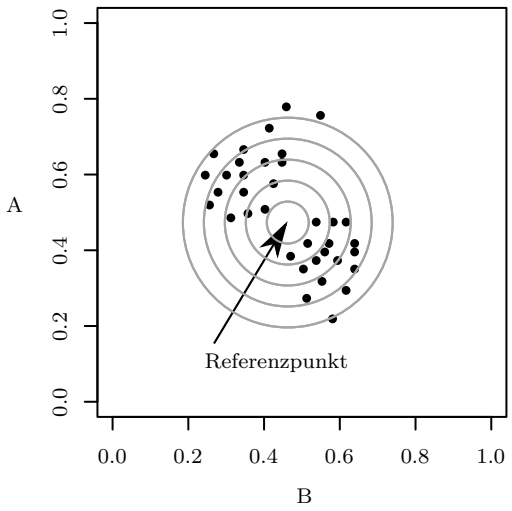
### 5.4.1 Modellierung des Merkmalsraums

Die Statistik bietet viele Werkzeuge, um einen Merkmalsraum zu modellieren und daraus Abnormitäten abzuleiten. Je nach Methode kann die Verteilung der Merkmalspunkte flexibler oder starrer beschrieben werden. Auf die Erkennung fehlerhafter Objekte hat die Modellierung großen Einfluss. Ein beliebig flexibles Verfahren (mit entsprechend vielen Parametern) kann jede Verteilung von Merkmalen durch ein Modell erklären. Befinden sich unter den untersuchten Merkmalen auch Merkmale fehlerhafter Geodatenobjekte, werden auch die eigentlich fehlerhaften Geodatenobjekte vom Modell berücksichtigt. Folglich gibt es keine ungewöhnlichen Objekte mehr und die Analyse wird bedeutungslos.

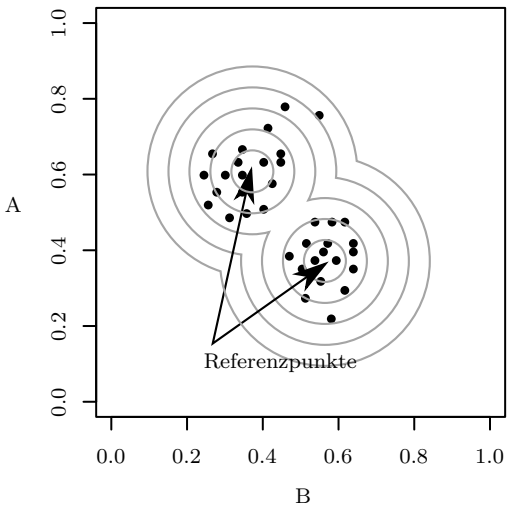
Abbildung 5.6 zeigt exemplarisch die Auswirkungen. Dargestellt sind Objekte in einem zweidimensionalen Merkmalsraum. Angenommen sei, dass es sich bei den Merkmalen um korrekte Geodatenobjekte handelt. Die Stichprobe ist also nicht gestört. Darstellung (a) zeigt die wohl einfachste Möglichkeit der Abnormitätsbestimmung. Es wird vorausgesetzt, dass sich alle betrachteten Objekte gleich verhalten. Folglich gibt es einen Punkt im Merkmalsraum, der das normalste Objekt beschreibt. Dieser wird als Referenzpunkt gewählt. Eine spezielle Abstandsbewertung wird nicht erwogen und daher der euklidische Abstand genutzt. Der einzige Parameter dieses Abnormitätsmodells ist daher die Wahl des Referenzpunktes.

Die Einschränkung auf einen eindeutigen Referenzpunkt für die Abnormitätsbewertung führt im Beispiel (a) dazu, dass der Referenzpunkt nicht in der Nähe eines Merkmals der Stichprobe liegt. Dadurch gibt es nur wenige normal eingeschätzte Objekte. Das widerspricht der Anschauung, nach der die Anzahl der Objekte mit zunehmender Abnormität abnehmen sollte. Die Konsequenz ist, dass fehlerhafte Objekte, die ein Merkmal aufweisen, das nahe dem Referenzpunkt liegt, nicht als ungewöhnlich eingestuft werden. Würde man statt einem zwei Referenzpunkte voraussetzen, könnte man mit der in (b) gezeigten Positionierung der Referenzpunkte plausiblere Ergebnisse erzielen. Bei mehr als einem Referenzpunkt wird als Abnormität für ein gegebenes Merkmal der Abstand zu dessen nächstgelegenen Referenzpunkt genutzt. Die sich ergebenden Abnormitäten sind wieder als Abstandslinien dargestellt. Je mehr Referenzpunkte angenommen werden, desto besser kann die Abnormitätsbewertung an die Stichprobe angepasst werden. Im Extremfall wird jeder Punkt der Stichprobe zum Referenzpunkt der Abnormitätsbetrachtung (Abbildung 5.6 (c)).

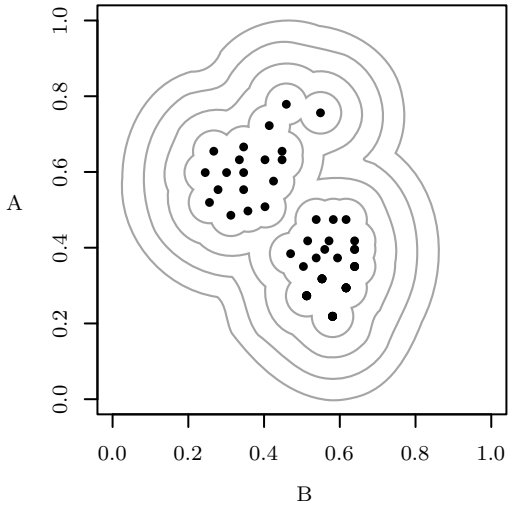
Auch die Bewertung der Abstände wirkt sich auf die Möglichkeiten zur Detektion fehlerhafter Geodatenobjekte aus. Die in (a), (b) und (c) genutzte euklidische Distanz führt



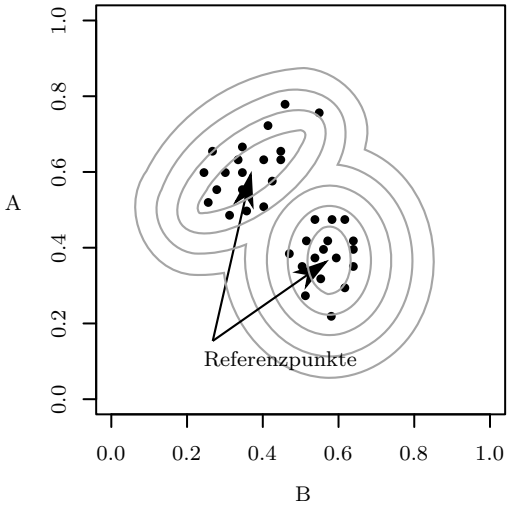

(a) 1 Referenzpunkt, euklidischer Abstand




(b) 2 Referenzpunkte, euklidischer Abstand




(c) Merkmale = Referenzpunkte, euklidischer Abstand




(d) 2 Referenzpunkte, Mahalanobis-Abstand

**Abbildung 5.6:** Abnormitäten einer Objektmenge mit zwei Merkmalen. Die Auswirkungen der Wahl der Referenzpunkte und die Bewertung der Abstände im Raum im Vergleich.

dazu, dass sich die Abnormität radial um die Referenzpunkte ausbreitet. Die Verbreitung der Abnormität hängt also nicht von den Merkmalen der Stichprobe ab. Entsprechend gibt es Bereiche mit geringer Abnormität, obwohl sich dort kein Stichprobenmerkmal befindet und gleichzeitig Bereiche mit zumindest einigen Merkmalen, aber hoher Abnormität. Darunter würde die Trennung von korrekten und falschen Geodatenobjekten leiden. Durch eine dynamische Bewertung der Abstände kann die Abnormitätsbewertung an die Stichprobe angepasst werden (Abbildung 5.6 (d)).

Der Einfluss der Bewertung der Abstände sinkt mit zunehmender Zahl an Referenzpunkten. Die Verteilung der Merkmalspunkte wird trotz der Nutzung der euklidischen Distanz in (c) besser erfasst als durch die angepasste Distanz in (d).

Die Möglichkeiten der Merkmalsraummodellierung unterscheiden sich im Detail um die Methoden, mit denen Referenzpunkte und Abstandsbewertungen realisiert werden.

### 5.4.2 Modellschätzung

Wie im vorigen Abschnitt gesehen ist eine gute Beschreibung der Verteilung der Merkmale im Merkmalsraum grundlegend für die Normalitätsbewertung von Geodatenobjekten. Je mehr Referenzpunkte und je detaillierter die Abstandsbewertung, desto besser kann die Verteilung der Merkmale angenähert werden.

Verfahren wie *Mahalanobis-Abstand*, *KMeans*, *Mean-Shift*, *Gaussian Mixture Models* oder *Ein-Klassen*-SVM stehen bereit, um eine solche Beschreibung zu ermitteln. Sie bieten unterschiedliche Möglichkeiten der Beschreibung von Merkmalsverteilungen. Von der Merkmalsverteilung ausgehend, kann schließlich die Abnormitätsbewertung definiert werden, wie in Abbildung 5.6 exemplarisch dargestellt.

Die Aufgabe der Modellschätzung ist die Auswahl eines geeigneten Verfahrens und die Parametrisierung des Verfahrens. Da die Einflussgrößen für eine Modellierung der Merkmalsverteilung unbekannt sind,müssen sie aus den gegebenen Daten statistisch geschätzt werden. Die Güte einer statistischen Schätzung ist abhängig von folgenden Faktoren:

- Das zugrunde gelegte statistische Modell. Die Schätzung kann nur so gut werden wie das gewählte Modell der Realität entspricht. Das Modell wird vorausgesetzt.
- Die Größe der Stichprobe, die zur Schätzung genutzt wird. Je größer die Stichprobe, desto besser können die Parameter des statistischen Modells geschätzt werden.
- Die Streuung der Stichprobe. Sind die Merkmale der Stichprobe über einen großen Bereich gestreut, benötigt man eine entsprechend große Stichprobe, um die Parameter zuverlässig zu schätzen.

Die Entscheidung für eine spezifische Methodik ist also eine Abwägung zwischen den Modellierungsmöglichkeiten (Abschnitt 5.4.1) und den Anforderungen der Parametrisierung.

Im Fall der Anwendung wird die Schätzung dadurch erschwert, dass von der Merkmalsmenge nicht bekannt ist, welche Merkmale von korrekten und welche von falschen Geodatenobjekten stammen.

Die für die Schätzung gewünschte Stichprobe entspricht den Merkmalen korrekter Geodatenobjekte. Die Positionen der Merkmale dieser *Basismenge* hängt von der Art der Merkmale und den Eingangsdaten ab. Die fehlerhaften Geodatenobjekte stören die Stichprobe.

Dies schließt die Nutzung eines zu flexiblen Modells für die Merkmalsverteilung aus. Ein zu allgemeines Modell kann alle Eventualitäten der Merkmalsverteilungen nachbilden und so nicht mehr zwischen gewöhnlichen und ungewöhnlichen Merkmalen unterscheiden.

Ein Modell mit wenigen Parametern ist robuster zu schätzen, aber auch vergleichsweise starr in seiner Anwendung. Das Risiko eines zu starren Modells liegt schließlich darin, dass Bereiche im Merkmalsraum als normal gewertet werden, in denen keine Merkmale diese These unterstützen. Ein Beispiel ist in Abbildung 5.6 (a) gezeigt, indem die Einschränkung auf ein einzelnes Cluster dazu führt, dass der normalste Bereich des Merkmalsraums keine Merkmale aufweist.

Die Prämisse der Merkmalsraummodellierung lautet daher, ein so flexibles Modell wie nötig, aber ein so starres Modell wie möglich zu wählen. Da mit der Flexibilität der Modellierung auch die Ansprüche an die statistische Schätzung der Parameter steigt, sind flexible Methoden nur bei Hinweisen auf relevante Verbesserungen sinnvoll.

#### Bewertung der Modellanpassung

Um einzuschätzen, ob das Modell zu starr ist, ist es plausibel, die Verteilung der Abnormitäten der Merkmalsmenge zu betrachten. Eine Darstellung findet sich in Abbildung 5.7. Zwei schon aus Abbildung 5.6 bekannte Beispiele für eine Anpassung sind herausgegriffen: die Beschreibung der Merkmalsmenge mit zwei Referenzpunkten (linke Spalte von Darstellungen) und mit einem Referenzpunkt (rechte Spalte). Die zwei Referenzpunkte beschreiben die Verteilung offensichtlich besser als der Fall mit einem Referenzpunkt. Unter den Darstellungen ist die *empirische* Dichte angegeben. Sie entspricht der Häufigkeit von Merkmalen für eine gegebene Abnormität. Die Darstellung zeigt nur ein schematisches Beispiel. Die endliche Menge an Merkmalen würde in der Praxis zu einer diskreten Funktion führen. Zur besseren Verdeutlichung wurde in der Abbildung allerdings eine kontinuierliche Approximation gewählt.

Die empirische Dichte zeigt im linken Beispiel mit zwei Referenzen einen guten Verlauf. Sie steigt sofort stark an, da sich viele Merkmale nahe der Referenzpunkte befinden. Im rechten Beispiel befindet sich das Maximum der empirischen Dichte deutlich weiter rechts, da es keine Merkmale in der Nähe des Referenzpunktes gibt. Da der Referenzpunkt aber genau in der Mitte der Merkmale liegt, gibt es trotzdem nur ein einzelnes ausgeprägtes Maximum.

Weil die empirische Dichte hinsichtlich einer endlichen Menge von Merkmalen bestimmt wird, liegt sie stets als diskrete Form vor. Um auch in diesem Fall eine genaue grafische Darstellung zu erhalten, wird die *empirische Verteilungsfunktion* genutzt. Ihr Wert für eine Abnormität $n$ entspricht der Anzahl der Geodatenobjekte mit einer mindestens $n$ großen Abnormität. Sie ist daher immer kontinuierlich. Im Beispiel ist die empirische Verteilungsfunktion jeweils unter den empirischen Dichten dargestellt.

### 5.4.3 Mahalanobis-Abnormität

Ein einfaches Modell des Merkmalsraums betrachtet die Entitäten des Merkmalsraum als homogene Gesamtmenge. Ein solcher Fall ist etwa in Abbildung 5.9 dargestellt. Erneut handelt es sich um den Merkmalsraum der 255 Waldobjekte. Die Darstellung erfolgt über die ersten zwei Hauptkomponenten einer PCA-Analyse (Abschnitt 3.1.2). Drei Merkmale

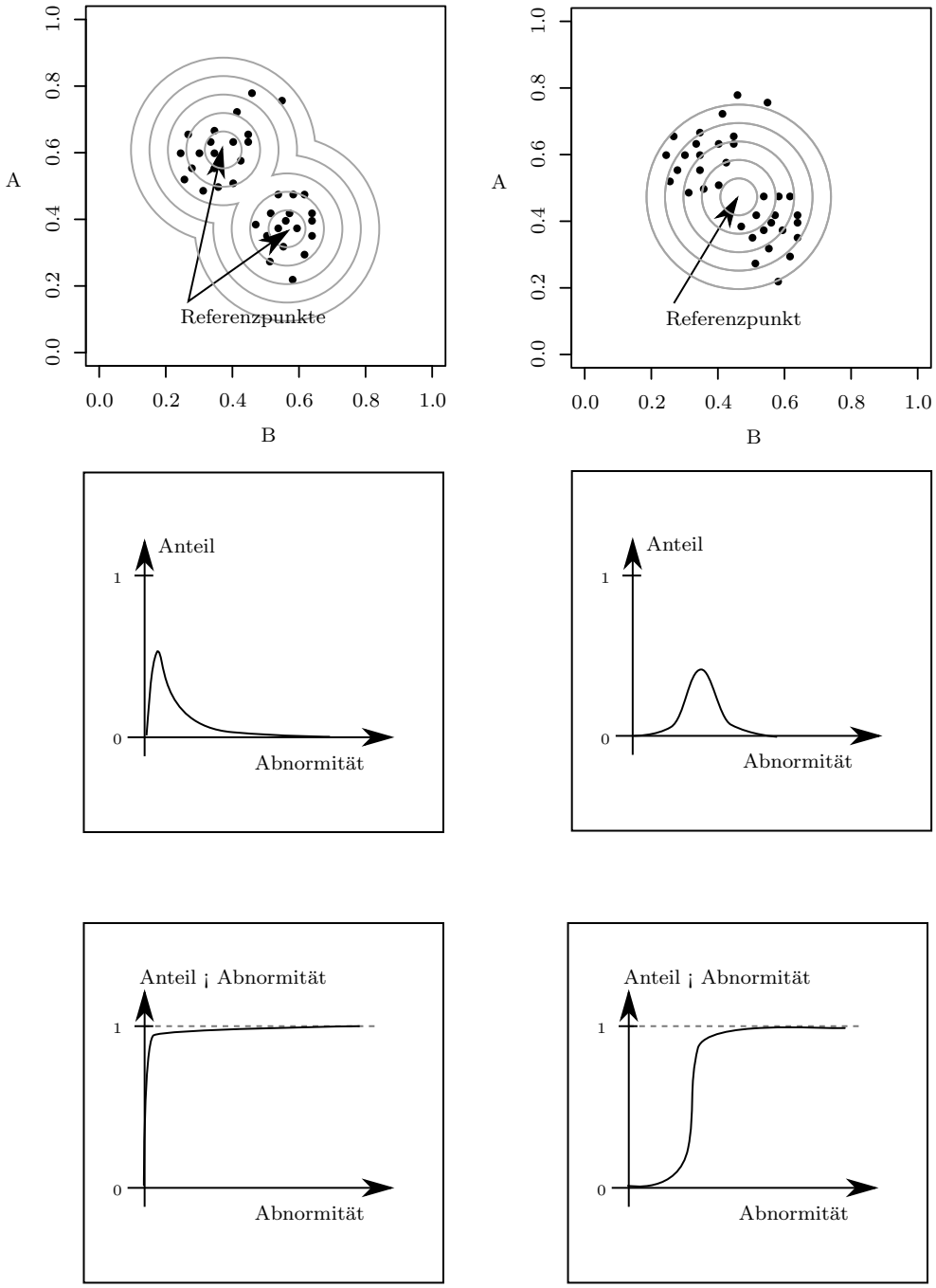



**Abbildung 5.7:** Beispiel für die Auswirkungen der Merkmalsraummodellierung auf die Abnormitätsverteilung. Im linken Fall beschreiben zwei Referenzpunkte die Punktmenge. Im rechten Fall verallgemeinert der einzelne Referenzpunkt die Verteilung zu stark. Die empirische Dichte ist jeweils unter den Merkmalsraumdarstellungen angegeben. Zuunterst finden sich die stattdessen häufig betrachteten empirischen Verteilungsfunktionen.


