Zur Qualität von objektstrukturierten Geodaten

Gerhard Joos

Vollständiger Abdruck der von der Fakultät für Bauingenieur- und Vermessungswesen der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines Doktor-Ingenieurs (Dr.-Ing.) genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. Wolfgang Reinhardt
Referent: Univ.-Prof. Dr.-Ing. habil. Wilhelm Caspary
Korreferent: o. Prof. Dr.-Ing. Wolfgang Möhlenbrink (Universität Stuttgart)

Diese Dissertation wurde am 16. September 1998 bei der Universität der Bundeswehr München eingereicht.

Tag der mündlichen Prüfung: 30. März 1999

Zur Qualität von objektstrukturierten Geodaten / Gerhard Joos – Schriftenreihe des Studienganges Geodäsie und Geoinformation der Universität der Bundeswehr München. Heft 66 / 2000 (€ 10,-)

1 Einleitung und Problemstellung

Die Entwicklung von Geoinformationssystemen (GIS) befindet sich an der Schwelle von teuren Einzelplatzsystemen, die nur von wenigen Experten bedient werden konnten, zu bedienungsfreundlichen, weit verbreiteten und vernetzten Anwendungen mit Raumbezug. Dadurch dienen GIS in immer mehr Bereichen als Grundlage für Entscheidungen. Sie stellen ein zentrales Werkzeug für die Gestaltung unserer Lebensumgebung dar. Informationen, die aus einem GIS abgeleitet werden, basieren auf Daten, welche Phänomene der realen Welt beschreiben. Nur wenn die Daten korrekt sind, kann das GIS auch vernünftige Ergebnisse liefern. Wann können Daten als korrekt bezeichnet werden, und wann sind sie für eine Anwendung nicht mehr geeignet? Während bestimmte Auswertungen noch vernünftige Werte liefern, können andere Anwendungen mit denselben Daten zu unbrauchbaren Ergebnissen führen. Die Daten bilden nicht nur das Kernstück für jede Anwendung, sondern verschlingen bei der Erfassung und Pflege auch die meisten Kosten. Sie stellen daher einen kritischen Faktor für den Erfolg eines GIS dar. Um den Aufwand beim Einsatz eines GIS zu rechtfertigen, müssen Informationen, die mit Hilfe eines GIS gewonnen werden, zuverlässig sein. Wenn die Daten veraltet sind, weil sie nicht mehr gepflegt wurden, oder wenn sie unvollständig und fehlerhaft sind, geht schnell die Akzeptanz bei den Anwendern verloren und das eigentliche Ziel, nämlich eine Erhöhung der Effektivität, wird ad absurdum geführt. Außerdem können Entscheidungen, die aufgrund von fehlerhaften Informationen gefällt werden, schwerwiegende Folgen haben. Schlimmstenfalls sind Leib und Leben oder die Umwelt gefährdet. Datenfehler können aber auch zu erheblichen finanziellen Mehraufwendungen und Folgekosten führen, z.B. bei Planungen auf der Basis von fehlerhaften Daten.

Aus den genannten Gründen muß auf die Qualität der Daten ein besonderes Augenmerk gelegt werden. Wenn die Eignung der Daten angesprochen wird, so ist dies immer von der gewünschten Anwendung abhängig. Geodaten insbesondere Geobasisdaten zeichnen sich aber gerade dadurch aus, daß sie für möglichst viele verschiedene Aufgaben geeignet sein sollen. Die Beschreibung der Qualität muß also möglichst umfassend erfolgen, so daß anhand der Qualitätsbeschreibung die Eignung für bestimmte Anwendungen abgelesen werden kann.

Der in dieser Arbeit verwendete Qualitätsbegriff geht auf die Definition der internationaler Norm in deutscher Übersetzung *DIN ISO 8402, 1991*, zurück. Danach ist Qualität die Gesamtheit von Eigenschaften und Merkmalen eines Produktes oder einer Dienstleistung, die sich auf deren Eignung zur Erfüllung festgelegter oder vorausgesetzter Erfordernisse beziehen. Aus dieser Definition ergeben sich für Geodaten die Fragen, was die Erfordernisse sind, wie sie beschrieben werden können und wie sichergestellt werden kann, daß die Geodaten auch den Erfordernissen entsprechen.

Daten eines Geoinformationssystems bezeichnet man dann als fehlerhaft, wenn ihr Informationsgehalt nicht der Situation vor Ort entspricht. Da aber die reale Welt in ihrer Komplexität nicht vollständig erfaßt werden kann, läßt sich der Zusammenhang zwischen den Daten und der Realität nur über eine Abstraktion herstellen. Die Wissenschaft verwendet zur Beschreibung von Phänomenen der realen Welt Modelle, welche die Realität in Abhängigkeit von einer Zielsetzung möglichst gut repräsentieren sollen. Weil ohne eine entsprechende Modellierung keine Daten erfaßt werden können, befaßt sich das Kapitel 1 mit Methoden zur Modellierung der realen Welt für Geoinformationssysteme.

Um Geodaten verstehen und damit richtig verwenden zu können, muß der Anwender das den Daten zugrundeliegende Modell kennen. Er braucht also zu den eigentlichen Geodaten noch weitere Angaben über diese Daten. Mit diesen Zusatzdaten, auch Metadaten genannt, beschäftigt sich das Kapitel 3. Das Modell reicht allerdings zur vollständigen Beschreibung der Daten nicht aus, weil der Anwender vor der Verwendung abklären sollte, ob die Daten verfügbar sind, wo sie herkommen, welches dokumentierte Qualitätsniveau sie besitzen und ob Referenzanwendungen bekannt sind.

Zur objektiven Beschreibung des Qualitätsniveaus bedarf es eines Qualitätsmodells. In Kapitel 4 wird ein System von Qualitätskriterien und Qualitätsmaßen aufgebaut, das eine vollständige und anwendungsunabhängige Taxierung der Qualität von Geodaten erlaubt. Die Qualitätskriterien werden nicht nur verbal, sondern auch mit Hilfe der Prädikatenlogik formuliert. Dadurch ist es möglich Kriterien abzuleiten, bei denen die Bedingungen verletzt sind. Um die Qualität verschiedener Datensätze vergleichen zu können, sind quantitative Angaben erforderlich. Diese setzen voraus, daß

neben den Qualitätskriterien auch Qualitätsmaße definiert sein müssen. Die Bezugseinheit ist für Qualitätsmaße eine wichtige Größe, da sich die Maße auf ganze Datensätze bzw. Teile daraus beziehen können, oder eingeschränkt auf eine oder mehrere Objektklassen, auf einzelne Objekte oder nur auf bestimmte Attribute einer Menge von Objekten oder eines einzelnen Objektes. Abhängig von der Bezugseinheit werden unterschiedliche Qualitätsmaße definiert.

Um aus den digitalen Geodaten die gewünschten Informationen abzuleiten, sind oft sehr komplexe Analysen erforderlich. Um die Qualität dieser abgeleiteten Informationen objektiv bewerten zu können, ist es anzustreben, aus den Qualitätsmaßen der Daten Qualitätsmaße der Analyseergebnisse abzuleiten. Dazu müßten aber dann Methoden entwickelt werden, die dies leisten. Dieser Aspekt der Auswertung von Qualitätsinformationen geht über die Zielsetzung dieser Arbeit hinaus.

In den Kapiteln 5, 6 und 7 werden Methoden erarbeitet, mit denen sichergestellt werden kann, dass Qualitätsziele eingehalten sind. Qualitätsziele sind dabei vor der Erfassung festgelegte Werte für Qualitätsmaße, die als untere Schranke des angestrebten Qualitätsniveaus dienen.

Das in Kapitel 5 beschriebene Qualitätsmanagement bei der Erfassung von Geodaten beruht auf einem organisatorischen Ansatz, der sich an der internationalen Normenfamilie ISO 9000ff orientiert. Durch festgeschriebene Verfahrensabläufe wird gewährleistet, daß eine Organisation zu jeder Zeit einen Überblick über die Qualität der erfaßten Geodaten hat und bei Bedarf in den Prozeß der Digitalisierung verbessernd eingreifen kann. Durch eine umfassende Dokumentation kann der Erfassungsprozeß zurückverfolgt werden.

Im Datenmodell verankerte eindeutige Regeln führen zu Konsistenzprüfungen (Kapitel 6). Durch eine Abstraktion der Regeln können Typen von Grundregeln gebildet werden, die bei einer formalen Festlegung in Prüfroutinen des GIS überführt werden können. Da durch den automatischen Prüfablauf alle Inkonsistenzen eines Datenbestandes ermittelt werden können, stellen Konsistenzprüfungen wesentliche Hilfsmittel beim Qualitätsmanagement dar.

Da nicht für alle Qualitätskriterien automatische Prüfroutinen aufgestellt werden können, ist bei vielen Kontrollen der menschliche Bearbeiter erforderlich, der durch visuellen Vergleich der Geodaten mit der als korrekt anerkannten Situation (z.B. Vergleich im Feld, mit Luftaufnahmen oder Original-erfassungsquellen) nach fehlerhaften Objekten sucht. Weil diese Methode sehr aufwendig ist, kann unter bestimmten Bedingungen nur eine Teilmenge der Objekte untersucht und daraus auf das Qualitätsniveau des gesamten Datensatzes geschlossen werden. Unter welchen Bedingungen dies möglich ist, und wie ein solcher Stichprobenplan mit statistischen Methoden aufgestellt werden kann, wird in Kapitel 7 behandelt.

Damit Qualitätsangaben von unterschiedlichen Datensätzen verglichen werden können, müssen sie auf vergleichbaren Qualitätsmodellen beruhen. Aus diesem Grund wurden bei verschiedenen internationalen Standards und Normentwürfen für Geoinformationen Qualitätsaspekte eingearbeitet. Kapitel 8 vergleicht eine Auswahl der derzeit aktuellen oder in Entwicklung befindlichen Normen und Standards.

Stand der Wissenschaft

Die Idee zu dieser Arbeit ergab sich aus dem Fehlen eines Modells, mit dem unabhängig von einer Anwendung die Qualität von Geodaten objektiv beschrieben werden kann. Der Bedarf an einem solchen Qualitätsmodell ist in der Wissenschaft schon bald erkannt worden, nachdem Geoinformationssysteme in der Praxis eingesetzt wurden (*Chrisman, 1982*). Zahlreiche internationle GIS-Konferenzen wurden veranstaltet, bei denen der Datenqualität-Problematik eine wichtige Rolle beigemessen wurde. In den Beiträgen wurde insbesondere der Aspekt zur Beschreibung der geometrischen Genauigkeit von Geoobjekten behandelt (*Shi, 1994, Scheuring, 1995, Glemser, 1996, Vauglin, 1999*).

Veregin, 1989, Caspary, 1993, und *Stanek und Frank, 1993*, geben fünf Parameter zur Kategorisierung von Qualitätsinformation an.

Positionsgenauigkeit	Attributgenauigkeit (bzw. thematische Genauigkeit)	Aktualität	Konsistenz	Vollständigkeit

Die Monographie „Elements of Spatial Data Quality“ (*Guptill and Morrison, 1995*), initiiert durch die internationale kartographische Assotiation (ICA), gibt eine Übersicht von sieben Qualitätselementen:

Lineage	*Positional accuracy*	*Attribute accuracy*	*Completeness*	*Logical consistency*	*Semantic accuracy*	*Temporal information*

Die Herkunft (*lineage*) und die Aktualität (*temporal information*) fungieren als indirekte Qualitätsmerkmale. Sie lassen auf die Qualität von Geodaten schließen, ohne sie direkt zu quantifizieren. Die Genauigkeit bezieht sich auf eine ein- oder mehrdimensionale Zufallsvariable, daher kann die Genauigkeit von geometrischen und attributiven Eigenschaften mit denselben Methoden behandelt werden. Aus diesem Grund ist es naheliegend, die Genauigkeit von quantitativen Eigenschaften von Geoobjekten mit einer einheitlichen Methode zu beschreiben. Das Element semantische Genauigkeit wurde mit der Intension eingeführt, nicht quantitative Eigenschaften zu beschreiben. Als Beispiele werden Misklassifizierungen bei der Zuweisung zu Objektarten oder zu Attributwerten angegeben.

Alle vorgeschlagenen Qualitätskriterien entbehren einer einheitlichen, insbesondere einer formalen Definition als Prädikate. Die gemeinsamen Aspekte zur Kontrolle und zur Verbesserung der Qualität von objektstrukturierten Geodaten werden nicht behandelt. Die genannten Arbeiten stellen aber die Grundlage zur Neustrukturierung von Qualitätskriterien in Kapitel 4 dar.

Durch die Modellierung wird die Referenz d. h. der Bezug zur Erfassung und Prüfung von Geodaten festgelegt. Verschiedene Aspekte der Modellierung auf konzeptioneller, logischer und physikalischer Ebene kommen dabei zum Tragen. Diese werden in GIS-Lehrbüchern (*Burrough and MacDonnell, 1998, Göpfert, 1991, Bill und Fritsch, 1994*, und *Bartelme, 1995*) und Veröffentlichungen unterschiedlich behandelt, aus diesem Grund wird in Kapitel 2 der Stand der Wissenschaft mit den Begriffen, wie sie in dieser Arbeit verstanden werden, eingeführt.

Zur Prüfung der Konsistenz haben *Kainz, 1995, Plümer, 1996a* und *1996b, Plümer und Gröger, 1996, Wise, 1998*, und andere Autoren wissenschaftliche Beiträge geleistet. Die Veröffentlichungen behandeln die Konsistenz auf logischer Modellierungsebene. Dabei werden Bedingungen zur Einhaltung der strukturellen Integrität von Geodaten formuliert und Verfahren zur Prüfung oder zur Sicherstellung entwickelt. Bedingungen, die sich aus der konzeptionellen Modellierung ableiten lassen, wurden aus wissenschaftlicher Sicht bisher nicht behandelt. Zur Formulierung von topologischen Beziehungen wird das 9-Intersection-Modell von *Egenhofer et al., 1994*, verwendet, das im Anhang A dargestellt wird.

Die Statistische Qualitätskontrolle ist in der industriellen Fertigung seit den 40er Jahren eine bewährte Methode zur Kontrolle von Erzeugnissen mit Hilfe von Stichproben bei der Produktion und bei der Übernahme von Gütern (*Wald, 1945*, und *Mace, 1964*). Durch die internationale Normierung der Stichprobenprüfung (*ISO 2859, 1991*, und *ISO 3951, 1992*) wurde die Theorie für die praktische Anwendung aufbereitet (*Uhlmann, 1982*, und *Schilling, 1982*). Neuere Arbeiten (*Tayi, 1995*) beschäftigen sich mit der Auswirkung von Just-in-time Produktion auf die Qualitätskontrolle, da unter dieser Randbedingung Losgrößen, Zeitpunkt der Kontrolle und Nachbearbeitung überdacht werden müssen. Spezielle Verfahren zur Reduzierung des mittleren Stichprobenumfangs werden in einer Dissertation vom *Müller, 1998* angegeben. Stichprobenprüfungen finden im GIS-Bereich unter anderem bei der Erfassung von Daten für Fahrzeugnavigationssysteme Anwendung. Allerdings fehlt eine wissenschaftliche Grundlage, die Normen zur Stichprobenprüfung auf Geodaten zu übertragen. Nach einer kurzen Darstellung des Standes der Wissenschaft zur statistischen Qualitätskontrolle wird in Kapitel 7 auf die Besonderheiten von Geodaten eingegangen. Ein Verfahren zur Bestimmung von Qualitätszielen nach Wirtschaftlichkeitsüberlegungen für Geodaten wird entwickelt.

In der europäischen (CEN) und der internationalen (ISO) Initiative zur Normierung von Geoinformation sind jeweils Arbeitspakete zur Behandlung der Qualität von Geodaten eingerichtet (*DIN V ENV 12656, 1999, ISO 191113, 1999, ISO 19114, 1999*). In den Normen wird ein Rahmen zum Verständnis des Konzeptes von Datenqualität gegeben. Es werden exemplarisch Qualitätskriterien aufgezeigt, aber keine Qualitätsmaße vorgegeben.

2 Modellierung der realen Welt

In einem Geoinformationssystem soll die reale Welt in Form von Daten repräsentiert werden, damit aus dem System Informationen über Phänomene, die einen Bezug zur Erde haben, ermittelt werden können. Dabei wird die reale Welt durch eine spezielle, meist fachgebundene Sicht der Wirklichkeit abstrahiert. Der Abstraktionsvorgang wird als Modellierung bezeichnet, das Ergebnis stellt ein Modell der Realität dar. Der Vorgang der Aufteilung und Klassifizierung von Information in ihre Bestandteile bezeichnet *Bartelme, 1995*, als Analyse. Das Formen von komplexen Gebilden aus den einzelnen Teilen wird *ebenda* Synthese genannt. Das Modell kann als eine Vorschrift betrachtet werden, die Objekte der realen Welt selektiert, sie benennt und festlegt, welche geometrischen und beschreibenden Merkmale die Objekte für die konzipierten Anwendungen charakterisieren.

Wenn von einem realen Objekt gesprochen wird, impliziert diese Ausdrucksweise schon Abgrenzungen und Aggregationen verschiedener Phänomene der Natur. Tatsächlich existiert die Kategorisierung nur durch Vergleich der Wahrnehmung des Menschen mit erlernten Mustern. Diese wirken bei der Erfahrung der Welt mit den Sinnen als Filter. Die Klassenbildung ist also durch Erfahrungen und durch das Wissen des Wahrnehmenden geprägt.

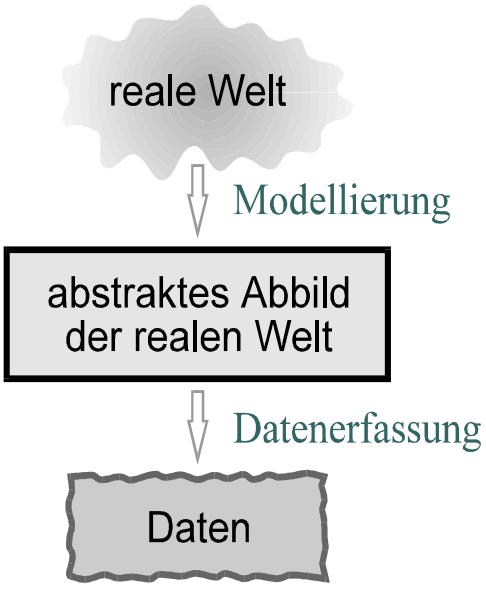


Abbildung 1: Von der realen Welt zu Daten über die reale Welt.

Wegen der fachspezifischen Wahrnehmungsgewohnheiten unterschiedlicher Disziplinen und um Fehlinterpretationen zu vermeiden, ist es erforderlich, die Klassen, in welche die realen Objekte eingeteilt werden, genau zu definieren. Das überführt bestimmte Phänomene in eine klare Struktur, die eine vollständige, eindeutig abgegrenzte Kategorisierung beinhaltet. Die Projektion einer durch das Datenmodell vorgegebenen Teilmenge der realen Welt wird als abstraktes Abbild der realen Welt bezeichnet. Die europäische Norm „Geoinformation – Datenbeschreibung – Qualität“ *(prEVN 12656, 1998)* verwendet dazu den Begriff „konzeptuelle Wirklichkeit“.

Die positivistische Sichtweise, daß alle Phänomene durch Regeln eindeutig und scharf abgegrenzt werden können, stößt bei einigen natürlichen Erscheinungen an Grenzen. Tatsächlich besteht ein enger Zusammenhang zwischen dem Stand der Technik von GIS-Software und dem abstrakten Abbild der realen Welt, weil sich die Modellierung immer an den Stand der Technik halten muß. Es macht keinen Sinn, Phänomene modellhaft zu beschreiben, die nicht in einem GIS abgebildet werden können.

Einige neuere Forschungsansätze versuchen, das abstrakte Abbild der realen Welt um Objekte mit unscharfen Grenzen zu erweitern (*Caspary, 1995*). Die Beschreibung von unscharfen Grenzen muß mathematisch exakt erfolgen, weil ein GIS nur mathematisch eindeutige Repräsentationen verwalten kann. In der zweiwertigen Booleschen Algebra kann eine Aussage über die Zugehörigkeit eines Elements zu einer Menge nur die beiden Werte „wahr“ oder „falsch“ annehmen. Die Erweiterung dieser Algebra, bei der beliebig viele Zwischenwerte angenommen werden können, wird als „fuzzy set theory“ bezeichnet. Die Modellierung von kontinuierlichen Übergangsfunktionen an den Nahtstellen zwischen unscharfen Objekten kann über die „fuzzy set theory“ erfolgen (*Molenaar, 1995*).

Das abstrakte Abbild der realen Welt muß unabhängig davon, wie komplex die Möglichkeiten zur Modellierung sind, immer ein eindeutiges, widerspruchsfreies Modell der Realität darstellen, denn dadurch wird festgelegt, was bei der Digitalisierung der Daten erfaßt werden soll.

Bevor eine Festlegung von Datenstrukturen erfolgen kann, muß bei der Konzipierung eines Systems erst über den Inhalt nachgedacht werden. Die Frage danach, welche Informationen aus dem System abgeleitet werden sollen, steht dabei im Vordergrund. *Tomlinson, 1997,* spricht in diesem Zusammenhang von Informationsprodukten, weil die Präsentation der Informationen berücksichtigt werden muß, um ein optimales Systemdesign zu erzielen. Erst wenn die Frage nach den Informationsprodukten hinreichend geklärt ist, kann überlegt werden, welche Methoden zur Gewinnung dieser Produkte benötigt werden, und welche Daten dazu erforderlich sind. Eine Gliederung vieler Lehrbücher nach den Punkten Hardware, Software und Daten in dieser Reihenfolge suggeriert einem Systemdesigner eine entgegengesetzte Vorgehensweise.

Als Grundlage für Fachinformationssysteme, GIS, die für bestimmte Fachanwendungen konzipiert und aufgebaut werden, stellen einige Institutionen (z.B. die Landesvermessungsämter) sogenannte Geobasisdaten zur Verfügung. Diese Daten zeigen entweder die charakteristisch vereinfachte Landschaft (topographische Geobasisdaten) oder die Eigentumsverhältnisse (Daten über das Liegenschaftskataster, Liegenschaftsbuch und Grundbuch). Die Geobasisdaten sind anwendungsübergreifend und so zu modellieren, daß sie den räumlichen Bezug für möglichst viele Anwendungen bilden können.

2.1 Beschreibung der verschiedenen Modellierungsebenen

Wenn geklärt ist, welche Informationsprodukte mit welchen Methoden aus welchen Daten abzuleiten sind, muß die Struktur der Daten festgelegt werden. Natürlich ist erst zu prüfen, ob die erforderlichen Daten an anderer Stelle verfügbar sind, und ob diese Daten den Anforderungen für die gewünschte Anwendung genügen. Letzteres ist ein zentraler Punkt dieser Arbeit und wird in den weiteren Kapiteln ausführlich diskutiert. Zuvor wird auf die Elemente des Datenmodells eingegangen.

Die Modellierung von Geodaten erfolgt auf verschiedenen Ebenen. In der Literatur werden die Ebenen unterschiedlich festgelegt und deswegen auch unterschiedlich bezeichnet.

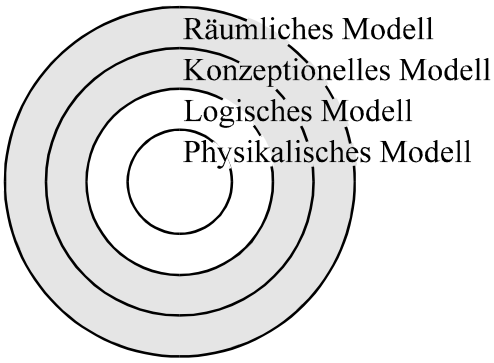


Abbildung 2: Das Vierschalen-Modell nach *Bill und Fritsch, 1994*

Bill und Fritsch, 1994, legen ein Vierschalen-Modell fest und verdeutlichen damit gleichzeitig die Interdisziplinarität, welche der Aufbau eines Geoinformationssystems erfordert. Während in den äußeren Schalen die Sicht eines Anwenders benötigt wird, ist mit zunehmender Annäherung an den Kern das Fachwissen eines Informatikers erforderlich. Der Geo-Informatiker soll das Bindeglied zwischen den beiden Spezialisierungen darstellen.

Im räumlichen Modell legt der Anwender das Objekt hinsichtlich seiner thematischen Ausdehnung und Abgrenzung fest. Im konzeptionellen Modell werden die vom Anwender vorgegebenen Objekte in festgelegte Strukturen abgebildet. Die eigentliche Datenverwaltung erfolgt in einem Datenbankverwaltungssystem auf der Basis des logischen Modells. Dieses setzt als physikalische Organisationsform auf Baum- und Listenstrukturen auf.

In Anlehnung an die Datenbanktechnologie unterscheidet *Bartelme, 1995,* zwischen drei Schemata. Ein Schema stellt dabei die formalisierte Beschreibung des Modells dar. Dazu kann eine lexikalische oder graphische Sprache verwendet werden. Bartelme unterscheidet zwischen

- dem externen Schema (anwendungsspezifisch)
- dem konzeptionellen Schema
- und dem internen Schema (logisches und physisches Schema).

Diese Aufteilung wird als Drei-Schema-Architektur bezeichnet. Dabei identifiziert und beschreibt das konzeptionelle Schema Objekte, deren Charakteristika und gegenseitige Beziehungen in einer allgemeingültigen und eindeutigen Form, die von konkreten Anwendungen unabhängig ist. Es bildet

den Unterbau für das externe Schema, das anwendungsspezifisch ausgerichtet ist. Das interne Schema stellt eine Vereinigung von logischem und physischem Datenschema dar. Das logische Schema enthält nach Bartelme Festlegungen wie etwa die Tabellenstruktur, Relationen, Such- und Sortierkriterien. Im physischen Schema ist beschrieben, wie das logische Schema auf das physische Speichermedium abzubilden ist.

Für das Design komplexer Systeme schlägt *Booch, 1996,* die objektorientierte Zerlegung vor, um gegenüber Veränderungen möglichst tolerant zu sein, und um Software mit Hilfe möglichst weniger Hilfsmittel realisieren zu können. Die objektorientierte Denkweise, bei der Daten und die damit verbundenen Methoden als Einheit gesehen werden, entspricht einer sehr modernen Sicht auf Systeme. Da die Geodaten, wie sie in der vorliegenden Arbeit betrachtet werden, für viele Anwendungen zur Verfügung stehen sollen, und deren Methoden beim Aufbau des Modells unter Umständen noch nicht bekannt oder zumindest nicht endgültig festgelegt sind, bleiben die Methoden bei der Beschreibung der Objekte im allgemeinen unberücksichtigt. Die Konzepte der objektorientierten Modellierung werden als eine mögliche Form der Beschreibung der realen Welt herangezogen.

Der Begriff digitales Geoobjekt, oder kurz Objekt, wie er in dieser Arbeit angewendet wird, soll nicht als Objekt der objektorientierten Programmierung verstanden werden.

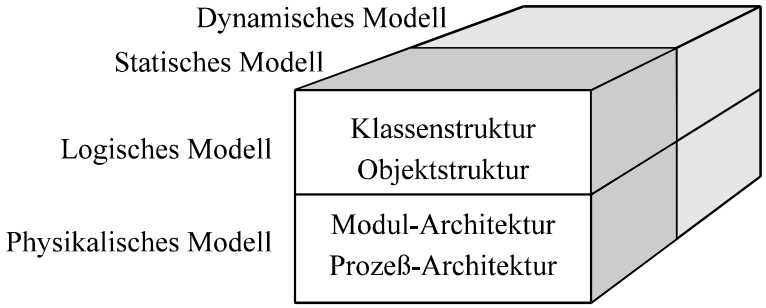


Abbildung 3: Die Methode der objektorientierten Modellierung nach Booch.

Die für Geoinformationsysteme relevanten Abstraktionsstufen, wie sie *Maguire und Dangermond, 1991*, verwenden, gehen auf *Peuquet, 1984*, zurück. In einer hierarchischen Ordnung lauten die Ebenen von oben nach unten:

- Datenmodell oder konzeptionelles Modell (*data model or conceptual model*)
- Datenstruktur oder logisches Modell (*data structure or logical model*)
- Dateienstruktur oder physikalisches Modell (*file structure or physical model*)

Diese Gliederung in Abstraktionsebenen der Modellierung und die damit verbundenen Modelle zur Beschreibung der realen Welt wird der vorliegenden Arbeit zugrunde gelegt und in den nächsten drei Abschnitten beschrieben.

Da das Datenmodell nicht die Daten modelliert, wie der Name assoziiert, sondern die reale Welt, oder auch die Informationen, die in der realen Welt enthalten sind, wird in der Informatik oft auch der Begriff **Informationsmodell** verwendet. Das logische und physische Modell modellieren tatsächlich die Daten, deshalb ist der Begriff **Datenmodell** an dieser Stelle durchaus angebracht. In der Literatur wird der Begriff Datenmodell allerdings oft für alle Stufen der Abstraktion von der realen Welt bis zur Speicherbelegung verwendet.

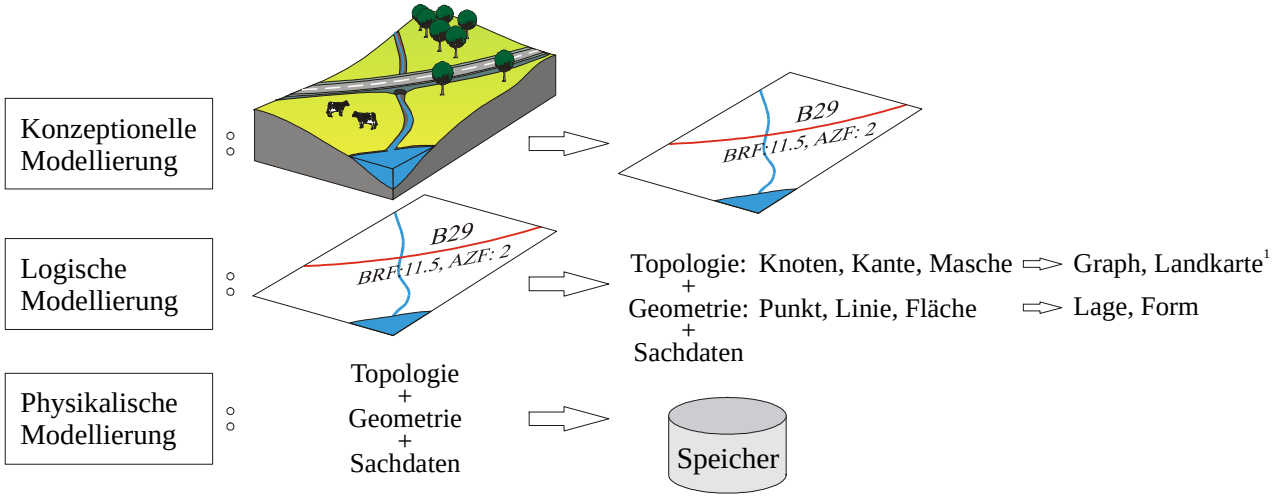


Abbildung 4: Die drei Ebenen der Modellierung und die damit verbundenen Abbildungen.

2.2 Konzeptionelle Modellierung

Durch die konzeptionelle Modellierung[2] wird festgelegt, welche Objekte der realen Welt wie und unter welchen Restriktionen erfaßt werden sollen. Der Modellierer legt somit den Inhalt, die Struktur und die Regeln eines Datenbestandes fest. Änderungen oder Ergänzungen können nur durch eine Modellanpassung erfolgen. Bevor mit der Digitalisierung begonnen wird, muss das Modell also feststehen. Existieren bei einer Modelländerung schon Daten, muss eine Migration erfolgen. Diese kann sehr aufwendig werden.

Der Anwender kann durch sorgfältiges Studieren des konzeptionellen Modells entscheiden, ob die Daten für eine Anwendung prinzipiell geeignet sind.

Zum Beispiel kann eine Straße auf sehr unterschiedliche Arten modelliert werden. Für Algorithmen zur Auffindung der kürzesten Verbindung in einem Straßennetz müssen die Daten eine Knoten - Kanten - Struktur aufweisen. Liegen die Straßen allerdings in einer flächenhaften Struktur vor, so muß die Straßenachse als Kante erst aus den Daten abgeleitet werden. In einer Flurstücksdatenbank , wie dem Automatisierten Liegenschaftskataster (ALK) der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland werden Straßen über die Straßenflurstücke als flächenhafte Objekte modelliert..

2.2.1 Inhalt

Bei der inhaltlichen Festlegung werden von den Objekten der realen Welt diejenigen ausgewählt, welche im Datensatz enthalten sein sollen. Um die Welt in eine sinnvolle und vertraute Struktur zu bringen, werden Objektklassen vereinbart, in denen alle Objekte mit gleicher Thematik zusammengefaßt werden.

2.2.1.1 Definition

Zur eindeutigen Zuordnung eines Objektes zu einer Klasse, muß die Objektklasse unmißverständlich definiert werden. Diese **Definition** ist insbesondere deshalb erforderlich, da unterschiedliche Fachdisziplinen verschiedene Vorstellungen von Objektklassen besitzen können. Die Definitionen sollten die Objektklassen in einen höheren Zusammenhang einordnen und nur die Charakteristika enthalten, die Objektklassen von anderen verwandten Klassen unterscheiden. Sie sollen weder interne Ringschlüsse enthalten, bei denen ein Wort mit einem Wort des gleichen Wortstammes definiert wird, noch externe Ringschlüsse, bei denen ein Wort auf eine andere Definition verweist, in der genau das

[1] Der Begriff Landkarte wird hier im Sinne der Graphentheorie verwendet (*Aigner, 1984, Volkmann, 1996*)

[2] Nach *DIN V ENV 12009, 1997* auch konzeptuelle Modellierung genannt

Wort wieder verwendet wird. In anderen Definitionen festgeschriebene Wörter können verwendet werden, ohne daß die Definition wiederholt werden muß. Außerdem sollen keine elliptischen Definitionen verwendet werden. Bei einer elliptischen Definition wird ein allgemeiner Begriff in einer sehr eingeschränkten Weise definiert, so daß der Eindruck entstehen kann, daß der Begriff nur in diesem Kontext existiert (*ISO/TC 211 N365, 1997*).

2.2.1.2 Erfassungskriterien

Nicht alle Objekte einer Klasse müssen notwendigerweise auch erfaßt werden. Daher ist es erforderlich im Modell **Erfassungskriterien** aufzustellen. Die Erfassungskriterien bilden Restriktionen. Sie können in drei Gruppen eingeteilt werden: geometrische Restriktionen, semantische Restriktionen und Auswahl nach kartographischen Gesichtspunkten. Als geometrische Erfassungskriterien gelten Mindestwerte für geometrische Größenangaben: Fläche, Länge, Breite und Höhe. Damit soll verhindert werden, daß zu viele unbedeutende Objekte in den Datenbestand übernommen werden (*Joos, 1996*). Die Bedeutung eines Objektes ist entweder von seinem tatsächlichen Gebrauch oder von einer semantischen Klassifizierung abhängig. Bei hoheitlichen Einteilungen wird häufig auch der Begriff Widmung verwendet. Diese Restriktion bezieht sich auf eine Eigenschaft von Objekten.