**Abbildung 5.8:** Eine Wiesenfläche befindet sich innerhalb von Objekt *B*. Objekt *A* besteht ausschließlich aus Bäumen.

sind zur bessere Orientierung markiert. Sie entstammen den Geodatenobjekten aus Abbildung 5.8. Das Merkmal $A + B$ entspricht der Vereinigung der beiden abgebildeten Geodatenobjekte.

In diesem Kapitel sollte man sich in Erinnerung rufen, dass die PCA-Analyse durch die Reduktion eine Vereinfachung des Merkmalsraums vornimmt. In den meisten Fällen dürfte eine Konzentration von Merkmalen im PCA-Raum auch auf eine Konzentration im eigentlichen Merkmalsraum hinweisen. Diese Schlussfolgerung und ihr Umkehrschluss sind jedoch nicht mathematisch garantiert.

Die Darstellung in Abbildung 5.9 legt nahe, dass sich die meisten Geodatenobjekte in der Nähe von Beispielobjekt $A$ befinden und dort eine Häufung bilden. Etliche Merkmale liegen zwar außerhalb, aber sie bilden keine eigene Häufung. Die Merkmale außerhalb der Häufung entsprechen nach dem Konzept den fehlerhaften Geodatenobjekten. Je weiter entfernt ein Merkmal von der Häufung liegt, desto eher wird es als fehlerhaftes Objekt eingeschätzt. Das fehlerhafte Objekt $B$ ist entsprechend weiter von der Häufung entfernt als das Objekt $A + B$.

Für den Fall einer einzelnen Häufung kann die in Abschnitt 3.2.3 beschriebene *Mahalanobis-Distanz* genutzt werden, um Abstände in diesem Merkmalsraum zu messen. Sie geht davon aus, dass sich die Verteilung der Merkmale durch eine durch alle Punkte geschätzen Mittelwert $\mu$ und eine Kovarianz $\Sigma$ zusammenfassen lassen.

Der Mittelwert steht für das Zentrum der Menge. Er kann als Referenzpunkt genutzt werden, um für jedes Geodatenobjekt $o$ den Abstand zur dominierenden Mehrzahl an Geodatenobjekten zu bestimmen:

$$abn_{\text{Mah}}(o) := d_{\text{Mah}}(o, \mu, \Sigma) \tag{5.3}$$

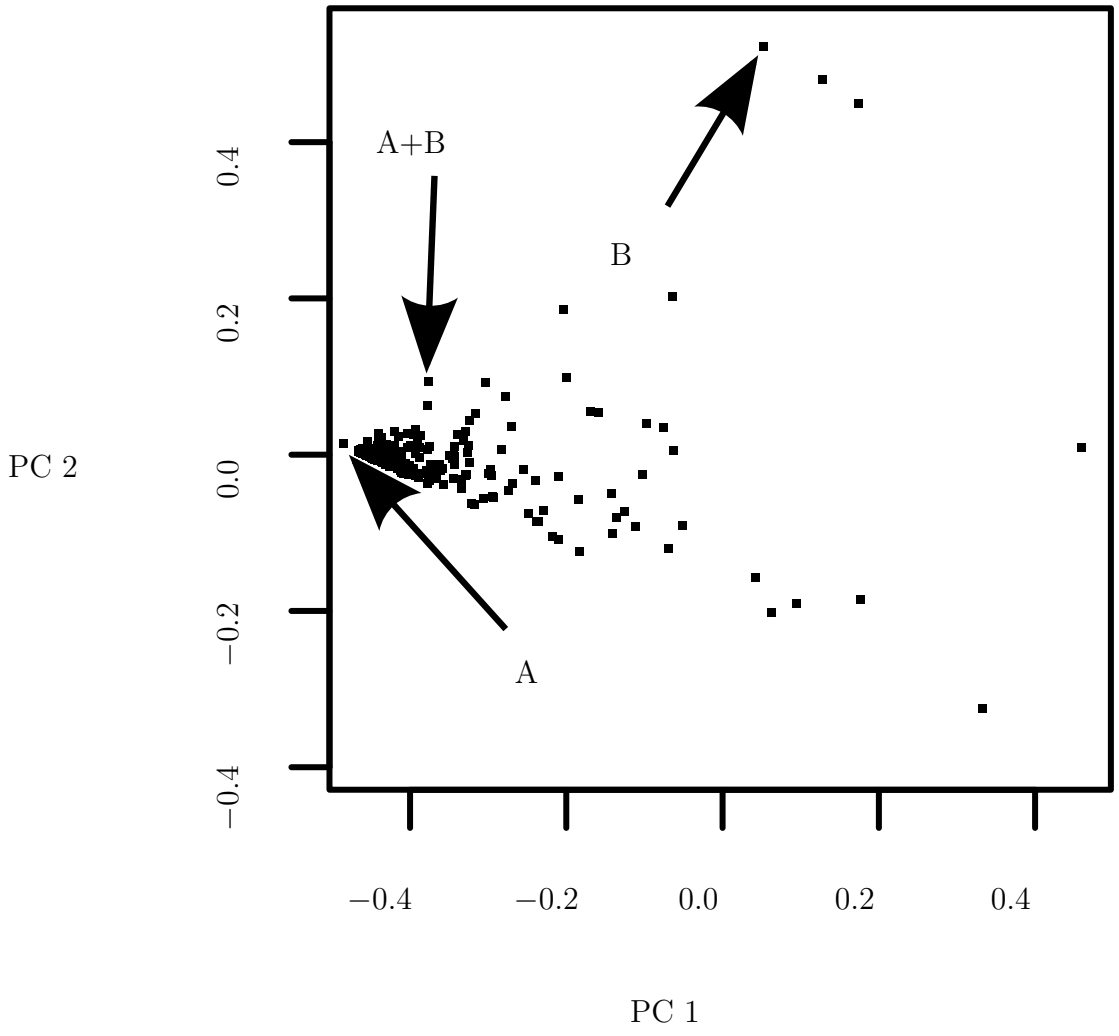



**Abbildung 5.9:** Eine PCA-Ansicht des Merkmalsraums für Waldobjekte mit dem Landbedeckungshistogramm als Merkmal. Die markierten Merkmale entsprechen den Objekten aus Abbildung 5.8.

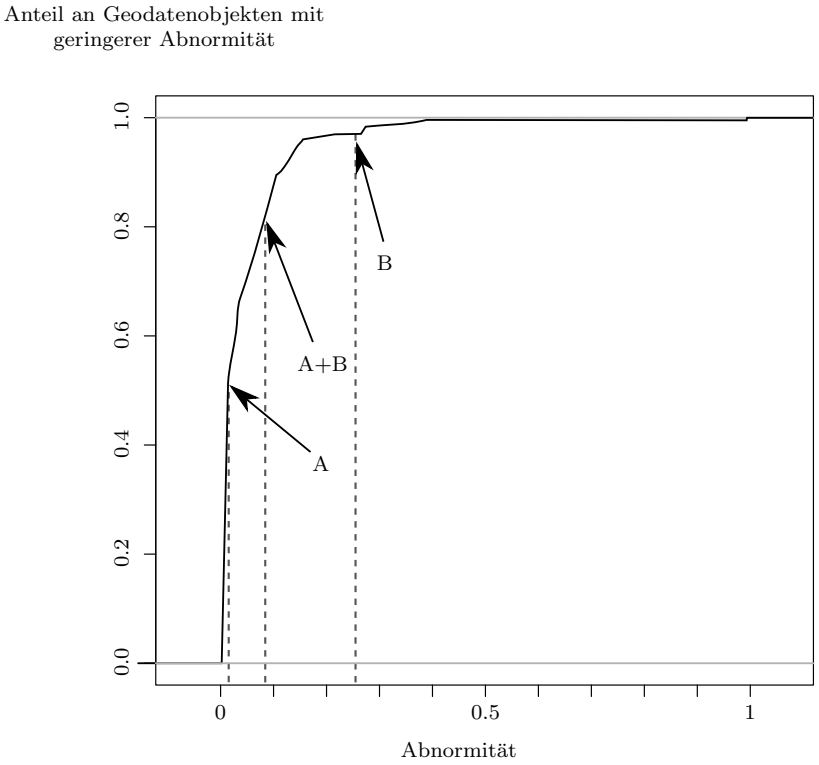



**Abbildung 5.10:** Empirische Verteilungsfunktion der normierten Abnormitätswerte aller 255 Wald-Geodatenobjekte. Für jeden Abnormitätswert ist der Anteil an Geodatenobjekten angegeben, deren Abnormität kleiner als der jeweilige Wert ist. Die Abnormitätswerte der drei Beispielobjekte sind markiert.

Es wird für die Geodatenobjekte also nicht entschieden, ob sie zum Cluster gehören. Für jedes Geodatenobjekt wird stattdessen der Abstand zum Mittelwert ermittelt.

Eine Einschätzung der Abnormität eines Geodatenobjekts kann nur relativ zu den Abnormitäten aller Geodatenobjekte erfolgen. Daher ist eine Normierung der Abnormitäten sinnvoll, sodass alle Abnormitäten zwischen 0 und 1 liegen. In Abbildung 5.10 ist die normierte Abnormität der schon bekannten 255 Waldobjekte dargestellt. Die Darstellung erfolgt über eine *empirische Verteilungsfunktion*. Auf der Abszisse ist die normierte Abnormität dargestellt. Gegen sie ist auf der Ordinate der Anteil der Geodatenobjekte aufgetragen, deren Abnormitäten geringer ist als der zugehörige Wert auf der Abszisse. Der Verlauf der Kurve steigt zunächst steil an. Dies rührt daher, dass sehr viele Objekte sich nahe des Mittelwertes befinden. Entsprechend viele Objekte haben eine sehr geringe Abnormität. Selbst das eigentlich tadellose Beispielobjekt $A$, das nur aus Bäumen besteht, hat eine Abnormität, die von der Hälfte der Geodatenobjekte übertroffen wird. Der Grund:

Viele Wald-Geodatenobjekte enthalten Wiesenanteile. Dies kann man gut in Abbildung 5.5 nachvollziehen. Daher ist ein Geodatenobjekt ganz ohne Wiese ungewöhnlicher als Geodatenobjekte, die zu einem kleinen Teil aus Wiese bestehen. Dieses Argument spricht für die Richtigkeit der Abnormitätsbewertung.

Die Mahalanobis-Abnormität ist auf einen einzigen Referenzpunkt beschränkt. Die Beschränkung erlaubt allerdings eine robuste Schätzung und auch die Bewertung der Abnormitäten.

### 5.4.4 Untersuchungen zu weiteren Modellierungsmethoden

Da die fehlerhaften Geodatenobjekte häufig dadurch falsch sind, dass sie teilweise eine falsche Fläche aufweisen, liegen die Merkmale der falschen Geodatenobjekte oft dicht an den Merkmalen normaler Geodatenobjekte. Eine sichere Unterscheidung von korrekten und falschen Geodatenobjekten ist somit nicht zuverlässig möglich. Damit verbietet sich eine genaue Optimierung der Modellanpassung über grobe Verbesserungen hinaus. Die in Abschnitt 5.4.2 dargestellte Möglichkeit zur Prüfung der Verteilung der Abnormitäten ist daher eine geeigneter Indikator für den Grad der Modellanpassung.

Wie in Kapitel 6 anhand umfangreicher Testdaten demonstriert, bringt die Mahalanobis-Abnormität bereits bei der Nutzung eines einzelnen Referenzpunktes überzeugende Ergebnisse. Die Anwendung flexiblerer Modellierungsverfahren zeigen trotz umfangreicher Untersuchungen dagegen keine besseren Ergebnisse als die Mahalanobis Abnormität.

## 5.5 Objektbasierte Fehlererkennung

Die in Kapitel 4 beschriebenen existierenden Ansätze markieren Geodatenobjekte in Folge ihrer Analyse als korrekt oder falsch. Die so als falsch markierten Geodatenobjekte werden zur Nachkontrolle oder Korrektur an eine (manuelle) Bearbeitung weitergereicht. Auch mit dem neuen System ist eine solche Objektmarkierung möglich.

Wie gesehen, kann jedem Geodatenobjekt durch das Abnormitätsmodell eine kontinuierliche Bewertung zugeordnet werden. Für die Weiterverarbeitung wird die Liste der Geodatenobjekte zunächst anhand der Bewertung von gewöhnlich zu ungewöhnlich sortiert. Schließlich müssen die Geodatenobjekte nach einer Strategie aufgrund ihrer Bewertung als (potentiell) fehlerhaft markiert werden.

Es wird ein Schwellwert $c$ mit $0 < c \leq 100$ vorgegeben und es werden die ungewöhnlichsten $c$ Prozent Geodatenobjekte einer Prüfung unterzogen. Da durch die Vorsortierung der Daten nach dem Gütemaß die zu erwartende Fehlerfindungsrate höher ist, als bei einer zufällig ausgewählten Stichprobe, kann $c$ kleiner gewählt werden, als die sinnvolle Stichprobengröße. Dadurch verringert sich die Zahl der zu prüfenden Geodatenobjekte bei gleichzeitig höherer Trefferquote für fehlerhafte Geodatenobjekte.

## 5.6 Weitere Eigenschaften der Abnormitätsanalyse

Die kontinuierliche Bewertung eines Szeneobjekts eröffnet neue Möglichkeiten der Anwendung. So können kontinuierliche Maße etwa genutzt werden, um Änderungen an einzelnen Geodatenobjekten zu bewerten und so iterativ die Szene zu optimieren. In diesem Zusammenhang könnte von Interesse sein, neben der eigentlichen Abnormität, die *Richtung* der Abnormität zu berücksichtigen, das *Verbesserungspotential*.

Abbildung 5.11 verdeutlicht das Prinzip. Die Abnormität eines Geodatenobjekts entspricht dem Abstand vom Merkmal zum Mittelpunkt $\mu$. Dies entspricht der Länge des Richtungsvektors $\mu - \mathbf{M}$ vom Merkmal $\mathbf{M}$ zum Mittelpunkt (in der Abbildung als Pfeil eingezeichnet). Der Richtungsvektor kann darüber hinaus komponentenweise betrachtet werden: In jeder Komponente ist jeweils der Anteil an Landbedeckung abzulesen, den das Geodatenobjekt zu viel oder zu wenig hat. In der Abbildung sind die Anteile von Wald ($\Delta$ Wald) und Wiese ($\Delta$ Wiese) gestrichelt eingezeichnet.

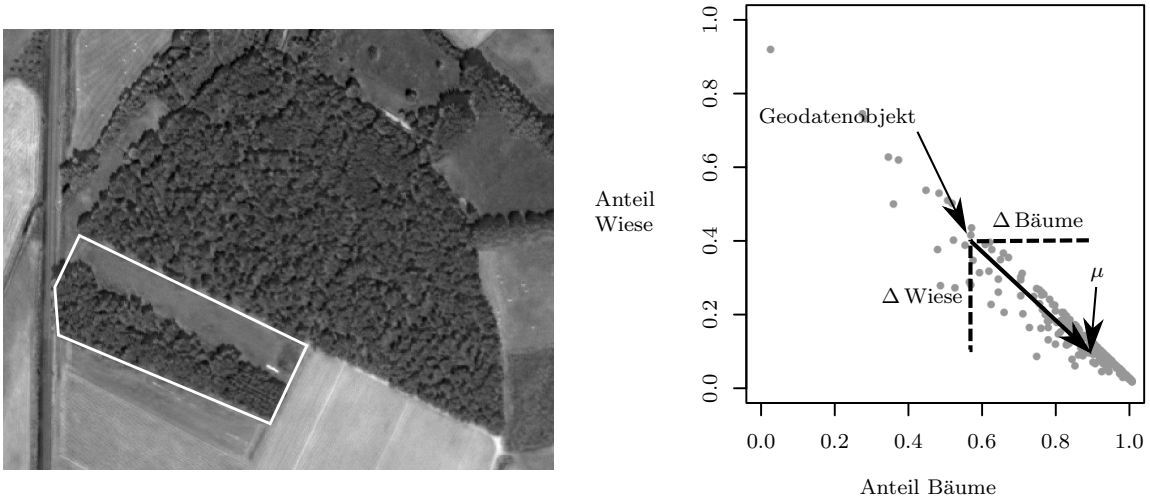



**Abbildung 5.11:** Im linken Bild eine Ansicht des im Merkmalsraum (rechts) markierten Wald-Geodatenobjekts (im Bild abgegrenzt durch eine durchgezogene Linie). Ein Vergleich des Merkmals mit dem Mittelwert aller Wald-Geodatenobjekte $\mu$ ergibt, dass das Geodatenobjekt zu wenig Wald und zu viel Wiese aufweist.