Wenn die Ausschlußkriterien zu restriktiv sind, weil die Charakteristik einer Gruppe von Objekten bei ausschließlicher Betrachtung der Individuen verloren geht, können auch kartographische Gesichtspunkte bei der Auswahl von Objekten herangezogen werden. Dieser Vorgang der qualitativen Generalisierung bei der Modellierung läßt sich nicht nach gemeingültigen, formalen Vorschriften festlegen, dadurch kann bei dieser Vorgehensweise kein objektives abstraktes Abbild der realen Welt im absoluten Sinn erzeugt werden (*Hake und Grünreich, 1994*). Ein Beispiel für eine Auswahl nach kartographischen Gesichtspunkten stellt eine Seenplatte dar, bei der jeder einzelne See die Mindestfläche unterbietet, aber das Gebiet stark durch die vielen Seen geprägt ist, so daß die Seenplatte nicht vernachlässigt werden darf. Eine Abhilfe ergibt sich dabei, wie in der Kartographie üblich, durch eine Zusammenfassung mehrerer Seen zu einem See, oder durch Erfassung von Seen, obwohl sie gegen die Regel verstoßen. Diese muß in den Metadaten über das konzeptionelle Modell erläutert werden.

Die genannten Erfassungskriterien sind hierarchisch voneinander abhängig. Die Entscheidungsfindung, ob ein Objekt erfaßt werden soll, läßt sich durch das Flußdiagramm in Abbildung 5 darstellen.

Ob ein Objekt unbedeutend ist, wenn es eine bestimmte Mindestgröße unterschreitet oder gewisse Eigenschaften nicht aufweist, ist stark von der Anwendung abhängig. Daher sind die Erfassungskriterien wichtige Informationsquellen zur Einschätzung der Eignung eines Datensatzes für eine bestimmte Anwendung.

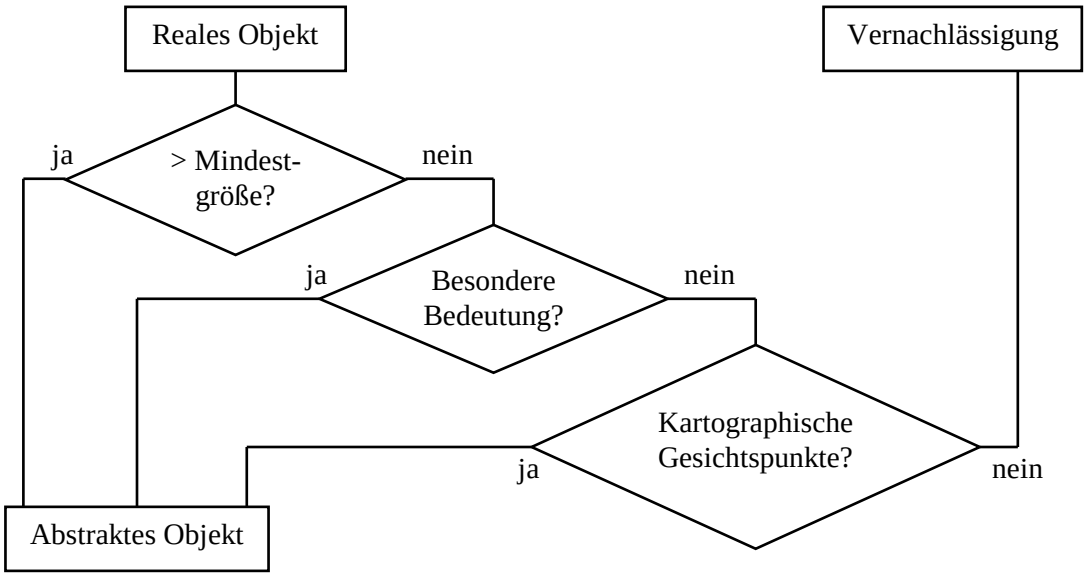


Abbildung 5: Flußdiagramm der Hierarchie von Erfassungskriterien.

2.2.1.3 Gebiet

Für die Beschreibung der räumlichen Ausdehnung von Daten wird ein **Gebiet** festgelegt. Alle Aufgaben in einem Geoinformationssystem beziehen sich auf ein bestimmtes Gebiet. In dieser Arbeit wird zur Bezeichnung eines Gebietes die Abkürzung *G* oder *g* eingeführt. Der Großbuchstabe bezieht sich dabei auf die Festlegung im abstrakten Abbild der realen Welt und das kleine *g* bezeichnet das Pendant in den digitalen Daten.

Das Gebiet ist räumlich beschränkt und im Normalfall einfach zusammenhängend, kann aber auch n-fach zusammenhängend sein ($n < \infty$). Beispiele für einfach zusammenhängende Gebiete sind durch Blattschnitt begrenzte Kartenflächen. Als dreifach zusammenhängendes Gebiet würde ein Landkreis mit einer Ex- und einer Enklave bezeichnet.

Abbildung 6: a) g_1:einfach zusammenhängendes Gebiet b) g_2: fünffach zusammenhängendes Gebiet c) g_3: fünffach zusammenhängendes Gebiet, topologisch äquivalent zu b).

Für eine zweidimensionale Modellierung, stellt die Abbildung 6 mehrere Möglichkeiten zusammenhängender Gebiete dar.

Das Gebiet kann geometrisch über die räumliche Ausdehnung definiert werden oder als Menge von Geoobjekten. Bei flächenhaften Geoobjekten ergibt sich das Gebiet über die Vereinigung aller Einzelobjekte. Bei punkt- und linienhaften Geoobjekten kann das Gebiet über die konvexe Hülle der Einzelobjekte festgelegt werden. Für viele Zwecke wird zur Definition des Gebietes ein umschreibendes Rechteck mit Seiten parallel zu den Koordinatenachsen benötigt. In diesem Fall reichen die Koordinatenpaare zweier diagonal liegender Ecken zur Beschreibung des Gebietes aus.

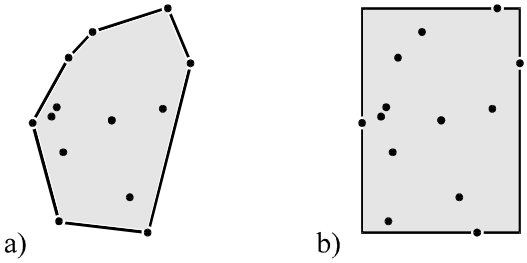


Abbildung 7: Das Gebiet festgelegt durch a) die konvexe Hülle eines Punkthaufens b) umschreibendes Rechteck derselben Punktmenge.

2.2.1.4 Attribute

Bei der Angabe des Inhalts des konzeptionellen Modells müssen die Eigenschaften der Objekte in Form von **Attributen** beschrieben werden. Zu jeder Objektklasse werden die zugehörigen Attribute vereinbart. Die Attribute müssen zuerst definiert werden. Dabei gelten dieselben Regeln wie bei der Definition von Objektklassen. Mit der Definition muß festgelegt werden, auf welche Einheiten sich die Attributwerte beziehen. Für physikalische Größen ist es anzustreben, die international eingeführten Einheiten des SI (Système International d‘Unités) zu verwenden.

Wenn nicht alle Attribute auf alle Instanzen der Objektklasse zutreffen, ist es sinnvoll zwischen erforderlichen und optionalen Attributwerten zu unterscheiden. Bei den erforderlichen Attributen muß für jede Instanz ein korrekter Wert eingetragen werden. Bei den optionalen Attributen erfolgt nur ein Eintrag, wenn das jeweilige Objekt eine Eigenschaft hat, die mit einem gültigen Wert beschrieben werden kann.

2.2.2 Struktur

Die abstrakten Objekte weisen eine Struktur auf, so daß Ordnungsprinzipien der realen Welt nachgebildet werden. Die Struktur gibt an, wie man aus atomaren Bestandteilen zu höherwertigen Komplexen kommt (*Bartelme, 1995*). Die Ordnungen haben entweder hierarchischen Charakter, indem Objekte bestimmter Klassen zu übergeordneten Objektklassen zusammengefaßt werden, oder sie strukturieren die geometrische Repräsentation der Objekte.

2.2.2.1 Hierarchische Gliederung

Zur besseren Übersicht und um gewisse Abfragen auf die Daten einfacher gestalten zu können, ist es hilfreich eine **hierarchische Gliederung** in das konzeptionelle Modell zu bringen. Dabei werden Ober- und Unterklassen gebildet. Diese Strukturierung kann auch als thematischer Baum oder Objekthierarchie bezeichnet werden (*Bill und Fritsch, 1994*).

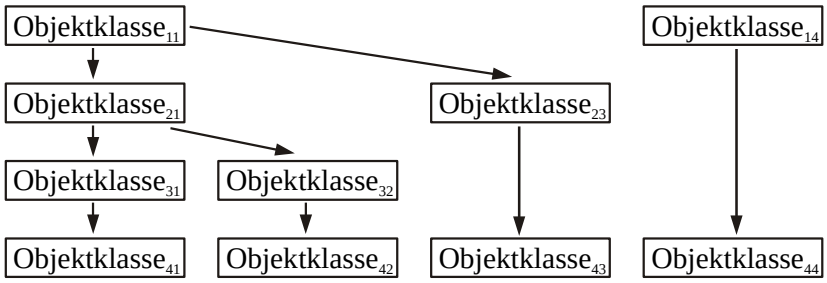


Abbildung 8: Hierarchische Gliederung mit unterschiedlicher Anzahl von Zwischenstufen.

Diese Vorgehensweise entspricht einer Gruppenbildung von Objektklassen. Es können komplexe Objekte vereinbart werden, die eine Verbindung mehrerer gleich- oder verschiedenartiger Objekte zu einem neuen Objekt darstellen. Um Objekten, bei denen bestimmte Eigenschaften gleich sind, nicht mehrfach denselben Attributwert zuweisen zu müssen, ist es sinnvoll Objekte und Teilobjekte einzuführen, wobei ein Objekt ein oder mehrere Objektteile besitzen kann. Dadurch wird eine redundante Speicherung von Attributwerten vermieden. Für viele Anwendungen und zur Pflege der Daten ist dies von Vorteil. Als Träger der geometrischen Information ist immer die unterste Stufe der Hierarchie anzusehen.

2.2.2.2 Indirekte Positionsangabe

Neben einer geometrischen Lageangabe, die üblicherweise mit Koordinaten erfolgt (*ISO 19111, 1999*), besteht die Möglichkeit, die Position eines Objektes auch indirekt zu beschreiben (*ISO 19112, 1999*). Mit einer Adressenangabe aus Ort, eventuell Postleitzahl, Straßenname und Hausnummer oder im Kataster mit Hilfe von Gemeinde, Gemarkung, Flur und Flurstücksnummer kann ein Objekt Haus oder Flurstück angesprochen werden. Allerdings erfolgt das Wiederfinden nur durch zusätzliche Informationen, wie Stadtplan oder Flurkarte, auf denen die Konvention zur Vergabe von Namen oder Nummern dokumentiert ist. Die eigentliche Positionsbestimmung erfolgt also wieder über Koordinaten im Kartenblatt oder über Nachbarschaften zu koordinierten Objekten.

Eine Visualisierung von solchen Objekten, deren Position indirekt beschrieben ist, kann nur erfolgen, wenn sich die Lage und unter Umständen auch die Form aus anderen Quellen in Koordinaten umwandeln läßt. Es ist zwar möglich, auf der Basis von indirekten Positionierungssystemen eine Topologie aufzustellen - so liegt zum Beispiel in den meisten deutschen Städten die Hausnummer 6 neben dem Haus mit der Nummer 8 - aber es existieren so viele Ausnahmen, daß bei räumlichen Analysen mindestens eine Topolgogie und in vielen Fällen auch eine Metrik erforderlich ist.

2.2.2.3 Dimension

Zur Modellierung der räumlichen Eigenschaften eines Objektes muß zuerst die **Dimension** der geometrischen Repräsentation festgelegt werden. Dazu sind zwei Begriffe von Dimensionalität zu unterscheiden. Die Dimension der zu betrachtenden Mannigfaltigkeit ist durch die Anzahl von Vektoren einer Basis festgelegt. Der Einbettungsraum einer Manigfaltigkeit kann die gleiche oder eine höhere Dimension besitzen.

In Geoinformationssystemen werden üblicherweise alle Objekte auf die Erdoberfläche bzw. deren Projektion auf eine ellipsoidische Approximationsfläche bezogen. Durch eine Basis von zwei Parametern kann die gesamte Mannigfaltigkeit abgebildet werden. Die Dimension dieser Bezugsfläche ist damit zwei, obwohl die Erde ein dreidimensionaler Körper im Raum ist. Der Einbettungsraum hat damit die Dimension drei. Wird jedem Punkt der Bezugsfläche genau ein Höhenwert zugeordnet, wie dies bei digitalen Geländemodellen geschieht (von Felsüberhängen einmal abgesehen), so ergibt sich wiederum eine zweidimensionale Mannigfaltigkeit. Da sich die Geländeoberfläche allerdings aus der „flachen“ Bezugsfläche heraushebt, also in den dreidimensionalen Einbettungsraum hineinreicht, spricht man in diesem Zusammenhang oft von einer 2,5 D-Modellierung (*Schmidt und Fritsch, 1996*).

Von einer echten dreidimensionalen Modellierung kann erst gesprochen werden, wenn Körper mit einem Volumen modelliert werden. Die wegen ihrer Internetfähigkeit weit verbreitete 3D-Modellierungssprache VRML (*virtual reality markup language - ISO/IEC 14772-1:1997*) bietet Ansätze zur Realisierung von 3D-GIS. Es gibt noch keine kommerziellen Geoinformationsysteme, die echte Körper verwalten können. Allerdings lassen die meisten Systeme eine Höhenmodellierung zu und stellen Werkzeuge bereit, das Gelände zu visualisieren.

2.2.2.4 Repräsentationsformen der Geometrie

Abhängig von der Dimension der räumlichen Daten stehen unterschiedliche Möglichkeiten der Repräsentation zur Verfügung. Prinzipiell lassen sich drei Arten der räumlichen Darstellung unterscheiden:

- Raster
- Gitter
- Vektor

Unter einem **Raster** versteht man die lückenlose Aufteilung einer Mannigfaltigkeit in gleichartige Teile.

Für eine ebene zweidimensionale Mannigfaltigkeit können die geometrischen Primitive Drei-, Vier- oder Sechseckmaschen verwendet werden (*Bartelme, 1995, S. 95*). Wegen der einfachen Matrixform haben sich in der Bildverarbeitung und in Geoinformationssystemen rechteckige Rasterzellen (Pixel) durchgesetzt.

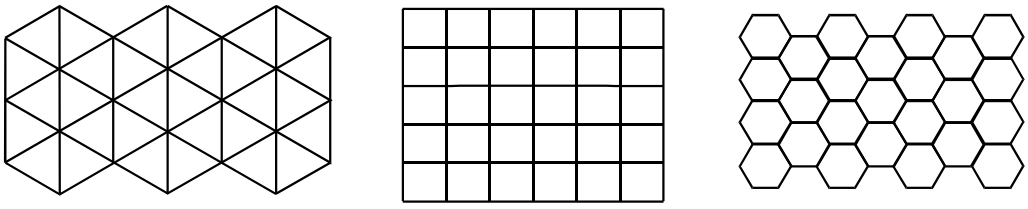
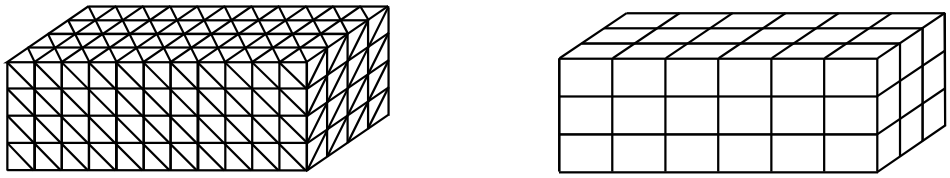


Abbildung 9: Unterschiedliche Rasterzellen mit jeweils gleicher Grundfläche.

Für eine euklidische, dreidimensionale Mannigfaltigkeit kann der Raum in gleichartige Volumina aufgeteilt werden. Als Elementarkörper steht dafür der Tetraeder (Abbildung 10a) zur Verfügung. Auch aus praktischen Erwägungen werden Volumenelemente häufig mit quaderförmigen Rasterzellen (Voxel, Abbildung 10b) als dreidimensionale Erweiterung der Pixel modelliert.



a) b)

Abbildung 10: a) Tetraeder als Elementarkörper für 3D-Raster b) Quader als 3D-Volumenelement.

Unter einem **Gitter** versteht man regelmäßig in der Mannigfaltigkeit angeordnete Punkte mit einem oder mehreren Attributwerten. Im Gegensatz zum Raster besitzen die einzelnen Elemente (Punkte) keine Länge, Fläche oder Volumen. Die Verarbeitungsmechanismen von Gitterstrukturen sind denen von Rasterrepräsentationen allerdings sehr ähnlich. Aus diesem Grund wird in vielen Lehrbüchern und auch in vielen kommerziellen Geoinformationssystemen nicht zwischen diesen beiden Grundstrukturformen unterschieden. Aus theoretischer Sicht besteht allerdings ein Unterschied zwischen den beiden Repräsentationen. Ein digitales Höhenmodell stellt z.B. ein Gitter dar, weil sich die einzelne Höheninformation auf genau den Gitterpunkt bezieht.

In **Vektormodellen** ist der Punkt der Träger der geometrischen Information. Alle höheren Strukturen wie Linien und Flächen bauen auf dem Punkt auf. Da das Verständnis von Geometrie, wie sie in Geoinformationssystemen verwendet wird, auf Euklid (ca. 300 v. Chr.) zurückgeht, sollen hier aus seinem I. Buch der Elemente die ersten sieben Definitionen wörtlich zitiert werden:

„**Definitionen.**

1. Ein Punkt ist, was keine Teile hat,
2. Eine Linie breitenlose Länge.
3. Die Enden einer Linie sind Punkte.
4. Eine gerade Linie (Strecke) ist eine solche, die zu den Punkten auf ihr gleichmäßig liegt.
5. Eine Fläche ist, was nur Länge und Breite hat.
6. Die Enden einer Fläche sind Linien.
7. Eine ebene Fläche ist eine solche, die zu den geraden Linien auf ihr gleichmäßig liegt.“

(Euklid, ca. 300 v. Chr., in einer Veröffentlichung seiner Elemente von *Thaer, 1975*)

2.2.2.5 Geodätisches Bezugssystem

Zur Festlegung der Position eines Punktes im Raum beziehungsweise in Bezug auf die Erde mit Koordinaten muß ein Bezugssystem vereinbart werden. Da viele unterschiedliche Referenzsysteme Anwendung finden, müssen für eine Umrechnung zwischen diesen Bezugssystemen die Parameter der Festlegung bekannt sein.

Als Approximationsfläche der Erde für Lagekoordinaten wird üblicherweise ein Rotationsellipsoid verwendet. Die Lage dieses lokal oder global bestanpassenden Ellipsoids wird mit sechs Parametern in Bezug auf den Massenmittelpunkt der Erde beschrieben. Sie bestehen aus drei Translations- und drei Rotationsparameter. Die Größe und Form des Rotationsellipsoids wird üblicherweise durch die Länge der großen Halbachse und durch die Abplattung festgelegt.

Die Landeskoordinaten ergeben sich durch Abbildung der Punkte der gekrümmten Referenzfläche in eine Ebene. Wenn die Abbildungsvorschriften bekannt sind, lassen sich Koordinaten in beliebige Systeme umrechnen.

2.2.2.6 Geometrische Primitive

Für eine zweidimensionale Modellierung ergeben sich als Elemente zur Darstellung von geometrischen Sachverhalten nach Euklid die geometrischen Primitive:

- Punkt
- Linie mit Interpolationsfunktion
- Fläche.

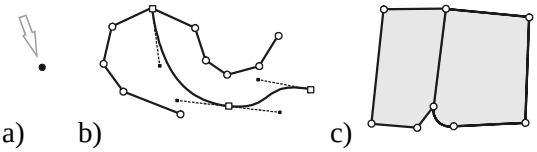


Abbildung 11: a) Punkt, b) Linien und c) Flächen mit unterschiedlichen Interpolationsfunktionen zwischen Stützpunkten.

2.2.2.7 Topologische Primitive

Die räumliche Beziehung zwischen Geoobjekten wird durch die Topologie beschrieben. Als Primitive zur Beschreibung von topologischen Zusammenhängen dienen:

- Knoten
- Kante (gerichtet oder ungerichtet)
- Masche.

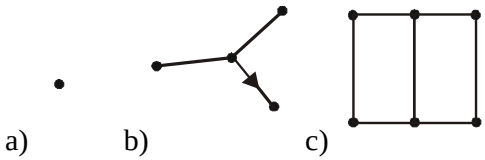


Abbildung 12: Topologische Primitive: a) Knoten, b) Kante und c) Masche

2.2.2.8 Auflösung

Zu digitalen Geodaten wird oft eine Maßstabsangabe gemacht. Ein Maßstab als solcher ergibt bei Vektordaten natürlich keinen Sinn. Was mit der Maßstabsangabe allerdings zum Ausdruck gebracht werden soll, ist der Detaillierungsgrad. Mit anderen Worten welches Detail kann in den Daten noch aufgelöst werden. Denn jede Modellbildung der realen Welt stellt auch eine Generalisierung dar. In analogen Karten ist der Grad der Generalisierung stark mit der Maßstabszahl korreliert.

Bei einer Rasterrepräsentation von sekundären digitalen Datenquellen (Orthophoto oder Satellitenbild) entspricht die Auflösung der Größe einer Rasterzelle im übergeordneten Koordinatensystem. Bei einer gescannten Karte ist die Detailierungstiefe in erster Linie vom Kartenbild abhängig, die Rasterweite bewirkt nur eine unterschiedliche Darstellungsqualität. Unterhalb einer bestimmten Pixelgröße entspricht der Informationsgehalt des Rasterbildes dem der analogen Vorlage.

In einer Vektorrepräsentation gibt die Auflösung an, ab welchem Abstand zwei räumlich getrennte Punkte als ein Punkt dargestellt werden dürfen, und wann bei einer Interpolationsfunktion weitere Stützpunkte eingefügt werden müssen, um die wahre Form des Objektes besser wiederzugeben. Wenn nicht explizit als Regel vereinbart, läßt sich über die Auflösung steuern, ab welcher Größe ein Wechsel zwischen geometrischen Primitiven erfolgen muß. Zum Beispiel ist es nicht sinnvoll bei einer geringen Auflösung, einen kleinen Bach als flächenhaftes Objekt zu modellieren. Auf diese Weise ist die Auflösung auch mit den Erfassungskriterien korreliert, da Objekte, deren Größe unterhalb einer bestimmten Auflösung liegen, u.U. weggelassen werden müssen.

Die Auflösung ist außerdem mit der Genauigkeit verknüpft. Obwohl eine sehr hohe Punktgenauigkeit mit einer geringen Auflösung verbunden sein kann und eine sehr geringe Genauigkeit mit einer extremen Detailtiefe, so ist es doch sinnvoll, Auflösung und Genauigkeit aufeinander abzustimmen. Wenn unter der Genauigkeitsangabe die Übereinstimmung des gesamten digitalen Objekts mit dem abstrakten Objekt gemeint ist (siehe dazu die Definition von Genauigkeit in Abschnitt 4.3.4), so ergibt sich die Auflösung, also z.B. der Abstand von Stützpunkten, auch aus den Genauigkeitsvorgaben.

Obwohl Auflösung und Genauigkeit miteinander korreliert sind, besitzen die Begriffe eine unterschiedliche Wertigkeit. Auflösung ist ein Begriff der Modellierung, da diese zur Festlegung des abstrakten Abbildes der realen Welt nötig ist, und Genauigkeit ist ein Kriterium zur Beschreibung der Qualität von Geodaten.

2.2.2.9 Attributtyp

Die Art der Attributwerte ist an die Typvereinbarung geknüpft. Dabei kommen alle durch das Datenbankmanagementsystem unterstützte Typen in Frage. Einfache Datentypen sind ganze oder rationale Zahlen mit unterschiedlichen Minimal- und Maximalwerten, Zeichenketten oder Boolesche Werte. Es kann auch eine Auswahl von vordefinierten, festen Attributwerten sein, die als Liste einen eigenständigen Attributtyp darstellen.

Es können allerdings auch Verweise auf beliebige andere Attribute vergeben werden. Die Entwicklungen der Multimediatechnologie stellt dazu neue Typen von Attributen zur Verfügung. So können Rastergrafiken, Klänge und Videosequenzen in beliebiger Kombination auch in Verbindung mit den klassischen Attributtypen verwaltet und wiedergegeben werden. Geoinformationssysteme, die mit solchen Attributtypen arbeiten, werden als Multimedia-GIS bezeichnet.

Durch die Internettechnologie und die damit verbundenen standardisierten Austauschformate spielen Multimediadaten eine steigende Rolle. Sie dienen allerdings nur zur Veranschaulichung von Sachverhalten. Analysen und komplexe Abfragen lassen sich bisher nur auf den klassischen Attributtypen ausführen.

2.2.3 Regeln

Die Objekte müssen bestimmte, im Modell definierte Regeln erfüllen. Diese Regeln können sich entweder auf das Objekt selbst oder auf seine Interaktion mit anderen Objekten beziehen.

2.2.3.1 Objektbildung

Die Abgrenzung eines abstrakten Objektes gegenüber anderen abstrakten Objekten derselben Klasse muß als Regel formuliert werden. Dabei sind die Fragen zu beantworten: wo fängt ein abstraktes Objekt an, und wo hört es auf, und wann muß ein abstraktes Objekt in kleinere abstrakte Objekte aufgeteilt werden? Oder in welchen Fällen ist ein komplexes abstraktes Objekt zu modellieren, falls dieses in der hierarchischen Gliederung vorgesehen ist? Dies kann auch als Aggregation von abstrakten Objekten zu einem abstrakten Objekt einer höheren Hierarchiestufe bezeichnet werden. Prinzipiell ist immer dann ein neues abstraktes Objekt zu bilden, wenn sich mindestens ein Attributwert ändert, da sonst die eindeutige Zuordnung zwischen Geometrie und Attribut nicht mehr gewährleistet ist.

Der ATKIS-Objektartenkatalog z.B. legt Regeln fest, nach denen die ATKIS-Objekte und -Teilobjekte gebildet werden müssen. Das Teilobjekt ist nach dieser Festlegung Träger der geometrischen Information, und das Objekt gehört zur abstrakten Objektklasse der nächst höheren Hierarchiestufe. Ein Objekt kann sich aus einem oder mehreren Teilobjekten zusammensetzen.

„Ein neues Objekt wird gebildet,
a) wenn Objekte verschiedener Objektarten aneinandergrenzen;
b) wenn sich ein Name ändert (z.B. eine Straße ändert in ihrem Verlauf ihren Namen);
c) wenn sich der Objekttyp ändert (z.B. eine Straße wird in ihrem Verlauf linienförmig und komplex modelliert);
d) bei Änderung des Wertes eines herausgehobenen Attributes;
e) an Landesgrenzen.

Es ist zweckmäßig, ein neues Objekt zu bilden
f) in individuellen objektabhängigen Fällen.

Regeln zu den Fällen d) und f) sind als besondere Objektbildungsregeln bei der betreffenden Objektart angegeben. Objekte können über Referenzen zu komplexen Objekten zusammengefaßt werden.“ (*AdV, 1995*)

„Objektteile werden gebildet,
a) wenn ein Attribut hinzutritt, wegfällt oder sich ein Attributwert ändert. Ändern sich im Verlauf eines Objektes Attributwerte kontinuierlich (z.B. Breitenangaben) so sind für die Objektteilbildung Größenklassen oder andere Kriterien festgelegt.
b) an Knoten topologischer Netze.

Objektteile können gebildet werden, wenn mehr als eine Überführungsreferenz in jeweils eine Richtung zu einem anderen Objektteil gebildet werden müßte.“ (*AdV, 1995*)

2.2.3.2 Objektschlüssel

Der Objektschlüssel oder auch Objektidentifikator realisiert den umkehrbar eindeutigen Zugriff auf ein individuelles Objekt und muß zur EDV-technischen Verarbeitung geeignt sein (*Bill und Fritsch, 1994*).

Die Geoinformationssysteme vergeben in der Regel für jedes Objekt, das eingelesen oder digitalisiert wird, einen systeminternen Objektidentifikator. Mit dieser Nummer oder Adresse kann ein Objekt zur Selektion oder Visualiserung angesprochen werden. Das System wacht über die Konsistenz der vergebenen Objektidentifikatoren, damit keine doppelten oder falschen Adressen Verwendung finden.

Beim Austausch von Geodaten zwischen unterschiedlichen Geoinformationssystemen geht diese Information allerdings in der Regel verloren und das Zielsystem vergibt beim Einlesen neue systeminterne Nummern. Aus diesem Grund ist es sinnvoll auf konzeptioneller Ebene benutzerdefinierte Objektschlüssel zu vergeben. Diese sollen dauerhaft mit den Objekten verbunden sein. Damit läßt sich unabhängig davon, in welchem System die Daten gerade verwendet werden, ein Objekt immer eindeutig identifizieren.

Der Anwender ist allerdings für die Konsistenz der Objektschlüssel selbst verantwortlich. So muß die Zuordnung zwischen realem Objekt und abstraktem Objekt eindeutig sein, d.h. der Anwender darf keinen Objektschlüssel doppelt vergeben, und es darf kein abstraktes Objekt geben, das keinen Objektschlüssel besitzt. Das Konzept der Vergabe von Objektschlüsseln muß durch das konzeptionelle Modell festgelegt werden, auch wenn die tatsächliche Zuweisung der Werte bei der Datenerfassung erfolgt.

Dadurch, daß der Objektschlüssel während der gesamten Lebensdauer eines abstrakten Objektes unverändert erhalten bleibt, ist der Objektschlüssel für die Fortführung von digitalen Daten von entscheidender Wichtigkeit. Anhand des Objektschlüssels kann einem Anwender mitgeteilt werden, welches abstrakte Objekt nicht mehr existiert, welches verändert wurde (mit Angabe der differentiellen Änderung) und ob ein neues abstraktes Objekt seit der letzten Datenabgabe hinzugekommen ist. Durch entsprechende Austauschformate, können die Änderungen automatisch in das Zielsystem eingelesen werden, ohne daß die restlichen, unveränderten Objekte betroffen sind. Die einheitliche Datenbankschnittstelle (EDBS) der AdV ist dafür konzipiert worden, solche differentiellen Änderungen an Nutzer des sogenannten Sekundärnachweises weiterzuleiten (*AdV, 1986 - EDBS*).

2.2.3.3 Beziehungen zwischen Objekten

Objekte stehen in Beziehungen zueinander. Diese Beziehungen werden durch Assoziationen modelliert. Assoziationen müssen bei der Erfassung zugewiesen werden. Topologische Beziehungen lassen sich aus der Geometrie ableiten und sind damit implizit in den Daten enthalten oder sie können in einer Struktur zur topologischen Darstellung explizit modelliert werden.

Weil bei der inhaltlichen Festlegung des konzeptionellen Modells die reale Welt aus mehreren Blickwinkeln betrachtet werden kann, ist es möglich, daß ein geometrisches Element gleichzeitig mehreren Objektklassen zugewiesen wird. Zum Beispiel wird eine Siedlung in ATKIS gleichzeitig Wohnbaugebiet, Ortslage und Gemeindefläche sein. Allerdings gibt es Objektklassen, die sich gegenseitig ausschließen, d.h. deren Instanzen mit Ausnahme des Randes keine gemeinsamen Punkte haben dürfen, z.B. kann eine Fläche nicht gleichzeitig Wald und bebautes Gebiet sein. Wenn zwei Objekte außer dem Rand weitere gemeinsame Punkte haben, spricht man von einer Überlagerung dieser Objekte. In topographischen Modellen dürfen sich z.B. administrative Einheiten, reale Flächennutzung und punkthafte Objekte bestimmter Objektklassen überlagern.

Außerdem gibt es Gruppen von Objektklassen, die dazu angelegt sind, jeden Punkt des Gebietes eindeutig einem Objekt aus der Gruppe von Klassen zuzuordnen. Es darf dann keine Teilfläche innerhalb des Gebietes geben, die nicht durch ein Objekt aus einer Klasse dieser Gruppe belegt ist. Als Beispiel für Objekte, die das Gebiet vollständig überdecken müssen, können administrative Einheiten wie „Gemeinde“ und „gemeindefreies Gebiet“ genannt werden. Da die zweite Objektklasse gerade als

Negation der erstgenannten definiert wurde, ist klar, daß die Vereinigung von beiden Mengen wieder die Gesamtfläche des Gebietes ohne Überschneidungen ergeben muß.

Bei einer zweidimensionalen Modellierung läßt sich für Schnittpunkte von Objekten nicht entscheiden, ob die Objekte niveaugleich miteinander verbunden sind, oder ob eine Über- oder Unterführung zwischen den Objekten existiert. Die räumliche Trennung der realen Objekte bezieht sich dabei auf die dritte Dimension. Dieser Zusammenhang zwischen den Objekten muß bei der konzeptionellen Modellierung beschrieben werden. Um festzulegen, daß keine Verbindung zwischen den sich schneidenden Objekten besteht, dürfen an dieser Stelle keinen topologischen Knoten gebildet werden. Diese einfache Art der Modellierung unter Verwendung der Möglichkeiten des logischen Modells ist für eine kartographische Darstellung oder für Analysen nicht ausreichend, bei denen die Information erforderlich ist, ob das Objekt oben oder unten liegt. Es liegt in diesem Fall kein planarer Graph vor und kann deshalb mit den logischen Modellen einiger GIS-Softwareprodukte nicht abgebildet werden.

Ob eine Über- oder Unterführung vorliegt, kann durch ein Attribut erfaßt werden. Bei einer kartographischen Darstellunge ist damit eindeutig, welches Objekt unterbrochen, und welches Objekt durchgezogen visualisiert werden muß.

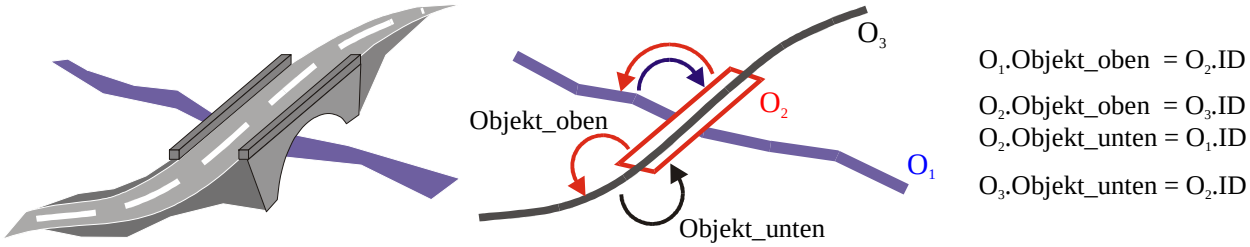


Abbildung 13: Referenzen zur Beschreibung von Beziehungen zwischen Objekten

Wenn für eine Anwendung die Information benötigt wird, welches Objekt über oder unter einem anderen Objekt geführt wird, muß diese Referenz zwischen den Objekten im konzeptionellen Modell abgebildet werden. Die Referenz kann dabei entweder als Assoziation zwischen Objekten modelliert werden, die miteinander über Unter- oder Überführungen verknüpft sind, als eigenständiges Referenzobjekt, das als Verweis zwischen den Objekten dient, oder über Attribute mit den eindeutigen Objektschlüsseln der Objekte als Attributwerte, auf die sich die Über- oder Unterführungen beziehen. Wieviel derartiger Referenzen ein Objekt haben kann, wird in der Informationstechnologie als Kardinalität bezeichnet (*Burkhardt, 1997*), und muß auf konzeptioneller Ebene als Regel vereinbart werden.