Bleibt man bei dem Beispiel, käme für eine Verbesserung der Abnormität des Geodatenobjekts die Abgabe der Wiesenfläche oder die Hinzunahme von Bäumen in Frage. Beides würde zu einer Verbesserung der Bewertung führen. Das Verbesserungspotential alleine ist allerdings nicht ausreichend um eine Korrektur eines Geodatenobjekts einzuleiten, wie man an folgendem Beispiel erkennen kann: Betrachtet sei ein Waldobjekt, das nur aus Bäumen bestehe. Da es viele Waldobjekte mit Lichtungen, also einem Wiesenanteil gibt, würde für dieses Objekt ein zu viel an Bäumen und zu wenig an Wiese erkannt werden. Hier würde die Abgabe von Bäumen allerdings nichts an der Position im Merkmalsraum ändern, da das Objekt auch dann noch vollständig aus Bäumen bestünde. Ergänzende Richtlinien scheinen also notwendig.

Eine Möglichkeit wäre, räumlichen Kontext zu berücksichtigen, indem benachbarte Geodatenobjekte und deren Bewertung in die Überlegungen mit einbezogen werden. So könnte etwa ein Mangel an Bäumen in einem Geodatenobjekt und ein Überschuss an Bäumen in einem anderen benachbarten Geodatenobjekt als Hinweis für sinnvolle Änderungen genutzt werden. Ein Vergleich ist aber schwierig, wenn sich die Objektklassen in ihrer Landbedeckung sehr ähneln. Dann ist auch das Verbesserungspotential wenig aussagekräftig.

Es bleibt festzuhalten, dass die Abnormitätsbewertung zusammen mit der Bestimmung des Verbesserungspotentials eine wichtige Grundlage für optimierende Verfahren darstellen kann. Allerdings wäre eine ergänzende Steuerung etwa durch ein regelbasiertes System notwendig.

# 6 Leistungsabschätzung und experimentelle Untersuchung des Verfahrens

In diesem Kapitel werden die entwickelten Algorithmen anhand von Testdatensätzen ausgewertet. Anhand der Testdatensätze wird überprüft, inwieweit die impliziten Annahmen der Ansätze zutreffen und wie die kontinuierliche Objektbewertung für die Fehlererkennung in Geodaten einzusetzen ist.

## 6.1 Testdaten

Um die Eigenschaften der entwickelten Verfahren zu untersuchen, ist es notwendig sie unter möglichst realistischen Bedingungen zu testen. Getestet werden daher Geodatenobjekte in tatsächlich existierenden Szenen. Eine Bewertung wird durch den Vergleich des ermittelten Ergebnisses mit einem Referenzergebnis ermöglicht.

Da es es keine öffentlich verfügbaren Testdatensätze, die die Bewertung einer objektweisen Geodatenprüfung ermöglichen würden, gibt wurden für die hier dargestellten Ergebnisse zwei ausführliche Testdatensätze erstellt. Tabelle 6.1 zeigt eine Übersicht der Datengrundlage der Testdatensätze.

### 6.1.1 Geodaten

Die Geodaten der Testszenen entstammen ATKIS Daten (Basis-DLM). Eine Darstellung der Geodaten ist Abbildung 6.1 zu entnehmen.

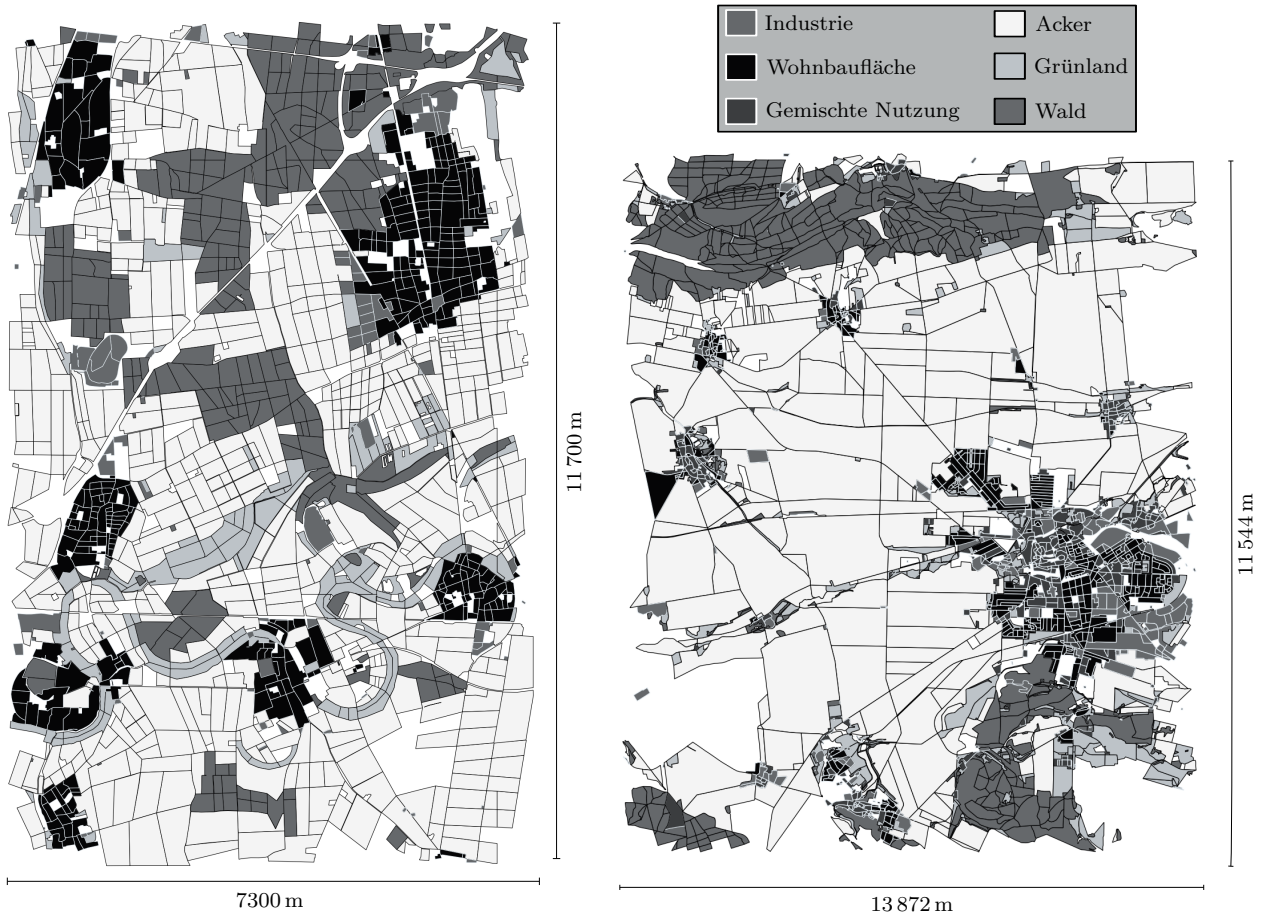
Die Szene *Weiterstadt* zeichnet sich durch einen bemerkenswert vielseitigen Szeneninhalt aus. Nahezu sämtliche Flächen werden wirtschaftlich genutzt. Acker- und Grünlandflächen werden unterbrochen durch Nutzwälder und mehrere Siedlungsgebiete mit Gewerbe und Industrie. Für den Test wurden die vorherrschenden Objektklassen *Wohnbaufläche* mit der Kennzeichnungsnummer 2111, *Industrie- und Gewerbefläche* (2112) (kurz: *Industrie*), *Ackerland* (4101) (kurz: *Acker*) und *Wald* (4107) genutzt. Geodatenobjekte der Objektartklasse *Grünland* (4102) wurden nicht zur Ergebnisdarstellung herangezogen, da sie von den Eigenschaften her sehr der Objektklasse Ackerland ähneln und in Tests sehr ähnliche Ergebnisse gezeigt haben.

Die Szene *Halberstadt* ist fast doppelt so groß wie die Szene Weiterstadt. Die Objektklassen wechseln sich verglichen mit der Weiterstadtszene deutlich weniger ab. Die Szene wird bestimmt durch große zusammenhängende Ackerflächen und Waldgebiete an den Rändern der Szene und der Stadt Halberstadt.

Zusätzlich zu den bekannten Objektklassen aus der Weiterstadt-Szene tritt hier auch die Objektklasse *Gemischte Nutzung* (2113) auf. Objekte dieser Objektklasse definieren sich

**Tabelle 6.1:** Übersicht über die Datengrundlage der Testszenen.

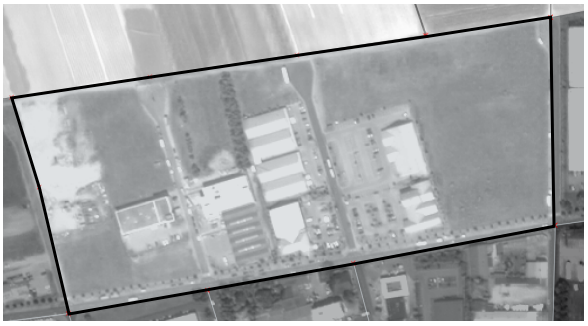
| Bezeichnung | Fläche | Objektklassen und Anzahl | | | |
|---|---|---|---|---|---|
| | | id | Bezeichnung | Anzahl | Fehler |
| Weiterstadt | 85 km$^2$ | 2111 | Wohnbaufläche | 353 | 20 |
| | | 2112 | Industrie | 102 | 29 |
| | | 4101 | Acker | 581 | 21 |
| | | 4107 | Wald | 268 | 13 |
| | | **Summe** | | **1304** | **81** |
| | | id | Bezeichnung | Anzahl | Fehler |
| Halberstadt | 160 km$^2$ | 2111 | Wohnbaufläche | 440 | 16 |
| | | 2112 | Industrie | 174 | 6 |
| | | 2113 | Gem. Nutzung | 434 | 16 |
| | | 4101 | Acker | 302 | 17 |
| | | 4107 | Wald | 522 | 12 |
| | | **Summe** | | **1871** | **54** |



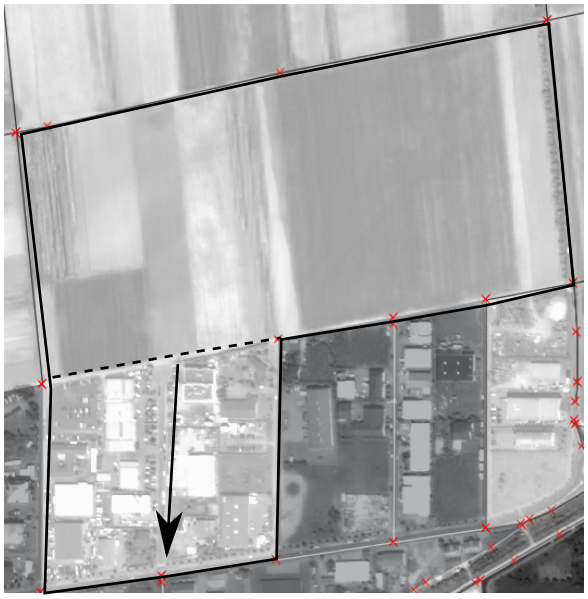

(a) Weiterstadt

(b) Halberstadt

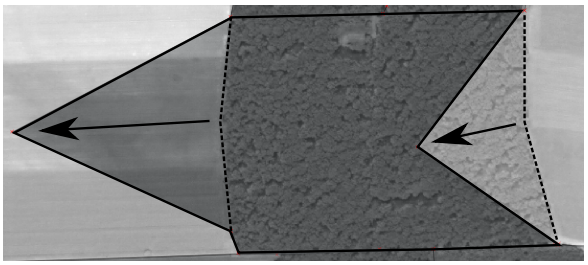
**Abbildung 6.1:** Darstellung der Geodaten der Testszenen.



(a) Fehlerhaftes Ackerobjekt in Originaldaten



(b) Manuelle Änderung an Ackerobjekt



(c) Manuelle Änderung an Wald- und Ackerobjekt

**Abbildung 6.2:** Präparierung der Testszenen durch manuelle Fehlereinbringung durch Nachahmung realer Fehler.

darüber, dass in dem Gebiet eine Vielzahl von Landnutzungen vorkommt, die sich jede für sich nicht als spezifischere Objektklasse ausweisen lässt. Diese treten in der Testszene fast ausschließlich im Stadtgebiet auf.

Im Originalzustand weisen die Geodaten der Testszene nur sehr wenige Fehler auf. Um eine aussagekräftigere Bewertung der Ergebnisse zu erlauben, wurden daher die Geodaten manuell weiter verfälscht. Um realistische Fehlerbilder zu erhalten, wurden hierzu einzelne Geodatenobjekte durch Verschiebungen der Grenzen abgewandelt. Das Vorgehen ist in Abbildung 6.2 dargestellt. Abbildung (a) zeigt einen tatsächlichen Fehler in den Daten (ohne manuelle Eingriffe). Auf dem Ackerobjekt befinden sich Gebäude, vermutlich Gewerbeflächen. Darstellungen (b) und (c) zeigen das Vorgehen für Veränderungen an realen Geodatenobjekten um fehlerhafte Geodatenobjekte zu erhalten. Die Anzahl an Gesamtfehlern pro Testszene und Objektklasse ist in Tabelle 6.1 einzusehen. Ohne manuelle Fehlereinbringung weisen die Szenen nur einstellige Fallzahlen auf.

**Tabelle 6.2:** Eigenschaften der IKONOS-Bilddaten. Die geometrische Auflösung bezieht sich auf eine senkrechte Sicht auf die Erdoberfläche.

| Kanal | Geometrische Auflösung | Spektralbereich |
|---|---|---|
| PAN | 0,82 m | 450 nm to 900 nm |
| Rot | 3,28 m | 632 nm to 698 nm |
| Grün | 3,28 m | 506 nm to 595 nm |
| Blau | 3,28 m | 445 nm to 516 nm |
| NIR | 3,28 m | 757 nm to 853 nm |

### 6.1.2 Fernerkundungsdaten

Der Satellit IKONOS (Orbitalhöhe: 681 km, sonnensynchrone Umlaufbahn) trägt mehrere Kameras für verschiedene geometrische Auflösungen und Spektralbereiche [10] (Tabelle 6.2). Die Daten wurden einer Vorverarbeitung unterzogen, sodass im Rahmen einer nicht näher bekannten *Pansharpening*-Filterung [2] die geometrische Auflösung aller Kanäle auf einheitliche 1 m gebracht wurde.

Einen Überblick über die Testszenen in den Fernerkundungsdaten gibt Abbildung 6.3. Einen Eindruck der Detailfülle gibt Abbildung 6.3(c).

### 6.1.3 Qualitätsmaße

Um Aussagen über die Erkennungsleistung der neuen Verfahren zu erhalten, werden die Ergebnisse der Algorithmen gegen Ergebnisse einer Referenz verglichen.

Die Fragestellung an die Referenz lautete für ein gegebenes Geodatenobjekt: *Ist dies ein fehlerhaftes Geodatenobjekt?* Die Antwort der Referenz mit der entsprechenden Entscheidung des Systems verglichen.

Vier Fälle inwieweit Antworten von Referenz und System überein stimmen, können unterscheiden werden:

| | | Laut Referenz | | Σ System |
|---|---|---|---|---|
| | | Fehlerhaft | Nicht fehlerhaft | |
| Laut System | Fehlerhaft | *tp* | *fp* | 👍$_S$ |
| | Nicht fehlerhaft | *fn* | *tn* | 👎$_S$ |

Die Kürzel *tp*, *fp*, *tn*, *fn* stehen dabei für die Bedeutungen *tatsächlich fehlerhaft* (engl. *true positives*), *irrtümlich fehlerhaft* (engl. *false positives*), *irrtümlich nicht fehlerhaft* (engl. *false negatives*) und *tatsächlich nicht fehlerhaft* (engl. *true negatives*).

Die in Kapitel 5 präsentierte objektbasierte Abnormitätsbewertung antwortet statt mit einer *Ja/Nein* Entscheidung mit einem kontinuierlichen Gütewert. In diesem Fall hängen die Werte von einem festzusetzendem Gütewert ab, der als Entscheidungsgrenze zwischen als richtig und fehlerhaft zu wertenden Geodatenobjekten dient: Sei $\mathcal{O}$ die Menge aller Geodatenobjekte und $k >= 0$ ein beliebiger festzusetzender Schwellwert, dann ist

$$👍_S(k) = |\{o \in \mathcal{O} | abn(o) \geq k\}| \quad \text{und} \quad 👎_S(k) = \left|\mathcal{O} \setminus 👍_S(k)\right|$$

Die Darstellung der Erkennungsleistung erfolgt über einen Graph, der die verschiedenen Kennzahlen für alle möglichen Schwellwerte darstellt.

(a) Weiterstadt

(b) Halberstadt

(c) Detailansicht

**Abbildung 6.3:** Die Testszenen in den Fernerkundungsdaten.

**Tabelle 6.3:** Bewertung der Ausgangslage, alle Geodatenobjekte als fehlerhaft anzunehmen.

| | Wohnbaufläche | Industrie | Acker | Wald | Gem. Nutzung |
|---|---|---|---|---|---|
| **Weiterstadt** | | | | | |
| *Precision* | 0,06 | 0,28 | 0,04 | 0,05 | |
| *Recall/tpr* | 1 | 1 | 1 | 1 | |
| *fpr* | 1 | 1 | 1 | 1 | |
| **Halberstadt** | | | | | |
| *Precision* | 0,03 | 0,03 | 0,03 | 0,05 | 0,02 |
| *Recall/tpr* | 1 | 1 | 1 | 1 | 1 |
| *fpr* | 1 | 1 | 1 | 1 | 1 |

Jeder der vier grundlegenden Werte *tp, fp, tn, fn* ist für sich genommen wenig aussagekräftig. Erst durch ein Abwägen der einzelnen Kennzahlen lassen sich sinnvolle Aussagen treffen.

**Precision/Recall** Eine übliche Vorgehensweise ist der Vergleich der Kennwerte *Precision* (engl. *für Genauigkeit*) und *Recall* (engl. *für Trefferquote*).

$$Precision = \frac{tp}{tp + fp} = \frac{tp}{\text{Laut System fehlerhaft}} \tag{6.1}$$

$$Recall = \frac{tp}{tp + fn} = \frac{tp}{\text{Laut Referenz fehlerhaft}} \tag{6.2}$$

Während *Precision* den Anteil der gefundenen tatsächlich fehlerhaften Geodatenobjekten im Verhältnis zur Gesamtzahl der als fehlerhaft eingeschätzten Geodatenobjekte angibt, drückt *Recall* den Anteil der gefundenen falschen Geodatenobjekte im Verhältnis zu allen fehlerhaften Geodatenobjekten aus.

Auch hier gilt wieder: Jeder Wert für sich betrachtet erlaubt keine sinnvolle Ergebniseinschätzung. Für einen maximalen *Recall*-Wert reicht es aus, alle Geodatenobjekte als falsch anzunehmen, der *Precision*-Wert dagegen betrachtet nicht, wie viele der fehlerhaften Geodatenobjekte vom System *nicht* gefunden wurden.

Die Kennzahlen lassen sich für jede Untermenge der Gesamtobjektmenge bestimmen, deren Geodatenobjekte als fehlerhaft angenommen werden. Ohne eine Vorauswahl von Geodatenobjekten durch ein System gilt jedes Geodatenobjekt als potentiell fehlerhaft.

Die Kenngrößen $\text{👍}_S$ und $\text{👎}_S$ lassen sich aus den Objektmengen aus Tabelle 6.1 herleiten: Die Werte für die Testszenen sind in Tabelle 6.3 aufgeführt. Wie man sieht, schwankt der *Precision*-Wert stark, da die Häufigkeit von fehlerhaften Geodatenobjekten sich unterscheidet. Dies erschwert den Vergleich von Ergebnissen.

**ROC: tpr/fpr** Eine Alternative zur *Precision/Recall*-Betrachtung ist statt der *Precision*, die *False Positive Rate* (*fpr*) zu nutzen. Der *Recall*-Wert wird in diesem Kontext auch *True Positive Rate* (*tpr*) genannt:

$$tpr = Recall = \frac{tp}{tp + fn} = \frac{tp}{\text{Laut Referenz fehlerhaft}} \tag{6.3}$$

$$fpr = \frac{fp}{tn + fp} = \frac{fp}{\text{Laut Referenz } \textit{nicht} \text{ fehlerhaft}} \tag{6.4}$$

Aus historischen Gründen wird die *tpr/fpr*-Gegenüberstellung auch als engl. *Receiver-Operating-Characteristic* (ROC) bezeichnet.

Der *tpr/fpr*-Vergleich ist unabhängig von der Häufigkeit des Auftretens der Fehlerklasse: Die Werte liegen für die Ausgangslage (Tabelle 6.3) jeweils bei 1. Da bei der Prüfung von Geodaten fehlerhafte Geodatenobjekte nur sehr selten vorkommen (Tabelle 6.1), ist er dem *Precision/Recall*-Vergleich überlegen [13].

Ausgehend von einer ROC-Kurve, bietet sich die Fläche unter der Kurve (engl. *Area under the curve*, AUC) als Zusammenfassung an.

Da in der Literatur häufig auf *Precision/Recall*-Werte zurück gegriffen wird, werden sie auch hier trotz der begrenzten Aussagekraft (begleitend zu den ROC-Kurven) genutzt.

### 6.1.4 Testvorgehen

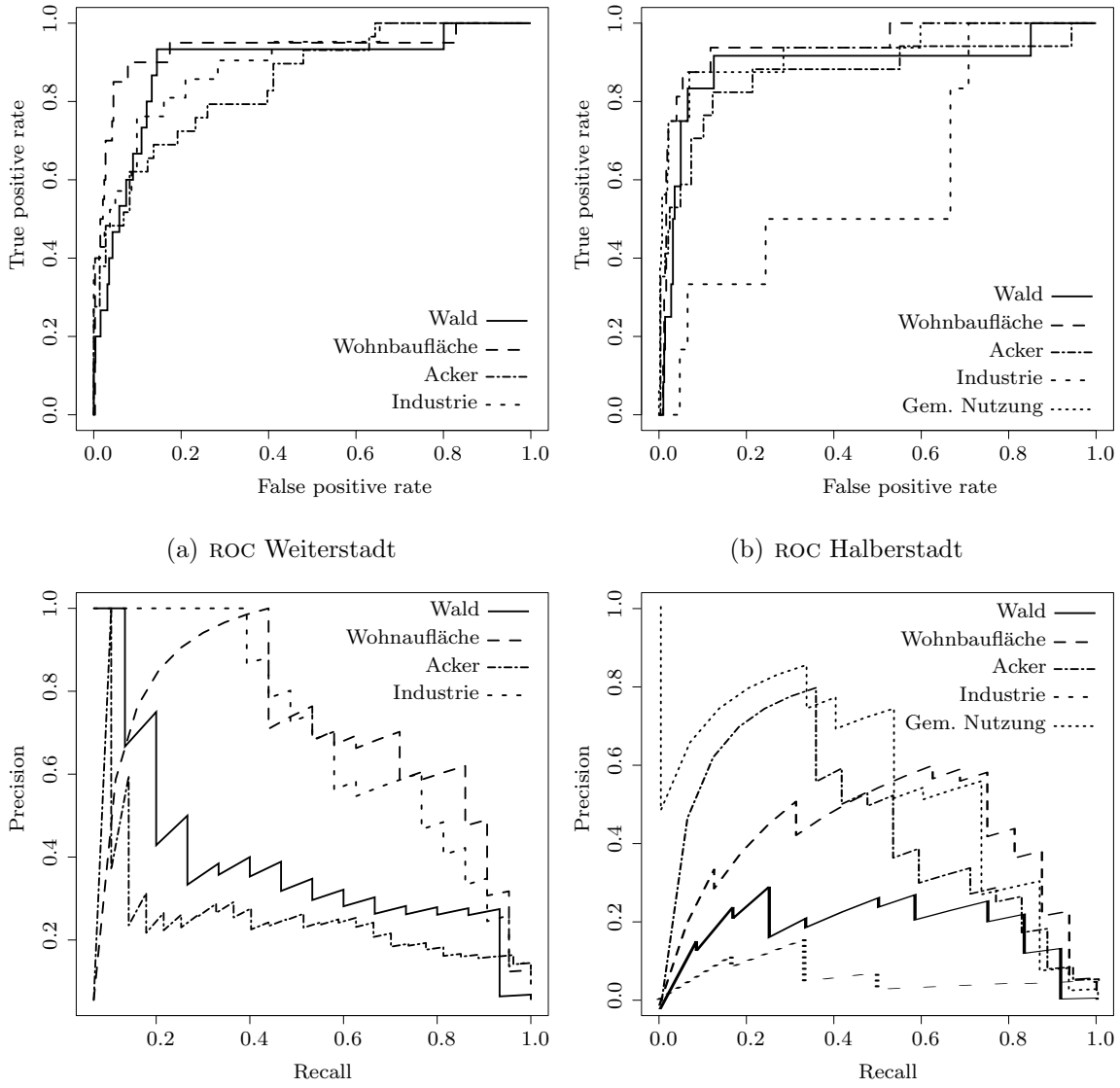
Das generelle Vorgehen für die Analyse einer Szene ist, die Satellitendaten mit einer SVM zu analysieren um die Landbedeckung zu bestimmen. Zuvor muss die SVM über eine Stichprobe manuell trainiert werden. Anschließend wird für jedes Geodatenobjekt das Landbedeckungshistogrammmerkmal bestimmt. Pro Objektklasse wird schließlich ein Abnormitätsmodell aufgestellt. Hierzu wird im Normalfall die Gesamtheit aller Geodatenobjekte der entsprechenden Objektklasse betrachtet. Mit dem Modell wird abschließend die Abnormität für jedes Geodatenobjekt bestimmt.

Die Untersuchungen der Erkennungsleistung des Systems konzentrieren sich auf folgende Bereiche:

- Einfluss der behandelten Objektklassen
- Einfluss der Korrektheit der Szene
- Eignung der Objektbewertung zur Detektion fehlerhafter Geodatenobjekte.

## 6.2 Erkennungsleistung des Systems

In Abbildung 6.4 sind ROC- und *Precision/Recall*-Kurven für die Szenen Weiterstadt und Halberstadt dargestellt. Die verschiedenen Objektklassen sind einzeln aufgeführt. Fast alle ROC-Kurven zeigen einen sehr ähnlichen Verlauf. Nur die *Industrie*-Kurve der Szene Halberstadt weicht deutlich vom Verlauf der anderen Kurven ab. Dies wird auch aus der Betrachtung der AUC-Werte deutlich (Tabelle 6.4). In der Halberstadt-Szene liegt der Industriewert mit 0,6 unter den anderen Kurven (0,87–0,93). In der Weiterstadt-Grafik fällt die Industriekurve dagegen nicht von den Werten ab (0,89 verglichen mit 0,71–0,92).




(a) ROC Weiterstadt

(b) ROC Halberstadt

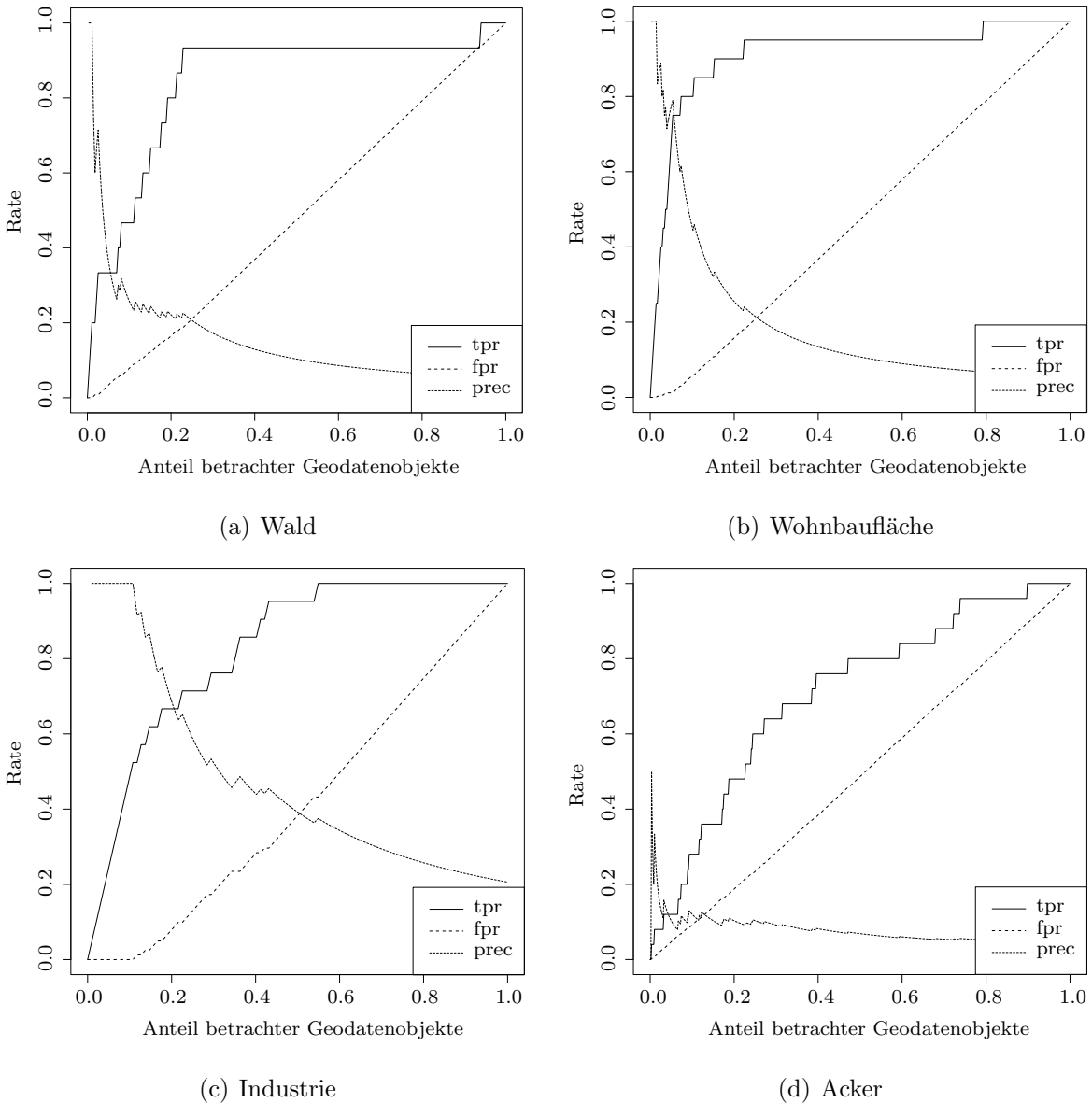
(c) *Precision/Recall* Weiterstadt

(d) *Precision/Recall* Halberstadt

**Abbildung 6.4:** Erkennungsleistung der verschiedenen Objektklasse bei vergleichbarer Landbedeckungsanalyse.

**Tabelle 6.4:** AUC-Werte der Testszenen

| Objektklasse | Weiterstadt | Halberstadt |
|---|---|---|
| Wald | 0,89 | 0,89 |
| Wohnbaufläche | 0,93 | 0,94 |
| Acker | 0,71 | 0,87 |
| Industrie | 0,89 | 0,6 |
| Gem. Nutzung | - | 0,93 |



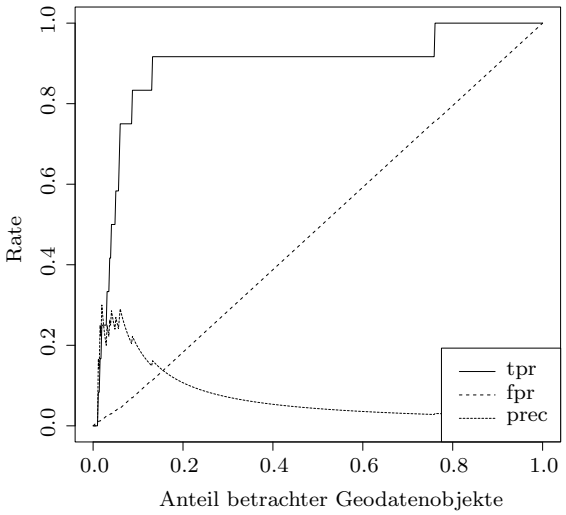

(a) Wald

(b) Wohnbaufläche

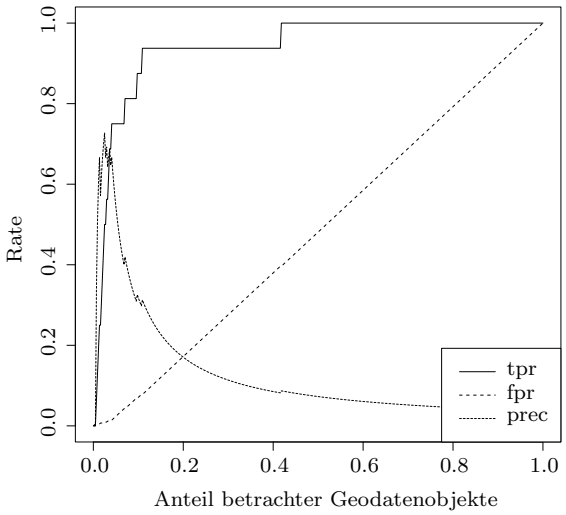
(c) Industrie

(d) Acker

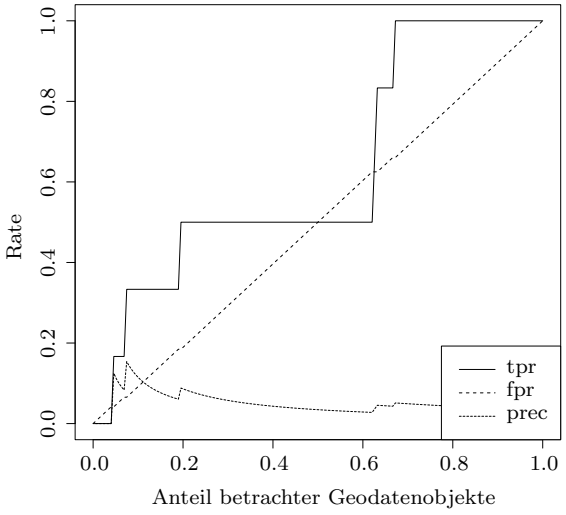
**Abbildung 6.5:** Szene Weiterstadt: Qualität der Ergebnismenge falls jeweils ein Anteil der am schlechtesten bewerteten Geodatenobjekte als fehlerhaft gewertet wird.