2.2.3.4 Wertebereich für Attribute

Für Attributwerte ist es in vielen Fällen nicht sinnvoll, alle durch den vereinbarten Typ möglichen Werte zuzulassen. Wenn es eine den Attributen immanente Einschränkung gibt, die aus Regeln oder Gesetzen abgeleitet werden kann, so sollte diese Einschränkung des Wertebereiches als konzeptionelle Regel festgeschrieben werden. Bei der Formulierung der Regel muß unbedingt die Grundlage, also das Gesetz oder Naturgesetz, angegeben werden, damit eine Änderung dieser Grundlage im konzeptionellen Modell nachgeführt werden kann.

Für quantitative Attribute kann die Angabe des Wertebereiches als geschlossenes Intervall angegeben werden. Wenn mehrere Bereiche möglich sind, so sind diese als Mengen von geschlossenen Intervallen anzugeben.

Für nicht quantitative Attribute, die als Zeichenkette angegeben werden, wie z.B. Namen, läßt sich der Wertebereich nicht einfach angeben. Da Namen aber von einer Institution vergeben werden, kann diese Institution in ihrem Zuständigkeitsbereich über die gültigen Namen von Objekten Auskunft geben. Diese Liste stellt dann den Wertebereich für das Attribut Name von den betroffenen Objektklassen dar. Auf diese Art kann der attributive Wertebereich auch für andere nicht quantitative Attribute bestimmt werden.

2.2.3.5 Zuordnung von geometrischen Primitiven in Abhängigkeit von der Form des Objekts

Die Zuordnung eines Objekts zu einer bestimmten geometrischen Repräsentation wurde schon im Abschnitt 2.2.2.8 angesprochen. Dabei wurde behandelt, ab welcher Breite eine linienhafte Repräsentation in Abhängigkeit von der Auflösung noch sinnvoll ist. Dieser Wert stellt allerdings eine untere Grenze für einen Wechsel zwischen geometrischen Primitiven dar und sollte nur angewendet werden, wenn keine anderen Angaben vorhanden sind.

Es besteht auch die Möglichkeit jeder Objektklasse genau eine Art der geometrischen Repräsentation zuzuordnen. Diese starre Einteilung wird den Objekten aber in vielen Fällen nicht gerecht. Prinzipiell besitzt jedes natürliche Objekt auch eine räumliche Ausdehnung. Jede Zuordnung zu einem geometrischen Primitivum stellt dadurch eine Vereinfachung dar. Wird ein natürliches Objekt als Linie oder Punkt erfaßt, so bleibt zwar die Information über die Lage des Objekts erhalten, aber die Information über die Form geht verloren. Aus diesem Grund muß es eindeutige Regeln geben, nach denen die Zuweisung von geometrischen Grundelementen in Abhängigkeit von der Form des Objektes stattfindet.

Vier geometrische Begriffe zur Beschreibung der Form des Objekts können herangezogen werden:

- Länge: Maximale Ausdehnung entlang des Objekts (auch gekrümmt) gemessen
- Breite: Maximale Ausdehnung senkrecht zur Länge gemessen
- Mindestbreite auf eine bestimmte Länge (MbbL): Mindestwert für eine Breite, der über die angegebene Länge oder einen bestimmten Prozentsatz der Gesamtlänge nicht unterschritten werden darf.
- Fläche: Flächeninhalt eines Objekts, das durch einen äußeren Rand und u.U. mehreren Binnenabgrenzungen festgelegt ist.

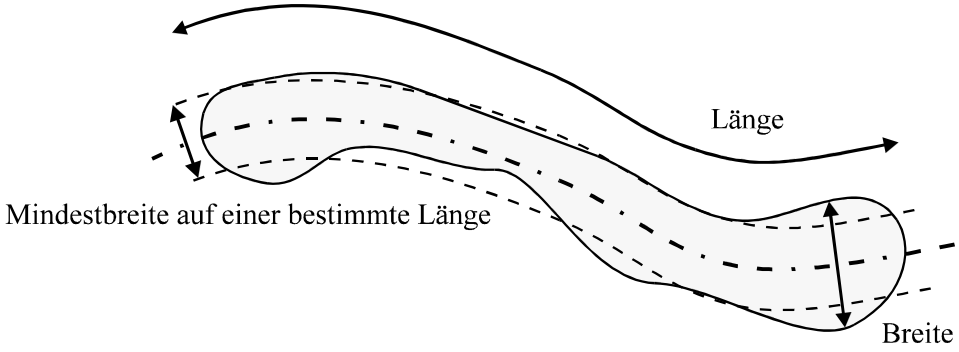


Abbildung 14: Beschreibung der Form eines Objekts.

Die einzelnen Kriterien müssen entweder unterschritten (<) oder überschritten (≥) sein, oder sie sind für die Festlegung nicht von Bedeutung (-). Durch die dreiwertigen Aussagen, die alle mit der logischen Konjunktion (∧) verknüpft werden, ergeben sich theoretisch 3^4=81 Konstellationen für die Zuweisung der drei geometrischen Primitive Punkt, Linie und Fläche. Davon sind aber nur eine Auswahl von Kombinationen sinnvoll. Voraussetzung ist, daß die Objekte das Erfassungskriterium erfüllen.

Eine Zuweisung der verknüpften Mindest- oder Maximalwerte kann über eine Tabelle erfolgen. Die aufgeführte Tabelle gibt eine sinnvolle Auswahl von 11 Kombinationen an.

Länge	< $Länge_0$	-	≥ $Länge_0$	≥ $Länge_0$	≥ $Länge_0$	≥ $Länge_0$	-	-	-	-	-
Breite	-	-	< $Breite_0$	< $Breite_0$	≥ $Breite_0$	≥ $Breite_0$	< $Breite_0$	-	≥ $Breite_0$	-	-
Mindestbreite auf einer bestimmte Länge (MbbL)	-	-	< $MbbL_0$	≥ $MbbL_0$	< $MbbL_0$	≥ $MbbL_0$	-	< $MbbL_0$	< $MbbL_0$	≥ $MbbL_0$	-
Fläche	-	< $Fläche_0$	-	-	-	-	≥ $Fläche_0$	≥ $Fläche_0$	≥ $Fläche_0$	≥ $Fläche_0$	≥ $Fläche_0$
geomtrisches Primitivum	Punkt	Punkt	Linie	Fläche	Linie	Fläche	Linie	Linie	Linie	Fläche	Fläche

2.3 Logische Modellierung

Bei der logischen Modellierung werden die im konzeptionellen Modell festgelegten Entitäten auf die Strukturen der GIS-Software abgebildet. *Bartelme, 1995*, betrachtet daher das zugehörige logische Schema als Teil des konzeptionellen Schemas. Da sich die kommerziellen GIS-Softwareprodukte in den Möglichkeiten der logischen Modellierung unterscheiden, ist es nicht immer möglich alle konzeptionellen Modelle auf alle Produkte direkt abzubilden. In manchen Fällen müssen Datenstrukturen emuliert oder durch Hilfskonstruktionen umgangen werden.

Die Spline-Interpolation ist ein Beispiel für eine Darstellung von linienhaften Objekten, die nicht in jedem Geoinformationssystem implementiert ist. Die Verwaltung von komplexen Objekten muß bei manchen Systemen über Hilfstabellen nachgebildet werden. Diese Beispiele sollen verdeutlichen, daß ein Nutzer bei der Auswahl von Geodaten für eine Anwendung die Möglichkeiten des Zielsystems mit der logischen Struktur der Daten vergleichen muß.

2.3.1 Raumbezug

Die logische Struktur der Geometriedaten wird mit den Primitiven zur Beschreibung von Geometrie und Topologie (Abschnitt 2.2.2.6 und 2.2.2.7) ausgedrückt. Die Knoten und Kanten können verschiedene Graphen im Sinne der Graphentheorie bilden (*Wilson und Watkins, 1990*). Sie können entweder als Spaghetti-Code oder in Form eines planaren oder nicht planaren Graphen oder Digraphen strukturiert sein. Bei einem ebenen Graph[3] wird die Ebene durch die Kurven in endlich viele zusammenhängende Gebiete zerlegt, die als Länder oder Maschen bezeichnet werden. Ein ebener Graph mit seinen Ländern heißt Landkarte (*Volkmann, 1996*).

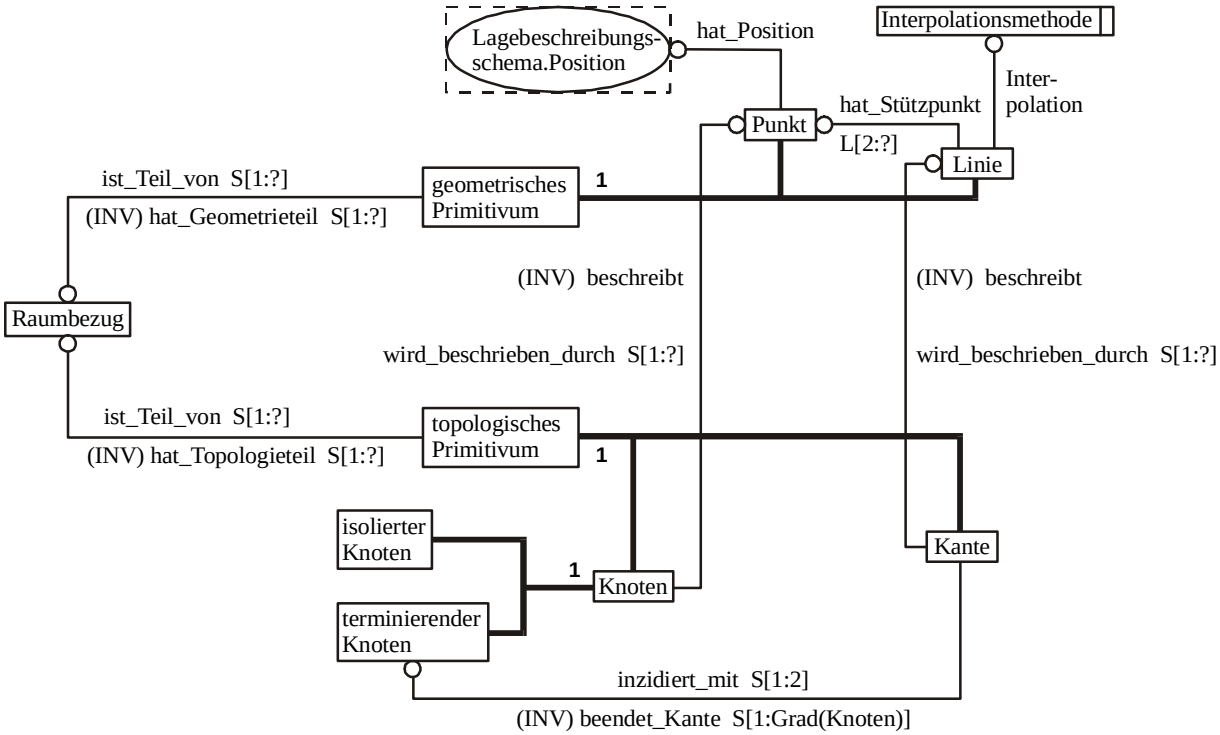


Abbildung 15: Logisches Modell für einen ungerichteten Graphen in EXPRESS-G Notation (in Anlehnung an *DIN V ENV 12160:1997*)

[3] Ein ebener Graph ist ein derart in die Ebene abgebildeter Graph, daß keine zwei Kanten bzw. die sie darstellenden Kurven einen Schnittpunkt haben. Ein Graph, der einem ebenen Graph isomorph ist, heißt planarer Graph (*Volkmann, 1996*)

Jede dieser Strukturen ist durch ihren spezifischen Zusammenhang zwischen den Elementen Knoten, Kante und Masche charakterisiert. Zur Beschreibung des Zusammenhangs kann entweder auf feststehende Strukturen, wie z.B. den planaren Graphen, zurückgegriffen werden, oder er muß explizit in Form von Diagrammen (Abbildung 15), z.B.: EXPRESS-G (*ISO/DIS 10303-11, 1992*), UML (*Larman, 1998*), Entity Relationship Modell-Diagramm, oder verbal ausgedrückt werden.

Für die Beschreibung der Geometrie sind Parameter zur Festlegung der jeweiligen Interpolationsfunktion erforderlich. Im einfachsten Fall, der geradlinigen Verbindung zwischen zwei Stützpunkten, werden außer den Koordinaten der Stützpunkte keine zusätzlichen Informationen benötigt. Für Kurven höherer Ordnung wie z.B. Splines, Kreisbögen, Klothoiden müssen weitere Angaben zu deren eindeutigen Festlegung gemacht werden. Die Auswahl der Parameter zur Bestimmung dieser Kurven für linienhafte Objekte oder zur äußeren oder inneren Abgrenzung von flächenhaften Objekten muß im logischen Modell festgelegt werden. Die Parameter können bei unterschiedlichen Implementierungen der logischen Sicht unterschiedlich verfügbar sein.

2.3.2 Sachdaten

Die Beziehungen zwischen den Objekten, die Zuordnung von Geometrie und Sachdaten und die Beziehungen zwischen den Sachdaten werden durch die logische Modellierung auf die GIS-Software abgebildet. Außerdem können noch Regeln zwischen den Sachdaten angegeben werden. Die Methoden zur Beschreibung der logischen Struktur von Sachdaten entsprechen denen des Raumbezugs. In Abhängigkeit davon, unter welcher GIS-Architektur das Modell implementiert werden soll, lassen sich diese Zusammenhänge besser als ERM (*entity relational model*) oder mit der objektorientierten Modellierungstechnik (*unified modelling language*, UML) darstellen. Allerdings sollte die logische Struktur der Daten von der Implementierung unabhängig sein.

2.3.3 GIS-Architektur

Die Verwaltung der Daten in der GIS Anwendungssoftware kann in unterschiedlicher Weise erfolgen. Am Anfang der Entwicklung von GIS-Anwendungsprogrammen standen, bedingt durch die geringe Leistungsfähigkeit der Hardware, Systeme mit einer proprietären Datenverwaltung. Diese Architektur entspricht dem Typ 1 nach der in Abbildung 16 dargestellten Einteilung von *Aronoff, 1989*. Die Fortschritte der Informationstechnik bei der Entwicklung von Datenbankmanagementsystemen (DBMS) haben dazu geführt, daß die Sachdaten mit kommerziellen DBMS in einer Attributdatenbank verwaltet werden, während die Geometriedaten unabhängig in einer separaten Datenbank oder in Dateien abgelegt sind. Typisch bei diesem Typ der GIS-Architektur (Typ 2) ist die Verwendung von Computer-Aided-Design-(CAD)-Programmen für die Behandlung der Geometriedaten. Eine komplette Neuentwicklung der GIS-Datenbank zur gemeinsamen Verwaltung von Geometrie- und Sachdaten (Typ 4) ist vor allem für objektorientierte Systeme erforderlich. Die Verwendung eines relationalen DBMS zur Speicherung der Geometrie- und Sachdaten in einer Datenbank kann erreicht werden, indem das System mit einer Schnittstelle für Topologie und Geometrie ergänzt wird (Typ 3). Diese Schnittstelle kann sowohl vom Hersteller der GIS Anwendungssoftware entwickelt, als auch vom Datenbankhersteller als Erweiterung zusammen mit der Basissoftware vertrieben werden. Die Tendenz in der kommerziellen Entwicklung von DBMS geht dahin, das rein relationale Datenmodell mit Zusatzfunktionen und neuen Datentypen zu erweitern, so daß auch raumbezogene und Multimedia-Daten im DBMS verwaltet werden können. Der ISO/IEC Standard zur Festlegung einer einheitlichen Datenbankabfragesprache (SQL), *ISO/IEC 9075, 1992*, wird mit der Version 3 Erweiterungen zur Verwaltung von Geodaten enthalten. Mit dieser Entwicklung lassen sich die Typen 3 und 4 der Einteilung von Aronoff nicht mehr unterscheiden.

In der Praxis werden immer noch alle vier Typen von GIS-Architekturen angetroffen. Es bestehen auch Mischformen.

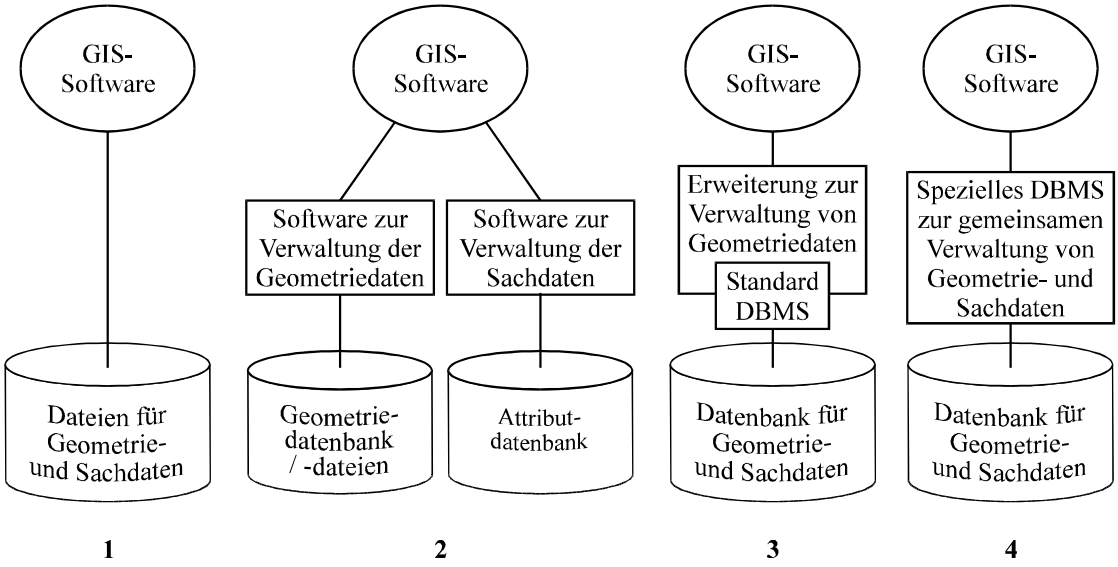


Abbildung 16: Vier prinzipielle GIS-Architekturen nach *Aronoff, 1989.*

2.3.4 Datenbankverwaltungssysteme

Das Datenbankverwaltungssystem (data base management system, DBMS) verwaltet den auf Dauer angelegten Datenbestand, hält ihn auf logischer Ebene konsistent und macht ihn berechtigten Nutzern zugänglich. Die Daten des konzeptionellen Modells müssen dazu in die logische Struktur des DBMS gebracht werden. Zur Manipulation der Daten und für Abfragen muß die Datenstruktur bekannt sein. Verschiedene Methoden zur Datenverwaltung sind möglich und kommen als DBMS in Geoinformationssystemen zur Anwendung.

Neben den klassischen Modellen, wie hierarchisches Modell, Netzwerkmodell und relationales Datenmodell (*Date, 1981*), sind die objektorientierten und objektrelationalen Datenbankverwaltungssysteme dabei den Schritt von der Forschung in die Praxis zu gehen (*Dogac et al., 1994, Date, 1995 und Stonebraker, 1996*). Zur Unterscheidung der Systeme und weil sie für die logische Datenmodellierung wichtig sind, werden deren Grundprinzipien in den nächsten fünf Abschnitten erläutert.

2.3.4.1 Hierarchisches Datenmodell

Das hierarchische Datenmodell kann als einfache Baumstruktur veranschaulicht werden. Mit Ausnahme der Wurzel hat jedes Element genau einen Vorgänger. Jedes Element kann keinen, einen oder mehrere Nachfolger haben. Die gewünschten Daten werden durch Navigation über die Hierarchiestufen abgerufen.

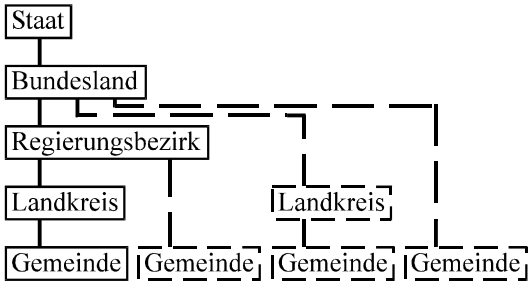


Abbildung 17: Hierarchische Modellierung der politischen Gliederung in Deutschland mit Berücksichtigung von kreisfreien Städten und von Bundesländern ohne Regierungsbezirke.

In einem Beispiel zur politischen Gliederung (Abbildung 17) können die Landkreise eines Regierungsbezirkes sehr schnell ermittelt werden. Da das Modell asymmetrisch ist, müssen zur Klärung der Frage zu welchem Regierungsbezirk ein bestimmter Landkreis gehört, alle Äste des Teilbaumes unterhalb von Regierungsbezirk durchsucht werden, bis der Landkreis gefunden wurde. Zur Abfrage, welche Landkreise zu einem Bundesland gehören ist eine zweistufige Suche erforderlich. Falls die hierarchische Struktur auf die Abfragen optimiert ist, können damit sehr schnell große Datenmengen durchsucht werden. Allerdings ist das Modell statisch und damit für andere Anfragen sehr unflexibel.

Weil ein Element nur einen Vorgänger haben darf, ist diese Art der logischen Modellierung in Geoinformationssystemen nicht geeignet. Kreisfreie Gemeinden lassen sich nur mit Erweiterung des Datenmodells darstellen. Damit müssen zur Abfrage, welche Gemeinden in einem Regierungsbezirk liegen, zwei Äste durchlaufen werden. Berücksichtigt man, daß nicht alle Bundesländer in Regierungsbezirke aufgeteilt sind, zeigt sich schnell die Komplexität der hierarchischen Modellierung.

2.3.4.2 Netzwerk-Datenmodell

Eine Erweiterung des hierarchisches Datenmodell bei dem ein Element auch mehrere Vorgänger haben kann, ist durch das Netzwerk-Datenmodell gegeben. Diese Struktur wird auch als CODASYL-Datenbank bezeichnet. Das Akronym CODASYL steht dabei für COnference on DAta SYstems Languages (*Bartelme, 1995, Date 1981*). Die Veranschaulichung der Struktur des Netzwerk-Datenmodells als Graph unterscheidet sich vom hierarchischen Modell dadurch, daß auch Zyklen (*Aigner, 1984*) auftreten können.

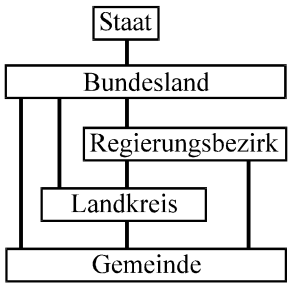


Abbildung 18: Politische Gliederung in Netzwerk-Struktur

Die Navigation zur Abfrage der Daten kann nicht nur in einer Richtung erfolgen. In dem Beispiel zur politischen Gliederung kann das Element Gemeinde nun entweder zu einem Landkreis gehören, oder einen Regierungsbezirk oder ein Bundesland als Vorgänger besitzen. Der Landkreis gehört entweder zu einem Regierungsbezirk oder direkt zu einem Bundesland. Damit wird die Redundanz vermieden, wie sie im hierarchischen Modell vorhanden ist. Die Nachteile der starren Struktur bleiben allerdings bestehen. Die Zusammenhänge zwischen den Elementen werden kompakter modelliert, damit steigt allerdings auch die Komplexität des Systems.

2.3.4.3 Relationales Datenmodell

Im relationalen Datenmodell werden die Daten in Tabellen, auch als Relationen bezeichnet, organisiert. In einer Zeile der Tabelle, diese wird auch Tupel oder Rekord genannt, werden die Eigenschaften einer einzigen Entität abgelegt. Und eine Spalte der Tabelle, oft auch mit Attribut bezeichnet, gibt die Eigenschaften der Entitäten an. Die Menge aller möglichen Attribute wird Domäne genannt. Im relationalen Datenmodell werden sowohl die Entitäten als auch die Relationen zwischen Entitäten in Tabellen abgelegt (*Date, 1995*).

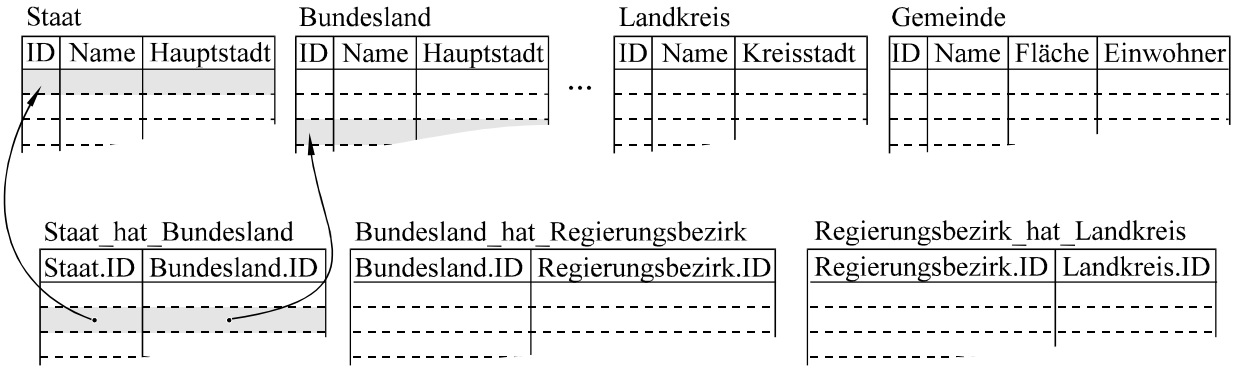


Abbildung 19: Politische Gliederung im relationalen Datenmodell

Durch eine Relationenalgebra können verschiedene Operationen auf den Daten ausgeübt werden. Die Sprachen zur Festlegung des logischen Datenmodells (*data definition language*, DDL) und zur Beschreibung der Methoden auf den Daten *(data manipulation language*, DML) sind Bestandteil der standardisierten Abfragesprache für relationale Datenmodelle (*structured query language*, SQL).

2.3.4.4 Objektorientiertes Datenmodell

Die Vorteile der objektorientierten Programmierung beim Design und der Implementierung von großen und komplexen Programmsystemen haben bewirkt, daß sich bei Neuentwicklungen auch im Bereich GIS die objektorientierten Programmiersprachen durchgesetzt haben. Zahlreiche Forschungsaktivitäten versuchen die Prinzipien der objektorientierten Programmierung auf Datenbankverwaltungssysteme zu übertragen (*Dogac et al., 1994, Loomis, 1995*).

Die Prinzipien der objektorientierten Modellierung lassen sich mit den Schlagworten Klasse, Objekt, Kapselung, Vererbung und Polymorphismus beschreiben, die im folgenden näher erläutert werden.

Klassen oder Objektklassen beschreiben prototypische Datenstrukturen, die Methoden und Schnittstellen zur Kommunikation beinhalten. Die Verbindung von Daten mit ihren Methoden wird als Kapselung bezeichnet. Die Kapselung und ein Verbieten von Zugriffen von außen außer über die definierten Schnittstellen führen zu einer hohen Modularität. Die Instanzen von Objektklassen sind Objekte. Durch das Konzept der Objekthierarchie und der Vererbung lassen sich Eigenschaften und Methoden von sogenannten Superobjekten auf hierarchisch tiefer stehende Objekte übertragen. Die Fähigkeit, daß dieselbe Botschaft an Objekte verschiedener Objektklassen gesandt unterschiedliche Aktionen auslösen kann, wird mit Polymorphismus bezeichnet.

Die Notwendigkeit der langfristigen Speicherung einer großen Anzahl von Objekten, bei der ihre Zustände einen Programmlauf überdauern, die sogenannte Persistenz, macht den Bedarf der Erweiterung des objektorientierten Paradigmas auf Datenbankverwaltungssysteme deutlich.

2.3.4.5 Objektrelationales Datenmodell

Eine Möglichkeit, die Vorteile einer objektorientierten Modellierung mit denen der relationalen DBMS zu verknüpfen, stellen die objektrelationalen Datenbankverwaltungssysteme dar. Weil für relationale DBMS eine spezielle Algebra existiert, kann eine Anfrage algebraisch optimiert werden. Außerdem wurden große Datenbestände aufgebaut, die nur schwer in ein neues Paradigma der objektorientierten Datenbanken übertragen werden können. Daher liegt es nahe, die relational vorliegenden Daten weiterhin mit den konventionellen DBMS zu verwalten und Anwendungen mit objektorientierten Methoden zu entwickeln, die über eine irgendwie geartete Schnittstelle auf die relationalen Daten zugreifen.

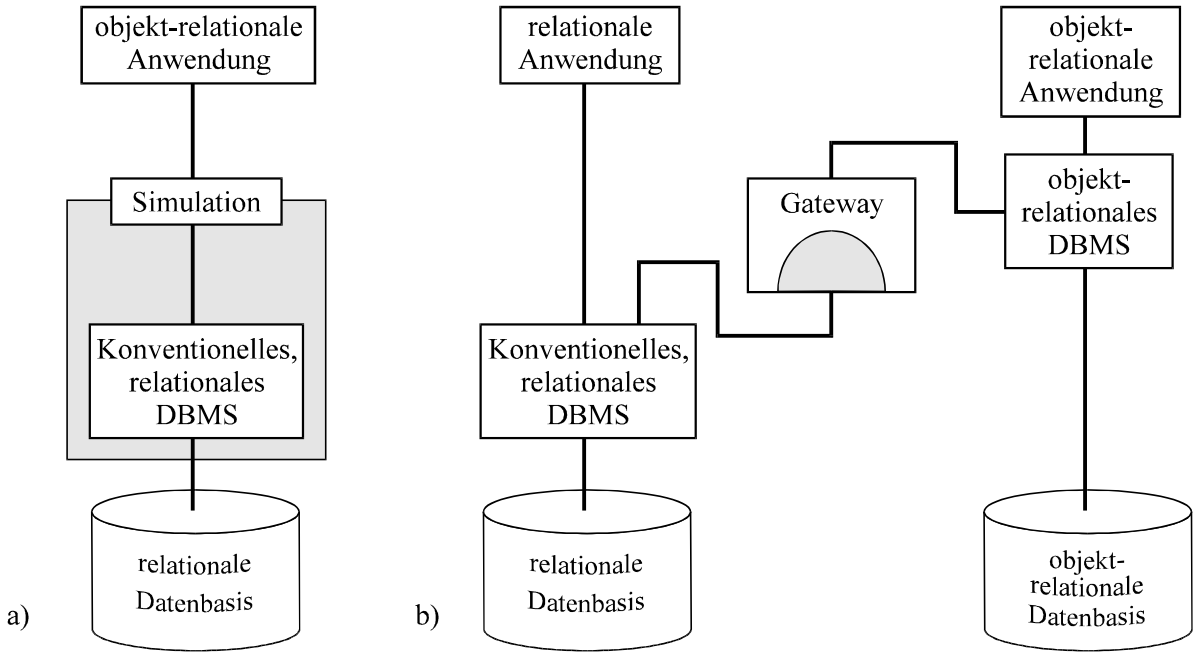


Abbildung 20: Strategien zum Aufbau eines objektrelationalen DBMS nach *Stonebraker, 1996.*

Stonebraker, 1996, gibt verschiedene Architekturen zum Aufbau von objektrelationalen Systemen an. In der Abbildung 20 sind zwei Beispiele dargestellt. Unter a) wird die objektrelationale Daten-

manipulation vollständig von einer zwischengeschalteten Schnittstelle simuliert. Nach der Architektur unter b) können Zugriffe auf bestehende relationale Datenbanken und auf objektrelationale Daten miteinander verknüpft werden.

SQL3 ist ein internationaler Standard, der Elemente der objektrelationalen Datenmodellierung beinhaltet. Bei genauer Betrachtung lassen sich die objektrelationalen DBMS mit den GIS-Architekturen aus Abbildung 16 vergleichen. In einigen kommerziellen GIS-Produkten hat der Anwender auf die Geometrie- und Sachdaten einen objektorientierten Zugang, obwohl die Daten in einer konventionellen, relationalen Datenbank abgelegt sind.

2.4 Physikalische Modellierung

Bei der physikalischen Modellierung (oft auch als physische Modellierung bezeichnet) werden die Daten auf Speicherplätze zugewiesen. Speichermedien können der Hauptspeicher, ein externer Massenspeicher (Festplatte, u.U. auch Backup-Medium wie Magnetband) oder Speichermedien mit ausschließlichem Lesezugriff wie z.B. CD-ROM sein. Bei Zugriffen auf die Daten in verteilten Systemen ist auch die Strukturierung der Daten beim Austausch in den Protokollen zu berücksichtigen.

Der Zugriff auf externe Daten (Festplatte, CD-ROM, Band oder Netz) ist gegenüber Hauptspeicherzugriffen erheblich zeitintensiver. Die Positionierung des Lesekopfes kostet im Verhältnis zum sequentiellen Lesevorgang viel Zeit, daher empfiehlt es sich, die Anzahl der Positionierungen zu minimieren. Für Geodaten ergibt sich daraus die Forderung, räumlich benachbarte Daten auch physikalisch benachbart abzuspeichern.

Zur Optimierung der Zugriffszeiten auf Geodaten wurden spezielle Verfahren der räumlichen Indizierung entwickelt. Die Quadtree-Struktur, die sowohl für Raster- als auch für Vektordaten geeignet ist, stellt die meist verbreitete Methode dar.

Bei der Betrachtung der Qualität von ganzen Geoinformationssystemen spielt das Antwortzeitverhalten bei Analysen oder bei der Visualisierung der Daten eine Rolle. Es ist daher wichtig, daß die physikalische Modellierung auf diese Anforderungen hin optimiert wird. Da in dieser Arbeit die Qualität von Geodaten unabhängig von Hardware oder Software betrachtet werden soll, wird die unterste Stufe der Modellierung von Daten in Geoinformationssystemen nicht weiter vertieft werden.

2.5 Datenschemata

Ein Modell beschreibt die zu erfassenden Daten für das Geoinformationssystem vollständig und konsistent. Die Form der Dokumentation wird dabei nicht festgelegt. Im allgemeinen wird das Modell prosaisch oder in Tabellen und mit vielen beispielhaften Skizzen beschrieben. Oft ist das Modell nur als Erfassungsanweisung erstellt worden. Schon die verschiedenen Landessprachen, in denen die Modelle der Basisdaten im europäischen Raum beschrieben sind, stellen ein Hindernis bei der Verbreitung von Geodaten außerhalb der nationalen Nutzerkreise dar.

Bei der Übernahme oder vor der Erfassung von Daten muß das Modell in die Definitionssprache der jeweiligen GIS-Software umgesetzt werden. Eine formale Beschreibung des Modells wird als Datenschema bezeichnet.

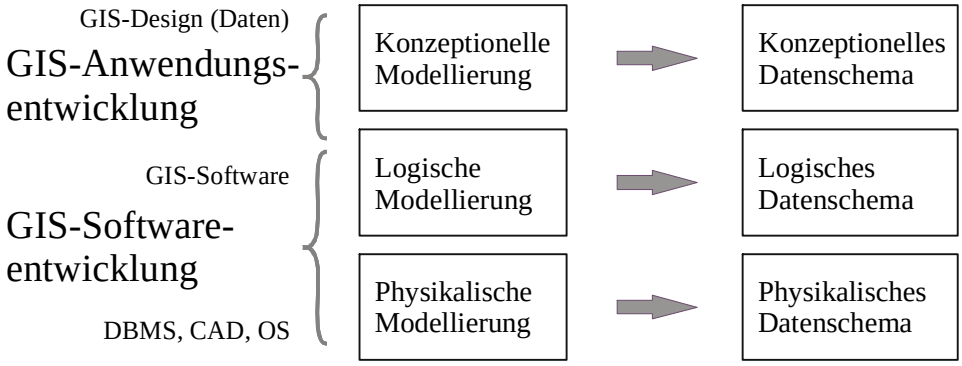


Abbildung 21: Die verschiedenen Ebenen der Modellierung spiegeln sich in unterschiedlichen Datenschemata wider.

Ein Datenschema kann entweder lexikalisch oder durch eine stringente graphische Notation festgelegt werden. Die Graphik stellt im allgemeinen eine Einschränkung der Möglichkeiten zur Darlegung der Sachverhalte dar. So ist es zum Beispiel schwierig, Regeln in einer graphischen Notation auszudrücken. Allerdings gibt eine Graphik einen sehr guten Überblick über die Zusammenhänge in den Geodaten.

Bei DBMS wird zwischen der formalen Sprache zur Deklaration des *Datenbankinhaltes (data definition language*, DDL) und der formalen Sprache zur Manipulation und Verarbeitung der Daten (*data manipulation language*, DML) unterschieden. Die Sprachkonstrukte sind im allgemeinen genormt, damit sie herstellerunabhängig verwendet werden können.

Eine Sprache zur formalen Beschreibung von geometrischen und konzeptionellen Modellen wurde von der materialverarbeitenden Industrie (vor allem der Automobilindustrie) mit dem Ziel entwickelt, CAD-Pläne über längere Zeit zu archivieren und mit Zulieferern Pläne digital auszutauschen. Die Sprache, die aus der Initiative STEP (*Product Data Representation and Exchange*) hervorgegangen ist, hat den Namen EXPRESS, das Pendant zur graphischen Formulierung des Datenschemas heißt EXPRESS-G (*ISO/DIS 10303-11, 1992*).

EXPRESS und EXPRESS-G können auch zur produktunabhängigen Formulierung von Schemata für Geodaten angewandt werden. Das technische Komitee „Geoinformation“ der europäischen Organisation für Normung (CEN TC 287) verwendet EXPRESS zur Beschreibung der Geodaten-modellierung für eine genormte Schnittstelle zum Datenaustausch.

Eine weitere Sprache, mit der Schemata systemunabhängig beschrieben werden können, stellt die *Unified Modelling Language* (UML) dar. Mit UML können Objektklassen, deren Eigenschaften und Assotiationen und, weil UML aus dem Bereich der objektorientierten Modellierung stammt, auch Methoden zu diesen Objektklassen festgeschrieben werden. Weil es Werkzeuge zum interaktiven Entwickeln von Schemata gibt, und diese in der Lage sind, verschiedene Beschreibungssprachen ineinander überzuführen, und zusätzlich Sourcecode in verschiedenen objektorientierten Programmier-sprachen erzeugen können, wird UML sowohl bei ISO im TC211 als auch vom OGC zur Beschreibung von Spezifikationen eingesetzt.

3 Metadaten

Ebenso wie eine Karte ohne Angaben von Maßstab, Inhalt (Überschrift und Legende), Urheber, Herausgabedatum, Zeitpunkt der Datenerhebung und eventuell weiteren Informationen unbrauchbar ist, müssen auch Anwender von Geodaten mit Informationen über die Daten versorgt werden. Diese Informationen braucht der Anwender für die Entscheidung, ob die Geodaten für seine Anwendung geeignet sind. Sie werden als Metadaten bezeichnet. Die Metadaten benötigen ein eigenes Modell, das Metadatenmodell.

Im Metadatenmodell wird festgelegt mit welchen Angaben die Daten zu beschreiben sind. Die erforderlichen Angaben lassen sich in fünf Gruppen einteilen, die sich auf folgende Punkte beziehen:

- Modell
- Herkunft der Daten
- Qualität
- Verfügbarkeit
- Referenzanwendungen

Die Metadaten stellen somit die Menge aller zur Beurteilung der Einsetzbarkeit relevanten Informationen dar. Sie beschreiben die Daten vollständig.

Große Datenbestände sind selten homogen. Es müssen dann differenzierte Metadaten angegeben werden, die sich nur auf einen Teil des Datenbestandes beziehen. Im Extremfall kann sich die Metainformation nur auf ein oder wenige Objekte der Datenbasis oder sogar nur auf spezielle Attributwerte beziehen. Die Beziehung zwischen den Daten und den Metadaten und wie diese Beziehung in das Modell abgebildet werden kann, wird in Kapitel 4.6 behandelt.

Für Recherchen nach Daten mit bestimmten Inhalten ist eine Volltextsuche wünschenswert, weil dann von einer strengen Struktur der Metadaten abgewichen werden kann. Da für viele Begriffe auch Synonyme verwendet werden, hilft ein Thesaurus bei der Recherche nach den gewünschten Daten. Der Thesaurus setzt die Suchbegriffe in die Standardbegriffe um und erleichtert damit das Auffinden von gewünschten Informationen. Der Thesaurus gibt entweder direkte Synonyme, Definitionen oder verwandte Begriffe aus.

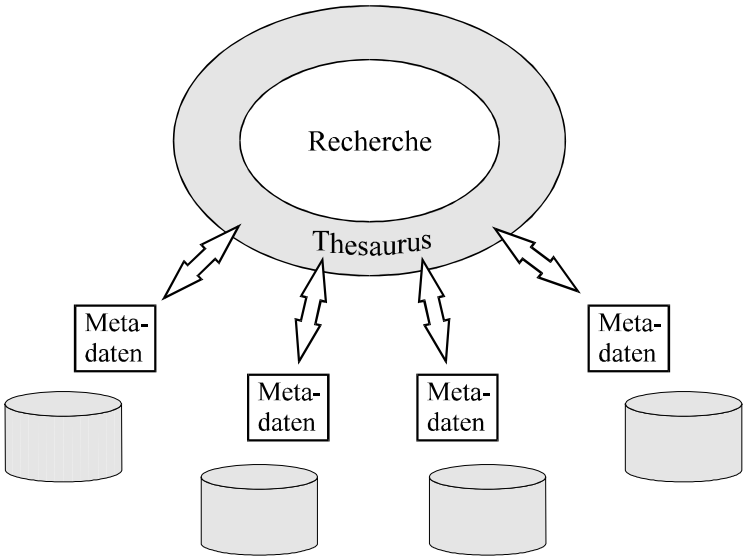


Abbildung 22: Thesaurus zum Auffinden von gewünschten Geodaten durch Recherche in den zugehörigen Metadaten.

Das Internet bietet Datenproduzenten die Möglichkeit, Ihre Metadaten einer sehr breiten Anwenderschicht online zugänglich zu machen. Durch die guten Vergleichsmöglichkeiten kann sich ein potentieller Anwender über die Verwendbarkeit von angebotenen Daten informieren. Auf diese Art kann sich ein Markt für Geodaten erst richtig etablieren. Für die Recherche muß es entweder Server geben, die Metadaten von verschiedenen Anbietern zentral abrufbar halten, oder Suchmaschinen müssen das gesamte Netz über sogenannte Robots nach Stichworten durchkämmen.

In Europa wurde das Projekt MEGRIN initiiert, bei dem ein zentraler Server, genannt „GDDD - *Geographical Data Description Directory*“, über verfügbare Geodaten informiert. Der Nachteil einer solchen zentralen Datenhaltung ist, daß die Daten von der Institution, die diesen Dienst anbietet,

permanent gepflegt werden müssen. Dabei ist diese Institution auf Angaben von den Datenanbietern angewiesen. Ein zentraler Metadatenserver kann auch nur Informationen weitergeben, die im System bekannt sind. So wird eine zentrale Informationsquelle nie vollständig sein. Im Fall von MEGRIN werden z.B. nur Informationen zu amtlichen Daten angeboten.

Wenn die Datenanbieter selbst Metadaten im Internet zur Verfügung stellen, so ist die Pflege der Daten erheblich einfacher. Allerdings werden Suchmaschinen benötigt, die Metadaten als solche erkennen. Außerdem sind die Datenanbieter angehalten, sich an Standards zu halten und den potentiellen Kunden vollständig über die Eigenschaften der Geodaten aufzuklären.

3.1 Das Modell der Daten

Das konzeptionelle Modell als inhaltliche Beschreibung der Daten gibt Auskunft, welche Objektklassen in den durch die Metadaten beschriebenen Daten zu finden und wie diese strukturiert sind. Damit ist das Modell ein wesentlicher Bestandteil der Metadaten.

Da die Beschreibung des Modells bzw. des Schemas sehr unterschiedlich sein kann, von einer freien textlichen Beschreibung bis hin zu einer proprietären Datendefinitionssprache, ist es sinnvoll neben dieser vollständigen Festlegung eine vereinfachte Beschreibung zu geben, die im Idealfall einer internationalen Norm oder einem Standard entspricht. Im Rahmen der Normungsaktivitäten werden vom europäischen (CEN) und internationalen (ISO) Komitee für Normung konzeptionelle Schemata für Metadaten erarbeitet (*prENV 12657, 1998, ISO 19115, 1999*).

Bei einer Kurzfassung des Modells der Daten müssen die Punkte Inhalt, Struktur und Regeln, wie sie im Abschnitt 2.2 aufgeführt wurden, angesprochen werden.

3.2 Die Herkunft der Daten

3.2.1 Urheber

Der Urheber ist der Schöpfer eines Werkes. Bei Geodaten ist als Urheber die Institution zu verstehen, die die Geodaten erfaßt bzw. den Auftrag zur Erfassung erteilt hat und damit die Kontrolle über das Verfahren besitzt, also aus dem abstrakten Abbild der realen Welt digitale Objekte erzeugt hat. Der Urheber hat die Urheberrechte (*Copyright*) an den Daten.

Für einen Anwender ist der Urheber von Interesse, weil er über ihn Abgabemodalitäten, weitere Dienstleistungen oder Informationen erfragen kann, die in den Metadaten unbeantwortet geblieben sind.

Dabei sind folgende Angaben wichtig:

- Name und Adresse der Institution
- Name, Funktion innerhalb der Institution, Telefon- und Faxnummer, e-mail Adresse eines Ansprechpartners

3.2.2 Datenquellen

Ein erfahrener Nutzer von Geodaten gewinnt aus der Information über Erfassungsquellen einen ersten Eindruck über die Verwendbarkeit der Daten. Fehler oder Ungenauigkeiten in den Datenquellen wirken sich direkt auf die digitalen Daten aus. Wenn nicht mehrere, unabhängige Quellen herangezogen und bei der Erfassung miteinander verglichen werden, so erfolgt die Erfassung von Geodaten ohne Redundanz. Dabei können keine falschen Klassifizierungen, fehlende Objekte oder Attributwerte aufgedeckt und auch keine Genauigkeitsmaße aus den Daten abgeleitet werden.

Als Erfassungsquelle für die Ojektgeometrie stehen entweder die reale Welt und deren Abbild oder durch Interpretation und Klassifizierung vorverarbeitete Produkte in analoger oder digitaler Form zur Verfügung.

Als einzige **originäre Datenquelle**, also Quelle die ungefiltert und nicht interpretiert ist, dient die reale Welt. Die Datenerhebung erfolgt dabei durch:

- Felderfassung: Datenquelle ist die reale Welt, wie sie sich einer Person vor Ort darstellt

Als **sekundäre Datenquelle**, kann eine Erfassungsquelle bezeichnet werden, die ein Abbild der realen Welt darstellt, in dem a priori keine Klassifizierung stattgefunden hat, die aber durch die Beschränkung bezüglich Auflösung, Bandbreite und Dimensionalität gefilterte Informationen enthält. Als Beispiele können aufgezählt werden:

- Luftbild: photographisches Abbild der realen Welt auf lichtempfindlichen Film oder durch Sensoren in verschiedenen Spektralbereichen des Lichtes
- Orthophoto: Orthometrische Entzerrung des Luftbildes mit Hilfe eines digitalen Geländemodells
- Laserscandaten: entlang von Profilen durch Laufzeitmessungen von am Boden reflektierten Laserimpulsen ermitteltes digitales Geländemodell
- Satellitenbild: vergleichbar mit Luftbild oder Orthophoto aber mit Besonderheiten bei der geometrischen Entzerrung. Neben Lichtsensoren werden häufig Sensoren in anderen Bereichen des elektromagnetischen Spektrums verwendet, z.B. Interferenzbilder im Radarbereich (SAR)

Zu den **tertiären Datenquellen**, also aus originären Datenquellen durch Interpretation abgeleitete Zwischenprodukte, zählen

- Topographische Karte
- Thematische Karte (Unterteilung nach *Hake und Grünreich, 1994*)

Für Sachdaten läßt sich die Liste der Erfassunsquellen beliebig erweitern. Teilweise lassen sich Attributwerte durch Interpretation oder Messung aus den aufgezählten Quellen ableiten. In vielen Fällen müssen allerdings Verzeichnisse, Register und andere fachliche Zusammenstellungen herangezogen werden. Der Bezug zu den Objekten wird dabei meistens durch eindeutige Identifikatoren hergestellt. Für Multimediaattribute ist die Auswahl von Erfassungsquellen eingeschränkt, weil bisher noch wenige zugängliche Archive zu Geoobjekten, bei denen der Raumbezug herstellbar ist, existieren.

Die Datenquellen müssen mit dem Namen der Quelle, sowie Angaben über Urheber, Maßstab und Datum der Aufnahme genau identifiziert werden. Für einzelne Verfahren müssen unter Umständen spezielle beschreibende Angaben gemacht werden, um die Datenquellen vollständig einschätzen zu können (z.B. die Flughöhe und die verwendete Kamera bei Luftbildern).

Um geometrische Informationen aus den Datenquellen entnehmen zu können, muß ein Raumbezug hergestellt werden. Der Raumbezug wird im allgemeinen durch eine Transformation des in der Datenquelle festgelegten Koordinatensystems in das übergeordnete System der Geodaten hergestellt. Die Bestimmung der Transformationsparameter erfolgt dabei über Paßpunkte. Die verwendeten Paßpunkte, ihre Koordinaten und die Herkunft dieser Koordinaten sind nachzuweisen.

Nicht alle Informationen zu einem digitalen Geoobjekt müssen aus derselben Datenquelle entnommen werden. Die Zuordnung zu einer Objektklasse, die Festlegung der Geometrie und die Zuweisung einzelner Attributwerte kann jeweils aus einer anderen Datenquelle stammen. Neben der Auflistung der verwendeten Datenquellen muß außerdem angegeben werden, welche Information aus welcher Quelle entnommen wurde.

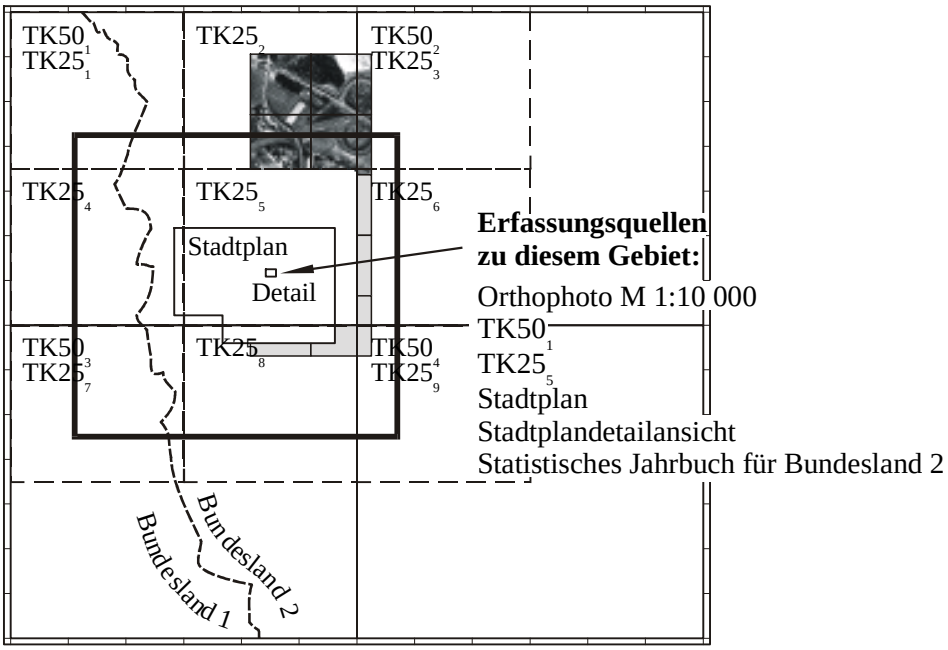


Abbildung 23: Überlagerung von Gebieten, die jeweils die Flächen der Datenquellen repräsentieren.

Jede einzelne Datenquelle deckt ein bestimmtes Gebiet ab. Durch Überlagerung der Datenquellen ergeben sich neue Gebiete mit homogener Quellenlage. Je mehr Datenquellen herangezogen werden, um so kleiner werden diese Gebiete.

3.2.3 Erfassungsmethoden

Abhängig von den Datenquellen lassen sich Geodaten mit unterschiedlichen Methoden erfassen. Damit der Entstehungsweg eines Datensatzes nachvollzogen werden kann, muß er offengelegt werden. Die Erfassung kann entweder manuell oder automatisiert ablaufen. Innerhalb des Erfassungsprozesses sind Kontrollen durchzuführen. Ist ein Qualitätsmanagement eingeführt, so ist dieses anhand des Qualitätsmanagementhandbuches und eventuell einer Zertifizierung nachzuweisen.

3.2.4 Vorverarbeitung und Transformationen

Alle Verarbeitungsschritte, die zwischen der Datenerfassung und der Verwendung der Geodaten für eine Aufgabe durchgeführt wurden, sind zu dokumentieren. Ein lückenloser Nachweis der Historie von Geodaten ist für eine Einschätzung der Verwendbarkeit von großer Bedeutung.

Da die Möglichkeiten zur Veredelung oder Anpassung der Geodaten mannigfaltig sind, kann eine Beschreibung der Vorverarbeitung oder Transformation nicht formalisiert werden. Wichtig ist, daß alle Algorithmen und Parameter vollständig dokumentiert sind.

3.2.5 Sprache der Sachdaten

Durch die Globalisierung des Geodatenmarktes und durch Anwendungen, die nationale Landesgrenzen überschreiten, ist die Sprache, in der Sachdaten beschrieben sind, für einen Anwender von Interesse. Während Eigennamen in der jeweiligen Landessprache (z.B. München), als Transliteration (z.B. Muenchen) oder als Transkription (z.B. Munich oder Monaco de Bavaria) dargestellt werden können, sind textliche Attribute in beliebige Sprachen übersetzbar. Es ist möglich, Sachdaten multilingual in der Datenbank zu verwalten. In den Metadaten ist anzugeben, in welcher Sprache die Daten ursprünglich erfaßt wurden, und in welchen weiteren Sprachen die Sachdaten abgegeben werden können.

Auch die Sprachen zur Beschreibung der konzeptionellen Modellierung und der Datenschemata sind für das Verständnis der Geodaten wichtig. Bei Übersetzungen in andere Sprachen ist auf die

Definitionen von Objektklassen und Attributen besonderen Wert zu legen, weil wörtlich übersetzte Begriffe vor unterschiedlichem kulturellen Hintergrund zu Fehlinterpretationen führen können. Als Beispiel sei die topographische Objektklasse „Gehölz“ angeführt. Für eine wörtliche Übersetzung ins Englische können die Begriffe „grove“ oder „wood“ verwendet werden. Der Begriff „grove“ hat im Deutschen aber noch die Bedeutung „Hain“ und für „wood“ können im Deutschen neben „Gehölz“ auch die Begriffe „Holz“ oder „Wald“ verwendet werden. Eine Abhandlung zu linguistischen Konzepten bei Geoobjekten in unterschiedlichen Sprachen findet sich bei *Ferrari, 1996*.

Unterschiedliche Sprachen verwenden unter Umständen verschiedene Schrift, so daß beim Austausch von Geodaten die zugrunde liegenden Zeichensätze berücksichtigt werden müssen (ASCII: *ISO 646, 1991*, das lateinische Alphabet der westeuropäischen Sprachen: *ISO 8859, 1988*, das vollständige lateinische Alphabet: *ISO 6937, 1994*, und ein universeller Zeichensatz, der alle Alphabete der Welt umfaßt: *ISO 10646, 1993*).

3.2.6 Aktualität

Bei originären und sekundären Datenquellen gibt das Datum der Aufnahme den tatsächlichen Stand des zu erfassenden abstrakten Abbildes der realen Welt wider. Die Aufnahme kann sich über einen Zeitraum von Bruchteilen von Sekunden (Luftbilder) über mehrere Minuten (Fernerkundungsdaten) bis zu mehreren Tagen bei der Feldaufnahme erstrecken. Ausschlaggebend ist der Zeitpunkt des Aufnahmeendes, in kritischen Fällen sind Anfang und Ende der Aufnahme anzugeben (wenn die Aufnahme sich z.B. über mehrere Monate oder Jahre erstreckt). Bei tertiären Datenquellen muß die Metainformation über die Aktualität der zugrunde liegenden primären oder sekundären Datenquelle bekannt sein. Aus einer dezidierten Quellenangabe läßt sich somit die Aktualität eines Objektes oder eines Attributwertes eines Objektes ableiten.

Mit dem Datum der Aufnahme wird dokumentiert, bis zu welchem Zeitpunkt Änderungen der realen Welt in der Quelle festgehalten sind. Zur Einschätzung der Aktualität müssen die Änderungszyklen von speziellen Geoobjekten beachtet werden. Die Änderungszyklen sind von der Dynamik eines abstrakten Objektes abhängig. Reale Objekte sind permanent sowohl thematischen als auch räumlichen Veränderungen unterworfen. Beispiele für unterschiedliche Dynamik von einzelnen abstrakten Objektklassen sind in der folgenden Tabelle aufgeführt.

	hohe zeitliche Variation	geringe zeitliche Variation
thematische Änderung	Wetterlage	Flächennutzung: Wald → Wiese → Wohnbau
geometrische Änderung	Lkw einer Speditionsfirma mit Navigationssystem für ein Flottenmanagement	geologische Gesteinsschichten (Stratigraphie)

Geodaten sind per se nie aktuell, da im Allgemeinfall eine zeitliche Verschiebung zwischen der tatsächlichen Änderung in der realen Welt und der Nachführung dieser Änderung in der Datenbasis besteht.

Wenn sich ein Objekt der realen Welt zu einem bestimmten Zeitpunkt geändert hat, kann dies auf den Datenbestand drei verschiedene Auswirkungen haben:

1. Die Änderung ist in der Erfassungsquelle und im Datenbestand enthalten.
2. Die Änderung ist nicht in der Erfassungsquelle, d.h. der Zeitpunkt der Aufnahme liegt vor dem Änderungszeitpunkt, und folglich auch nicht im Datenbestand enthalten.
3. Die Änderung ist in der Erfassungsquelle aber nicht im Datenbestand enthalten.

In den Fällen 1 und 2 entspricht der Stand der Geodaten auch dem Stand der Erfassungsquelle und des abstrakten Abbildes der realen Welt zum Aufnahmezeitpunkt. Im 2. Fall ist der Datenbestand zwar nicht aktuell, weil sich seit der Aufnahme eine Änderung ergeben hat, aber die Metadaten geben einem Nutzer keine falsche Information. Mit der Datumsangabe erhält der Nutzer den Hinweis, daß er sich,

falls er aktuelle Daten für die Anwendung benötigt, neuere Daten besorgen oder falls diese nicht verfügbar sind, über Änderungen der realen Objekte in diesem Gebiet anderweitig informieren muß.

Im 3. Fall wird dem Nutzer eine Aktualität der Geodaten suggeriert, die de facto nicht vorhanden ist. Die digitalen Daten stimmen nicht mit ihrer Entsprechung im abstrakten Abbild der realen Welt überein. In diesem Fall liegt ein Fehler in den Daten vor, der zur Kategorie Unvollständigkeit der Datenbasis (Abschnitt 4.3.1) zählt.

Die Aktualität von Daten kann erst unter Einbeziehung der Dynamik von Objekten oder Attributen eingeschätzt werden.

Das Datum der Aufnahme besteht entweder aus der Angabe von Tag, Monat und Jahr oder mit zusätzlicher Angabe der Uhrzeit in Stunden, Minuten und eventuell auch Sekunden. Die Jahresangabe sollte immer mit vierziffriger Angabe erfolgen, da ansonsten mit der Jahrtausendwende Probleme bei der Berechnung von Zeitintervallen entstehen. Die Datumsangabe ist in Datenbankmanagement-systemen ein Standarddatentyp und wird in der internationalen Norm *ISO 8601, 1988*, behandelt.

3.2.7 Fortführung

Als Fortführung bezeichnet man die Erfassung von Änderungen des abstrakten Abbildes der realen Welt in der Datenbasis. Die reale Welt ist permanent Änderungen unterworfen, allerdings sind nur diejenigen Änderungen von Interesse, die sich auf das abstrakte Abbild der realen Welt auswirken. Das Problem der Fortführung besteht in der Bekanntmachung von Änderungen an den Betreiber des Geoinformationssystems. Für verschiedene Objektklassen und verschiedene Anwendungen bestehen unterschiedliche Anforderungen an die Fortführungszyklen. Es gibt zwei Methoden, die Änderungen des abstrakten Abbildes der realen Welt in den digitalen Datenbestand aufzunehmen:

- ereignisorientierte Fortführung: für Objekte der sogenannten Spitzenaktualität ist eine Änderung unmittelbar nach Kenntnisnahme in der Datenbasis nachzuführen. Der Zeitraum zwischen Änderung und Kenntnisnahme ist durch geeignete Maßnahmen, wie z.B. Meldedienste, möglichst kurz zu halten.
- periodische Fortführung: In regelmäßigen zeitlichen Abständen sind alle Objekte, für die eine Fortführung ansteht, auf Änderungen hinsichtlich Existenz, Eigenschaften und Relation zu anderen Objekten zu prüfen, und gegebenenfalls nachzuführen. Mit der Kenntnis des Zeitpunktes der Aufnahme und des Zyklus der Fortführung kann ein Nutzer den Zeitpunkt einer Revision aller Objekte dieser Objektklasse berechnen.

Es besteht die Möglichkeit, innerhalb eines Datenbestandes für verschiedene Objektklassen unterschiedliche Fortführungsmodi einzuführen. Wichtig ist, daß der Nutzer erfährt, wann er spätestens nach einer Änderung in der realen Welt mit fortgeführten Daten rechnen kann. Diese Information ist in den Metadaten zu führen.

3.3 Angaben zur Qualität der Daten

Damit ein potentieller Anwender vor oder spätestens beim Einsatz der Daten einschätzen kann, ob das Qualitätsniveau der Daten den Erfordernissen seiner Anwendung gerecht wird, benötigt er Angaben über die Qualität der Daten. Diese Angaben setzen ein Modell voraus, in dem festgelegt ist, wie die Qualität beschrieben und mit welchen Methoden die tatsächliche Qualität von konkreten Daten ermittelt werden kann. Da Qualitätsangaben zusätzliche Informationen zu den Daten darstellen, sind sie Teil und aufgrund ihrer tragenden Rolle sogar wesentlicher Bestandteil der Metadaten.

Das Qualitätsmodell wird in Kapitel 4 ausführlich behandelt.

3.4 Verfügbarkeit

Ein Anwender muß anhand von Metadaten erkennen können, ob die benötigten Daten auch verfügbar sind. Die Verfügbarkeit wird in den folgenden Abschnitten unter räumlichen, thematischen, formalen, finanziellen, rechtlichen und organisatorischen Aspekten betrachtet.

3.4.1 Flächendeckung

Im Zusammenhang mit der Verfügbarkeit von Geodaten versteht man unter Flächendeckung das Gebiet, wo Geodaten mit den beschriebenen Eigenschaften verfügbar sind. Die Gebietsangabe erfolgt entweder durch Koordinaten des begrenzenden Polygons oder über eine politische oder naturräumliche Gliederung.

Sind Datenbestände erst im Aufbau, oder werden vorhandene Datenbestände ergänzt oder aktualisiert, so soll der Planungshorizont angegeben werden, also zu welchem Zeitpunkt welches Gebiet verfügbar sein wird. Kann die Planung nicht eingehalten werden, müssen die Metadaten über die Verfügbarkeit aktualisiert werden. Sobald die Daten erfaßt sind und alle Qualitätsprüfungen durchlaufen haben, sind sie in der Metadatenbank als freigegeben einzutragen.

Über die Flächendeckung erhalten die Metadaten einen räumlichen Bezug. Es bietet sich daher an, zur Navigation durch Metadaten ein Geoinformationssystem einzusetzen.

3.4.2 Abgabeeinheiten

Mit der Abgabeeinheit ist der kleinste Auszug aus der Datenbank bezeichnet, der an Abnehmer weitergegeben wird. Mögliche Einheiten sind in der folgenden Aufzählung zusammengestellt.

- die gesamte Datenbasis
- ein thematischer Auszug aus der Datenbasis
- ein Teilgebiet der Datenbasis (entweder vordefinierte oder kundenspezifische Teilgebiete)
- individuelle Objekte
- nur geometrische Information der Datenbasis (mit oder ohne Signaturen)
- Änderungen einer Datenbasis gegenüber einem bestimmten (vordefinierten oder kundenspezifischen) Zeitpunkt

Die aufgezählten Einheiten können in beliebigen Kombinationen zusammengefaßt werden, wobei nicht alle möglichen Kombinationen sinnvoll sind.

Je stärker sich die Abgabeeinheit an den Bedürfnissen des Kunden orientiert, um so größer ist der Aufwand bei der Datenbereitstellung. Die Extraktion der gewünschten Daten muß über gezielte Abfragen erfolgen. Wenn die Datenbasis nicht in einer gemeinsamen Datenbank sondern z.B. in einzelnen Dateien abgelegt ist, so müssen diese Abfragen in allen Einzeldateien durchgeführt werden, und als Ergebnis kann die Bereitstellung auch wieder in Einzeldateien oder durch eine Zusammenführung der Teile in einer gemeinsamen Datei erfolgen.

3.4.3 Abgabeformate

Die Abgabeformate können sich von den originären Formaten der Datenverwaltungskomponente des Geoinformationssystems unterscheiden und hängen auch von den Abgabeeinheiten ab. Werden z.B. nur die Geometriedaten benötigt, so können die Daten in einem einfachen CAD (Computer Aided Design) Austauschformat wie z.B. DXF (Digital Exchange Format) der Firma Autodesk übermittelt werden.

Sollen sowohl Geometrie- als auch Sachdaten abgegeben werden, so müssen komplexere Schnittstellen verwendet werden. Als primäres Abgabeformat wird üblicherweise das herstellerspezifische Austauschformat angeboten, damit gewährleistet ist, daß alle Nutzer mit der gleichen GIS-Software die Daten problemlos übernehmen können. Die meisten GIS-Softwareentwickler gehen dazu über, die systemspezifischen Formate der führenden Hersteller lesen und teilweise auch schreiben zu können.

Herstellerunabhängige Austauschformate befinden sich in der Entwicklung. Im Rahmen der europäischen Normungsaktivitäten wird die Sprache EXPRESS (*ISO/DIS 10303-11, 1994*) als Austauschformat vorgegeben. EXPRESS hat den Vorteil, daß sowohl das konzeptionelle und logische Schema als auch die Geodaten verschlüsselt werden können. Die Austauschstruktur verwendet eine Klartextverschlüsselung und wird als sequentielle Datei realisiert.

Ein ursprünglich für den militärischen Bereich entwickeltes Austauschformat stellt DIGEST dar (*Digital Geographic Information Exchange Standard*; *DGIWG, 1997*). Die Geodaten in Form von Raster-, Gitter- oder Vektorrepräsentation werden in eine sequentielle Dateistruktur gebracht, und können so zwischen Nutzern ausgetauscht werden.

Zur Abgabe und zur Fortführung von Geobasisdaten der öffentlichen Vermessungsverwaltungen in Deutschland, also sowohl für ATKIS als auch ALK, wurde die Einheitliche Datenbankschnittstelle (EDBS) entwickelt. EDBS ermöglicht nicht nur die Abgabe von Geometrie und Sachdaten in objektstrukturierter Form, sondern unterstützt auch eine Abgabe von Änderungsdatensätzen für den sogenannten Bezieher von Sekundärdaten.

Neue Ansätze, die auf den Möglichkeiten der verteilten Datenhaltung in objektorientierten Systemen beruhen, werden von OGC verfolgt. Dabei steht nicht der Austausch von Daten in Dateiform im Vordergrund, sondern ein direkter Zugriff auf einen vernetzten Rechner z.B. bei der Institution, bei der die Daten erfaßt und gepflegt werden. Außerdem werden bei dieser Zugriffsmethode keine fest vorgegebenen Inhalte transportiert, sondern über einen Auruf von in Schnittstellenspezifikationen festgelegten Methoden können dynamisch Abfragen gestartet werden, die bis zu komplexen Analysen gehen können, bei denen nur die Ergebnisse übermittelt werden. Wo sich die Daten physikalisch befinden, spielt für die Anwendung keine Rolle. Aus diesem Grund sind in einem solchen interoperablen Umfeld Metadaten und davon insbesondere Qualitätsangaben besonders wichtig.

Die Internettechnologie stellt eine neue Herausforderung für die breite Anwendung von Abfragen mit Geobezug dar. Herkömmliche Browser unterstützen nur einfache Grafiken im Rasterformat und ohne Georeferenzierung. Für GIS-Anwendungen sind aber komplexere Datenstrukturen erforderlich, die in vielen Fällen eine Vektorrepräsentation voraussetzen. Welche Funktionalitäten dabei auf der Seite des Clients und welche auf der des Servers ablaufen sollen, hängt von verschiedenen Faktoren ab. Einige der Faktoren sind:

- Leistungsfähigkeit des Browers in der Verarbeitung von Geodaten
- Datenvolumen, das übertragen werden muß, und Bandbreite, die zur Verfügung steht
- Das Datenvolumen ist abhängig von der Anwendung, die durchgeführt werden soll
- Rechtliche Aspekte, wie Nutzungsrecht an den Daten oder Abrechnungsmöglichkeiten für Dienstleistungen

Welches dabei die optimale Architektur darstellt, ist noch weitgehend unerforscht. Es gibt allerdings schon Vorschläge, wie die Datenbeschreibungssprache für das *World Wide Web (www)* im Hinblick auf Geodaten erweitert werden kann.

3.4.4 Kosten

Beim Verkauf oder bei einer Nutzungsüberlassung von Geodaten fallen in der Regel Kosten an. Ein wichtiges Kaufkriterium für einen potentiellen Nutzer ist der Preis der Daten. Die Preisangaben beziehen sich auf Abgabeeinheiten. Falls bei der Abnahme größerer Datenmengen oder für bestimmte Anwendergruppen Rabatte gewährt werden, sind diese bei den Kosten anzugeben, um eine Preiskalkulation zu ermöglichen.