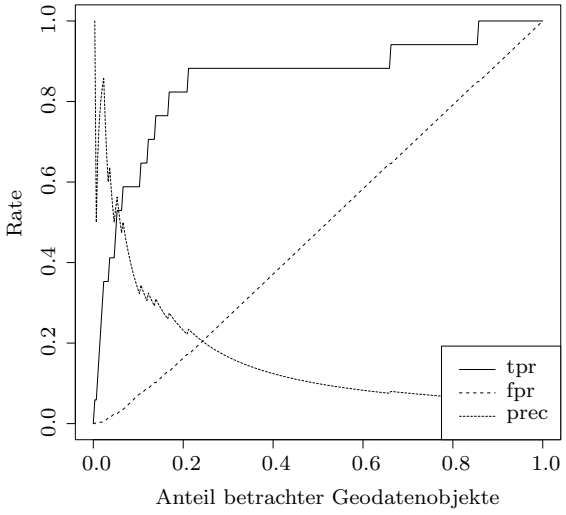

(a) Wald



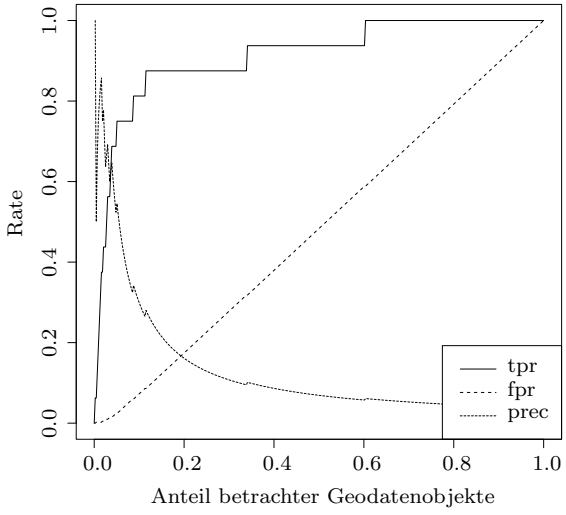

(b) Wohnbaufläche




(c) Industrie




(d) Acker




(e) Gemischte Nutzung

**Abbildung 6.6:** Szene Halberstadt: Qualität der Ergebnismenge falls jeweils ein Anteil der am schlechtesten bewerteten Geodatenobjekte als fehlerhaft gewertet wird.

**Tabelle 6.5:** Ergebnisse des WIPKA-Systems auf den Testszenen.

| | Wohnbaufläche | Industrie | Acker | Wald | Gem. Nutzung |
|---|---|---|---|---|---|
| **Weiterstadt** | | | | | |
| *Recall/tpr* | 0,65 | 0,86 | 0,68 | 0,13 | |
| *fpr* | 0,01 | 0,43 | 0,15 | 0,01 | |
| *Precision* | 0,81 | 0,34 | 0,17 | 0,4 | |
| **Halberstadt** | | | | | |
| *Recall/tpr* | 0,75 | 1 | 0,29 | 0,08 | 0,7 |
| *fpr* | 0,03 | 0,97 | 0 | 0,07 | 0,16 |
| *Precision* | 0,5 | 0,04 | 0,83 | 0,2 | 0,14 |

Abbildungen 6.5 und 6.6 zeigen eine ergänzende Darstellung. Hier sind nun die einzelnen Kennwerte *tpr*, *fpr*, *Precision* über den Anteil an Geodatenobjekten aufgetragen, der betrachtet werden muss, wenn man nur die am schlechtesten bewerteten Geodatenobjekte betrachtet. Auf allen Grafiken sieht man, dass der *fpr*-Wert mit zunehmenden Anteil an betrachteten Geodatenobjekten nahezu linear ansteigt. Dies ist zwangsläufig der Fall, wenn wie hier die Anzahl der betrachteten Geodatenobjekte die Menge an fehlerhaften Geodatenobjekten übersteigt. Gleichzeitig steigt aber deutlich stärker der Anteil gefundener fehlerhafter Geodatenobjekte mit dem Anteil betrachteter Geodatenobjekte (*tpr*). Bei den Objektklassen Wald/Weiterstadt, Wohnbaufläche/Weiterstadt, Wald/Halberstadt, Wohnbaufläche /Halberstadt und auch Gemischte Nutzung/Halberstadt, können rund 90 % der fehlerhaften Geodatenobjekte gefunden werden, wenn nur die 20 % der am schlechtesten bewerteten Geodatenobjekte betrachtet werden. Gemischte Nutzung/Halberstadt und Acker/Halberstadt liegt mit knapp unter 90 % auf ähnlich gutem Niveau.

Nur bei Acker/Weiterstadt und Industrie/Halberstadt fallen die Ergebnisse ab. Acker in Weiterstadt erfordert, dass man die Hälfte der Geodatenobjekte betrachtet, um 80 % der fehlerhaften Geodatenobjekte zu finden. Betrachtet man hier nur 20 % der Geodatenobjekte, werden aber immer noch etwa 35 % der fehlerhaften Geodatenobjekte gefunden. Nur bei Industrie/Halberstadt eignet sich das Ergebnis kaum für die Prüfung der Geodaten.

**Vergleich mit dem Wipka-System** Zum Vergleich wurden die Testdaten auch mit dem WIPKA-System bewertet. Dabei wurden die üblichen Konfigurationswerte genutzt (Abschnitt 4.1.1). Die Ergebnisse sind in Tabelle 6.5 dargestellt. Die *tpr*-Werte liegen mit 0,65, 0,86, 0,68 and 0,13 (Weiterstadt) bzw. 1, 0,75, 0,7, 0,29 and 0,08 (Halberstadt) meist unter den Ergebnissen des neu entwickelten Systems. In den Fällen, in denen die Werte das neue System übertreffen (etwa Industrie in Halberstadt) ist die *fpr*-Rate sehr hoch (0,97). Folglich müssen hier (fast) alle Geodatenobjekte geprüft werden. Das schlechte Abschneiden hat vor allem zwei Ursachen:

- Die Parameter entsprechen Erfahrungswerten und sind demnach nicht optimal für die Testszenen. Bessere Parameter könnten etwa die *fpr*-Werte der Halberstadt Industrieobjekte verbessern, selbst wenn sich dann die *tpr*-Werte ebenfalls verringern werden.

- Die Parameter prüfen die Korrektheit anhand der in Abschnitt 4.1.1 dargestellten Methodik. Für Objekte einer Objektklasse wird eine Landbedeckungsklasse als positiv, die jeweils anderen als negativ bewertet. Wird für ein Geodatenobjekt ein festgesetzter Anteil an positiv gewerteter Landbedeckung nicht erreicht, wird das Geodatenobjekt als fehlerhaft bewertet. Dieses Kriterium ist deutlich primitiver als das für die Abnormitätsanalyse genutzte Landbedeckungshistogramm. Dadurch ist die Erkennungsleistung des WiPKA-Systems stark eingeschränkt. Komplexere Kriterien werden für die regelbasierte Auswertung nicht genutzt, da sie schwierig zu konfigurieren sind.

## 6.3 Unterschiedliches Abschneiden der verschiedenen Objektklassen

Es stellt sich die Frage, warum die Erkennungsleistung für die Objektklassen Industrie (Halberstadt) und Acker (Weiterstadt) von der der übrigen Objektklassen und teilweise auch zwischen den Testszenen abweicht.

### 6.3.1 Industrie (Halberstadt)

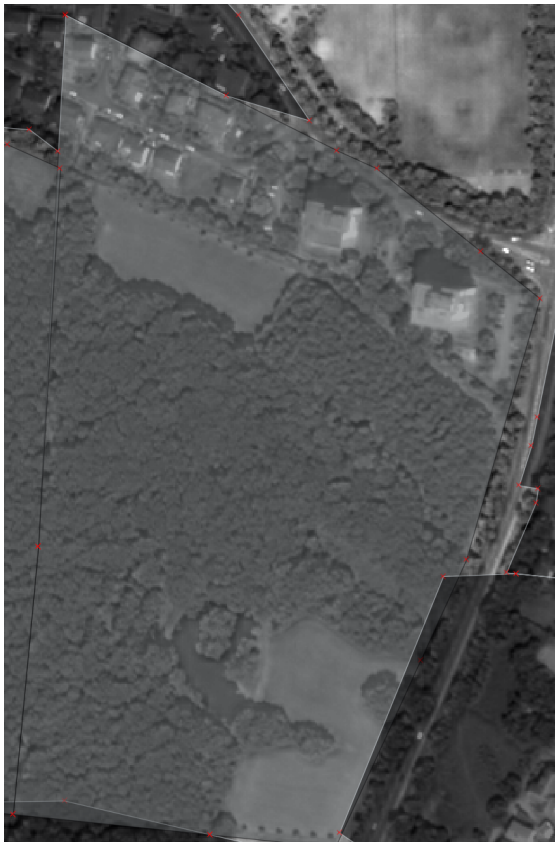
In Abbildung 6.8 sind Satellitenbilder beider Szenen mit den Industrieobjekten überlagert dargestellt. In beiden Szenen gibt es sowohl größere Gebiete zusammenhängender Industrieobjekte, umgangssprachlich Gewerbegebiete, als auch Kleinbetriebe in der Fläche (Bildbeispiele siehe Abbildung 6.9).

Eine Möglichkeit für ein schlechtes Abschneiden der Klasse *Industrie* ist, dass die Modellierung des Merkmalsraums die Komplexität der Erscheinungsweise nicht ausreichend wiedergeben kann. Würde also eine differenzierte Betrachtung im Merkmalsraum bessere Ergebnisse ermöglichen?

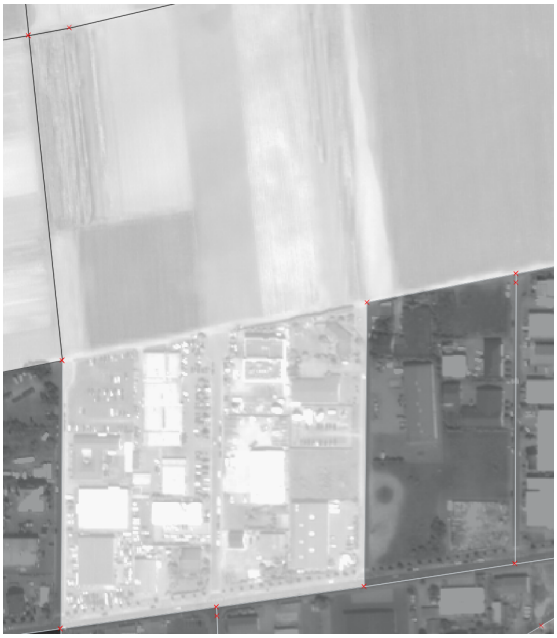
Zur Überprüfung dieser These ist es sinnvoll, den Merkmalsraum zu betrachten. Hierzu sind in Abbildung 6.10 die ersten beiden Hauptkomponenten einer PCA-Analyse der Industriemerkmalsräume der Szenen Weiterstadt (a) und Halberstadt (b) dargestellt. Zum Vergleich sind darunter die Waldmerkmalsräume dargestellt. Die Objektklasse Wald steht hierbei stellvertretend für die Objektklasse mit guten Ergebnissen.

In den Abbildungen wird unterschieden zwischen Merkmalen von Geodatenobjekten, die von der Referenz als fehlerhaft (+) bzw. nicht fehlerhaft ($\cdot$) eingeschätzt wurden. Die Merkmale scheinen sich in allen Fällen im Merkmalsraum als Dreiecke abzuzeichnen. Dies rührt daher, dass die Merkmalsdimensionen, die durch das Landbedeckungshistogrammmerkmal (Abschnitt 5.3.1) gebildet werden, zusammenaddiert 1 ergeben.

Zunächst alleine den Halberstadtmerkmalsraum (b) betrachtend wird deutlich, dass sich die Merkmalsverteilung der Industrieobjekte auch bei einem detaillierteren Modell für den Merkmalsraum nicht besser hätte beschreiben lassen. Zwar häufen sich die Punkte nicht



(a) Wald

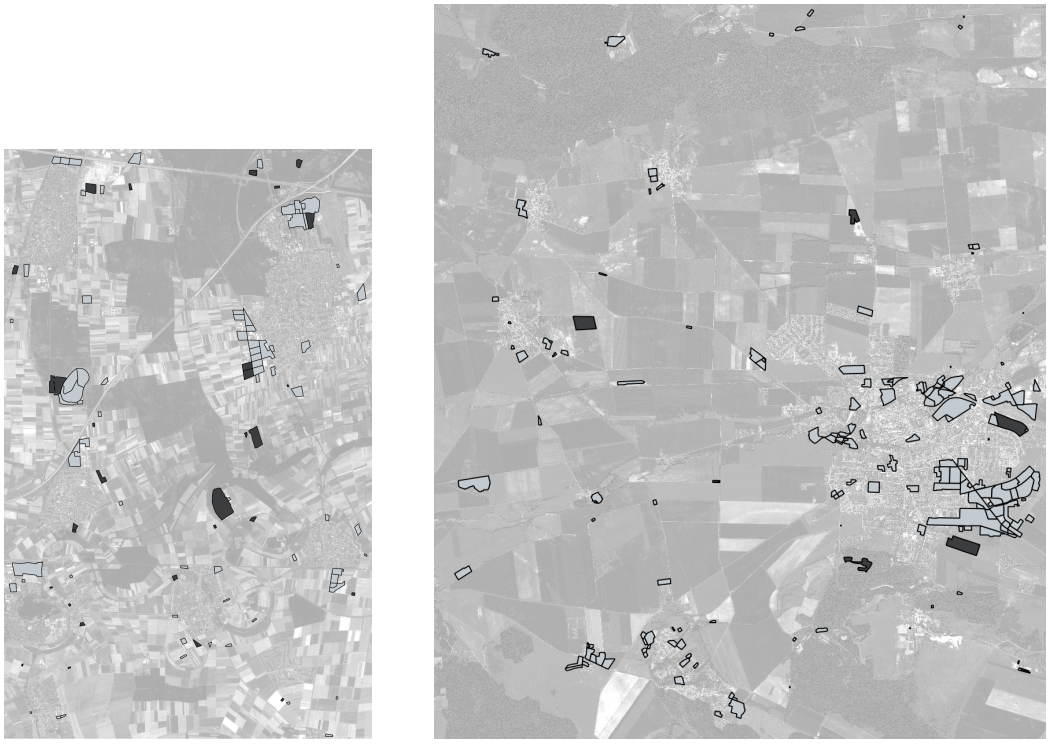


(b) Acker

(c) Wohnbaufläche

(d) Industrie

**Abbildung 6.7:** Nicht erkannte fehlerhafte Geodatenobjekte. Geodatenobjekte sind umrandet und zur besseren Unterscheidung semitransparent überblendet.



(a) Weiterstadt, 90° gedreht (b) Halberstadt

**Abbildung 6.8:** Dargestellt sind die Industrieobjekte beider Testszenen. Schwarz umrandet: *Nicht fehlerhaft* laut Referenz, schwarz ausgefüllt: *fehlerhaft* laut Referenz. Im Hintergrund: IKONOS-Bilder der Szenen.