3.4.5 Abgabebeschränkung

Wenn die Abgabe der Geodaten auf bestimmte Nutzergruppen eingeschränkt ist, so muß diese Information einem potentiellen Nutzer zugänglich gemacht werden. Nationale oder auch militärische Schranken waren lange Zeit ein Hindernis für uneingeschränkten Zugriff auf bestehende Datenbestände. Im Rahmen einer zunehmenden Globalisierung und einer Betrachtung von großen Datenbasen unter marktwirtschaftlichen Aspekten weitet sich die Abgabebeschränkung zugunsten eines globalen Zugangs zu Geodaten auf.

3.4.6 Nutzungsrechte

Mit der Abgabe der Daten können bestimmte Einschränkungen der Nutzung verbunden werden. Insbesondere wird dies die Weitergabe an Dritte betreffen. Indem digitale Daten sehr einfach und ohne Qualitätsverlust vervielfältigt werden können und ein Kopierschutz, wie er in vielen Fällen bei Software eingesetzt wird, schwer zu realisieren ist, muß eine entsprechende Nutzungsbeschränkung vertraglich abgesichert werden.

3.4.7 Haftung

Bei einer Abgabe der Daten muß der rechtliche Haftungsanspruch geklärt sein. Wie ist zwischen den Vertragspartnern zu verfahren, wenn in den Daten Fehler enthalten sind, die zu einem Schaden führen? Das Ausmaß eines solchen Schadens hängt von der Anwendung ab, kann aber z.B. bei Planungsfehlern beachtliche Dimensionen annehmen.

Im Allgemeinfall wird bei der Abgabe ein Haftungsausschluß angegeben, der durch einen Verweis auf die Qualität der Daten zu rechtfertigen ist. Wenn die Qualitätsangaben keine falschen Erwartungen an die Daten suggerieren, wird der Nutzer die erforderliche Sorgfalt beim Umgang mit den Daten walten lassen, um Schäden zu vermeiden.

3.4.8 Vertrieb

Wenn die vertreibende Institution sich vom Urheber unterscheidet, ist die Kontaktstelle anzugeben, bei der die Geodaten erworben werden können. Die Bestell- oder online-Bezugsmodalitäten sind anzugeben. Falls keine Gebühren beim Vertrieb von Geodaten erhoben werden, oder sobald die Abrechnung beim Bezug über das Internet geklärt ist, kann ein direkter Zugriff auf die Daten in einem Server erfolgen.

3.5 Referenzprojekte

Üblicherweise wurden in der Vergangenheit Geodaten in einem Projekt für einen bestimmten Zweck erfaßt. Dieser Anwendungszweck gibt Aufschluß über die Art der Modellierung und läßt einen potentiellen Nutzer auf weitere mögliche Anwendungen schließen.

Da Datenerfassung und Datenpflege mit einem sehr großen Aufwand verbunden sind, wurde dazu übergegangen, Basisdaten zu erfassen, die in einer Vielzahl von Projekten Anwendung finden können und gegebenenfalls mit themenspezifischen Fachdaten ergänzt werden.

In den Metadaten soll die Information geführt werden, welche Projekte mit den zu beschreibenden Geodaten durchgeführt wurden. Dabei sind auf die ausgeführten Analysen und Auswertungen einzugehen und darauf, welche Produkte aus diesen Daten abgeleitet wurden. Es sind nicht nur die erfolgreichen Projekte aufzuführen, sondern auch die Aufgaben, die mit minderem oder ohne Erfolg angegangen wurden. In diesem Fall sollen insbesondere die aufgetretenen Probleme angegeben werden und wenn möglich Hinweise dazu, wie sie umgangen werden können.

Unterscheidet sich die Institution, die ein Referenzprojekt mit Hilfe der Geodaten mit oder ohne Erfolg bearbeitet hat, vom Urheber der Geodaten, so sind der Name, Anschrift und Ansprechpartner dieser Institution anzugeben, vorausgesetzt diese Institution ist mit einer Veröffentlichung dieser Daten einverstanden.

3.6 Öffentliche und interne Metadaten

Nicht alle Metadaten sind zur Weitergabe an einen Nutzer von Geodaten bestimmt. Wird für die Datenerfassung ein Qualitätsmanagement eingeführt, so werden im Sinne einer Rückverfolgbarkeit des Produktionsablaufes die Namen des Erfassers und der Kontrolleur einer Erfassungseinheit gespeichert. Zum Schutz von persönlichen Daten müssen Namen im Metainformationssystem vertraulich behandelt und der Zugriff nur einer autorisierten Person erlaubt werden. Metadaten dieser Art können als interne Metadaten bezeichnet werden.

Öffentliche und interne Metadaten können entweder in getrennten Datenbanken verwaltet werden, oder sie liegen in einer gemeinsamen Datenbank. Im letzteren Fall muß ein Filter die internen Informationen vor einer Recherche durch einen externen Nutzer schützen. Die gemeinsame Datenhaltung vermeidet aber eine redundante Haltung von Metadaten und vermindert damit eine Gefahr von Inkonsistenzen. Zum Beispiel wird ein Datensatz nach Durchführung einer Endkontrolle als verfügbar freigegeben. Bei einer getrennten Datenhaltung muß diese Information vom Betreiber des Metadaten-informationssystems in beiden Datenbanken eingetragen werden.

4 Qualitätsmodell

Das Qualitätsmodell dient der Festlegung von Kriterien und Methoden zur vollständigen Beschreibung der Übereinstimmung zwischen Daten und ihrem Pendant in der realen Welt. Es ist ein wesentlicher Bestandteil der Metadaten. Es beschreibt zum einen die Qualität des Datenmodells, das wiederum einen Bestandteil der Metadaten darstellt und zum anderen die Qualität der Daten. Das Qualitätsmodell setzt sich somit aus den Komponenten Modellqualität und Datenqualität zusammen.

Will ein Anwender entscheiden, ob ein Datensatz für die gewünschte Anwendung geeignet ist, muß er zuerst ein abstraktes Abbild der realen Welt entwerfen, mit dem er festlegt, wie die Daten für die Anwendung beschaffen sein müssen (Anwendungsmodell). Durch einen Vergleich mit dem abstrakten Abbild, das den Daten zugrunde liegt, kann er entscheiden, ob die Daten prinzipiell den Erfordernissen entsprechen. Dieser Vergleich ist in Abbildung 24 als Forderung (1) dargestellt.

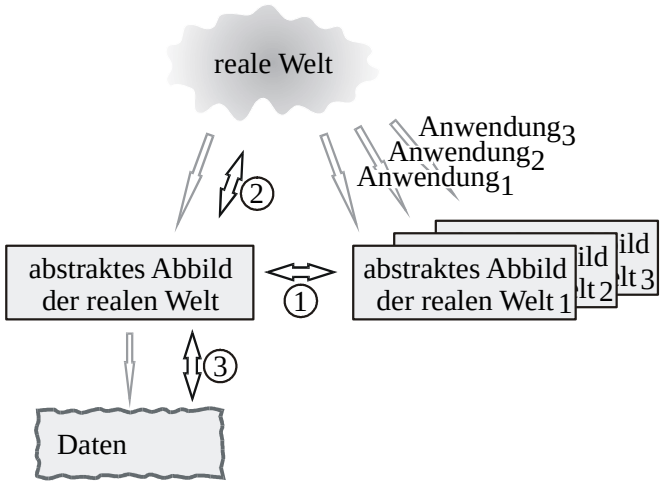


Abbildung 24: Eignung von Geodaten für eine Anwendung.

Voraussetzung Nummer (2) verlangt, daß das Datenmodell die reale Welt widerspruchsfrei repräsentiert. Dieser Punkt wird durch die Beschreibung der Modellqualität (Abschnitt 4.1) dargelegt. Als dritte Voraussetzung (3) müssen die Daten ihrer Spezifikation entsprechen. Der Grad dieser Entsprechung wird als Datenqualität bezeichnet (Abschnitt 4.2).

Erst wenn der Anwender sichergestellt hat, daß das Datenmodell mit dem Anwendungsmodell vereinbar ist, daß das Datenmodell detailliert genug ist, um keine Spielräume bei der Interpretation zuzulassen, die dann mit dem Anwendungsmodell nicht mehr vereinbar sind, und daß Datenfehler, die nie vollständig vermeidbar sind, innerhalb von tolerablen Grenzen liegen, sollte er die Daten für seine Anwendung einsetzen.

Die Qualitätsmetadaten sollen einem Anwender Informationen geben, anhand derer er entscheiden kann, ob Daten für seine Anwendung geeignet sind (englisch: *„fitness for use“*). Dieses Ziel resultiert aus der allgemeinen Definition von Qualität in *DIN ISO 8402, 1991*.

Mit dem Qualitätsmodell sollen demnach mehrere Ziele verfolgt werden:

- Qualitätsmerkmale strukturieren
- Qualitätsmaße festlegen
- Qualitätsmaße ermitteln
- Produktion von Geodaten steuern, damit Qualitätsziele eingehalten werden
- Nutzer über die Qualität der Daten informieren

4.1 Modellqualität

Nach Abbildung 1 werden bei der Datenerfassung zwei Abstraktionsstufen unterschieden. Stufe 1, die Modellierung, erzeugt aus der realen Welt ein abstraktes Abbild. Dieses Abbild muß eindeutig aus den realen Objekten hervorgehen. Wenn diese Zuordnung Mehrdeutigkeiten oder Widersprüche erzeugt, oder die Anforderungen nicht komplett abdeckt werden, dann ist das Datenmodell fehlerhaft oder schlimmstenfalls ungeeignet. Die Eignung des Datenmodells, die reale Welt für die Belange der Datenerfassung in GIS wiederzugeben, wird als Modellqualität bezeichnet.

Die Modellqualität kann auf verschiedene Arten beschrieben werden, die folgenden Kriterien sollten aber immer zur Bewertung eines Datenmodells herangezogen werden.

- Sind alle Elemente zur vollständigen Beschreibung des Datenmodells entsprechend der Zusammenstellung in Kapitel 2 enthalten?

Werden Elemente bewußt weggelassen, so ist dies in der Dokumentation zu vermerken und gegebenenfalls zu begründen. Ein unvollständiges Modell gibt dem Datenerfasser größere Freiheiten. Eigenschaften, die im Modell nicht vorgeschrieben sind, können frei interpretiert werden. Dies führt dazu, daß Daten unterschiedlicher Struktur entstehen können, obwohl sie nach demselben Datenmodell erfaßt wurden, wenn unterschiedliche Datenerfasser diese Freiheit unterschiedlich auslegen.

Bei ATKIS hat dies zu einer Reihe von länderspezifischen Besonderheiten geführt, auf die im folgenden näher eingegangen wird.

Unterschiedliche Abgabegebinde: Da die Abgabeeinheiten in der ATKIS-Dokumentation nicht vereinbart sind, geben manche Bundesländer die Daten im Bezug auf den Blattschnitt einer Topographischen Karte im Maßstab 1:10 000 (TK10) ab, andere im regelmäßigen und unregelmäßigen Zusammenschluß von mehreren Einheiten bezogen auf den Blattschnitt der Deutschen Grundkarte M 1:5 000 (DGK5). Die Blattschnitte dieser Kartenwerke beziehen sich auf die Koordinatenlinien unterschiedlicher Koordinatensysteme. Während die TK10 nach geographischen Koordinaten geschnitten ist, bezieht sich die Begrenzung der DGK5 auf Gauß-Krüger-Koordinaten.

Unterschiedliche Objektbegrenzung am Rand der Abgabeeinheit: Objekte werden in den Bundesländern entweder am Rand des Erfassungsgebietes scharf begrenzt oder die Objektbegrenzung erfolgt ausschließlich nach den ATKIS-Objektbildungsregeln, und bei der Datenabgabe werden die Objekte selektiert, die ganz oder teilweise innerhalb des Blattschnittes liegen. Diese Form der Datenhaltung und –abgabe hat auch Auswirkung auf die automatische Prüfung der Konsistenz von Geodaten (siehe Abschnitt 6.3).

Unterschiedliches Objektverständnis bezüglich der Objektbegrenzung: da die Objektbildungsregeln nicht immer eindeutig anwendbar sind, entstehen willkürliche Objektgrenzen. Diese unterschiedliche Ausgelegung der Objektbildungsregeln in den Bundesländern führt zu nicht dokumentierten Zusatzregeln. Während die einen versuchen, möglichst große Objekte zu bilden, teilen die anderen die Objekte in kleinere Einheiten auf. Ein großer Fluß, wie z.B. der Rhein, stellen nach dem ATKIS-OK ein einziges flächenhaftes Objekt dar. Aus praktischen Gründen läßt sich dieser aber so nicht erfassen. Diese Situation ist durch die Objektbildungsregeln nicht abgedeckt.

Unterschiedliche Referenzen: Während die einen Bundesländer bei planfreien Kreuzungen nur je eine Referenz nach oben oder unten pro Teilobjekt zulassen, sind bei anderen Mehrfachreferenzen zulässig. Der Objektartenkatalog macht dazu keine Aussage.

- Gibt es Widersprüche oder Mehrdeutigkeiten bei der Objektmodellierung?

Die Zuordnung zwischen den realen Objekten und den abstrakten Objekten muß eindeutig sein. Selbstverständlich werden nicht alle Phänomene der realen Welt abgebildet werden, aber wenn ein reales Objekt modelliert werden soll, dann darf es nicht mehr als eine Entsprechung im Datenmodell haben. Außerdem muß die Information zu einem Objekt wieder eindeutig seiner Entsprechung in der realen Welt zugeordnet werden können.

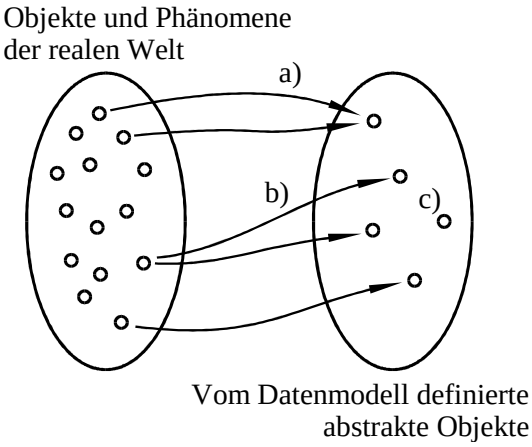


Abbildung 25: Mengendiagramm zur Modellqualität

Eine Mehrdeutigkeit als Mangel der Modellqualität stellt dieses Beispiel aus dem Objektartenkatalog (*Feature Attribute Coding Catalogue*, FACC) des ***DI****gital* ***G****eographic Information* ***E****xchange* ***ST****andard*, DIGEST, dar (*DIGEST, 1994*). Das folgende Beispiel entspricht der Kategorie b) in Abbildung 25.

Ein Hubschrauberlandeplatz kann nach FACC auf zwei verschiedene Arten modelliert werden:

Hubschrauber-landeplatz	*GB035 Heliport*	
	GB005 Airport/Airfield	mit Attribut *APT Airfield Type, APT = 9 Heliport*

Bei einer Abfrage aller Hubschrauberlandeplätze im Umkreis von einem bestimmten Radius müssen also immer zwei Objektklassen berücksichtigt werden, wobei eine durch ein Attribut eingeschränkt werden muß. Ein Anwender, der diese Mehrdeutigkeit des Datenmodells nicht kennt, wird nur einen Teil der Hubschrauberlandeplätze angezeigt bekommen, da sich einige Erfasser für die eine und andere Erfasser für die andere Lösung entscheiden werden. In Gefahrensituationen ist dieser Umstand sehr kritisch.

Ein weiteres Beispiel für Inkonsistenzen bei der Datenmodellierung stammt aus dem ATKIS-OK. Dieses Beispiel kann der Kategorie a) in Abbildung 25 zugeordnet werden:

Für die Objektart 7101 Verwaltungseinheit sind unter anderen die beiden Attribute EWZ und FLA vereinbart.

7101 Verwaltungseinheit	EWZ	Einwohnerzahl
	FLA	Fläche in ha

Die Attribute haben den Attributtyp ganze Zahlen. Als Attributwerte sind die tatsächlichen Werte für die jeweiligen Attribute einzutragen. Nach dem ATKIS-OK Teil D0 ist es vorgesehen, daß Attributwerte, die nicht zutreffen, deren Wert bei der Erfassung nicht bekannt ist, oder deren Attributwert zwar bekannt aber nicht in der Liste der möglichen Attributwerte aufgelistet ist, mit bestimmten Zahlenwerten verschlüsselt werden. Die Zahlen für die Attributwerte mit dieser besonderen Bedeutung sind in der folgenden Tabelle aufgeführt.

9997	Attribut trifft nicht zu
9998	Nach Quellenlage derzeit keine Zuordnung möglich
9999	Sonstiges

Wie hat ein Anwender den Zahlenwert zu interpretieren, wenn bei einem der Attribute EWZ_Einwohnerzahl oder FLA_Fläche_in_ha der Objektart Verwaltungseinheit der Wert 9998 eingetragen ist? Daß eine Gemeinde 9998 Einwohner oder eine Fläche von 9998 ha hat, ist nicht auszuschließen, nicht einmal sehr unwahrscheinlich. Da die besonderen Attributwerte 9997, 9998 und

9999 im Wertebereich für die Attribute liegen, kommt es an diesem Punkt zu einem Widerspruch des Datenmodells.

- Sind Objektklassen festgelegt worden, die in der realen Welt nicht vorkommen?

Dieses Kriterium ist vielleicht nur von akademischem Interesse. Es entspricht der Kategorie c) in Abbildung 25. Beispiele in existierenden Datenmodellen können nicht angeführt werden.

- Entspricht das Datenmodell den Anforderungen durch die geplante Anwendung?

Werden Informationen erfaßt, die redundant sind, oder für die Anwendung nicht gebraucht werden, oder deren Nutzen die Kosten bei der Erfassung nicht rechtfertigt? Diese Frage hat einen sehr praktischen Hintergrund. In der Euphorie der Einführung von Geoinformationssystemen werden oft sehr ambitionierte Datenmodelle aufgestellt. Diese Modelle mögen zwar in sich konsistent und vollständig sein, sie müssen ihre Praktikabilität aber trotzdem unter Beweis stellen. Allgemeingültige Kriterien lassen sich zur Prüfung der Durchführbarkeit nicht finden. Es sind letztendlich wirtschaftliche oder politische Erwägungen, die angeben, welcher Aufwand getrieben werden darf, um die Tiefe des Modells mit Daten zu füllen und diese auch zu pflegen.

Wichtig ist, daß der Anwender sich über seine Anforderungen an die Daten im klaren ist. Unangemessene Forderungen nach einer hohen Genauigkeit wirken sich auf die Erfassungsgrundlagen und –methoden aus, und somit auch auf die Kosten. Die Auflösung der Daten (siehe Abschnitt 2.2.2.8) wirkt sich auf das Datenvolumen aus. Bei bestimmten Analysen gibt es Grenzen für die Datenmenge, die noch bei akzeptablem Zeitverhalten analysiert werden kann. Gerade räumliche Analysefunktionen zeigen oft ein Zeitverhalten, das quadratisch mit der Anzahl von Objekten wächst. Die Frage, in welcher Auflösung die Daten gebraucht werden, sollte also mit der Anwendung abgestimmt werden.

4.2 Datenqualität

Da sich Geodaten, insbesondere Geobasisdaten, dadurch auszeichnen, daß sie für viele Anwendungen verwendbar sind, ist eine pauschale Aussage über „geeignet“ oder „nicht geeignet“ nicht akzeptabel. Die Qualität der Daten soll nach objektiven, anwendungsunabhängigen Kriterien bewertet und dem Anwender zugänglich gemacht werden.

Zu diesem Zweck werden in der Literatur zwei Ansätze diskutiert. Ein Ansatz beruht auf einer Beschreibung und Offenlegung der Ergebnisse von durchgeführten Qualitätsuntersuchungen (englisch: *„truth in labeling“*). Der andere Ansatz beruht auf der Festlegung und Dokumentation von Qualitätszielen, verbunden mit einer Erklärung, daß alle Daten mindestens den gesteckten Zielen entsprechen (englisch: *„threshold concept“*). Während der erste Ansatz eher auf eine nachträgliche, unabhängige Kontrolle und Qualitätsbewertung abzielt, erfordert der letztere begleitend zur Datenerfassung ein Qualitätsmanagement, das sicherstellt, daß die Ziele überall eingehalten werden.

Die Definition der Datenqualität nach dem „*truth in labeling*“-Konzept wird in der europäischen Vornorm *prENV 12656, 1998*, eingeführt und auch für den Entwurf des internationalen Normierungskomitees zur Beschreibung der Datenqualität *ISO 19113, CD, 1999*, verwendet. Das „threshold“-Konzept wird vor allem bei der Erfassung von Daten für die Fahrzeugnavigation eingesetzt. Es findet sich daher auch in der internationalen Norm für *Geographic Data Files* (GDF), *prENV ISO 14825, 1996*, wieder.

In der vorliegenden Arbeit kommen beide Konzepte zum Tragen. Sie sind auch miteinander vereinbar. Beide benötigen objektive Qualitätskriterien (Abschnitt 4.3) und zugehörige Qualitätsmaße (Abschnitt 4.5). Während das „truth in labeling“-Konzept die Qualität nur beschreibt, wird beim „treshold“-Konzept schon eine Entscheidung über Ablehnung oder Akzeptanz getroffen. Insbesondere bei der statistischen Qualitätskontrolle (Kapitel 7) werden Grenzwerte benötigt. Diese Grenzwerte sind natürlich wieder anwendungsbezogen.

4.2.1 Motivation für das Festlegen von Qualitätskriterien

Zur Identifizierung und Unterscheidung verschiedenartiger Datenfehler müssen Kriterien aufgestellt werden, die durch ihre Definition idealerweise ein orthogonales System bilden, d.h. alle Fehler eindeutig und vollständig einer Kategorie zuordnen.

Ein solches System kann mit den vier Kriterien **Vollständigkeit, Richtigkeit, Konsistenz** und **Genauigkeit** gebildet werden. Für die meisten Datenfehler ist die Zuordnung zwar eindeutig, aber nicht unbedingt offensichtlich, deshalb müssen einige Fälle anhand von Fallstudien zur klaren Einordnung erläutert werden. Es kann auch Datenfehler geben, die gleichzeitig in mehr als eine Kategorie fallen. Dies ist kein Widerspruch zu der angestrebten Orthogonalität, da Mehrfachfehler in Zusammenhang mit einem oder mehreren Objekten nicht ausgeschlossen werden. Da die Zuordnung eines Fehlers zu genau einem der Qualitätskriterien wichtig ist, muß bei Qualitätsuntersuchungen auf die Verwendung in Spezialfällen eingegangen werden.

Es muß Ziel eines jeden Geoinformationssystems sein, das den Anspruch hat, die reale Welt zu einem definierten Zeitpunkt modellhaft zu repräsentieren, alle vier genannten Kriterien überall vollständig zu erfüllen. Daß der damit verbundene Aufwand aus finanziellen, zeitlichen oder systemimmanenten Gründen bei manchen Datenbeständen nicht betrieben wird, führt dazu, daß Geodaten unterschiedliche Qualitäten besitzen. Trotzdem ist beim Aufbau einer Geodatenbasis anzustreben, die Daten vollständig, richtig, konsistent und so genau wie nötig zu erfassen. Der Erfüllungsgrad dieser Zielvorgabe ist durch Qualitätsmaße anzugeben, und es ist Aufgabe eines funktionierenden Qualitätsmanagementsystems, die Einhaltung von Grenzwerten für Qualitätsmaße zu gewährleisten.

Die Festlegung der Datenqualität als Grad der Übereinstimmung zwischen dem abstrakten Abbild der realen Welt und dem Datensatz läßt diejenigen Einflüsse unberücksichtigt, die durch das Modell hervorgerufen werden. Diese sind die Auflösung der Daten, die erschöpfende Darstellung der realen Welt, also die Geschlossenheit des Modells, und die richtige Darstellung. Die Aktualität der Daten wird durch die Erfassungsquellen limitiert, welche in den Metadaten unter dem Begriff „Herkunft“ geführt werden.

Abbildung 26 zeigt beispielhaft die Umsetzung der abstrakten Welt in die Repräsentierung in das Datenmodell eines Geoinformationssystems. Die dabei aufgetretenen Fehler werden im einzelnen diskutiert, was zu einer Zuordnung zu den vier genannten Kriterien führt.

4.2.2 Einführendes Beispiel

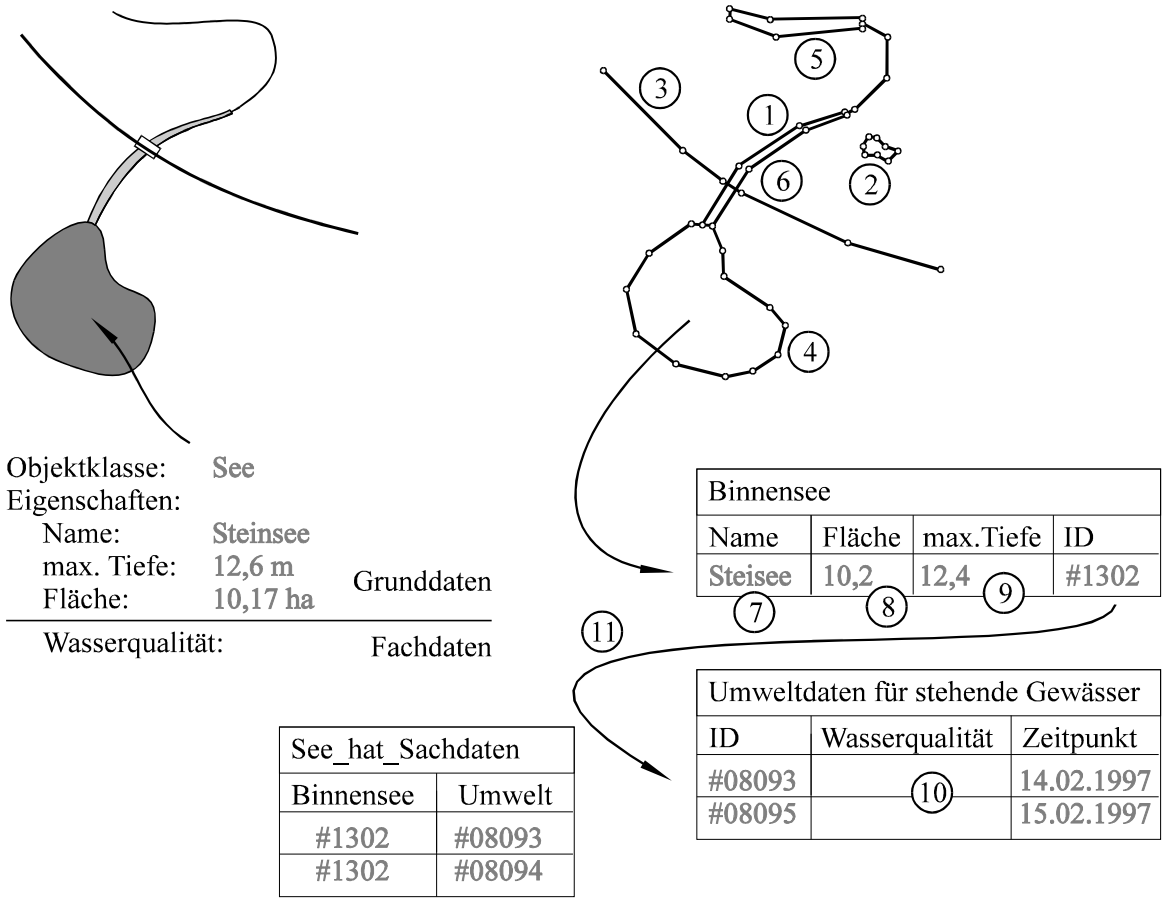


Abbildung 26: Das abstrakte Abbild der realen Welt und dessen Umsetzung in einen konkreten Datenbestand mit numerierten Datenfehlern.

Die dargestellte Situation zeigt eine abstrakte Welt mit einem See, in den ein anfangs schmaler, zur Mündung hin breiter Fluß führt. Im schmalen Bereich wird der Fluß durch die Mittelachse als Linie und im breiten Bereich als flächenhaftes Objekt modelliert. Der Fluß wird gekreuzt von einer Straße, deren Achse als Linie erfaßt werden soll. Im Kreuzungspunkt befindet sich eine Brücke. Da anhand der zweidimensionalen Modellierung nicht nachvollzogen werden kann, welches Objekt oben und welches unten liegt, wird dieser Sachverhalt durch Referenzen angegeben. Bei der Konstellation Straße mit Fluß ist diese Beziehung zwar bis auf wenige Ausnahmen eindeutig, aber bei der Überführung von Straße mit Straße oder Straße mit Schiene, läßt sich der Zusammenhang ohne Angabe von Referenzen weder anhand der Daten noch durch Erfahrungswerte feststellen. Der See ist durch seine Eigenschaften charakterisiert. Als topographische Eigenschaften sind entsprechend dem Modell Werte für „Namen“, „maximale Tiefe“ und „Fläche“ anzugeben. Die Eigenschaften werden in Form von Attributen modelliert. Die Objektklassen und die Attribute müssen im Datenmodell genau definiert sein. Bei der Objektklassendefinition wird angegeben, wann eine nichtfließende überirdische Wasseransammlung als See bezeichnet wird, der den Erfassungskriterien entspricht. Die Attribute sind ebenfalls so zu definieren, daß eine eindeutige Wertzuweisung möglich ist. So muß bei der maximalen Wassertiefe angegeben werden, auf welchen Wasserstand sich die Angabe beziehen soll. Es wird festgelegt wieviele signifikanten Stellen anzugeben sind und was als Seegrund bezeichnet wird. Dies ist besonders bei schlammigen Gründen sehr schwierig, deshalb stößt diese Art der scharfen Modellierung in vielen Fällen an ihre Grenzen. Zusätzlich werden mit den Grunddaten weitere Fachdaten verknüpft. Für den Bereich der Umweltdaten kann dies wie in der gegebenen Situation für das Objekt „See“ das Attribut „Wasserqualität“ sein, dessen Wertebereich im Modell der Fachdaten festgelegt ist.

Bei der Umsetzung dieser abstrakten Wirklichkeit in die Datenbasis sind entsprechend der Abbildung Fehler aufgetreten, die numeriert sind und im einzelnen behandelt werden.

(1) Die Brücke, welche die Straße über den Fluß führt, wurde nicht digitalisiert. Sie fehlt im Datensatz und es gibt auch kein anderes Objekt an dieser Stelle. Fehlerursache ist hierbei entweder unvollständige Digitalisierung und ein Versagen der nachgestellten Kontrolle oder das Problem der unvollständigen Datenabgabe bei Datenbankauszügen.

(2) Der kleine See, der sich im Datensatz befindet, gehört nicht zum abstrakten Abbild der realen Welt, und ist trotzdem digitalisiert worden. Der Grund kann z.B. bei der Nichteinhaltung von Mindestgrößen, in der Fehlinterpretation von Luftbildern oder in Fehlern bei der Aktualisierung der Daten liegen.

(3) Die Straße ist zwar vorhanden, liegt aber geometrisch an der falschen Stelle. Ursache dafür kann entweder eine z.B. durch Verdrängung bei der Generalisierung von kleinmaßstäbigen Karten geometrisch ungenau gewordene Erfassungsquelle sein, ein Fehler beim Einpassen der Erfasssungsquelle oder, daß Daten zusammengeführt wurden, ohne deren unterschiedliche geodätische Bezugssysteme zu berücksichtigen.

(4) Das flächenhafte digitale Objekt „See“ entspricht nicht der Form des abstrakten Objektes. Die Stützpunkte sind in ihrer Lage falsch, und der interpretierte Rand zwischen den Stützpunkten gibt nicht den wahren Verlauf der Uferlinie wieder. Aufgrund ungenauer Erfassungsquellen, nicht präziser Erfassungsmethoden, einer unzureichenden Anzahl von Stützpunkten oder falscher Interpolation ergeben sich die geometrischen Diskrepanzen.

(5) Im oberen Verlauf ist der Fluß einer falschen Objektklasse zugeordnet. Obwohl das Objekt als Linie abstrahiert wurde, ist es teilweise als Fläche erfaßt worden. Damit gehört es zu einer falschen Klasse von geometrischen Primitiven. Die thematische Zuordnung zu einer Objektklasse kann in gleicher Weise falsch sein. Ursache hierfür sind im wesentlichen Fehlinterpretationen bei der Auswertung von primären Erfassungsquellen, Mißklassifizierungen bei der automatischen Verarbeitung von Satellitenbildern oder auch Verwechslungen.

(6) Die Referenz zwischen Straße und Fluß fehlt. Das kann mit dem Fehlen der Brücke zusammenhängen. Je nachdem, ob die Referenzen als eigenständige Objekte modelliert werden, die den vertikalen Zusammenhang zwischen sich kreuzenden Objekten ohne Knotenbildung angeben, oder ob die Referenzen als Attribute der Objekte dargestellt werden, ist dieser Fehler unterschiedlich zu interpretieren. Referenzen stellen Informationen dar, die Beziehungen zwischen Objekten festlegen. Das Fehlen der Referenzen kann durch ein unzureichendes Datenschema, durch ein Versäumnis beim Digitalisieren oder durch unzulängliche Austauschformate verursacht werden.

(7) Der Attributwert für das alphanumerische Attribut „Name“ ist fehlerhaft eingetragen. Freie Texteinträge für Attributwerte sind besonders fehleranfällig. Sie müssen üblicherweise bei der Datenerfassung über die Tastatur eingetippt werden und führen daher häufig zu Buchstabenverwechslungen, Buchstabendrehern, Auslassungen oder Doppeleingaben. Für Recherchen ist die Richtigkeit der Einträge wichtig, denn sonst können die Objekte, bei denen ein Attribut falsch zugewiesen wurde, nicht gefunden werden. Wenn Straßennamen beispielsweise für jedes Straßensegment (von Knotenpunkt zu Knotenpunkt) redundant verwaltet werden, läßt sich die komplette Straße nicht mehr zusammenführen, wenn die Zugehörigkeit eines Segments zu einer Straße über den Namen erfolgt, mindestens ein Wert falsch ist und die Zugehörigkeit nicht anderweitig vom Modell unterstützt wird.

(8) Die Fläche eines Objektes steht in engem Zusammenhang mit dessen Geometrie. Wird der Wert für die Fläche dynamisch über die Analysefunktionen des GIS errechnet und durch das System eingetragen, wirken sich Lagefehler der Stützpunkte über das Fehlerfortpflanzungsgesetz auf die Genauigkeit des Attributwertes aus. Ist der Wert für die Wasserfläche aus anderen Quellen übernommen, weil er beispielsweise so im Liegenschaftskataster eingetragen ist, läßt sich der Grad der Übereinstimmung mit dem wahren Wert nicht aus der Objektgeometrie ableiten. Jedoch kann ein Vergleich der beiden Werte zur Aufdeckung grober Fehler durchgeführt werden. Gibt das System bei der Ausgabe von Attributwerten weniger Dezimalstellen an als das Datenmodell dies

für die Genauigkeit fordert, liegt ein Fehler bei der Umsetzung des Datenmodells auf das Geoinformationsmanagementsystem vor. Dieser Fehler muß unter dem Begriff Modellqualität (4.1) betrachtet werden.

(9) Die Tiefe eines Gewässers ist ein numerischer Wert (reelle Zahl), der bei einer zweidimensionalen Modellierung nicht aus der Objektgeometrie abgeleitet werden kann und somit aus anderen Quellen übernommen werden muß. Grundlage der Bestimmung des Wertes ist eine Messung, die als Zufallsexperiment mit einer bestimmten Verteilungsfunktion aufgefaßt werden kann. In der Fehlertheorie und Stochastik wird dieser Fehlertyp ausführlich und hinreichend behandelt (Literatur z.B. *Menges-Skala, 1973*).

(10) Für den Attributwert der Wasserqualität existiert in den Fachdaten bei diesem See kein Eintrag. Ist dieser Wert für Umweltanalysen unbedingt erforderlich, liegt hier ein Datenfehler vor. Ist dieser Wert hingegen optional, wird er nur eingetragen, wenn er verfügbar ist. Analysen müssen auf den mit Daten belegten, gegebenenfalls repräsentativen Seen durchgeführt werden. Zur Beurteilung der Signifikanz solcher Analysen ist es erforderlich, den Grad der Verfügbarkeit des Attributwertes zu kennen.

(11) Der Schlüssel zur Identifizierung eines Objektes muß eindeutig und richtig sein. Die Zuordnung von Sach- und Fachdaten zu einem geometrischen Objekt erfolgt über diesen Schlüssel. Die Vergabe des Schlüssels wird im Datenmodell über eine Regel vereinbart. Bei der Fortführung von Geodatenbasen ist der Schlüssel von zentraler Bedeutung.

4.2.3 Klassifizierung der Fehler

Auf einer höheren Abstraktionsstufe können die aufgeführten Fehler aufgrund gemeinsamer Kennzeichen in Gruppen eingeteilt werden. Die Mengendiagramme verdeutlichen diesen Zusammenhang. Dazu werden folgende Bezeichnungen eingeführt:

$\boldsymbol{A}$ Menge aller Objekte des abstrakten Abbildes der realen Welt. Diese Menge ist die Zielmenge bei der Erfassung. Alle abstrakten Objekte mit allen ihren Eigenschaften sollen nach allen Regeln in digitale Objekte überführt werden.

$\boldsymbol{a}$ Menge aller digitaler Objekte des Datensatzes. Die Eigenschaften der Objekte werden als Attribute gespeichert.

O Objekt der Menge $\boldsymbol{A}$. Abstraktes Element, das durch das konzeptionelle Modell aus der realen Welt herausgefiltert wird.

o Objekt der Menge $\boldsymbol{a}$. Digitales Geoobjekt.

$\boldsymbol{OK_i}$ Eine bestimmte Objektklasse des abstrakten Abbildes der realen Welt. Kann in diesem Fall auch als Menge aller existierenden, abstrakten Objekte einer bestimmten Klasse aufgefaßt werden.

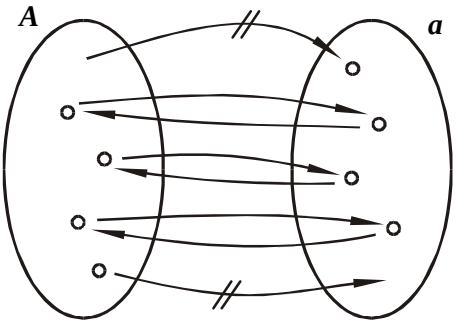
$\boldsymbol{ok_i}$ Abbild der abstrakten Objektklasse im Datenschema.

$O.E_i$ Eine bestimmte Eigenschaft des abstrakten Objektes O.

$O.\boldsymbol{E_i}$ Menge aller möglichen Werte für die i-te Eigenschaft des abstrakten Objektes O.

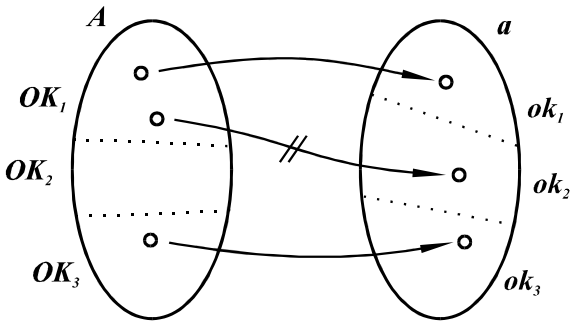
$O.E$ Alle Eigenschaften des abstrakten Objektes O.

$O.\boldsymbol{E}$ Menge aller möglichen Werte für $O.E$.



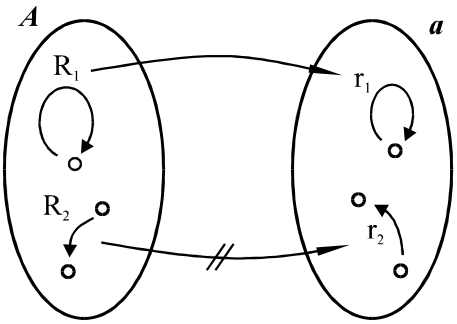
Existenz: Zwischen den Objekten des Abbildes der realen Welt und des Datensatzes muß eine bijektive Abbildung bestehen. Als Frage formuliert: ist jedem Element der Menge ***A*** genau ein Element der Menge ***a*** zugewiesen? Oder, wenn in ***A*** ein Element existiert, ist diesem dann auch genau ein Element in ***a*** zugeordnet, und gibt es für jedes Element in ***a*** auch das Urbild in der Menge ***A***?

Diese Bedingung wurde offensichtlich in der gezeigten Situation für die Punkte (1), (2) und (10) verletzt. Unter dem Aspekt, daß ein Referenzobjekt fehlt, oder daß Referenzattribute nicht vergeben wurden, fällt auch der Punkt (6) in diese Gruppe.



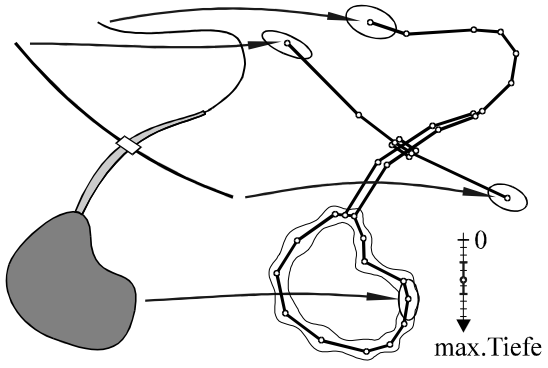
Klassifizierung: Entspricht die Zuordnung zu Objektklassen im Datenbestand der Zuordnung im abstrakten Abbild der realen Welt? Die Umkehrung braucht nicht untersucht zu werden, weil die Bijektivität unter der Existenzbedingung gefordert wurde.

Ein Verstoß gegen die Klassifizierungsbedingung ist unter dem Punkt (5) dargestellt. Wird unter einer Objektklasse die Menge aller möglichen oder sinnvollen Einträge zusammengefaßt, so fällt das Beispiel Punkt (7) unter diese Bedingung.



Einhaltung des Regelwerks: Gelten die Regeln, die auf den Elementen der Menge ***A*** oder zwischen den Elementen dieser Menge vereinbart wurden in gleicher Weise auf oder zwischen den Elementen der Menge ***a***?

Die Referenzen zwischen Objekten werden als Regel im Datenmodell vereinbart. Ein Verstoß gegen diese Regeln, wie unter Punkt (6) gezeigt, führt zur Verletzung dieser Bedingung.



Übereinstimmung von metrischen Größen mit ihren Sollwerten: Stimmen die zugewiesenen Werte aller quantitativen Größen mit ihren wahren Werten hinreichend genau überein? Die quantitativen Größen können dabei sowohl die Koordinaten zur Festlegung der Lage eines digitalen Objektes im Raum (nicht nur die Knoten- und Stütz- sondern auch die interpolierten Punkte) als auch numerische Attributwerte sein. Wie bei allen durch Messungen bestimmten empirischen Größen, können die Werte systematische, stochasische oder grobe Fehler aufweisen.

Unter den Punkten (3), (4), (8) und (9) wurde diese geometrische und attributive Bedingung verletzt.

Diese vier mit Graphiken verdeutlichten Bedingungen führen zu den am Beginn des Kapitels aufgeführten Kriterien. Sie können formal definiert werden und dienen als Grundlage zur Festsetzung von Qualitätsmaßen und Qualitätszielen. Die einzelnen Kriterien können sich situationsbedingt auf einzelne Attributwerte, individuelle Objekte, Objektklassen, Datensätze oder ganze Datenbasen beziehen.

4.3 Verbale und formale Definition der Qualitätskriterien

Die in den Definitionen verwendeten Ausdrücke müssen allgemeingültig sein, d.h. sie müssen die Wahrheitsfunktion identisch „wahr“ repräsentieren. Solche Ausdrücke werden in der mathematischen Logik als Tautologien bezeichnet (*Bronstein et al., 1995*).

4.3.1 Vollständigkeit

Definition: **Vollständigkeit der Objekte** ist dann und nur dann erreicht, wenn jedem abstrakten Objekt genau ein digitales Objekt mit allen Attributwerten und Beziehungen und wenn gleichzeitig jedem digitalen Objekt ein abstraktes Objekt mit allen Attributwerten und Beziehungen zugeordnet ist.

$$\forall O \in \boldsymbol{A}\ O \rightarrow o \ \wedge\ \forall o \in \boldsymbol{a}\ o \rightarrow O$$

Definition: **Vollständigkeit der Attribute** ist genau dann erreicht, wenn für jede Eigenschaft aller abstrakten Objekte den entsprechenden digitalen Objekten genau ein Attributwert zugewiesen wird, und wenn jeder Attributwert eines digitalen Objektes auch einer Eigenschaft des zugehörigen abstrakten Objekts entspricht. Dabei wird die Vollständigkeit der Objekte vorausgesetzt.

$$\forall O \in \boldsymbol{A}\ \forall O.E \in O.\boldsymbol{E}\quad O.E \rightarrow o.e \ \wedge\ \forall o \in \boldsymbol{a}\ \forall o.e \in o.\boldsymbol{e}\quad o.e \rightarrow O.E$$

4.3.2 Richtigkeit

Definition: **Richtigkeit der Objektklassifizierung** gibt an, daß für jedes abstrakte Objekt des Gebietes, das einer abstrakten Objektklasse angehört, das zugehörige digitale Objekt der entsprechenden digitalen Objektklasse angehört.

$$\forall O \in \boldsymbol{A}\ O \in \boldsymbol{OK_i} \rightarrow o \in \boldsymbol{ok_i}$$

Definition: **Richtigkeit der Attributwerte** gibt an, es gilt für alle abstrakten Objekte und für alle ihre Eigenschaften, daß den entsprechenden digitalen Objekten gültige und richtige Attributwerte zugewiesen wurden.

$$\forall O \in \boldsymbol{A}\ \ \forall O.E_i \in O.\boldsymbol{E}_i\ \ O.E_i \in O.\boldsymbol{E}_i \rightarrow o.e_i \in o.\boldsymbol{e}_i \quad \wedge\ \ O.E_i \mathrel{\hat{=}} o.e_i$$

4.3.3 Konsistenz

Definition: **Konsistenz** ist erreicht, wenn alle Objekte des Gebietes den Regeln des Daten- und Informationsmodells entsprechen.

$$\forall O \in \boldsymbol{A}\ \forall o \in \boldsymbol{a}\ \ R(O) = r(o)$$

Die Regeln können sich auf Objekte, Attribute und Beziehungen beziehen. Entsprechend den drei hierarchischen Ebenen der Datenmodellierung (siehe Abschnitt 2.1) gehören diese Regeln zu einer dieser drei Ebenen. Daher wird zwischen

- Konsistenz bezüglich der konzeptionellen Modellierung (konzeptionelle Konsistenz)
- Konsistenz bezüglich der logischen Modellierung (logische Konsistenz)
- Konsistenz bezüglich der physikalischen Modellierung (physikalische Konsistenz)

unterschieden.

4.3.4 Genauigkeit

Der Begriff Genauigkeit bezieht sich auf quantitative Merkmale. Ein quantitatives Merkmal ist ein Merkmal, dessen Werte einer Skala zugeordnet sind, auf der Abstände definiert sind, deshalb wird sie auch metrische oder Kardinalskala genannt (*DIN 55350-12, 1989*). Der Vorgang zur Bestimmung des Skalenwertes ist eine Messung. Als Meßgröße bezeichnet man eine Zufallsgröße, deren Werte durch Messung ermittelt werden.

Es wird vorausgesetzt, daß der wahre Wert für die Meßgröße existiert. Der wahre Wert bezeichnet den tatsächlichen Merkmalswert unter den bei der Ermittlung herrschenden Bedingung.

Definition: **Genauigkeit** als Zielvorgabe ist erreicht, wenn die wahren Werte der zu dem Objekt gehörenden Meßgrößen mindestens mit einer vorgegebenen Wahrscheinlichkeit P = (1-α) innerhalb der um die Erwartungswerte gebildeten Vertrauensbereiche (*VB*) liegen.

$$o.\xi \in VB(X) \text{ mit } P \geq 1 - \alpha$$

X ist dabei eine Zufallsvariable, die zur Meßgröße des Objektes o gehört. ξ ist der zugehörige wahre Wert, der nicht exakt ermittelt werden kann, weil jede Messung ein Zufallsexperiment mit einer gewissen Meßunsicherheit darstellt.

Die Zufallsvariable kann sich dabei auf verschiedene Eigenschaften eines Objektes beziehen

a) auf die Lage im Raum, ausgedrückt in Koordinaten x_i, y_i und bei dreidimensionaler Modellierung auch z_i aller Punkte P_i eines Objektes[4]

b) auf gemessene quantitative Attributwerte $o.e_j$ des Objekts

c) auf Angaben von Zeitpunkten oder Zeitintervallen für das Objekt

4.3.4.1 Meßunsicherheit

Da die Abweichung eines Meßwertes von dem wahren Wert der Meßgröße systematische und stochastische Anteile besitzt, wird als Maß für die Meßunsicherheit die Quadratwurzel aus der mittleren quadratischen Abweichung der Beobachtungsreihe vom wahren Wert (englisch: *root mean squarred error, RMSE*) verwendet. Er bezieht sich auf den wahren Wert ξ einer Meßgröße X und setzt sich aus der zufälligen Abweichung $\varepsilon_{xi} = x_i - \bar{x}$ und der unbekannten systematischen Abweichung $\delta_x = \bar{x} - \xi$ zusammen (*H. Schmidt, 1994 und 1997, DIN 1319-1, 1995*):

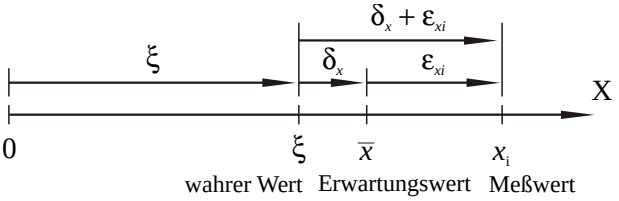


Abbildung 27: Zufällige und systematische Abweichung einer Beobachtung x_i vom wahren Wert der Meßgröße X und dem Erwartungswert $\bar{x}$.

[4] Die Lage des Objekts bezieht sich dabei auf alle Punkte, also sowohl die Stützpunkte als auch die interpolierten Punkte

Die Berechnung des RMSE erfolgt nach folgenden Formeln, in denen $\boldsymbol{E}[\bullet]$ für den Erwartungswert von [•] steht.

$$MSE_{xx} := \boldsymbol{E}[(X-\xi)^2] = \sigma_x^2 + \delta_x^2 \text{ mit } \sigma_x^2 = \boldsymbol{E}[(X-\bar{x})^2] = \boldsymbol{E}[\varepsilon_x^2] \text{ und } \boldsymbol{E}[\varepsilon_x] = 0, \ \boldsymbol{E}[\delta_x] \neq 0$$
$$RMSE_x := \sqrt{MSE_{xx}}$$

Während der zufällige Anteil oft durch Wiederholungsmessungen geschätzt werden kann, geben die Meßwerte keine Information über den systematischen Anteil. Dieser muß in der Regel als Erfahrungswert, gestützt auf eine genaue Analyse des Meßverfahrens, eingeführt werden.

4.3.4.2 Vertrauensbereich

GIS-Anwender ohne wahrscheinlichkeitstheoretische Fachkenntnisse verstehen häufig unter einer Genauigkeitsangabe, daß der wahre Wert maximal um diesen Betrag von dem gespeicherten Wert entfernt liegen darf. Da aber die Erfassung von numerischen Größen für ein GIS meistens durch eine Messung erfolgt, und dieser Meßprozeß als Zufallsexperiment betrachtet werden muß, kann immer nur ein Bereich angegeben werden, innerhalb dessen der wahre Wert mit einer bestimmten Wahrscheinlichkeit liegt. Dieser Bereich wird in der Meßtechnik durch die untere Fehlergrenze (G_u) und obere Fehlergrenze (G_o) (*DIN 1319-1, 1995*) festgelegt. In der Statistik wird dieser Bereich Vertrauensbereich oder Konfidenzintervall genannt. Die Wahrscheinlichkeit, daß der unbekannte wahre Wert zwischen den Konfidenzgrenzen liegt, wird als Sicherheitswahrscheinlichkeit (Vertrauensniveau, Konfidenzzahl) 1-α bezeichnet (*Kreyszig, 1973*).

Zur Berechnung der Konfidenzgrenzen (G_u, G_o) aus gegebenen Stichprobenwerten muß die Verteilung der als Zufallsvariablen verstandenen Meßgröße bekannt sein. Nach dem zentralen Grenzwertsatz der Statistik nähert sich die Verteilung einer Summe von unabhängig verteilten Zufallsgrößen, die dieselbe Verteilung besitzen, mit wachsender Zahl der Summanden einer Normalverteilung. (Satz von Lindeberg-Feller, siehe *Bauer, 1978*). Da bei einer Messung sehr viele unbekannte, unabhängige und gleichartige Einflüsse additiv auf die Ablesung einwirken, können aus gutem Grund die Meßgrößen als normalverteilt aufgefaßt werden.

Ist die Meßunsicherheit für eine Meßgröße zum Beispiel aus Herstellerangaben oder durch theoretische Überlegungen (z.B. Unsicherheit der Datenquelle plus Unsicherheit der Einpassung der Datenquelle in das übergeordnete Koordinatenystem plus Unsicherheit bei der geometrischen Modellierung plus Unsicherheit bei der Digitalisierung) bekannt, können die Konfidenzgrenzen näherungsweise aus der Normalverteilungsfunktion ermittelt werden.

$$\mathrm{P}\left(\bar{x} - u_{1-\alpha/2} \cdot RMSE_x \le \xi \le \bar{x} + u_{1-\alpha/2} \cdot RMSE_x\right) = 1-\alpha$$
$$\mathrm{P}\left(G_u(X) \le \xi \le G_o(X)\right) = 1-\alpha$$

Wegen der zweiseitigen Betrachtung muß für die Berechnung des Quantils $u_{1-\alpha/2}$ der standardisierten Normalverteilung die Wahrscheinlichkeit $1-\alpha/2$ herangezogen werden.

Übliche Signifikanzniveaus und die zugehörigen zweiseitigen Quantile sind in der folgenden Tabelle zusammengestellt.

Signifikanzniveau 1-α	90.0%	95.0%	99.0%	99.9%
Quantil $u_{1-\alpha/2}$	1,645	1,960	2,576	3,290

Für die Bestimmung der Lagekoordinaten eines Objektes bedeutet dies, wenn in maximal 5% aller Fälle die wahre Lage außerhalb einer Toleranz von +/-3m liegen darf, muß die Meßunsicherheit einen Wert von unter 3m/1,96 = 1,53m betragen. Diese Meßunsicherheit beinhaltet sowohl unbekannte systematische als auch stochastische Beiträge.

Ist die Meßunsicherheit für die zu erfassende Meßgröße unbekannt, so muß sie aus der gegebenen Stichprobe abgeleitet werden. Das Problem dabei ist, daß aus Wiederholungsmessungen nur die empirische Standardabweichung, also der stochastische Teil der Meßunsicherheit, geschätzt wird. Bei

geeigneter Meßanordnung kann oft der größere Teil des systematischen Einfluß eliminiert werden, der verbleibende unbekannte systematische Einfluß verhält sich wie eine stochastische Größe. Bei einer Ausgleichung wirkt eine Systematik in den Beobachtungen auf die Residuen und somit auch auf das stochastische a posteriori Modell. Die Auswirkung ist von der Meßanordnung und vom Auswertemodell abhängig. Der Schätzwert für die Meßunsicherheit wird umso zuverlässiger, je höher die Redundanz bei der Bestimmung der Meßgröße ist. Zur Bei der Angabe der Grenzen des Konfidenzintervals werden die Residuen mit der geschätzten Standardabweichung normiert. Aus diesem Grund müssen zur Berechnung der Konfidenzgrenzen die Quantile der Student-Verteilung (t-Verteilung) mit Freiheitsgrad f verwendet werden.

$$P(\bar{x} - t_{\alpha/2,f} \cdot rmse_x \le \xi \le \bar{x} + t_{\alpha/2,f} \cdot rmse_x) = 1 - \alpha$$

Die Redundanzen bei der Erfassung von Geoobjekten sind üblicherweise sehr gering. Bei der Georeferenzierung von Rasterdaten oder der Einpassung von analogen Datenquellen am Digitalisiertisch werden oft nicht mehr als vier Passpunkte verwendet. Je nach Transformationsgleichungen bedeutet dies einen Freiheitsgrad von f = 4 bis f = 0. Die Digitalisierung der Stützpunkte bei Geoobjekten wird im Normalfall ohne Redundanz durchgeführt. Aus diesem Grund besteht auch keine interne Kontrolle zum Aufdecken von groben Fehlern. Eine Überprüfung der Erfassungsgenauigkeit muß daher in einem separaten Verfahrensschritt durchgeführt werden (Kapitel 5.1).

Die Quantile der Student-Verteilung sind für kleine Redundanzen in der folgenden Tabelle aufgelistet.

1-α	90,0%	95,0%	99,0%	99,9%
$t_{\alpha/2,f}$ f = 10	1,81	2,23	3,17	4,59
$t_{\alpha/2,f}$ f = 5	2,02	2,57	4,03	6,87
$t_{\alpha/2,f}$ f = 4	2,13	2,78	4,60	8,61
$t_{\alpha/2,f}$ f = 3	2,35	3,18	5,84	12,92
$t_{\alpha/2,f}$ f = 2	2,92	4,30	9,92	31,60
$t_{\alpha/2,f}$ f = 1	6,31	12,71	63,66	636,58

Wenn also die wahre Lage eines Objektes in maximal 5% aller Fälle außerhalb des vorgegebenen Toleranzbereiches von +/-3m sein darf, muß die mit einer Redundanz von 3 bestimmte empirische Meßunsicherheit einen Wert von unter 3m/3,18 = 94cm betragen. Wird ein Signifikanzniveau von 99% angesetzt, verringert sich der Wert auf 3m/5,84 = 51cm. Um möglichst zuverlässige Aussagen zu machen, ist es erforderlich, die Bestimmung der Meßunsicherheit mit einer hohen Redundanz und einer günstigen Geometrie für die überbestimmte Berechnung von Transformationsparametern durchzuführen. Bei der Schätzung der Genauigkeit von Attributwerten und Zeitangaben sind vergleichbare Überlegungen anzustellen.

4.4 Verletzung der Qualitätskriterien

Die formale Beschreibung der Qualitätskriterien kann dazu verwendet werden, herauszufinden, wann die einzelnen Kriterien verletzt sind. Dies ist dann der Fall, wenn die Ausdrücke nicht wahr, also falsch sind, und damit ihre Negationen wieder eine Tautologie bilden.

Mit den Gesetzen der Prädikatenlogik lassen sich die Verneinungen der einzelnen Ausdrücke soweit umformen, bis sie zu Aussagen führen, die wieder interpretiert werden. Die daraus resultierenden Aussagen geben nur an, daß ein Qualitätskriterium verletzt wurde, aber sie machen keine Aussage darüber, wo oder für welches Objekt oder Attribut dieser Verstoß vorliegt.

4.4.1 Vollständigkeit

Wann eine Grundgesamtheit unvollständig ist, kann aus der Negation der Definitionsaussage zur Vollständigkeit abgeleitet werden. Bezogen auf die Vollständigkeit der Objekte lautet die Negation

$$\neg(\forall O \in \boldsymbol{A}\ O \rightarrow o\ \wedge\ \forall o \in \boldsymbol{a}\ o \rightarrow O).$$

Durch Anwendung der de Morganschen Regel läßt sich die Negation der Konjunktion auflösen. Dies führt zu der äquivalenten Aussage

$$\neg(\forall O \in \boldsymbol{A}\ O \rightarrow o)\ \vee\ \neg(\forall o \in \boldsymbol{a}\ o \rightarrow O).$$

Durch Anwendung der in der Prädikatenlogik geltenden Tautologie zur Verneinung von Ausdrücken mit Quantoren ergibt sich ein Ausdruck, der interpretiert werden kann

$$\exists O \in \boldsymbol{A}\ \neg(O \rightarrow o)\ \vee\ \exists o \in \boldsymbol{a}\ \neg(o \rightarrow O).$$

In Worten bedeutet diese Aussage, eine Grundgesamtheit ist genau dann unvollständig, wenn es ein abstraktes Objekt gibt, dem kein digitales Objekt zugeordnet ist, oder wenn mindestens ein digitales Objekt existiert, das keine Entsprechung im abstrakten Abbild der realen Welt hat. Einfacher ausgedrückt heißt dies, die Vollständigkeit ist verletzt, wenn die Objekte im Datensatz nicht mit den Objekten des abstrakten Abbildes der realen Welt übereinstimmen. Mit anderen Worten, es gibt mindestens ein Objekt, das entweder zuviel oder zuwenig erfaßt wurde.

4.4.2 Richtigkeit

Zur Ermittlung der Fälle, daß in der Menge der zu bewertenden Objekte die Richtigkeitsbedingung verletzt ist, wird die Definitionsaussage negiert.

$$\neg(\forall O \in \boldsymbol{A}\ (O \in \boldsymbol{OK}_i \rightarrow o \in \boldsymbol{ok}_i\ ))$$

$$\Updownarrow$$

$$\exists O \in \boldsymbol{A}\ \neg(O \in \boldsymbol{OK}_i \rightarrow o \in \boldsymbol{ok}_i\ )$$

Wenn die Implikation im hinteren Teil des Termes nach der Regel $A \rightarrow B = \neg A \vee B$ aufgelöst wird, ergibt sich

$$\exists O \in \boldsymbol{A}\ \neg(\neg O \in \boldsymbol{OK}_i \vee o \in \boldsymbol{ok}_i\ )$$

$$\Updownarrow \qquad .$$

$$\exists O \in \boldsymbol{A}\ O \in \boldsymbol{OK}_i \wedge o \notin \boldsymbol{ok}_i$$

Das heißt, es gibt mindestens ein abstraktes Objekt, das einer anderen Klasse angehört als das zugehörige digitale Objekt.

4.4.3 Konsistenz

In welchen Fällen liegen Inkonsistenzen vor?

$$\neg(\forall O \in \boldsymbol{A}\ \forall o \in \boldsymbol{a}\ R(O) = r(o))$$

$$\Updownarrow$$

$$\exists O \in \boldsymbol{A}\ \exists o \in \boldsymbol{a}\ R(O) \neq r(o)$$

Es liegen Inkonsistenzen vor, wenn es mindestens ein abstraktes und ein digitales Objekt gibt, deren Regeln sich widersprechen.

4.4.4 Genauigkeit

Die Verletzung des Genauigkeitskriteriums ist äquivalent mit der Aussage, daß die Wahrscheinlichkeit des Ereignisses „der wahre Wert einer Zufallsvariablen liegt zwischen einer vorgegebenen unteren und oberen Schranke“ größer als das erwartete Signifikanzniveau ist.

$$o.\xi \in VB(X) \text{ mit } P \geq 1-\alpha$$
$$\Updownarrow$$
$$P(G_u(X) \leq o.\xi \leq G_o(X)) \geq 1-\alpha$$

Die Negation dieser Ungleichung kann auf zwei Arten interpretiert werden.

$$\neg\left(P(G_u(X) \leq o.\xi \leq G_o(X)) \geq 1-\alpha\right)$$
$$\Updownarrow$$
$$P(G_u(X) \leq o.\xi \leq G_o(X)) < 1-\alpha$$
$$\Updownarrow$$
$$P(o.\xi < G_u(X) \vee o.\xi > G_o(X)) \geq \alpha$$

mit $G_o - \bar{x} = -G_u + \bar{x} = q \cdot RMSE$ und $q \ldots$ Quantil der jeweiligen Verteilung gilt dann die Beziehung

$$RMSE > RMSE_0 .$$

Die erste Möglichkeit zur Interpretation dieses Zusammenhangs bezieht sich auf das Genauigkeitsniveau aller Objekte bzw. aller Meßgrößen der Objekte. Wenn die untersuchten Zufallsvariablen zur selben Grundgesamtheit gehören, muß bei allen Objekten die Meßunsicherheit über dem vorgegebenen Schwellenwert liegen. D.h. die Objekte haben insgesamt ein zu niedrigeres Genauigkeitsniveau.

Die zweite Möglichkeit bezieht sich darauf, daß mehr Objekte den Schwellenwert übersteigen als nach der Wahrscheinlichkeit zulässig ist, weil sie nicht zur selben Grundgesamtheit gehören. Mit anderen Worten die Objekte des Gebietes haben kein einheitliches Genauigkeitsniveau. Die Inhomogenität kann durch einzelne Objekte mit groben Fehlern verursacht sein oder durch Erfassung auf Basis von Erfassungsquellen mit unterschiedlicher Genauigkeit oder anderer Erfassungsverfahren.

4.5 Qualitätsmaße

Zur Bewertung von Daten anhand der aufgestellten Qualitätskriterien müssen Maße eingeführt werden, die Aussagen über die Beschaffenheit der Daten quantifizieren. Dabei ist der Gültigkeitsbereich dieser Maße für ihre Festlegung sehr wichtig. Mögliche Bezugsgrößen sind das individuelle Objekt, der einzelne Attributwert von einem Objekt und eine Aggregation von räumlichen, zeitlichen oder thematischen Einheiten.

4.5.1 Fehlermaße für individuelle Objekte oder Attribute

Die ausführlichste Form der Dokumentation von Datenqualität liegt vor, wenn zu jedem Objekt und zu jedem Attribut ein Qualitätsmaß angegeben wird. Diese Daten können als zusätzliche Attribute mit dem Objekt gespeichert werden. Diese Option muß im Datenschema vorgesehen werden. Die Konsequenz ist ein erhebliches Anwachsen des Datenvolumens. Bei einer Angabe von Meß-unsicherheiten für Koordinatenwerte steigt die angegebene Zahl von Qualitätsmaßen linear mit der Anzahl von Stützpunkten von linien- oder flächenhaften Objekten. Sollen auch Kovarianzen angegeben werden, so wächst das Datenvolumen quadratisch.

Nicht nur das Datenvolumen steigt bei einer individuellen Qualitätserhebung, sondern auch der Aufwand, der mit der Bestimmung von Qualitätsmaßen für jedes einzelne Datum eines Geo-informationssystems verbunden ist. Wenn Qualitätsmaße unerfaßt bleiben oder mit voreingestellten Werten belegt werden, ist der Sinn einer individuellen Angabe von Qualitätsmaßen auf Objektebene fragwürdig. Wird jeder Wert für die Qualitätsmaße mit zusätzlichem Aufwand bestimmt und eingetragen, so ist das auch mit erheblichen Zusatzkosten bei der Datenerfassung verbunden. Die Anwendung muß diesen Aufwand rechtfertigen.

Eine andere Form der Angabe von Qualitätsmaßen auf Objektebene sind Fehlerlisten, in denen die Objekte, welche als fehlerhaft eingestuft wurden, über eindeutige Identifikatoren angesprochen werden. Als Identifikatoren kommen externe Objektnummern oder interne Objektnummern in Betracht, wenn

gewährleistet ist, daß sie sich während der Lebenszeit der Objekte nicht ändern. Eine andere Möglichkeit zur Identifikation von Objekten ist über die Koordinaten und die Objektklasse gegeben. Diese Variante ist aber nicht immer eindeutig. Wenn ein Objekt z.B. aus Versehen zweimal mit exakt identischer Geometrie erfaßt wurde, und eines dieser Objekte hat korrekte Attributeinträge und das andere nicht, dann ist die Beschreibung des Datenfehlers nach der zweiten Variante nicht ausreichend.

Zur Bereinigung von Datenfehlern müssen Objekte individuell angesprochen werden können. Wenn also Qualitätsprüfung und Qualitätsverbesserung durch Fehlerbeseitigung in zwei Arbeitsschritten durchgeführt werden, ist ein eindeutiger Schlüssel erforderlich.

4.5.1.1 Vollständigkeit, Richtigkeit und Konsistenz

Wenn als Bezugsgröße für die Qualitätsangabe ein Objekt herangezogen wird, dann sind die Maße für die Qualitätskriterien Vollständigkeit, Richtigkeit und Konsistenz vom Typ Boolesch. Ein Objekt kann entweder vorhanden sein oder fehlen, es kann der richtigen Objektklasse zugewiesen sein oder nicht und es kann konsistent bezüglich aller festgelegten Regeln sein oder inkonsistent.

4.5.1.2 Genauigkeit

Für die Genauigkeitsangabe von Koordinaten (Punktfehler) sind in der Geodäsie viele Qualitätsmaße entwickelt worden (*Schmidt, 1994*). Die Beschreibung der Genauigkeit von linien- und flächenhaften Objekten ist in den Arbeiten von *Shi, 1994, Scheuring, 1995*, und *Bethge, 1997*, ausführlich diskutiert. Als Fehlermaße werden Fehlerbänder, sogenannte ε-Bänder, die Fläche zwischen ursprünglicher und digitalisierter Kurve, Richtungs-, Längen- und Krümmungsfehler vorgeschlagen. Diese Maße stellen eine Aggregierung von Genauigkeitsmaßen der Punkte entlang des Objektes dar und können teilweise aus den Genauigkeitsangaben der Koordinaten abgeleitet werden.

Ein Qualitätsmaß in Anlehnung an die Maße für die Kriterien Vollständigkeit, Richtigkeit und Konsistenz mit Booleschem Typ kann durch Einführung eines Grenzwertes für die geometrische Unsicherheit eines Objektes eingeführt werden. Eine Verletzung der Genauigkeitsforderung liegt vor, wenn Teile des Objektes aus dem durch die Vertrauensgrenzen festgelegtenen Pufferbereich um das abstrakte Objekt herausragen. Als Qualitätsmaß wird angegeben, ob der Grenzwert überschritten ist oder nicht. In diesem Sinne kann von einem fehlerhaften Objekt bezüglich des Qualitätskriteriums Genauigkeit gesprochen werden.

Genauigkeitsmaße für Attribute sind vom Typ der Attributwerte abhängig. Durch Einführung von Schwellenwerten lassen sich aber alle Genauigkeitsangaben in Boolesche Variable überführen. Die Schwellwerte müssen natürlich als wichtige Metadaten zur Qualität der Daten offengelegt werden. Sie sind von der Erfassungsmethode abhängig.

4.5.2 Ein Gebiet als Bezugsgröße für Fehlermaße

Wenn nicht der einzelne Fehler oder das einzelne fehlerhafte Objekt oder Attribut von Interesse ist, sondern die Bewertung eines Datensatzes angegeben werden soll, der sich immer auf ein begrenztes Gebiet bezieht, müssen Fehlermaße, die sich auf ein Gebiet beziehen eingeführt werden. Diese Qualitätsmaße haben globalen Charakter. Sie werden als relative **Zuverlässigkeitswerte** angegeben und so konzipiert, daß ein Wert von 100% eine absolute obere Schranke darstellt, und jeder Fehler zu einer Absenkung des Wertes führt. Damit bilden sie ein Komplement zu der **Fehlerrate**, die mit dem Anteil von fehlerhaften Objekten ansteigt. Die Fehlerrate gibt die relative Anzahl von fehlerhaften Objekte bezogen auf die Gesamtzahl der abstrakten Objekte eines Gebietes an.