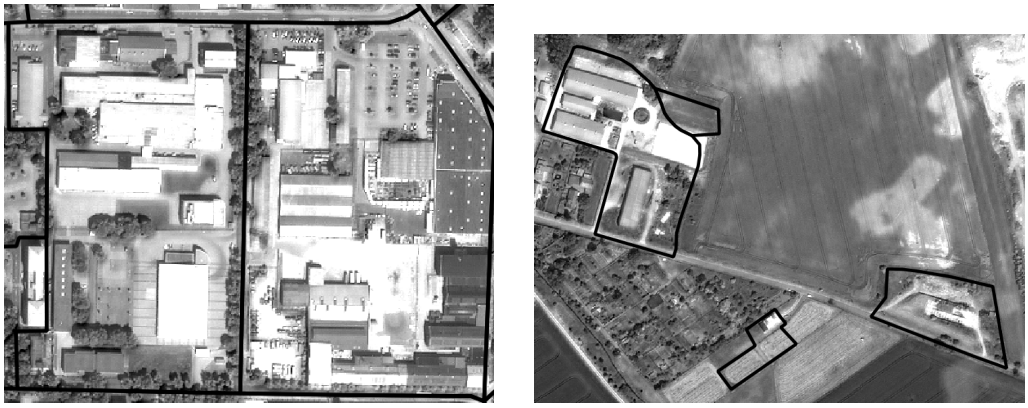
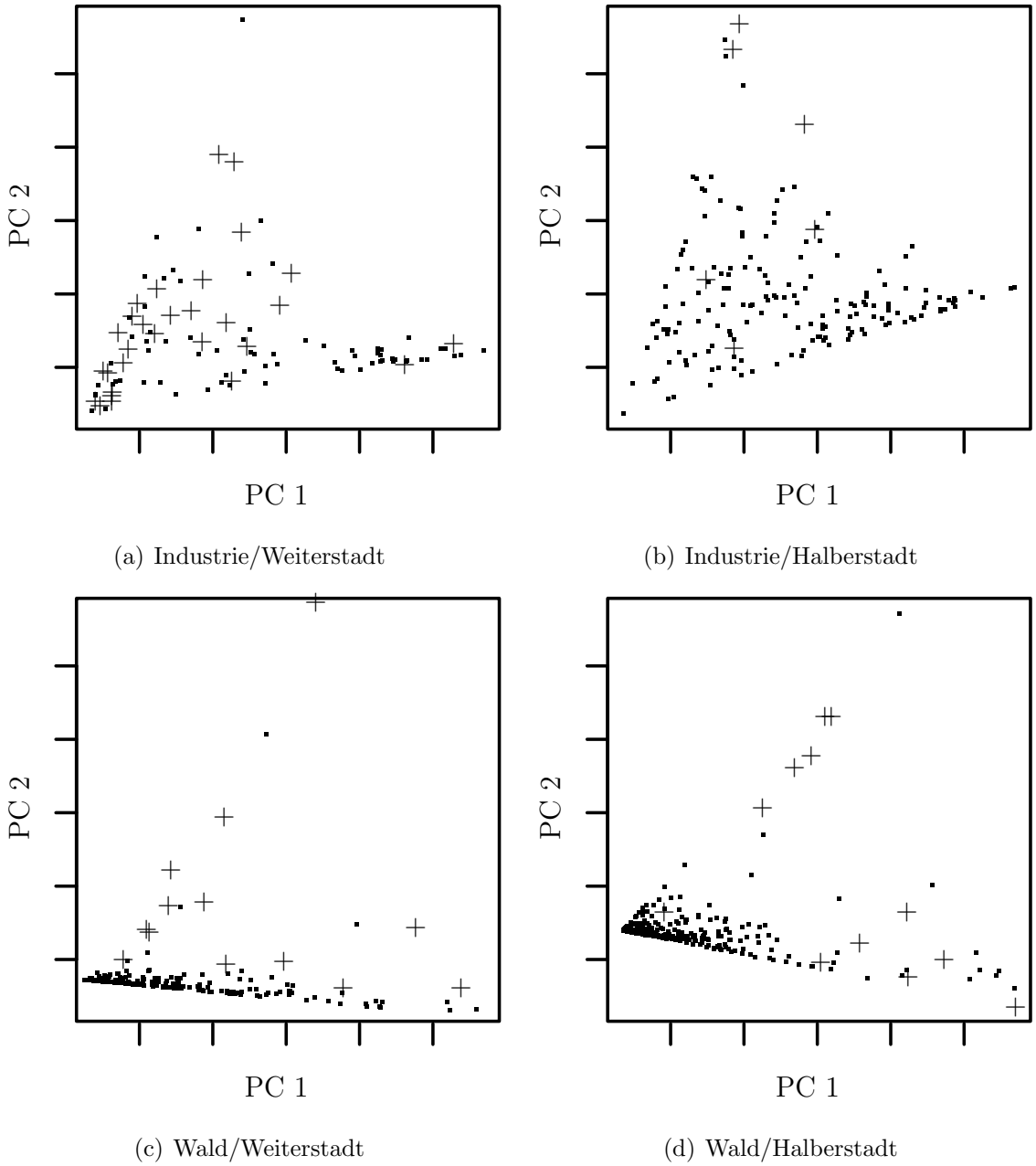


**Abbildung 6.9:** Industrieobjekte (schwarz begrenzt) aus der Halberstadt-Szene.




(a) Industrie/Weiterstadt

(b) Industrie/Halberstadt

(c) Wald/Weiterstadt

(d) Wald/Halberstadt

**Abbildung 6.10:** PCA Merkmalsraumdarstellungen, · nicht fehlerhaft laut Referenz, + fehlerhaft laut Referenz

deutlich an einem Ort, doch eine Aufteilung auf mehrere Bereiche ist auch nicht festzustellen. Die verschiedenen Arten von *Industrie* und *Gewerbegebieten* einerseits und einzeln stehenden Kleinbetrieben andererseits, sind also nicht im Merkmalsraum nachzuvollziehen. Die Merkmale unterscheiden sich nicht, da das Landbedeckungshistogrammmerkmal die Landbedeckungen mit der Objektgröße normiert.

Bei einem Vergleich zu den Industrieobjekten aus Weiterstadt (a) fällt die sehr ähnliche Verteilung der Merkmale auf. In beiden Szenen sind die Merkmale über einen großen Bereich des Merkmalsraums gestreut. Fehlerhafte Geodatenobjekte setzen sich nicht von den anderen Geodatenobjekten ab. Dies ist in beiden Szenen der Fall, jedoch gibt es in der Weiterstadt-Szene weit mehr als fehlerhaft eingestufte Geodatenobjekte. Daher fällt die Bewertung durch die ROC-Darstellung in Weiterstadt besser aus als in der Halberstadt-Szene.

Die Besonderheit der lockeren Verteilung der Merkmale im Fall Industrie wird insbesondere deutlich, wenn man zum Vergleich die Merkmalsräume der Waldobjekte (dargestellt in (c) und (d)) betrachtet. Hier häufen sich die Punkte deutlich. Eine Hauptkomponente reicht offenbar, um die Bandbreite der meisten Geodatenobjekte zu beschreiben. In Richtung PC2 heben sich nur vereinzelt (und überwiegend fehlerhafte) Geodatenobjekte von der Masse an Geodatenobjekten ab.

Dass sich die Merkmale bei den Industriemerkmalsräumen über einen sehr großen Bereich verstreuen, ist in erster Linie der weit gefassten Objektklassendefinition für Industrie im zugehörigen ATKIS-Objektartenkatalog geschuldet [3]:

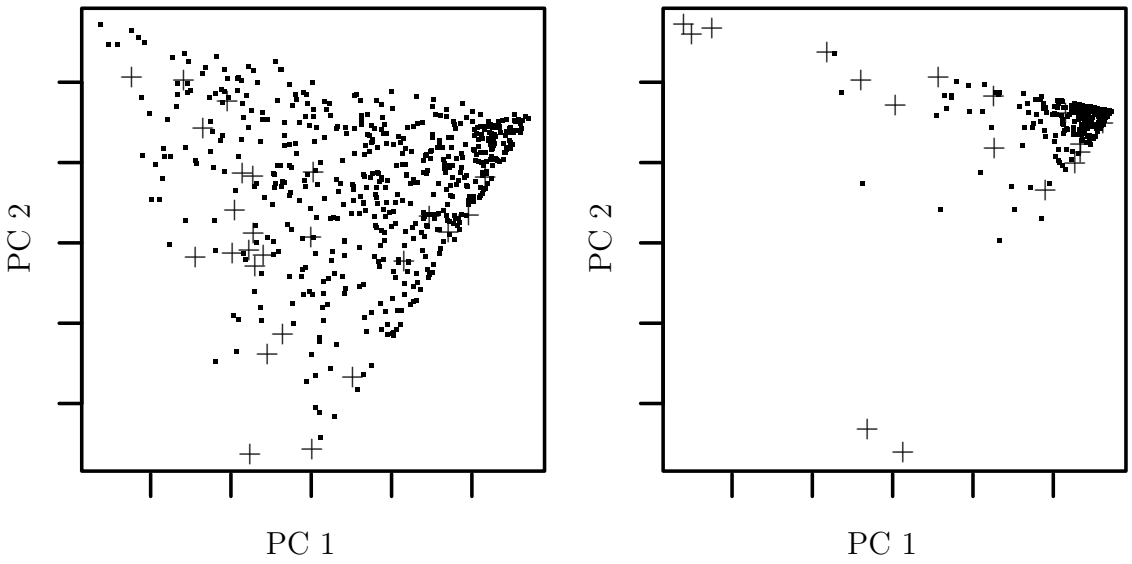
> Eine baulich geprägte Fläche, die ausschließlich oder vorwiegend der Unterbringung von Gewerbe- und Industriebetrieben dient. Dazu zählen auch z.B. Einkaufszentren, Lager/Depots, großflächige Handelsbetriebe, Ver- und Entsorgungsbetriebe, Messeeinrichtungen. Die Grenze zwischen einer Industrie- und Gewerbefläche und benachbarten Flächen wird in der Regel durch die Grenzen der bebauten Grundstücke unter Einbeziehung der Hofraumflächen gebildet.

Wie der Beschreibung zu entnehmen ist, haben sich die Grenzen der Geodatenobjekte nach Grundstücksgrenzen, also Besitzgrenzen zu richten. Da sich, wie auch in Abbildung 6.9 zu erkennen ist, Industrieobjekte sowohl im ländlichen wie städtischen Raum befinden, können sich auf dem Gelände sehr viele unterschiedliche Arten von Bodenbedeckung befinden, ohne offensichtlich gegen die Objektdefinition zu verstoßen. Dies wird gut durch die breite Streuung im Merkmalsraum nachvollzogen. Die Prüfung der Korrektheit eines Geodatenobjekts aus Satellitenbildern ist entsprechend problematisch. Dies könnte vielleicht auch das scheinbar unterschiedliche Verhalten der Referenz zwischen den Testszenen erklären.

### 6.3.2 Acker (Weiterstadt)

Neben der Industrie fällt auch die Bewertung der Objektklasse Acker in Weiterstadt ab. Tatsächlich verteilen sich auch in diesem Fall die Merkmale im Merkmalsraum erneut sehr weiträumig (Darstellung (a) in Abbildung 6.11).

Interessant ist, dass der Merkmalsraum der Ackerobjekte aus der Halberstadtszene (b) sich aber deutlich von dem aus Weiterstadt unterscheidet. Der Merkmalsraum aus der Halberstadtszene ist deutlich weniger gestreut. Dadurch dass sich zudem viele fehlerhafte




(a) Acker/Weiterstadt (b) Acker/Halberstadt

**Abbildung 6.11:** PCA Merkmalsraumdarstellungen, · nicht fehlerhaft laut Referenz, + fehlerhaft laut Referenz

Geodatenobjekte relativ isoliert im Merkmalsraum befinden, ist die Erkennungsleistung für Acker in der Halbstadtszene (AUC 0,87) deutlich besser als in Weiterstadt (AUC 0,71).

Die Definition für Acker im ATKIS-Objektartenkatalog lautet:

> Fläche für den Anbau von Feldfrüchten (z.B. Getreide, Hülsenfrüchte, Hackfrüchte) und Beerenfrüchten (z.B. Erdbeeren).

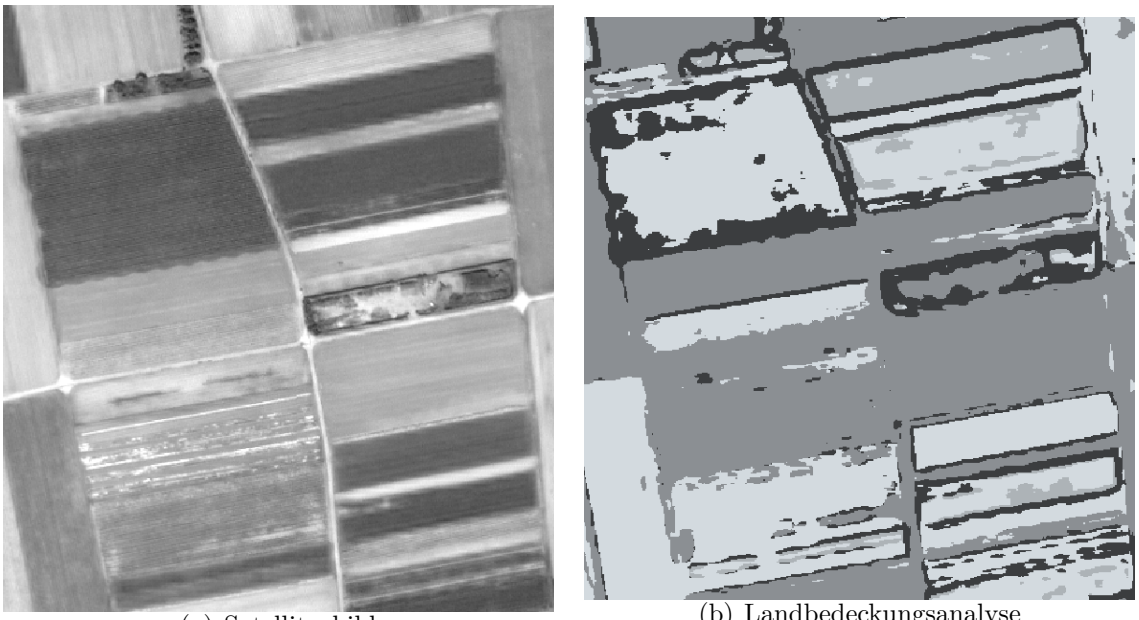
Die Objektdefinition schließt also, im Unterschied zur Definition von Industrie, keine umgebenden Flächen ausdrücklich mit ein.

Bereits aus Abbildung 6.3 ist zu erkennen, dass die Ackerflächen in Weiterstadt verglichen zu Halberstadt deutlich kleinteiliger aufgebaut sind. Auch wirkt das Erscheinungsbild der Weiterstadtobjekte komplexer. Abbildung 6.12 zeigt eine Detailansicht (a) eines Ausschnitts und die zugehörige Landbedeckungsanalyse. Die gesamte Fläche ist aus Ackerobjekten zusammengesetzt. Wie man sieht, werden Objekte mit sehr heterogener Landbedeckung analysiert. Entsprechend weit gestreut sind schließlich die Merkmale mit Merkmalsraum.

Dieses Beispiel verdeutlicht gut die Abhängigkeit des Bewertungsverfahrens vom eingesetzten Landbedeckungsanalyseverfahren. In diesem Fall ist das Erscheinungsbild von Ackerobjekten zu vielfältig, als dass fehlerhafte Geodatenobjekte zuverlässig erkannt werden könnten.

## 6.4 Einfluss der Güte der Geodaten

Eine für die Praxistauglichkeit der Lösung relevante Fragestellung ist, wie stark der Grad der Korrektheit der Geodatenobjekte Einfluss auf die Qualität der Ergebnisse hat. Hierzu werden die Testdaten der Szene Weiterstadt betrachtet. Um Aussagen in Abhängigkeit des



(a) Satellitenbild (b) Landbedeckungsanalyse

**Abbildung 6.12:** Darstellung von Ackerflächen (mehrere Geodatenobjekte) aus der Weiterstadtszene. Die Bedeutung der Farbwerte der Landbedeckung: von hell nach dunkel: Gras, Bäume, Hallen, Häuser

Fehlergrades der Daten zu erhalten, wird der Fehlergrad der Daten kontrolliert verändert und die Auswirkungen auf die Ergebnisqualität beobachtet. Die Anzahl der fehlerhaften Geodatenobjekte ist in jeder Beobachtung gleich, doch die Anzahl nicht fehlerhafter Geodatenobjekte wird nach folgender Anleitung variiert.

1. Die laut Referenz fehlerhaften und nicht fehlerhaften Geodatenobjekte werden getrennt.

2. Von den nicht fehlerhaften Geodatenobjekten wird eine Teilmenge ausgewählt. Die Teilmengen entsprechen von der Größe einem festzusetzenden Anteil an der Gesamtmenge. Die betrachteten Anteile liegen zwischen 0 % und 100 %, mit Zwischenabstufungen von 6 %.

   Die Auswahl einer Teilmenge erfolgt zufällig und erfolgt „*mit zurücklegen*“: Es wird jeweils aus der Gesamtmenge von nicht fehlerhaften Geodatenobjekten so lange ein Geodatenobjekt ausgewählt, bis die Anzahl dem gewählten Anteil entspricht. Vor jeder Ziehung wird das vorher gewählte Objekt wieder der Gesamtmenge hinzugefügt, sodass das gleiche Geodatenobjekt mehrfach ausgewählt werden kann. Das gleiche Auswahlverfahren wird für die fehlerhaften Geodatenobjekte durchgeführt. Die Anzahl an fehlerhaften Geodatenobjekten entspricht dabei immer der vollen Anzahl fehlerhafter Geodatenobjekte in der Szene. Pro Anteil werden auf diese Weise 10 000 Teilmengen bestimmt. Die hohe Zahl an Wiederholungen und die Auswahl von Geodatenobjekten jeweils aus der Gesamtmenge erhöht die Robustheit der Auswertung, da der Einfluss einzelner Geodatenobjekte auf das Ergebnis reduziert wird.

3. Für die Auswertung werden jeweils eine ermittelte Teilmenge der nicht fehlerhaften und der fehlerhaften Geodatenobjekte zu einer *Testmenge* vereinigt. Auf dieser Test-

menge führt das System wie gehabt eine Abnormitätsanalyse durch und bestimmt die Abnormität der Geodatenobjekte.

4. Analog zu den bisherigen Auswertungen wird mit Hilfe der Referenz eine Auswertung der Korrektheit der Ergebnisse mit der jeweiligen Testmenge durchgeführt und der AUC-Wert bestimmt.

In den Abbildungen 6.13 und 6.14 sind die Ergebnisse für die einzelnen Objektklassen und Testszenen als *Boxplots* dargestellt. Jeder Eintrag in einem Boxplot steht für einen festgesetzten Anteil der richtigen Geodatenobjekte. Interessanter und deshalb hier angegeben ist stattdessen jedoch der Anteil an fehlerhaften Geodatenobjekten an der gesamten Testmenge. Da die Anzahl der fehlerhaften Geodatenobjekte zwischen den Objektklassen variiert, ist auch die Beschriftung der Boxplots nicht einheitlich. Ebenso ist die erzielbare Qualität bei den Objektklassen unterschiedlich (vgl. Tabelle 6.4), sodass die Ordinaten ebenfalls eine unterschiedliche Skalierung aufweisen.

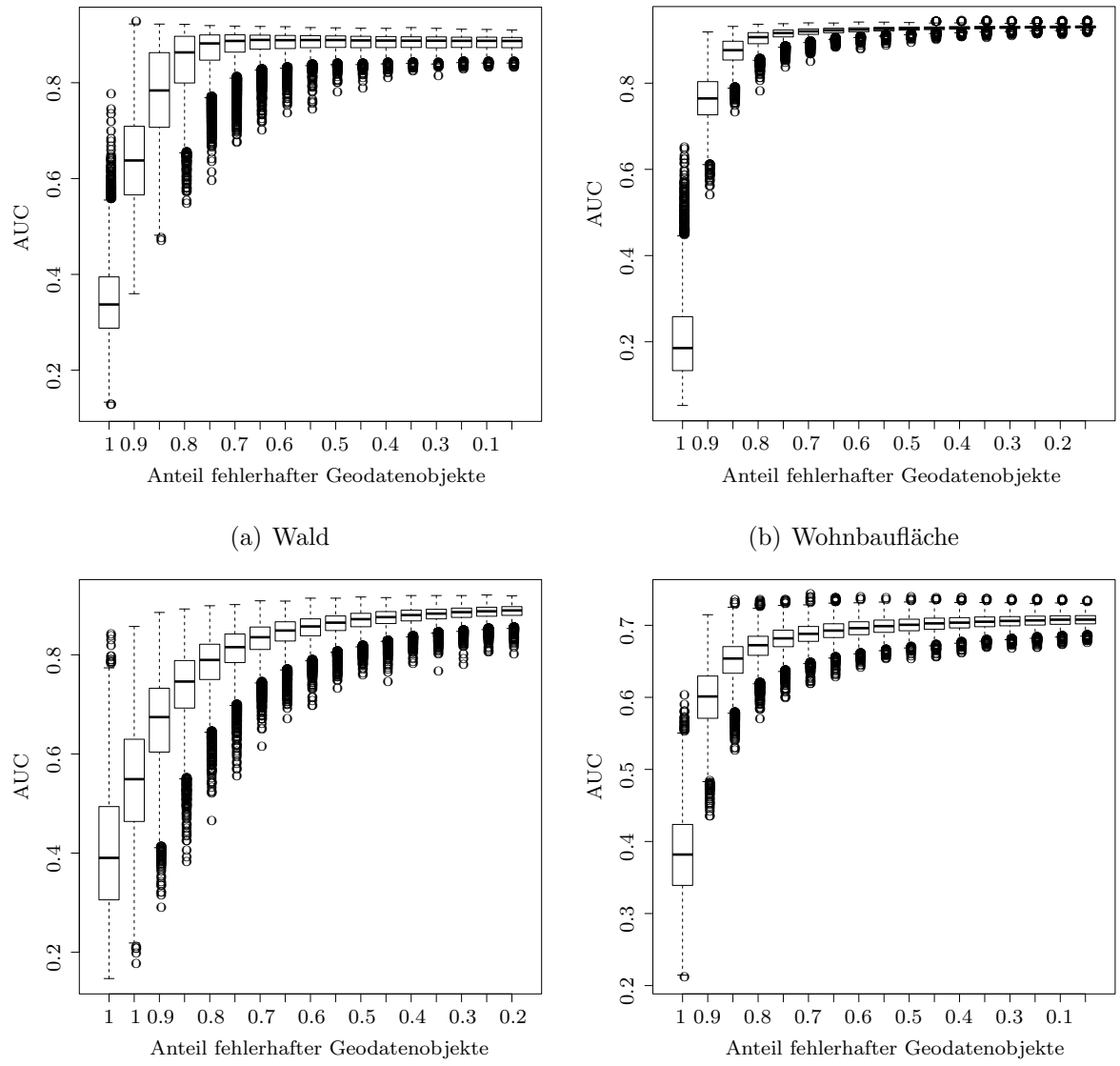
Die einzelnen Einträge zeigen, wie stark sich die jeweils 10 000 ermittelten AUC-Werte unterscheiden. Die schwarze Markierung innerhalb des gezeichneten Kastens entspricht dem Median. Als Kastengrenzen dienen die 25 und 75 % Prozentile. Die sich von der Box nach oben und unten erstreckenden Linien entsprechen dem 1,5–fachen Abstand zwischen Median und dem entsprechenden 25 bzw. 75 % Prozentil. Werte, die sich außerhalb des so markierten Bereichs befinden, sind als Ausreißer einzeln markiert.

Die Boxplots in den Darstellungen (a) - (d) bzw. (e) stellen die Ergebnisse für die verschiedenen Objektklassen dar. Man beachte, dass der geringste Anteil an fehlerhaften Geodatenobjekten (jeweils der letzte Wert) dabei der Situation entspricht, wie sie in Tabelle 6.1 dargestellt ist: Der Anteil von fehlerhaften Geodatenobjekten entspricht der Zusammensetzung der Testszene. Im Unterschied zu dem Test, dessen Ergebnis in Darstellung 6.4 (a) und (c) dargestellt ist, sind hier für das Ergebnis aber nicht einfach alle Geodatenobjekte gemeinsam betrachtet worden. Stattdessen wurden auch hier 10 000 Testmengen zufällig zusammengestellt. Somit zeigt der entsprechende Boxplot an, dass das in Abbildung 6.4 (a) und (c) dargestellte Ergebnis wenig abhängig von einzelnen Geodatenobjekten ist.

Betrachtet man die Einträge in dem Boxplot einer Objektart als Stützstellen einer Kurve, sieht man wie sich die Erkennungsleistung mit abnehmendem Fehleranteil wie erwartet steigert. Interessant ist, wie schnell die Erkennungsleistung ansteigt. Bei Wald- und Wohnbauflächenobjekten ist bereits bei einem Fehleranteil von etwa 70 % eine gute Erkennungsleistung gewährleistet. Bei Industrie- und Ackerobjekten ist eine stabile Erkennungsleistung bei höchstens 50 % fehlerhaften Geodatenobjekten möglich.

Die in der Praxis gesammelten Erfahrungen lassen in einer realen Situation einen Fehlergrad von nur wenigen Prozenten erwarten. Die Tests zeigen nun, dass selbst wenn jedes zweite oder sogar zwei von drei Geodatenobjekten falsch sind, noch gute Ergebnisse zu erwarten sind.

Dass selbst bei grob falschen Geodaten noch gute Ergebnisse erzielt werden, ist dadurch zu erklären, dass die fehlerhaften Geodatenobjekte einen Cluster bilden müssten, um die Ergebnisse stark zu verändern. Wie bereits in Abbildung 6.10 dargestellt ist das in der Praxis nicht der Fall. Nicht dargestellt, aber aus den Abbildungen nachvollziehbar ist der Umstand, dass selbst Testmengen ohne fehlerhafte Geodatenobjekte kaum bessere Ergebnisse ermöglichen. Auf eine manuelle Auswahl einer Trainingsmenge kann also verzichtet werden.



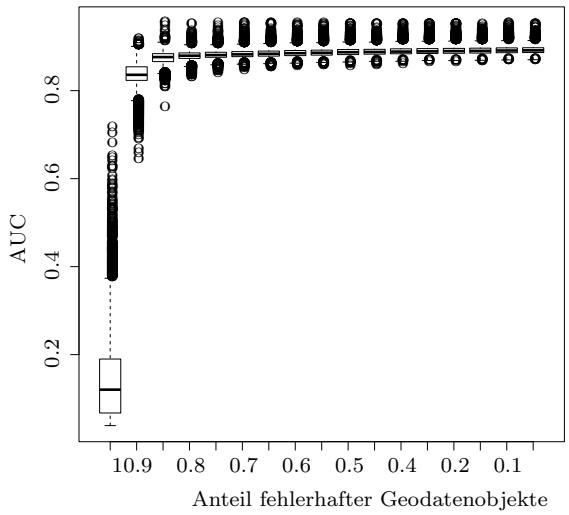

(a) Wald

(b) Wohnbaufläche

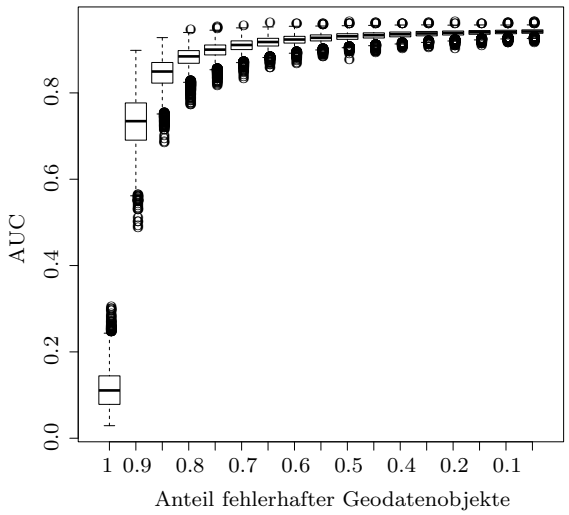
(c) Industrie

(d) Acker

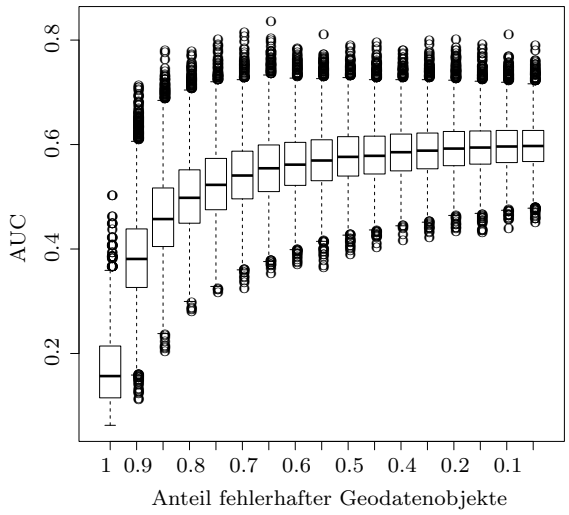
**Abbildung 6.13:** Robustheitsergebnisse der Szene Weiterstadt, Boxplots der einzelnen Objektklassen.



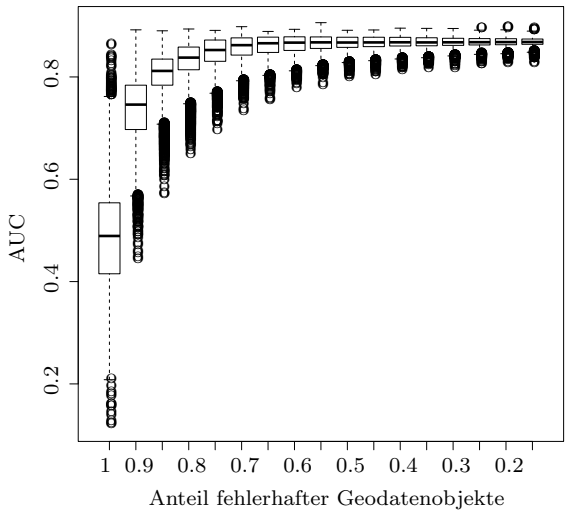

(a) Wald



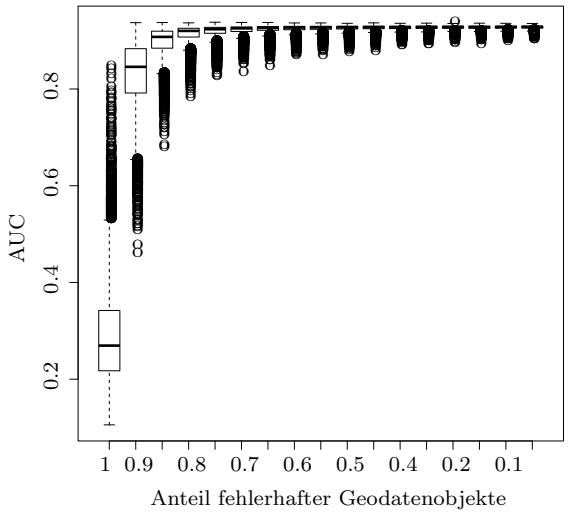

(b) Wohnbaufläche




(c) Industrie




(d) Acker




(e) Gemischte Nutzung

**Abbildung 6.14:** Robustheitsergebnisse der Szene Halberstadt, Boxplots der einzelnen Objektklassen.

# 7 Zusammenfassung und Ausblick

Geodaten sind eine unverzichtbare Grundlage für eine Vielzahl von Anwendungen. Die Verlässlichkeit der Daten ist allerdings durch ihre aufwendige Kontrolle nur schwierig einzuschätzen.

Das Ziel dieser Arbeit war die Entwicklung einer Methode, um Geodaten möglichst autonom auf Änderungen zu überprüfen. Neben den Geodaten werden für die Prüfung nur aktuelle Fernerkundungsdaten benötigt. Demonstriert wurde das Verfahren auf optischen Satellitendaten, eingeschränkt ist das System hierauf jedoch nicht. Sind durch die Analyse erst einzelne Problembereiche identifiziert, kann in einem dem System nachgeschalteten Arbeitsschritt die Korrektur der Geodaten vorgenommen werden. Die Korrektur wird sich abhängig von dem Verwendungszweck auf mehr als Fernerkundungsdaten stützen und womöglich Ortsbesichtigungen einschließen. Nur durch eine entsprechend starke Einschränkung der zu korrigierenden Fläche ist solch ein Vorgehen durchführbar.

Die bestehenden Lösungen (Kapitel 1.2) sind hierfür nur zum Teil geeignet. Einerseits gibt es eine Reihe von Verfahren, die alleine aus den Fernerkundungsdaten heraus einem Pixel (oder einer Gruppe von Pixeln) eine Objektklasse zuweisen. Diese Verfahren sind auf Spezialfälle von Objektklassen eingeschränkt. Sie sind aber geeignet, die Landbedeckung aus Fernerkundungsdaten zu ermitteln.

Auf der anderen Seite stehen objektbasierte Verfahren, bei denen Erkenntnisse statt auf Pixeln auf ganze Geodatenobjekte bezogen werden. Dies erlaubt den Umgang mit Geodatenobjekten von Objektklassen, die eine Landnutzung erfassen und sich eher durch eine Kombination von Landbedeckungen (z.B. Vegetation und Bebauung in Siedlungsgebieten) auszeichnen. Das Vorgehen erlaubt darüber hinaus die Berücksichtigung von Generalisierungen.

Das WiPKA-System (Abschnitt 4.1) führt hierzu eine regelbasierte Auswertung ein, die eine manuelle Konfiguration erfordert. Allerdings ist die richtige Parametrisierung abhängig von vielen Faktoren (Ergebnisse der Landbedeckungsanalyse, Generalisierungen), sodass die realisierbare Komplexität der Analyse praktisch begrenzt ist.

Das System von Walter (Abschnitt 4.2) dagegen erreicht einen Grad der Automatisierung wie er auch für den neuen Ansatz angestrebt wurde. Die Korrektheit eines Geodatenobjekts über eine Neubestimmung der Objektklassenzuordnung zu prüfen, hat aber entscheidende konzeptionelle Nachteile. Die Erkennung beschränkt sich durch das Vorgehen auf Geodatenobjekte mit korrekter Umgrenzung, aber falscher Objektklassenzuordnung. Dieses Fehlerbild entspricht wenig der Situation, die man sich bei einer sich graduell verändernden Szene über die Zeit vorstellt.

Das neu entwickelte System erreicht eine Bewertung von Geodatenobjekten durch einen Vergleich mit allen anderen Objekten der jeweils gleichen Objektklasse. Das Ziel ist also eine Auswertung der Unterschiede zwischen vorgeblich ähnlichen Geodatenobjekten. Die Analyse von Objekten einer Objektklasse ist daher unabhängig von den Eigenschaften der anderen Objektklassen.