Die Angabe hat einen statistischen Charakter und kann zum direkten Vergleich der Qualität von Datensätzen herangezogen werden. Außerdem können sie in einem Qualitätsmanagementsystem als Grenzwerte für die Akzeptanz eines Datensatzes verwendet werden.

Da die Fehlerrate von Objektklasse zu Objektklasse unterschiedlich ausfallen kann, und je nach Fehlerursache auch unterschiedlich sein wird, ist es oft notwendig, Angaben zu Fehlerraten der einzelnen Objektklassen oder zu einzelnen Attributen zu machen. Pauschalaussagen über alle

Objektklassen eines Datenbestandes werden den Unterschieden hinsichtlich der Erfassung und der Wichtigkeit nicht gerecht. Eine zu starke Aggregation von Qualitätsmaßen kann die Entscheidung über die Verwendbarkeit eines Datensatzes für bestimmte Projekte verhindern oder zu falschen Entscheidungen führen insbesondere, wenn die Fehlerraten bezogen auf Objekte bestimmter Objektklassen große Schwankungen zeigen.

Um einen Vergleich der Qualität von Datensätzen zu ermöglichen, ist es erforderlich, die Fehlermöglichkeiten zu sinnvollen Einheiten zusammenzufassen. Dazu wird der Satz von unabhängigen Qualitätskriterien aus Kapitel 4.3 verwendet und für jedes Kriterium ein globales Qualitätsmaß eingeführt bezogen auf ein Gebiet bzw. den gesamten Datensatz.

Beispielsweise wurde für den Datenbestand GEODATA TOPO-250K (der topographische Grunddatenbestand für Australien bezogen auf den Maßstab 1:250.000) von der zuständigen Institution, der Australian Surveying and Land Information Group (AUSLIG), eine Validierung durchgeführt. Nach einer Phase der Datenanalyse wurden Grenzwerte für Fehlerraten festgesetzt. Für jede erdenkliche Fehlermöglichkeit wurde ein separates Maß eingeführt. Die Akzeptanzgrenze der einzelnen Fehlerarten lag zwischen 0.5% und 5% (*Lawford, 1995*).

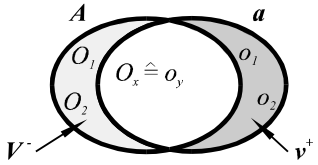
4.5.2.1 Vollständigkeit

Die Menge aller Objekte, die dem Vollständigkeitskriterium widersprechen, wird mit $\boldsymbol{V}$ bezeichnet. Sie stellt die Vereinigung der Menge aller Objekte *O* der abstrakten Realität dar, die keine Entsprechung in einem Objekt *o* des Datensatzes haben, mit der Menge aller Objekte *o*, die im Datensatz vorhanden sind, aber keine Entsprechung in der abstrakten Wirklichkeit haben. $\boldsymbol{V}$ enthält dadurch alle Objekte, die fehlen oder überflüssig sind.

Die zuviel erfaßten Objekte können zu einer Menge $\boldsymbol{v}^+$ und die zuwenig erfaßten Objekte entsprechend zu einer Menge $\boldsymbol{V}^-$ zusammengefaßt werden.

Alle Objekte *O* eines Gebietes *G* bilden eine Menge, die mit $\boldsymbol{A}$ bezeichnet wird. Alle Objekte *o* eines Datensatzes bilden eine Menge, die im folgenden mit $\boldsymbol{a}$ bezeichnet wird. Dieser Datensatz soll entsprechend dem Datenmodell genau das Gebiet *G* abdecken.

Der Zusammenhang zwischen den Mengen der erfaßten und zu erfassenden Objekte mit den Mengen der bezüglich des Vollständigkeitskriterium fehlerhaften Objekte läßt sich formal über die folgenden Beziehungen herstellen.



$$\boldsymbol{V}^- = \boldsymbol{A} \setminus \boldsymbol{A} \cap \boldsymbol{a} \qquad \boldsymbol{V} = \boldsymbol{V}^- \cup \boldsymbol{v}^+$$
$$\boldsymbol{v}^+ = \boldsymbol{a} \setminus \boldsymbol{A} \cap \boldsymbol{a}$$

Der Quotient aus der Anzahl der falsch erfaßten Objekte zu den zu erfassenden Objekten gibt eine relative Fehlerhäufigkeit an. Als Bezugsgröße wird die Anzahl der Objekte der abstrakten Wirklichkeit verwendet. Diese ist allerdings im allgemeinen nicht bekannt. Die Anzahl der erfaßten Objekte läßt sich im Geoinformationssystem ermitteln und so kann über die bekannten Größen die Bezugsgröße abgeleitet werden.

$$|\boldsymbol{A}| = |\boldsymbol{a}| + |\boldsymbol{V}^-| - |\boldsymbol{v}^+|$$

Rein intuitiv wird man eine Vollständigkeit von 100% als komplett erfaßtes Gebiet bezeichnen. Daher wurde das Maß für die Vollständigkeit so festgelegt, daß sich bei einer Mächtigkeit der Mengen der zuviel und zuwenig erfaßten Objekte von null ein Vollständigkeitsmaß (VM) von eins bzw. 100% ergibt.

$$\mathrm{VM}_D := \frac{|\boldsymbol{A}| - |\boldsymbol{V}|}{|\boldsymbol{A}|} = 1 - \frac{|\boldsymbol{V}|}{|\boldsymbol{a}| + |\boldsymbol{V}^-| - |\boldsymbol{v}^+|} = 1 - \frac{|\boldsymbol{V}^-| + |\boldsymbol{v}^+|}{|\boldsymbol{a}| + |\boldsymbol{V}^-| - |\boldsymbol{v}^+|} = \frac{1 - \frac{|\boldsymbol{v}^+|}{|\boldsymbol{a}| - |\boldsymbol{v}^+|}}{1 + \frac{|\boldsymbol{V}^-|}{|\boldsymbol{a}| - |\boldsymbol{v}^+|}}$$

Wertebereich für das Vollständigkeitsmaß bezogen auf ein Gebiet VM_D

Der pathologische Fall, daß kein Objekt, das erfaßt wurde, eine Entsprechung in der abstrakten Wirklichkeit hat, wird für die weiteren Betrachtungen ausgeschlossen, weil damit die Differenzen in beiden Nennern des Doppelbruchs zu null würden und ein unbestimmter Ausdruck entstünde. Die Menge der zuviel erfaßten Objekte ist immer eine echte Teilmenge von ***a***. Dadurch ist die Differenz $|\boldsymbol{a}|-|\boldsymbol{v}^+|$ immer eine positive, ganze Zahl und der Nenner des übergeordneten Bruches ist größer oder gleich eins. Die Differenz im Zähler des übergeordneten Bruches kann negativ und damit auch das Vollständigkeitsmaß kleiner als null werden.

Für die Interpretation dieses Qualitätsmaßes ist interessant zu untersuchen, wann es Werte annimmt, die kleiner als null sind.

$$1 - \frac{|\boldsymbol{v}^+|}{|\boldsymbol{a}| - |\boldsymbol{v}^+|} \le 0 \Rightarrow |\boldsymbol{a}| > |\boldsymbol{v}^+| \ge \tfrac{1}{2}|\boldsymbol{a}|$$

Das Maß für die Vollständigkeit eines Datensatzes bezogen auf ein bestimmtes Gebiet nach der gegebenen Definition nimmt negative Werte an, wenn mehr als die Hälfte aller erfaßten Objekte zuviel, d.h. nicht in der abstrakten Wirklichkeit, definiert durch das Datenschema, enthalten sind. Es können aber nicht mehr Objekte zuviel sein als insgesamt Objekte erfaßt worden sind.

Durch Betrachtung des ersten Termes in der Definitionsgleichung für das Vollständigkeitsmaß ergibt sich eine weitere Nullstelle, die der intuitiven Anschauung entspricht. Wenn keines der zu erfassenden Objekte in den Datenbestand übernommen wurde, so ist das Maß für die Vollständigkeit null.

VM kann nur Werte annehmen die kleiner als oder gleich eins sind. Damit sind die geforderten 100% eine obere Schranke und es gibt keine Konstellation, bei der Werte höher als 100% angenommen werden können.

Für verschiedene Anwendungen erscheint es sinnvoll, zwischen den zuviel und den zuwenig erfaßten Objekten zu unterscheiden, und für diese beiden Betrachtungen unterschiedliche Vollständigkeitsmaße anzugeben. Diese beiden Fälle können als Spezialfall für die angegebene Definition des Vollständigkeitsmaßes betrachtet werden, indem zum einen die Anzahl der überflüssigen Objekte und zum anderen die Anzahl der vergessenen Objekte zu null gesetzt wird. Zur Unterscheidung werden die beiden Begriffe Übervollständigkeitsmaß (ÜVM) und Untervollständigkeitsmaß (UVM) eingeführt.

1. Untervollständigkeitsmaß (UVM): $|\boldsymbol{v}^+| = 0 \Rightarrow \mathrm{UVM}_D = \dfrac{1}{1 + \frac{|\boldsymbol{V}^-|}{|\boldsymbol{a}|}}$

2. Übervollständigkeitsmaß (ÜVM): $|\boldsymbol{V}^-| = 0 \Rightarrow \mathrm{ÜVM}_D = \dfrac{|\boldsymbol{a}| - 2|\boldsymbol{v}^+|}{|\boldsymbol{a}| - |\boldsymbol{v}^+|}$

Abbildung 28 zeigt die Graphen der Vollständigkeitsfunktion für den Fall, daß, bezogen auf den untersuchten Datensatz , 0 bis 50% der Objekte fehlen oder überzählig sind. Man erkennt, daß zwischen 0% und 10% die Kurven nahezu gleich verlaufen, ab dann aber das Maß für die Übervollständigkeit sehr schnell gegen null geht. Die erste Ableitung beider Funktionen besitzt im Nullpunkt einen Wert von -1. D.h. die Abnahme der Vollständigkeit verläuft anfänglich proportional zu

diesem Quotienten, danach ist die Untervollständigkeit flacher und die Übervollständigkeit steiler. Die Abweichungen von einem linearen Verlauf resultieren aus der Wahl der Bezugsgröße für das Verhältnis, nämlich daß immer die geschätzte Anzahl der zu erfassenden Objekte herangezogen wird.

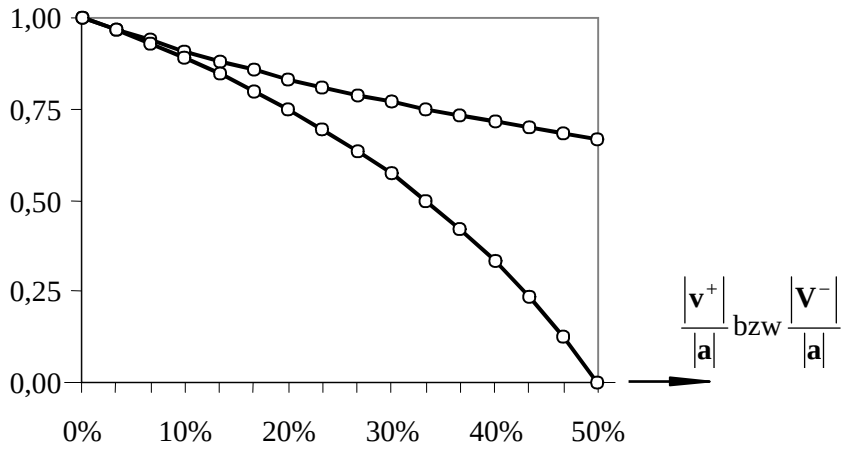


Abbildung 28: Diagramm zur Darstellung der Qualitätsmaße für Über- und Untervollzähligkeit in Abhängigkeit von der Fehlerrate.

4.5.2.2 Richtigkeit (Objekt und Attribut)

Wird ein Objekt der realen Welt erfaßt, so entsteht ein digitales Objekt in der Datenbasis. Beiden gemeinsam ist der geometrische Ort bezogen auf ein einheitliches Bezugssystem. Die Kategorisierung des Objekts wird durch die Zuordnung zu einer Objektklasse erreicht. Wenn die Objektklasse, zu welcher das Objekt in der abstrakten Welt gehört, und die Objektklasse des Datenschemas übereinstimmen, so ist das Objekt richtig erfaßt. Wenn sich die beiden Objektklassen unterscheiden, so ist das Objekt falsch erfaßt worden. Abhängig davon, ob die tatsächliche Objektklasse zum Datenschema zählt, wird der Fehler dem Qualitätskriterium Richtigkeit oder Vollständigkeit zugeordnet.

Wenn das fälschlicherweise erfaßte Objekt einer Klasse angehört, die nicht im Datenschema enthalten ist, so ist das digitale Objekt überzählig und muß dem Maß zur Bestimmung der Übervollständigkeit zugerechnet werden. Liegt aber eine Verwechslung der Objektklasse innerhalb des Datenschemas vor, so ist das Objekt bei der einen Klasse zuviel und bei der anderen zuwenig. Trotzdem ist das Objekt vorhanden und fällt daher nicht unter das Vollständigkeitskriterium.

Die einer falschen Klasse zugeordneten Objekte werden in einer Menge zusammengefaßt. Diese Menge wird mit ***r*** bezeichnet. Ein Maß für die Richtigkeit der Zuordnung zu Objektklassen eines Gebietes wird durch die Anzahl der falschen Kategorisierungen festgelegt, welche zu der Anzahl der Objekte im Datensatz in das Verhältnis gesetzt wird. Mit den Mengenbegriffen, die im Abschnitt zum Vollständigkeitsmaß eingeführt wurden, kann der Zusammenhang formal dargestellt werden. Pro falsche Zuordnung wird die Anzahl der Elemente in den Mengen zuviel und zuwenig erfaßter Objekte um eins erhöht. Tritt diese Erhöhung paarweise auf, dann ist das Objekt nicht der Menge ***V***, sondern der Menge ***r*** zuzuordnen.

$$\left(\left|\boldsymbol{V}_{OAx}^{-}\right|+1 \wedge \left|\boldsymbol{v}_{OAy}^{+}\right|+1\right) \Rightarrow |\boldsymbol{r}|+1$$

Mit dieser Festlegung kann ein neues Maß für die Richtigkeit von Objekten eines Datensatzes eingeführt werden. Das Maß wird mit RM bezeichnet.

$$\mathrm{RM}_D = 1 - \frac{|\boldsymbol{r}|}{|\boldsymbol{a}|}$$

Wertebereich für das Maß der Richtigkeit bezogen auf ein Gebiet RM_D

Die maximale Anzahl von Objekten, die nicht richtig zugeordnet wurden entspricht der Anzahl von Objekten im Datensatz. Dadurch ergibt sich für das Qualitätsmaß RM ein Wertebereich zwischen 0 und 1. Also 0%, wenn alle Objekte der falschen Klasse zugeordnet sind und 100%, wenn alle Klassen der Objekte des Datensatzes den Klassen der realen Welt entsprechen.

4.5.2.3 Konsistenz

Konsistenz ist ein Qualitätskriterium, das sich typischerweise auf ein Objekt oder eine Konstellation von mehreren Objekten bezieht, die genau lokalisiert und benannt werden können. Zu statistischen Zwecken ist es aber auch sinnvoll, die Konsistenz bezogen auf einen Datensatz oder ein Gebiet anzugeben, z.B. bei der Festlegung, wieviele Objekte, die den Konsistenzbedingungen widersprechen, ein Datensatz enthalten darf, damit er für eine bestimmte Anwendung noch herangezogen werden kann. Die Angabe kann entweder absolut oder relativ zu den vorhandenen Objekten in Prozent erfolgen. Die Abweichung von 100% der relativen Anzahl von Konsistenzverletzungen wird in Analogie zum Vollständigkeitsmaß mit KM abgekürzt.

Die Menge der Konsistenzverletzungen wird mit $\boldsymbol{k}$ bezeichnet. Das kleine k steht für das Qualitätskriterium Konsistenz, und der Kleinbuchstabe wurde gewählt, weil Inkonsistenzen nur bei den erfaßten Daten vorkommen können. Die Definitionsgleichung für das Maß der Konsistenz eines Gebietes lautet

$$\mathrm{KM}_D = 1 - \frac{|\boldsymbol{k}|}{|\boldsymbol{a}|}.$$

Wertebereich für das Maß der Konsistenz bezogen auf ein Gebiet KM_D

Die Anzahl der Inkonsistenzen in einem Datensatz kann die Anzahl der Objekte überschreiten, wenn jedes Objekt mehreren Konsistenzbedingungen widersprechen kann. Das bedeutet für die Mächtigkeit der Menge $\boldsymbol{k}$, daß sie größer als die der Menge $\boldsymbol{a}$ sein kann. Damit wird KM im Extremfall auch negative Werte annehmen. Werte größer als 100% können nicht angenommen werden. Der Verlauf der Funktion mit zunehmender Anzahl von Inkonsistenzen ist linear.

4.5.2.4 Genauigkeit

Nach der Definition für das Qualitätskriterium Genauigkeit (4.3.4) gibt es zwei Arten, die Genauigkeit der Objekte für ein bestimmtes Gebiet anzugeben.

Möglichkeit 1 besteht darin, für die interessierenden Meßgrößen (Koordinaten, Attribute oder Zeit) einen Mittelwert der Meßunsicherheit aus allen Einzelwerten zu ermitteln. Damit reduziert sich die Angabe für die Genauigkeit aller Objekte auf einen einzelnen Wert. Dieser kann dann mit einem Sollwert verglichen werden. Nach dieser Methode bleiben allerdings Schwankungen der Meßunsicherheit unberücksichtigt. Wenn die Genauigkeit im Interessengebiet nicht konstant ist, kann durch eine Berechnung der mittleren Abweichung von diesem Mittelwert ein Maß für die Inhomogenität eines Gebietes ermittelt werden. Wenn also der Mittelwert für die Meßunsicherheit nur knapp unter ihrem Sollwert liegt und die Schwankung der Meßunsicherheit innerhalb eines Gebietes sehr hoch ist, kann davon ausgegangen werden, daß es Objekte gibt, die den gestellten Anforderungen an die Genauigkeit der Geometrie, Attribute oder Zeit nicht entsprechen. Allerdings hat man nach dieser Methode eine quantitative Angabe über die Genauigkeit der Objekte in diesem Gebiet. Wenn die Meßunsicherheiten um Größenordnungen unter dem geforderten Wert liegen, so kann ein potentieller Anwender auch entscheiden, diese Daten für Analysen mit höheren Anforderungen an die Genauigkeit zu verwenden.

Möglichkeit 2 besteht darin, Analogien zu globalen Fehlermaßen der Kriterien Vollständigkeit, Richtigkeit und Konsistenz zu bilden. Diese Möglichkeit wird durch die Art der Definition in Abschnitt 4.3.4 von Genauigkeit nahegelegt, da die Genauigkeit all jene Objekte besitzen, deren Meßunsicherheit eine vorgegebene Schranke unterschreitet. Alle Objekte, bei denen die Meßunsicherheit der Koordi-

naten, der Attribute oder der Zeitangabe größer als ein Schwellwert ist, werden zu einer Menge $\boldsymbol{g}$ zusammengefaßt. Die Mächtigkeit dieser Menge gibt die Anzahl der ungenauen Objekte an. Mit dieser Angabe kann ein globales Fehlermaß gebildet werden, das analog zu den anderen Fehlermaßen aufgebaut ist:

$$\mathrm{GM_D} = 1 - \frac{|\boldsymbol{g}|}{|\boldsymbol{a}|}.$$

Diese Prozentangabe ist zwar für Genauigkeitsangaben in der Geodäsie atypisch, bringt aber bei Sensitivitätsanalysen gewisse Vorteile. Zum Beispiel wenn keine Fehlerfortpflanzung gerechnet wird, sondern eine Abschätzung erfolgt, was passiert, wenn der angegebene Prozentsatz die Vorgabe nicht erfüllt. Muß eine Fehlerfortpflanzung gerechnet werden, so steht der vorgegebene Sollwert zur Verfügung. In Abhängigkeit von der globalen Einhaltung der maximalen Meßunsicherheit ist das Ergebnis aus dieser Berechnung eher zu konservativ. Dies bedeutet, bei einem hohen Prozentsatz ist das Risiko gering, zu optimistische Angaben über die Genauigkeit von abgeleiteten Informationen zu machen.

Um einen guten Eindruck von der Genauigkeit der Objekte eines Gebietes zu erhalten, können die Angaben aus beiden vorgestellten Möglichkeiten kombiniert werden. Bei beiden Methoden wird eine mögliche, für digitalisierte Punkte wahrscheinlich sogar hohe Korrelation der Meßwerte vernachlässigt.

Geometrische Genauigkeit

Durch Zählen der Objekte, die teilweise außerhalb eines Puffers liegen, der durch das Vertrauensgebiet um die Sollage des Objektes gebildet wird, ergibt sich die Mächtigkeit der Menge $\boldsymbol{g}$.

Beispiel

Im ATKIS Objektartenkatalog Teil A wird eine Aussage über die anzustrebende Genauigkeit gemacht. Die Angabe lautet, daß die wesentlichen linearen Objekte des DLM 25 eine dort unpräzise eingeführte Modellgenauigkeit von $\pm 3\,\mathrm{m}$ aufweisen müssen (*AdV, 1995*). Die wesentlichen, linienhaft zu erfassenden Objektklassen werden mit Straßen, Fahrbahnen, Schienenbahnen, Bahnstrecken und Gewässer, topologische Knoten im Netz der Straßen und Schienenbahnen sowie deren niveaugleiche Kreuzungen benannt. Da die Modellgenauigkeit nicht weiter spezifiziert und weder in der Geodäsie noch in der Statistik ein eingeführter Begriff ist, gibt diese Angabe nur einen vagen Anhalt über die geometrische Genauigkeit von ATKIS-Daten.

Attributive Genauigkeit

Die Attribute eines Objektes oder gleichlautende Attribute verschiedener Objektklassen können unterschiedlich definiert sein, verschiedene Einheiten haben und auf verschiedenen Wegen erfaßt worden sein. Ein solcher Datenbestand ist aus der Sicht einer konsistenten Modellierung nicht sinnvoll, kann aber durch Zusammenführung verschiedener Datensätze entstehen. Eine globale Angabe für die Genauigkeit aller Attribute ist dabei nicht sinnvoll und sollte thematisch eingeschränkt werden.

4.5.3 Fehlerdichte

Als weitere Bezugsgröße neben der Anzahl von Objekten innerhalb des Bezugsgebietes kann die Fläche des Bezugsgebietes angegeben werden. Da der Quotient aus Anzahl von fehlerhaften Objekten zu der Fläche des Gebietes eine Aussage darüber macht, wie dicht die Fehler im Durchschnitt beieinander liegen, kann dieses Qualitätsmaß als Fehlerdichte bezeichnet werden.

Da die Fehlerdichte von der Dichte der Objekte innerhalb des Gebietes abhängt, ist dieses Qualitätsmaß von der Charakteristik des Gebietes abhängig. Für einige Anwendungen mag die Angabe über Fehler pro Quadratkilometer, oder auch artverwandte Angaben wie Fehler pro Kilometer Leitungslänge von Interesse sein.

4.6 Speicherung von Qualitätsdaten und Metadaten

Je nach Bezug der Qualitätsmaße und nach Verwendung der Metadaten für die Anwendung sind unterschiedliche Verknüpfungen zwischen den Daten und ihren Metadaten erforderlich. Die in Kapitel 3 eingeführten Gruppen von Metadatenelementen können verschiedenen Ebenen zugeordnet werden, auf die sich die Elemente beziehen können. In der folgenden Tabelle ist eine Auflistung der möglichen Bezugsgrößen angegeben.

	gesamte Datenbasis	räumliche Relevanz	zeitliche Relevanz	Objektklassen	Attributarten	Geometrie
Modell	●	○	○	○	○	○
Herkunft	●	●	●	●	●	●
Qualität	●	●	●	●	●	●
Verfügbarkeit	●	●	●	●	●	●
Referenz-anwendungen	●	●	●	●	●	●

● Bezug ist möglich

○ Bezug ist bedingt möglich

Das Modell bezieht sich im allgemeinen auf die gesamte Datenbasis, es ist aber auch möglich, Datenschemata zu erzeugen, die auf mehreren Modellen aufsetzen. Einfachstes Beispiel ist hierfür die hybride Geodatenverarbeitung. Dabei werden Daten aus verschiedenen logischen Datenmodellen zusammengeführt. Es ist aber auch möglich und in vielen Anwendungen sinnvoll, Daten unterschiedlicher konzeptioneller Modelle in einem Schema zusammenzuführen, so daß sie gemeinsam analysiert werden können. Ein typisches Beispiel hierfür ist die Zusammenführung von ATKIS- und GDF-Daten (*Walter und Fritsch, 1998*). Da GDF-Daten hauptsächlich im urbanen Bereich verfügbar sind, können sie im ländlichen Bereich durch Objekte der Objektklasse Straße aus ATKIS ergänzt werden. Dadurch kommt auch ein räumlicher Bezug zustande. Durch die Möglichkeiten einer derartigen Interoperabilität von Geoinformationssystemen lassen sich sehr komplexe Verknüpfungen zwischen Geodaten erzeugen. Über die Metadaten muß nachvollziehbar sein, auf welches Modell sich die Daten beziehen.

Da für alle Elemente der Tabelle Beispiele und Anwendungsfälle aufgezeigt werden können, ist ein Speicherungskonzept erforderlich, das allen unterschiedlichen Bezugsmöglichkeiten gerecht wird. Metadaten können entweder in das Datenschema integriert werden, so daß die Verwaltung zusammen mit den Geodaten erfolgt, oder sie werden in einem getrennten Informationssystem geführt. Eine Integration kann entweder über Relationen der Metadaten zu den jeweiligen Bezugsgrößen erfolgen, oder durch deren räumlichen Bezug. Die verschiedenen Konzepte werden in den folgenden Abschnitten vorgestellt und diskutiert.

4.6.1 Verwaltung der Metadaten getrennt von dem Datenbestand

Die Metadateninformationssysteme, die heutzutage eingesetzt werden, werden üblicherweise separat von den Daten geführt. In Aktenordnern oder als kleine Datenbank werden Informationen über den Stand der Erfassung festgehalten. Die Bezugsgröße ist dabei üblicherweise eine Erfassungseinheit, also ein bestimmtes Erfassungsgebiet. Die Trennung macht die Pflege dieser Daten schwierig, und sie ist nur durch die Disziplin von einzelnen Verantwortlichen gewährleistet. Eine Recherche beim Umgang mit den Daten, oder eine Einbeziehung in Analysen ist nicht möglich. Auch eine Abgabe an Nutzer der Daten ist im allgemeinen nicht vorgesehen und erfolgt oft nur bei Nachfrage.

Wenn zwischen externen und internen Metadaten keine Trennung besteht (vgl. 3.6) ist eine Abgabe dieser Informationen kritisch, da interne Metadaten auch persönliche Daten, z.B. über die erfassende Person, enthalten können, und dadurch unter den Datenschutz fallen.

Bei der Separation von Geodaten und ihren Metadaten ist der Bezug immer auf ein Gebiet, das durch eindeutige Identifikatoren angesprochen werden kann. Als Identifikatoren werden üblicherweise Blattnummern verwendet, weil die Erfassungseinheiten sich meist auf Blattschnitte von analogen Erfassungsquellen beziehen. Der Bezug auf einzelne Objekte ist durch diese Art der Metadatenverwaltung nur möglich, wenn auf die eindeutigen Objektschlüssel verwiesen wird. Dann ist auch eine Verknüpfung zwischen Objekten und ihren Objektattributen in der Datenbank möglich, und es kann nicht mehr von einer getrennten Verwaltung gesprochen werden.

4.6.2 Objektbezogene Metadatenhaltung

Wenn zu jedem Objekt Qualitätsmaße und weitere Metadaten angegeben werden, führt dies zu einer erheblichen Aufblähung des Datenvolumens. Außerdem sind die Ermittlung und Eintragung dieser Informationen, wie in Abschnitt 4.5.1 diskutiert, mit einem zusätzlichen Erfassungsaufwand verbunden.

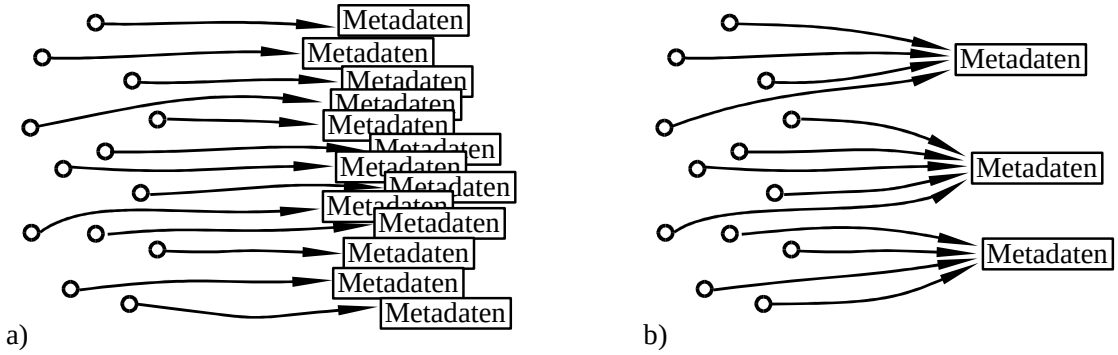


Abbildung 29: Relationen zwischen Objekten und Metadaten a) bezogen auf individuelle Objekte und b) bei Aggregation von gleichartigen Objekten.

Wenn die Metadaten auf Objektebene erforderlich sind, dann kann die Speicherung entweder durch Erweiterung der Liste von Objektattributen erfolgen oder durch Aufbau einer weiteren Tabelle der Metadaten, die über die Objektschlüssel eins zu eins mit den jeweiligen Objekten verknüpft werden. Eine Verallgemeinerung dieser Verwaltungsmethode ist möglich, wenn Objekte mit gleichen Metainformationen mit demselben Eintrag in der Matadatentabelle verknüpft werden (Abbildung 29).

4.6.3 Berücksichtigung des Raumbezuges von Metadaten

Gleiche Metadaten bei unterschiedlichen Objekten ergeben sich dadurch, daß gleichartige Objekte nach derselben Erfassungsmethode auf Basis derselben Erfassungsquelle und von derselben Person oder Institution erfaßt wurden. Als Konsequenz daraus kann eine Verwaltungsmethode der Metadaten abgeleitet werden, bei der die Qualitätsmaße immer in Bezug auf das Gebiet abgespeichert werden, in dem sie gültig sind. Die Gebiete, in denen gleichlautende Metadaten zu erwarten sind, ergeben sich durch Verschneidung aller Gebiete, die von den jeweiligen Datenquellen abgedeckt werden, und der Erfassungseinheiten. Die aus dieser Verschneidung resultierenden kleinsten Flächen, sind die Einheiten, in denen homogene Metadaten vorhanden sind. Ein Speichern der Metadaten an diesen flächenhaften Objekten ergibt die Variante der Metadatenverwaltung mit dem kleinsten Datenvolumen und dem geringsten Erfassungsaufwand.

Durch Einbeziehung von räumlichen Operatoren bei der Durchführung von beliebigen Analysen können die Metadaten und insbesondere die Qualitätsangaben berücksichtigt werden. Da keine direkte Verknüpfung zwischen den interessierenden Objekten und den Metadaten besteht, muß diese Beziehung im Bedarfsfall über räumliche Operatoren hergestellt werden. Eine entsprechende Abfrage

kann so formuliert werden: „Gebe die Qualitätsattribute des Metadatenobjektes aus, in dem das interessierende Objekt ganz oder teilweise enthalten ist.“

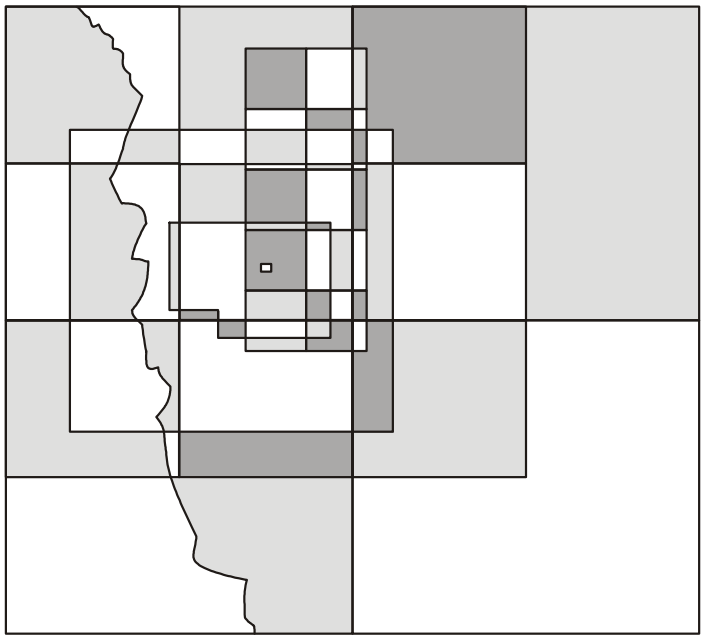


Abbildung 30: Die Verschneidung aller Gebiete der Datenquellen und Erfassungseinheiten führt zu den Einheiten, die homogene Geodaten bezüglich ihrer Metadaten beinhalten. Die Gebiete entsprechen dem Beispiel auf Seite 28.

Die übersichtliche, wartungsfreundliche und weniger speicherintensive Verwaltung der Metadaten über homogene Gebiete hat den Nachteil, daß Abfragen über räumliche Operatoren längere Antwortzeiten benötigen als relationale Verknüpfungen, und daß die Handhabung bei Anwendungen komplexer wird.

5 Qualitätsmanagement bei der Datenerfassung

5.1 Methoden zur Einhaltung von Qualitätszielen

Jeder Erfassungsvorgang, also Erstdigitalisierung und Fortführung von Geodaten, muß darauf ausgerichtet sein, die Qualitätsziele zu erreichen. Damit das nicht dem Zufall überlassen bleibt, bedarf es Methoden, die den Prozeß der Erfassung bezüglich der eingeführten Kriterien überwachen.

Ein System, bei dem der Unternehmensleitung die Verantwortung und aktive Mitgestaltung für die Erfüllung der Qualitätsziele übertragen, jeder Mitarbeiter in den Qualitätsprozeß eingebunden und durch Prozeßlenkung die Erfüllung gesichert wird, bezeichnet *Wittig, 1993,* als Qualitätsmanagement-system. Die ISO-Normenreihe 9000 bietet eine Grundlage für die wesentlichen Qualitätselemente zur Erarbeitung eines Qualitätsmanagement-Handbuches für Produktions- und Dienstleistungs-unternehmen[5]. Die Produktion von Geodaten kann sowohl als Produktionsprozess als auch als Dienstleistung verstanden werden.