Die Geodatenobjekte werden durch das Landbedeckungshistogrammmerkmal beschrieben. Es erfasst, aus welchen Anteilen an Landbedeckungen sich das Gebiet eines Geodatenobjekts zusammen setzt. Da sich unter den Geodatenobjekten auch fehlerhafte Geodatenobjekte befinden können, wird das Auftreten von Merkmalen durch ein statistisches Modell abstrahiert. Die Modellierung des Objektauftretens erfolgt über die Ermittlung des Schwerpunkts der Merkmale und der Kovarianzmatrix (Kapitel 5). Mit Hilfe der Mahalanobisdistanz wird für jedes Geodatenobjekt die Entfernung zum Schwerpunkt ermittelt. Die Entfernung wird als Abnormität interpretiert und zur Objektbewertung genutzt.

Wie die experimentelle Überprüfung des Systems in Kapitel 6 zeigt, stellt die Abnormitätsbewertung von Geodatenobjekten ein wirkungsvolles Mittel dar, um Geodatenobjekte zu prüfen. Wird eine manuelle Nachprüfung auf die 20 % ungewöhnlichsten Geodatenobjekte beschränkt, sind dennoch Fehlerkennungsraten von 80 % to 90 % die Regel. Somit ist eine Verfünffachung der Untersuchungsleistung möglich. Verglichen mit den Ergebnissen des Systems WiPKA mit statischer Parametrisierung zeigt das neue Verfahren eine deutlich bessere Erkennungsleistung.

Die Modellierung scheint der Problemstellung angemessen, Hinweise auf Vorteile durch eine komplexere Modellierung konnten nicht gefunden werden. Ein Vorteil der einfach gehaltenen Modellierung der Merkmalsräume ist die hohe Robustheit. Wie in Abschnitt 6.4 gezeigt, ist selbst wenn mehr als die Hälfte der Geodatenobjekte fehlerhaft ist, ein sinnvoller Einsatz möglich. Auch gelingt die automatische Anpassung an die verschiedenen Testszenen mit unterschiedlicher Landbedeckung. Das System stößt an seine Grenzen, wenn die Objektklassendefinitionen Aspekte mit einschließen, die sich nicht (statistisch) aus den Fernerkundungsdaten prüfen lassen. Ein Beispiel sind funktionale Details von Industrieanlagen.

Eine wichtige Grundlage für die Analyse der Geodatenobjekte ist eine Vorverarbeitung der Fernerkundungsdaten durch ein Verfahren des maschinellen Lernens. Im Rahmen dieser Arbeit wurde hierfür eine SVM (Abschnitt 5.2) eingesetzt. Die SVM stellt ein allgemein anerkanntes Verfahren für die Landbedeckungsanalyse dar. Durch die Nutzung des Landbedeckungsergebnisses für das Abnormitätsbestimmungsverfahrens in dem neuen System sind die Klassifikationsanforderungen jedoch offener als üblich (Abschnitt 5.2). Das System ermittelt das statistische Abnormitätsmodell unabhängig von einer manuellen Parametrisierung. Es kann sich der Landbedeckungsanalyse also selbstständig anpassen. Dadurch wird beispielsweise die Nutzung eines Landbedeckungsanalyseverfahrens möglich, bei der sich nicht wie bei der SVM nur die Parametrisierung der manuellen Stichprobe anpasst. Stattdessen könnte ein Landbedeckunganalyseverfahren etwa die Stichprobe in für die Erkennung optimale Landbedeckungen selbst unterteilen.

Für weitere Entwicklungen würde sich auch eine Nutzung der Abnormitätsbewertung für Optimierungsverfahren anbieten. Hierfür ist die Entwicklung weiterer Verfahren zur Steuerung eines Optimierungsansatzes nötig. Zunächst könnte aber auch ein regelbasiertes System genutzt werden.

# Literaturverzeichnis


[1] OpenGIS Implementation Specification for Geographic information - Simple feature access. In: *Open Geospatial Consortium Inc.* (2011)

[2] Alparone, L.. ; Wald, L.. ; Chanussot, J.. ; Thomas, C.. ; Gamba, P.. ; Bruce, L.M.: Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data-Fusion Contest. In: *IEEE Transactions on Geoscience and Remote Sensing* 45 (2007), Oktober, Nr. 10, S. 3012–3021. – ISSN 01962892

[3] Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (Veranst.): *Amtliches Topographisch-Kartographisches Informationssystem.* 2011. – URL `http://www.atkis.de/dstinfo/dstinfo2.dst_gliederung`

[4] Becker, Christian ; Ostermann, Joern: Impacts of a resolution pyramid on Gibbs random field classification. In: *ISPRS Hannover Workshop 2009 High-Resolution Earth Imaging for Geospatial Information.* Hannover, Germany : ISPRS, 2009

[5] Bossard, M ; Feranec, J ; Otahel, J ; Steenmans, Chris: CORINE land cover technical guide – Addendum 2000 / European Environment Agency. 2000 (40). – Forschungsbericht

[6] Bottou, L. ; Cortes, C. ; Denker, J.S. ; Drucker, H. ; Guyon, I. ; Jackel, L.D. ; LeCun, Y. ; Muller, U.A. ; Sackinger, E. ; Simard, P. ; Vapnik, V.: Comparison of classifier methods: a case study in handwritten digit recognition. In: *Proceedings of the 12th IAPR International Conference on Pattern Recognition* Bd. 2, IEEE Comput. Soc. Press, 1994, S. 77–82. – ISBN 0818662700

[7] Bowen, Zachary H. ; Waltermire, Robert G.: Evaluation Of Light Detection And Ranging (LIDAR) For Measuring River Corridor Topography. In: *Journal of the American Water Resources Association* 38 (2002), Februar, Nr. 1, S. 33–41

[8] Büschenfeld, Torsten ; Ostermann, Jörn: Edge Preserving Land Cover Classification Refinement Using Mean Shift. Rio de Janeiro - Brazil., 2012, S. 242–247

[9] Curlander, John: *Synthetic aperture radar : systems and signal processing.* New York : J. Wiley, 1991. – ISBN 9780471857709

[10] Dial, Gene ; Bowen, Howard ; Gerlach, Frank ; Grodecki, Jacek ; Oleszczuk, Rick: IKONOS satellite, imagery, and products. In: *Remote Sensing of Environment* 88 (2003), November, Nr. 1-2, S. 23–36. – ISSN 00344257

[11] Duda, Richard O. ; Hart, Peter E. ; Stork, David G.: *Pattern Classification.* 2. Wiley, 2001. – 654 S. – ISBN 9755031030

[12] Esch, Thomas ; Schorcht, Gunther ; Thiel, Michael: *UmweltSpezial Flächensparen: Satellitengestützte Erfassung der Bodenversiegelung in Bayern*. Augsburg : Bayrisches Landesamt für Umwelt, 2007. – ISBN 9783940009388

[13] Fawcett, Tom: An introduction to ROC analysis. In: *Pattern Recognition Letters* 27 (2006), Juni, Nr. 8, S. 861–874. – ISSN 01678655

[14] Fischer-Stabel, Peter ; Backes, Josef: Erfassung der räumlichen Lage versiegelter Flächen in Rheinland-Pfalz: Verfahrensentwicklung und Anwendung. In: *Statistische Monatshefte Rheinland-Pfalz* (2006), Nr. 11, S. 707–716

[15] Flechl, Barbara: *Geomultimedia und deren Markteinfluss auf kartographische Produkte*. Diplomarbeit, Universität Wien. Fakultät für Geowissenschaften, Geographie und Astronomie, September 2011

[16] Fortier, MFA ; Ziou, D ; Armenakis, C: Automated correction and updating of road databases from high-resolution imagery. In: *Canadian Journal of Remote* 27 (2001), Nr. 1, S. 76–89

[17] Fortin, Marie-Josée ; Dale, Mark R. T.: *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, 2005. – 392 S. – ISBN 0521009731

[18] Friedman, J: Another approach to polychotomous classifcation. (1996)

[19] Gerke, M. ; Heipke, C.: Image-based quality assessment of road databases. In: *International Journal of Geographical Information Science* 22 (2008), August, Nr. 8, S. 871–894. – ISSN 13658816

[20] Golla, B.: Y2K Update für CORINE Land Cover - oder wieviele Bodennutzungsdaten braucht das Land. In: *Landschaftsplanung.NET* (2000), Nr. 01, S. 1–11. – ISSN 14399954

[21] Gupta, Sudhir ; Rajan, K S.: Temporal Signature Matching for Land Cover Classification. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science* XXXVIII (2010), S. 492–497

[22] Haralick, R M. ; Dinstein, I ; Shanmugam, K: Textural features for image classification. In: *Ieee Transactions On Systems Man And Cybernetics* 3 (1973), Nr. 6, S. 610–621

[23] Haunert, J.-H.: *Aggregation in Map Generalization by Combinatorial Optimization*, Gottfried Wilhelm Leibniz Universität Hannover, Germany, Dissertation, 2009

[24] Helmholz, Petra: *Verifikation von Ackerland- und Grünlandobjekten eines topographischen Datensatzes mit monotemporalen Bildern*, Gottfried Wilhelm Leibniz Universität Hannover, Dissertation, 2012

[25] Helmholz, Petra ; Becker, Christian ; Breitkopf, Uwe ; Büschenfeld, Torsten ; Busch, Andreas ; Braun, Carola ; Grünreich, Dietmar ; Müller, Sönke ; Ostermann, Jörn ; Pahl, Martin ; Rottensteiner, Franz ; Vogt, Karsten ; Ziems, Marcel ; Heipke, Christian: Semi-automatic Quality Control of Topographic

Data Sets. In: *Photogrammetric Engineering & Remote Sensing* 78 (2012), Nr. 9, S. 959–972

[26] Helmholz, Petra ; Becker, Christian ; Breitkopf, Uwe ; Büschenfeld, Torsten ; Busch, Andreas ; Grünreich, Dietmar ; Heipke, Christian ; Müller, Sönke ; Ostermann, Jörn ; Pahl, Martin ; Vogt, Karsten ; Ziems, Marcel: Semiautomatic Quality Control of Topographic Reference Datasets. In: *ISPRS Commission 4 Symposium*, November 2010

[27] Jacobsen, Karsten: High resolution satellite imaging systems - overview. In: *PFG Photogrammetrie, Fernerkundung, Geoinformation* 2005 (2005), Nr. 6

[28] Kahneman, Daniel: *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011. – ISBN 0374275637

[29] Keuchel, Jens ; Naumann, Simone ; Heiler, Matthias ; Siegmund, Alexander: Automatic land cover analysis for Tenerife by supervised classification using remotely sensed data. In: *Remote Sensing of Environment* 86 (2003), August, Nr. 4, S. 530–541. – ISSN 00344257

[30] Koecher, Max: *Lineare Algebra und Analytische Geometrie (Springer-Lehrbuch) (German Edition).* Springer, 1992. – 286 S. – ISBN 3540556532

[31] Konecny, G.: Alternatives for mapping from satellite imagery. In: *Acta Astronautica* 17 (1988), März, Nr. 3, S. 355–358. – ISSN 00945765

[32] Konecny, G.: Hochauflösende Fernerkundungssensoren für kartographische Anwendungen in Entwicklungsländern. In: *ZPF, Bd. 64(2)*, 1996, S. 39–51

[33] Krickel, Bernd: Informationserhebung zur Aktualisierung von ATKIS® und Freizeitkataster in Nordrhein-Westfalen. In: *zfv* (2010), S. 240–246

[34] Kumar, Pavan ; Rani, Meenu ; Pandey, PC ; Majumdar, Arnab: Monitoring of deforestation and forest degradation using remote sensing and GIS: A case study of Ranchi in Jharkhand (India). In: *Report and Opinion* 2 (2010), Nr. 4, S. 14–20. – ISSN 15539873

[35] Le Moigne, Jacqueline (Hrsg.) ; Netanyahu, Nathan S. (Hrsg.) ; Eastman, Roger D. (Hrsg.): *Image Registration for Remote Sensing.* Cambridge : Cambridge University Press, 2011. – ISBN 9780511777684

[36] Lin, Chin-Jen: Comparison of Methods for Multiclass Support Vector Machines. In: *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 13 (2002), Januar, Nr. 4, S. 1026–7. – ISSN 10459227

[37] Mc Garigal, K ; Marks, B J.: *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure.* Department of Agriculture, Forest Service, Pacific Northwest Research Station, 1995

[38] Müller, Sönke ; Liedtke, Claus-Eberhard: *Extraktion baulich geprägter Flächen aus Fernerkundungsdaten zur Qualitätssicherung flächenhafter Geobasisdaten* . Ibidem; Auflage: 1, 2007. – 152 S. – ISBN 3898218597

[39] Ojala, Timo: A comparative study of texture measures with classification based on featured distributions. In: *Pattern Recognition* 29 (1996), Nr. 1, S. 51–59. – ISSN 00313203

[40] Platt, John C. ; Cristianini, Nello ; Shawe-Taylor, John: Large Margin DAGs for Multiclass Classification. In: *Advances in Neural Information Processing Systems* 12 (2000), S. 547 – 553

[41] Poulain, Vincent ; Inglada, Jordi ; Spigai, Marc ; Tourneret, Jean-Yves ; Marthon, Philippe: High resolution optical and SAR image fusion for road database updating. In: *IEEE International Geoscience and Remote Sensing Symposium*, IEEE, Juli 2010, S. 2747–2750. – ISBN 9781424495665

[42] Reis, Selçuk: Analyzing Land Use/Land Cover Changes Using Remote Sensing and GIS in Rize, North-East Turkey. In: *Sensors* 8 (2008), Oktober, Nr. 10, S. 6188–6202. – ISSN 14248220

[43] Saadat, Hossein ; Adamowski, Jan ; Bonnell, Robert ; Sharifi, Forood ; Namdar, Mohammad ; Ale-Ebrahim, Sasan: Land use and land cover classification over a large area in Iran based on single date analysis of satellite imagery. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (2011), September, Nr. 5, S. 608–619. – ISSN 09242716

[44] Schowengerdt, Robert A.: *Remote Sensing, Third Edition: Models and Methods for Image Processing*. Academic Press, 2006

[45] Tan, Pang-Ning: *Introduction to Data Mining*. Addison Wesley; Auflage: 1st International edition, 2005. – 769 S. – ISBN 0321420527

[46] Thiemann, Frank ; Sester, Monika ; Bobrich, Joachim ; Data, Large V.: Automatic derivation of land-use from topographic data. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XXXVIII (2010), S. 1–6

[47] Vapnik, Vladimir: *The Nature of Statistical Learning Theory*. Berlin : Springer, 2000. – ISBN 0387987800

[48] Vewaltungsgerichtshof Baden-Würtemberg: Urteil zur Abwasserberechnung. 2010 (2010), Nr. 1

[49] Vogt, Karsten ; Scheuermann, Björn ; Becker, Christian ; Büschenfeld, Torsten ; Rosenhahn, Bodo ; Ostermann, Jörn: Automated Extraction of Plantations from Ikonos Satellite Imagery using a Level Set Based Segmentation Method. In: *ISPRS Technical Commission VII Symposium* Bd. 38, 2010, S. 275–280

[50] Walter, V.: Object-based classification of remote sensing data for change detection. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 58 (2004), Nr. 3-4, S. 225–238. – ISSN 09242716

[51] Young, Kenneth R. ; Aspinall, Richard: Kaleidoscoping Landscapes, Shifting Perspectives. In: *The Professional Geographer* 58 (2006), November, Nr. 4, S. 436–447. – ISSN 00330124

[52] Zhang, Chunsun: Towards an operational system for automated updating of road databases by integration of imagery and geodata. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 58 (2004), Januar, Nr. 3-4, S. 166–186. – ISSN 09242716

[53] Zhang, Jianjun ; Fu, Meichen ; Qi, Wenzhang ; Yuan, Chun ; Tian, Di: A Comparison of Land Cover Classification Methods Based on Remote Sensing and GIS Technologies. In: *2009 International Conference on Information Engineering and Computer Science* (2009), Dezember, S. 1–6. ISBN 97814244-49941

[54] Ziems, M ; Breitkopf, U ; Heipke, C ; Rottensteiner, F: Multiple-model based verification of road data. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* I-3 (2012), Juli, S. 329–334

# Lebenslauf

**Christian Becker**
geboren am 23.04.1978
in Hannover

## Beruflicher Werdegang

| | |
|---|---|
| 05/2004 –04/2013 | *Wissenschaftlicher Mitarbeiter* am Institut für Informationsverarbeitung (TNT) der Gottfried Wilhelm Leibniz Universität Hannover |
| seit 05/2013 | System- und Algorithmenentwicklung bei der HaCon Ingenieurgesellschaft mbH in Hannover |

## Ausbildung

| | |
|---|---|
| 08/1998 – 04/2004 | *Studium der Mathematik Studienrichtung Informatik* mit Abschluss Dipl.–Math. an der Universität Hannover |
| 08/1984 – 06/1997 | *Schulische Ausbildung* mit Abschluss Abitur an der IGS Mühlenberg in Hannover |