Das Qualitätsmanagement bedient sich durchgreifender Prüfverfahren, damit eine Abweichung von den Qualitätszielen aufgedeckt werden kann. Die Prüfungen der digitalen Daten können entweder automatisch nach vorgegebenen Regeln oder durch visuelle Vergleiche erfolgen. Die Prüfungen erstrecken sich entweder auf den gesamten Datenbestand oder auf Stichproben, mit denen auf den gesamten Datenbestand geschlossen werden kann. Regelmäßige Kontrollen und Schulungen sollen zur kontinuierlichen Verbesserung der Qualität von Geodaten führen. Die Qualitätserfüllung schließt eine Kosten- und Terminerfüllung mit ein.

Das Qualitätsmanagement von Geodaten ist sehr stark von der Erfassungsmethode und von den Qualitätszielen abhängig. Trotzdem lassen sich spezifische Elemente des Qualitätsmanagements für Geodaten aufstellen.

5.2 Qualitätsziele

Durch Einführung von Schwellwerten für Qualitätsmaße können Qualitätsziele für die Akzeptanz von Datensätzen festgelegt werden. Die Qualitätsziele sind damit Vorgaben für die bei der Erfassung mindestens zu erreichenden Qualitätsmaße. Aufgabe des Qualitätsmanagement ist es, dafür zu sorgen, daß keine Daten produziert werden, die diesen Vorgaben nicht entsprechen. Für die statistische Qualitätskontrolle dienen Qualitätsziele als annehmbare Qualitätsgrenzlage (siehe Abschnitt 7.7.2).

Die Qualitätsziele sind von den Daten unabhängig. Sie können und müssen vor der Erfassung von Geodaten festgelegt werden. Die Anwender von Geoinformationssystemen sollten überlegen, welche Qualitätsmaße für ihre Anwendungen erforderlich sind. Dadurch können die Qualitätsziele nach Absprache zwischen Datenproduzent und Anwender vereinbart werden.

---

[5] Nach ISO/DIS 8402 kann „Produkt oder Dienstleistung“ folgendes bedeuten:

- Das Ergebnis von Tätigkeiten oder Prozessen (ein materielles Produkt, ein immaterielles Produkt wie eine Dienstleistung, ein Computerprogramm, ein Design, eine Gebrauchsanweisung), oder
- Eine Tätigkeit oder einen Prozess (wie das Erbringen einer Dienstleistung oder die Ausführung eines Produktionsprozesses).

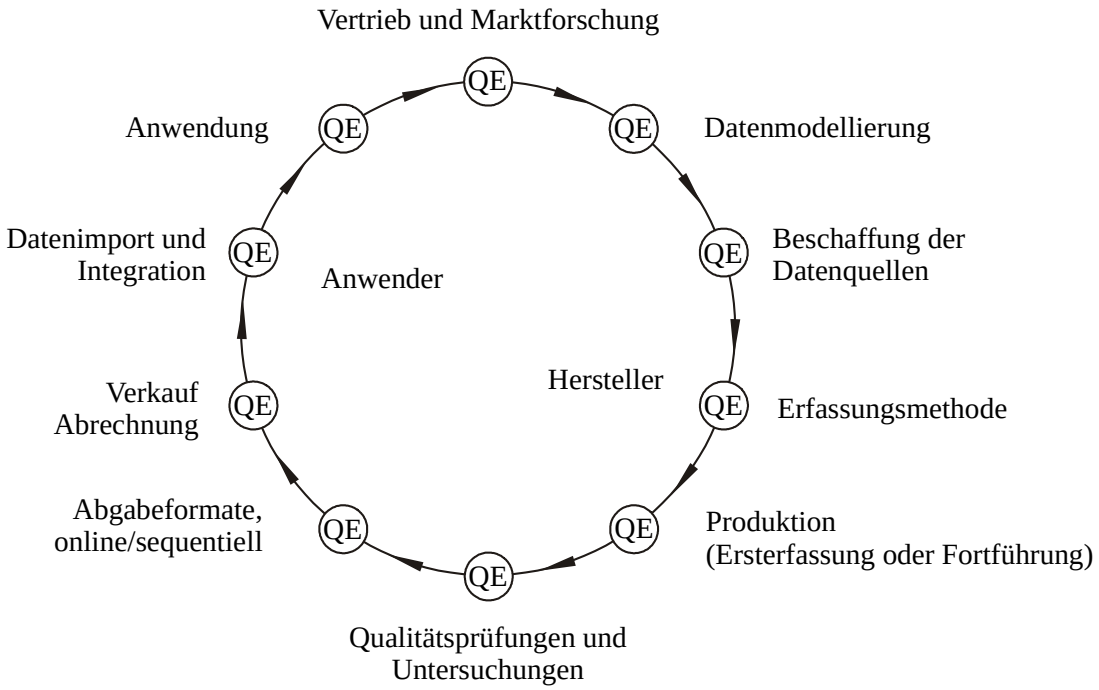


Abbildung 31: Die Qualitätselemente bilden einen geschlossenen Qualitätskreis

Ein Qualitätsmanagementsystem bezieht sich auf alle Qualitätselemente. Diese sind die Hauptaktivitäten, die im Qualitätskreis einer Institution nacheinander angeordnet sind (siehe Abbildung 31). Zwischen den einzelnen Elementen muß es Rückkopplungen geben. Bei der Erfassung von Geodaten sind die Schlüsselelemente Vertrieb und Marktforschung, weil da die Kundenforderungen und Produktmerkmale festgelegt werden, und die Qualitätsprüfungen, weil da sichergestellt wird, daß die Daten den Produktspezifikationen und damit den Kundenwünschen entsprechen.

Im folgenden wird ein Bündel aus drei Methoden zur Einhaltung von Qualitätszielen vorgestellt.

- Methode 1 ist das Qualitätsmanagement. Sie stellt ein organisatorisches Werkzeug dar, das vor allem die Leitung einer Institution anspricht und Richtlinien zur Führung von Mitarbeitern gibt, indem Arbeitsabläufe vollständig dokumentiert und deren Einhaltung ständig kontrolliert wird.
- Methode 2 ist die Konsistenzprüfung. Sie kann nur auf ein bestimmtes Qualitätskriterium, nämlich die Konsistenz angewandt werden, ist dort aber besonders effektiv, weil alle Objekte eines Datenbestandes automatisch geprüft werden können. Die Konsistenzprüfung stellt eine softwaretechnische Maßnahme zur Einhaltung der Qualitätsziele dar.
- Als Methode 3 wird die statistische Qualitätskontrolle eingeführt. Unter bestimmten Annahmen kann von einer Stichprobe auf die Qualität der gesamten Einheit geschlossen werden. Dieses statistische Werkzeug ist vor allem geeignet zur Abschätzung, ob die Daten den Anforderungen genügen, wenn eine 100%-Prüfung aus Zeit- oder Kostengründen nicht durchgeführt werden kann.

5.3 Qualitätsmanagement

Ein Qualitätsmanagementsystem besteht im wesentlichen aus Strukturen, Verantwortlichkeiten, Prozeduren, Prozessen und Ressourcen. Die ISO 9000-Standards zur Normierung von Qualitätsmanagement und Qualitätssicherung sind sehr allgemein gehaltene Richtlinien. Sie geben an, **was** zu tun ist, nicht **wie** es getan werden muß. Die 9000er Qualitätsnormen sind folglich Verhaltens- und Prozeßnormen, jedoch keine Normierungen der Qualität für Produkte oder Dienstleistungen. Europäische und internationale Normen zur Beschreibung der Qualität von Geodaten werden zur Zeit entwickelt (*prENV 12656, 1998, ISO 19113 CD, 1999, ISO 19114 CD, 1999*).

In den folgenden Abschnitten werden die Richtlinien der ISO 9000 auf das Verfahren zur Erfassung von Geodaten übertragen. Die Gliederung orientiert sich an den Forderungen an das Qualitäts-

sicherungssystem (*DIN ISO 9000-2, 1992*). Die Darlegung eines Qualitätssicherungssystems besteht dort aus 20 QS-Elementen. Ihre Anwendung auf die Erfassung von Geodaten wird in den folgenden Abschnitten 5.3.1 bis 5.3.20 diskutiert.

5.3.1 Verantwortung der obersten Leitung

Die oberste Leitung einer Institution, die Daten zum Einsatz in Geoinformationssystemen erfaßt, soll ihre Verpflichtung zur Qualität, die Qualitätszielsetzungen und die Qualitätspolitik festlegen und in einer leicht verständlichen Sprache dokumentieren. Die Qualitätspolitik soll zu der Organisation, den Erfassungsmethoden und zu den betroffenen Menschen passen. Die Zielsetzungen sollten anspruchsvoll aber erreichbar sein.

Durch ihr Handeln soll die oberste Leitung ihre Verpflichtung zur Qualität demonstrieren, indem sie sicherstellt, daß das Personal die Qualitätspolitik versteht und in der täglichen Arbeit umsetzt, und indem sie Abweichungen nicht akzeptiert. Jeder in der Organisation sollte sich für das Erreichen der Qualitätsziele verantwortlich fühlen.

Die Institution sollte so organisiert sein, daß sich alle Mitarbeiter des Umfangs, der Verantwortung und der Befugnis ihrer Tätigkeit und ihres Einflusses auf die Qualität der Daten bewußt sind. Bei einer manuellen Erfassung erzeugt letztendlich der Mitarbeiter am Digitizer oder am Bildschirm die Qualität der Daten. Nachgeschaltete Schritte können immer nur Abweichungen vom Soll feststellen und punktuell nacharbeiten. Eine Selbstkontrolle der Mitarbeiter bei der Erfassung ist daher ständig erforderlich. In Zweifelsfällen müssen sie entscheidungsbefugte Ansprechpartner haben. Daten dürfen niemals ohne Prüfung nach außen gegeben werden. Die Entscheidung über die Freigabe fällt zwar in der Regel der Prüfer bei der Endkontrolle, doch die Verantwortung trägt die oberste Leitung einer Institution. Aus diesem Grund müssen die Zuständigkeiten klar definiert und bekannt gegeben werden.

Die Leitung muß für das Erlangen der Qualitätsziele angemessene Ressourcen bereitstellen. Diese Ressourcen umfassen

- Personal, das die Verifizierung ausführt
- Kenntnis vorhandener Normen
- Schulung
- Ausreichende Zeit zur Durchführung der Arbeit
- Produktionspläne, die Zeit für Tätigkeiten wie Prüfungen und Verifizierungen einräumen
- Prüfmittel (z.B. Prüfsoftware, siehe Abschnitt 6.2)
- Verfahrensanweisungen
- Zugang zu Qualitätsaufzeichnungen
- Ein Umfeld, das einen Geist von Objektivität und Zusammenarbeit all derer fördert, die sich mit der Verifizierung befassen.

Bei der Ausschreibung von Digitalisieraufträgen muß eine entsprechende Qualitätsklausel integriert werden. Bei der Kalkulation von Projekten muß das Aufbringen der qualitätsbezogenen Ressourcen einberechnet werden.

5.3.2 Qualitätssicherungssystem (QS-System)

„Das QS-System wird üblicherweise mit Hilfe eines Qualitätssicherungs-Handbuches dokumentiert. Das QS-Handbuch kann ein Dokument sein, das von mehreren Dokumentensätzen unterstützt wird, wobei jeder nachfolgende Satz zunehmend detaillierter wird. So kann es z.B. ein Gesamtsystem-Handbuch und ein oder mehrere spezielle Verfahrens-Handbücher geben. Diese Dokumente legen gemeinsam das vollständige QS-System fest. (*DIN ISO 9000-2, 1992*)“

Sämtliche die Qualität von Daten beeinflussenden Vorgänge müssen in die Erfassungsverfahren integriert, in einem QS-Handbuch beschrieben und auf alle Daten angewandt werden. Das Qualitätsmodell mit seinen Qualitätskriterien (Abschnitt 4.3) und den Qualitätszielen (Abschnitt 5.2)

gehören genauso zum QS-Handbuch wie die Prüfverfahren (Abschnitte 6 und 7) und Qualitätsaufzeichnungen. Die Prüfverfahren sind für die unterschiedlichen Methoden detailliert zu dokumentieren.

5.3.3 Vertragsüberprüfung

Die Aufgabenstellung muß so eindeutig formuliert sein, daß sich Auftraggeber und Auftragnehmer vor Abschluß eines Leistungsvertrages zur Erfassung und Fortführung von Geodaten über die Anforderungen vollständig im klaren sind. Dabei müssen Angebot und Auftrag nicht nur formal übereinstimmen, sondern beide Vertragspartner müssen auch inhaltlich dieselben Vorstellungen über die Beschaffenheit der Daten besitzen.

Der Auftragnehmer muß prüfen, ob er die Aufgabenstellung nach sachlichen, organisatorischen und kapazitiven Kriterien erfüllen kann. Außerdem muß er prüfen, ob die Qualitätsforderungen angemessen festgelegt sind.

5.3.4 Designlenkung

Ein fehlerhaftes Design kann eine Hauptursache für Qualitätsprobleme sein. Wenn das Datenmodell unvollständig, inkonsistent (widersprüchlich) oder mehrdeutig ist (siehe Abschnitt 4.1: Modellqualität), treten bei der Datenerfassung Fälle auf, die nicht oder nur fehlerhaft umgesetzt werden können. Da ein Erfasser versuchen wird, die Phänomene, die er in der Erfassungsgrundlage erkennt, dem Datenmodell anzupassen, kommt es zu Fehlinterpretationen. Objektklassen, die im konzeptionellen Datenschema nicht berücksichtigt aber für eine Anwendung zwingend erforderlich sind, müssen mit großem Mehraufwand nachträglich erfaßt werden. Die damit verbundenen Anpassungen von Datenschema, Prüf- und Plotroutinen, Abgabeschnittstellen und Absprachen mit den Kunden lassen die damit verbundenen Kosten unverhältnismäßig steigen.

Bei der Entwicklung von neuen Produkten steigen die Qualitätsabweichungskosten von der Produktidee (Marketing) bis zur Marktreife (Kundendienst) überproportional an. Bei jedem Schritt sind für die gleiche Abweichung, die beseitigt werden muß, Kostenerhöhungen um das zehnfache zu erwarten (Abbildung 32 und Abbildung 33).

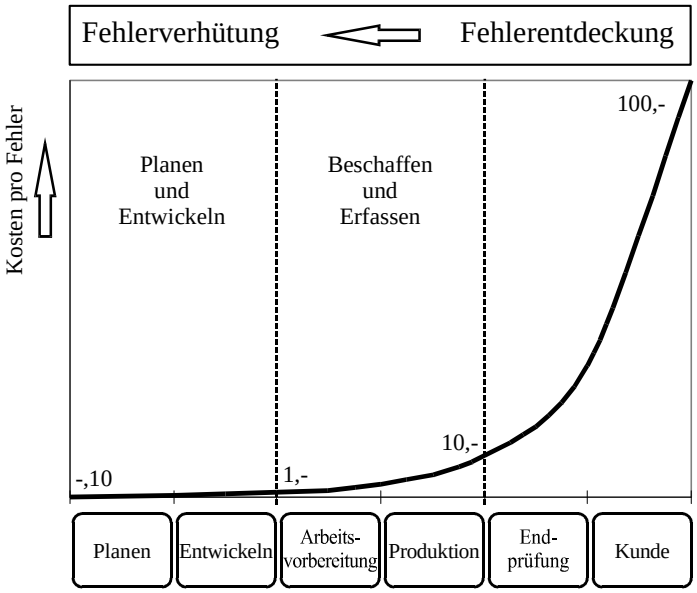


Abbildung 32: Zehnerregel der Qualitätsabweichungskosten (*Wittig, 1993*).

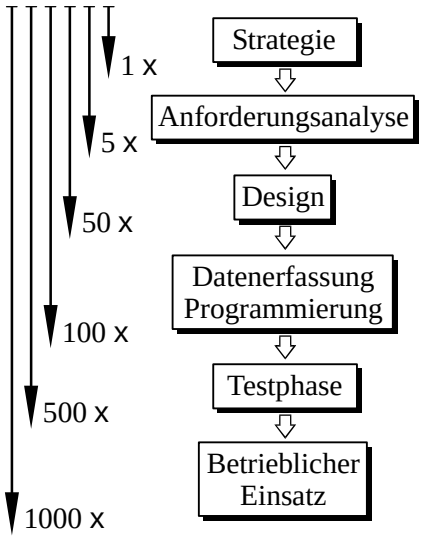


Abbildung 33: Anfallende Kosten für die Behebung von Fehlern in den einzelnen Phasen (*Klemmer und Spranz, 1997*).

Die hohen Kosten bei einer späten Aufdeckung von Fehlern zeigen das hohe Risikopotential durch mangelnde Planung und Vorbereitung. Aus diesem Grund ist die Designlenkung auch für die Qualität von Geodaten ein kritisches Qualitätselement.

5.3.4.1 Design- und Entwicklungsplanung

Design und Entwicklungsplanung von Geodaten erstreckt sich über alle Phasen der Produktdefinition. Dies sollte ein dynamischer Vorgang sein, bei dem alle Zwischenschritte und –ergebnisse zwischen dem Bedarfsträger und der Daten erfassenden Institution abgeglichen und kontinuierlich verbessert werden.

Als Ergebnis der Design- und Entwicklungsplanung entsteht ein Projektplan. Die Frage: „Wer soll was wie auf Basis welcher Grundlagen und mit welchen Qualitätszielen erfassen?“ soll dabei erschöpfend beantwortet werden. Zeitliche Vorgaben und Abhängigkeiten müssen in die Planung einfließen.

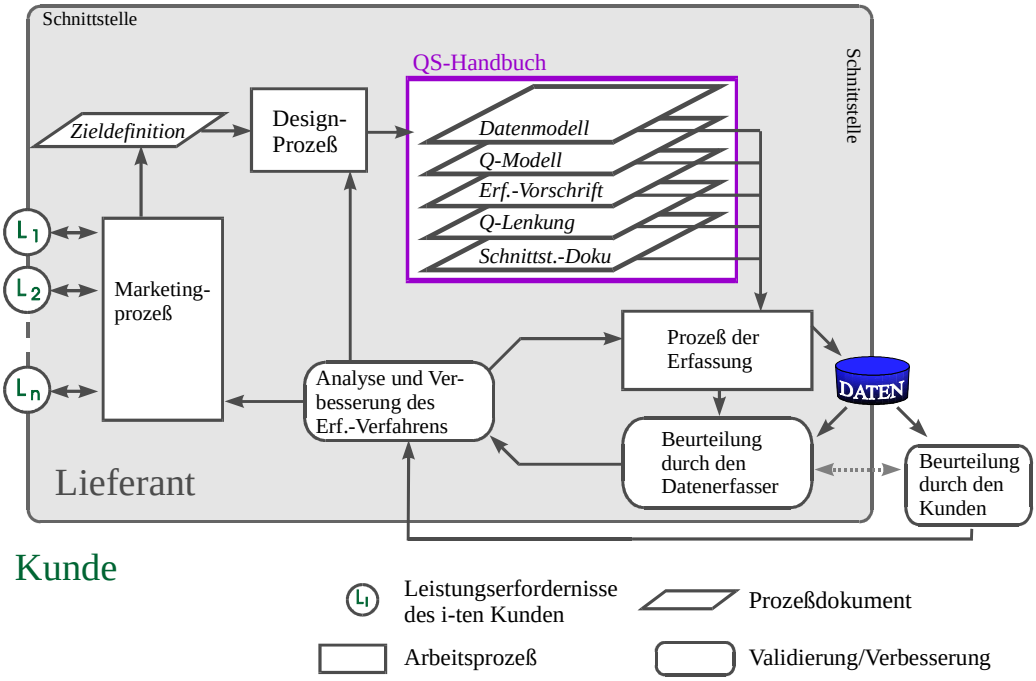


Abbildung 34: Qualitätslenkung bei der Datenerfassung als Dienstleistung (*Caspary, 1993*).

5.3.4.2 Designvorgaben

Designvorgaben haben typischerweise die Form von Spezifikationen für die Produktausführung. In ihnen wird das Zielprodukt mit allen Eigenschaften und Abhängigkeiten beschrieben. Für Geodaten sind das die Anforderungen, die aus einer oder mehreren Anwendungen heraus an die Daten gestellt werden. Wenn sich im Zuge der Designplanung Änderungen an den Vorgaben ergeben, so müssen diese in der Spezifikation fortgeschrieben werden, so daß dieses Dokument als definitive, aktuelle Referenz bis zum Abschluß der Designphase dienen kann.

5.3.4.3 Designergebnis

Ergebnisse des detaillierten Designs sind die endgültigen technischen Dokumente, wie sie für die Beschaffung, die Produktion, die Installierung, die Prüfung und die Datenpflege verwendet werden. Aus den Designergebnissen sollte ersichtlich sein, wie die Designvorgaben umgesetzt wurden.

5.3.4.4 Designverifizierung

ISO 9001 beschreibt vier Maßnahmen zur Designlenkung (Design-Reviews, Qualifikationsprüfungen, alternative Berechnungen, Vergleich mit bewährtem Design). Zwei oder mehr Maßnahmen können und sollen nach ISO 9000 miteinander kombiniert werden. Ein Review sollte von anderem Personal durchgeführt werden, als das, welches für die Designarbeit verantwortlich ist, idealerweise durch eine externe, autorisierte Organisation.

Bei einem Review sollte unter anderem auch geprüft werden, ob genormte oder standardisierte Komponenten Verwendung finden, falls solche existieren. Für einen langfristigen und breiten Einsatz der Daten ist diese Forderung sehr wichtig, da man davon ausgehen muß, daß die Daten über viele Versionen und unterschiedliche GIS-Software hinweg migrieren.

5.3.4.5 Designänderung

Es gibt viele Gründe, warum das Design von Geodaten Änderungen unterzogen werden muß. Mögliche Gründe sind, daß Fehler im Datenmodell erst im Laufe der Erfassung erkannt wurden, daß sich die Anforderungen an die Daten z.B. durch Gesetzesänderung geändert haben, daß das Erfassungsverfahren nicht praktikabel oder wirtschaftlich ist, oder daß die Designverifizierung eine Änderung fordert.

Wird das Design während der Erfassung geändert, entstehen Daten, die aus unterschiedlichen Spezifikationen hervorgegangen sind. Diese Information muß zumindest in den Metadaten abrufbar sein. Eine mögliche Konsequenz ist, die schon vorhandenen Daten mit der geänderten Spezifikation nachzuarbeiten, um einen homogenen Datenbestand zu erhalten. Kunden, die Daten der ursprünglichen Spezifikation schon geliefert bekommen haben, müssen informiert werden, und gegebenenfalls diese Änderungen durch differentielle Übernahme oder durch vollständigen Austausch nachvollziehen.

5.3.5 Lenkung der Dokumente

Da die Designlenkung ein dynamischer Vorgang ist, entstehen sehr viele Dokumente, die ständig überarbeitet, an vielen Stellen gelesen, kommentiert, mitgezeichnet und verabschiedet werden. Um Verwirrungen und Verwechslungen zu vermeiden ist ein Dokumentenmanagement erforderlich. Dabei ist es egal, ob die Dokumente in gedruckter oder digitaler Form verteilt werden. Durch Führung einer Dokumentenliste kann der Genehmigungsstatus, die Verteilung und der Überarbeitungsstatus nachvollzogen werden. Dabei sind die Dokumente auch vor unberechtigter Herausgabe zu schützen.

Es soll sichergestellt werden, daß nur genehmigte und aktuelle Dokumente benutzt, und daß Änderungen nur von einer autorisierten Stelle durchgeführt werden.

5.3.6 Beschaffung

Da die verwendeten Werkzeuge (Software, Hardware und vor allem Peripheriegeräte wie Digitizer, Scanner oder Plotter) und Unterlagen (Datenquellen) zur Qualität der Geodaten beitragen, ist die Qualität von Zulieferungen sicherzustellen. Es sind nur Lieferanten auszuwählen, die über genügend Kapazitäten verfügen, termingerecht die geforderte Qualität zu einem angemessenen Preis zu liefern. Die Beurteilung kann sich z.B. auf frühere Leistungen bei Lieferung ähnlicher Produkte oder Dienstleistungen beziehen, oder auf die Darlegung des Qualitätssicherungssystem durch den Lieferanten.

5.3.7 Vom Auftraggeber beigestellte Produkte

Wenn die Datenerfassung vergeben wird, so stellt der Auftraggeber üblicherweise die Datenquellen zur Verfügung. Diese analogen oder gescannten Erfassungsgrundlagen sind Produkte, die im Eigentum des Auftraggebers sind und an den Datenproduzenten in Erfüllung der Forderungen des Vertrages zur Verwendung geliefert werden. Der Datenproduzent akzeptiert bei der Lieferung die Verantwortung für die Schadensvermeidung und für die sachgerechte Lagerung, Handhabung und Verwendung, solange diese beigestellten Produkte im Besitz des Auftragnehmers sind.

Die erfassende Institution sollte deshalb folgende Vorkehrungen treffen:

- Eingangsprüfung der Datenquellen, um die erhaltene Menge (Vollständigkeit) und ihre Identität (Richtigkeit) zu prüfen und um Transportschäden festzustellen
- Sachgerechte Lagerung damit die Erfassungsgrundlagen, die oft Unikate sind, keinen Schaden z.B. durch Wassereinwirkung, Knicke oder Verbleichung aufgrund starker Sonneneinstrahlung nehmen

- Identifikation der Erfassungsquellen, um Verwechslungen vorzubeugen (insbesondere Rasterdaten, da sie nur durch ihren Dateinamen gekennzeichnet werden können)
- Verwahrung der Datenquellen, um unbefugten oder unsachgemäßen Zugriff vorzubeugen.

5.3.8 Identifikation und Rückverfolgbarkeit von Produkten

Durch den räumlichen Bezug oder durch aussagekräftige Bezeichnungen sind die Erfassungseinheiten eindeutig identifizierbar. Alle durchgeführten Bearbeitungsschritte sollen in Begleitdokumenten und Übersichten protokolliert und die Bearbeiter namentlich erwähnt werden. Es ist auch möglich, automatisch zu jedem Objekt den Benutzernamen desjenigen als Attribut zu speichern, der dieses Objekt erzeugt und zuletzt modifiziert hat. Damit ist jederzeit nachvollziehbar, wo sich die jeweiligen Unterlagen befinden, und wer welchen Arbeitsabschnitt durchgeführt hat. Dies ist vor allem bei einer Untersuchung der Prüfergebnisse auf Systematiken erforderlich.

Anhand einer Übersicht soll jederzeit eine Aussage über den Stand der Bearbeitung gemacht werden können.

5.3.9 Prozeßlenkung

Wenn Fehler durch eine Steuerung des Prozesses vermieden werden können, ist diese Methode einer Prüfung und der damit verbundenen Nachbearbeitung von Geodaten vorzuziehen.

In einer Verfahrensdokumentation werden alle Schritte der Bearbeitung dargelegt. Durch eine Erklärung anhand von Skizzen und Beispielen und durch die Behandlung von Sonderfällen, ist dieses Dokument besonders anschaulich zu gestalten. Dabei wird besonders auf die kritischen Qualitätsmerkmale des Produktes hingewiesen. Je nach Anwendung liegen diese in einem oder einer Kombination von den genannten Qualitätskriterien, sie können sich entweder auf alle oder eine Auswahl von Objektklassen, bestimmte Attribute oder Relationen für den gesamten Datenbestand oder Teilgebiete daraus.

Bearbeitungsschritte deren Fehler in nachgeordneten, darauf aufbauenden Schritten nicht mehr aufgedeckt werden können, sind für die Prozeßlenkung besonders wichtig.

Die Produktionsmittel, insbesondere der Digitizer, müssen in regelmäßigen Abständen überprüft werden, um die geforderte Digitalisiergenauigkeit garantieren zu können. Für den Datenerfassungsprozeß muß ein durchgreifendes Konzept zur Datensicherung existieren.

Können Arbeitsabläufe durch organisatorische oder technische Veränderungen verbessert werden, dann ist das Erfassungsverfahren und dessen Dokumentation dahingehend zu ändern. Alle in den Prozeß involvierten Personen sind über die Änderung zu unterrichten und gegebenenfalls neu einzuweisen.

5.3.10 Prüfungen

5.3.10.1 Eingangsprüfungen

Die Vollständigkeit und der Zustand der zur Erfassung notwendigen Unterlagen sind vom Auftragnehmer unmittelbar nach Erhalt zu kontrollieren, um frühzeitig fehlende Unterlagen nachfordern oder Transportschäden beim Spediteur geltend machen zu können. Die Bestandsaufnahme ist zudem zur Arbeitsplanung und zur projektbegleitenden Überwachung notwendig.

Sollen Daten fortgeführt werden, muß vor Einarbeitung von Änderungen in den Datensatz, eine Prüfung durchgeführt werden. Diese Prüfung bezieht sich vor allem auf die Konsistenz der Daten (siehe Abschnitt 6), da nur intakte Daten weiter verwertet werden können. Außerdem kann diese Prüfung automatisch durchgeführt und damit auf den gesamten Datenumfang angewandt werden.

5.3.10.2 Zwischenprüfungen

Zwischenprüfungen erlauben ein frühes Erkennen von Fehlern und eine rechtzeitige Behandlung fehlerhafter Produkte. Eine Identifikation von Fehlern vor Erreichen des Stadiums der Endprüfung

erhöht die Effektivität der gesamten Arbeit und verhindert die Fehlerfortpflanzung im weiteren Bearbeitungsprozeß.

In allen Produktionsschritten sollen Zwischenprüfungen durchgeführt werden. Zwischenprodukte dürfen nur weitergeleitet werden, wenn sie den Anforderungen entsprechen, sie entsprechend gekennzeichnet sind und ihr Qualitätsstand dokumentiert ist. Wenn der Digitalisiervorgang von Geodaten in mehreren Stufen durchgeführt wird, indem z.B. zuerst die Datenquellen aufbereitet und Erfassungsvorlagen erstellt werden, oder die Erfassung in getrennten Arbeitsschritten für unterschiedliche Objektklassen durchgeführt wird, dann hat vor dem Beginn der nächsten Erfassungsstufe eine Zwischenprüfung zu erfolgen.

5.3.10.3 Endprüfungen

Der Datenproduzent darf keine Daten an Dritte weitergeben, die nicht alle Prüfabschnitte vollständig durchlaufen haben. Diese Maßnahme fördert zum einen das gegenseitige Vertrauen von Anwender und Datenerfasser, zum anderen schützt es den Produzenten vor Rückweisungen bei der Eingangskontrolle des Anwenders und vor möglichen Produkthaftungen aufgrund grob fahrlässiger Fehler.

5.3.11 Prüfmittel

Die Tauglichkeit eingesetzter Prüfmethoden ist sicherzustellen, einschließlich der zugehörigen Software. Dabei ist besonders zu hinterfragen, wie durchgreifend die Prüfungen zum Aufdecken der kritischen Kriterien sind. Werden neue Fehlertypen entdeckt, muß untersucht werden, ob und wie eine Prüfung dieses Fehlertyps durchgeführt werden kann.

5.3.12 Prüfstatus

Mit dem Prüfstatus soll angezeigt werden, ob ein Datensatz

- noch nicht geprüft (Voreinstellung)
- geprüft und akzeptiert
- geprüft und bis zu einer Entscheidung zurückgehalten oder
- geprüft und rückgewiesen

worden ist. Der Status kann als Metainformation, digital oder in Begleitpapieren zu den Daten geführt werden. Die prüfende Person, die verwendete Prüfmethode und das Datum der Prüfung sind mit dem Prüfstatus zu führen. Der Prüfer zeichnet sich durch seine physische oder digitale Unterschrift für die ordnungsgemäße Durchführung der Prüfung verantwortlich.

5.3.13 Lenkung fehlerhafter Produkte

Wurde festgestellt, daß irgendein Zwischen- oder Endprodukt die technische Spezifikation nicht erfüllt, muß die unbeabsichtigte Verwendung vermieden werden. Wenn bei der Eingangsprüfung ermittelt wird, daß die Erfassungsvorlagen nicht ausreichen, um den Qualitätsforderungen zu entsprechen, weil sie z.B. nicht aktuell genug, oder über schlecht bestimmte Paßpunkte georeferenziert wurden, oder weil sie zu viele Fehler aufweisen, dürfen diese Datenquellen nicht zur Erfassung von Geodaten herangezogen werden. Die digitalen Objekte können nicht genauer, vollständiger oder korrekter werden als die Grundlagen, auf denen sie beruhen.

Eine Weiterverarbeitung von noch nicht freigegebenen Zwischenprodukten soll durch eine besondere Kennzeichnung und durch eine räumlich getrennte Lagerung verhindert werden. Digitale Daten, die noch nicht alle Prüfungen durchlaufen haben oder deren Korrekturen noch nicht eingearbeitet wurden, sind in getrennten Unterverzeichnissen zu speichern und durch entsprechende Namensgebung vor Verwechslung mit den fertiggestellten Daten zu schützen. Dies ist nur möglich, wenn die Erfassungseinheiten in separaten Dateien abgelegt sind. Werden die Daten direkt in einer Produktionsdatenbank gespeichert, so ist durch Einführung von unterschiedlichen Alternativen eine Trennung zwischen geprüften und freigegebenen Daten und ungeprüften oder zurückgewiesenen zu

gewährleisten.

Es ist nützlich, Verfahren zum Umgang mit Fehlern einzuführen, die erst entdeckt werden, wenn die Daten schon an Kunden weitergegeben wurden. Durch Führung einer Kundenkartei kann nachvollzogen werden, wann welche Daten an wen abgegeben wurden. Sollten zu einem späteren Zeitpunkt gravierende Fehler in den Daten festgestellt werden, können die betroffenen Kunden benachrichtigt und mit einem bereinigten Datensatz beliefert werden. Bei der Fortführung der Daten können die registrierten Abnehmer in gleicher Weise über Änderungen informiert und mit aktualisierten Daten beliefert werden.

5.3.14 Korrekturmaßnahmen

Die Ursachen für entdeckte oder potentielle Fehler sollten unverzüglich identifiziert werden, um gegebenenfalls eine Korrekturmaßnahme zur Verhinderung des Wiederauftretens zu entwickeln. Dabei ist nicht nur die vordergründige Ursache von Belang, sondern durch wiederholtes Hinterfragen ist der eigentliche Grund ausfindig zu machen. Ursachen können sein:

- Fehlbedienung der Erfassungssoftware, möglicherweise bedingt durch unzureichende Schulung oder mangelnde Motivation der Erfasser, deren Ursache wiederum innerhalb (z.B. unangemessene Arbeitsbedingungen) oder außerhalb der Institution zu suchen ist
- Nichtbefolgung von Verfahren, möglicherweise verursacht durch unangemessene Prozeßlenkung

5.3.15 Handhabung, Lagerung, Verpackung und Versand

In allen Phasen der Datenerfassung ist mit dem angelieferten Material, dem Material in Bearbeitung und mit den digitalen Daten so umzugehen, daß Schäden, Datenverluste oder Beeinträchtigungen der Lesbarkeit vermieden werden. Zur Lagerung von digitalen Daten gehört zum einen ein durchgreifendes Datensicherungssystem, mit mehreren Versionen der gesicherten Daten und einer räumlichen Trennung der Speichermedien. Gleichzeitig ist aber auch die physikalische Konsistenz der Daten erforderlich. Die Daten sollten in keinem proprietären Format archiviert werden, damit sie auch von anderen GIS-Softwareprodukten gelesen werden können.

Ergänzend zu der Produktion von haptischen Produkten ist es bei digitalen Geoobjekten möglich, über Rechnernetze direkt auf die Daten zuzugreifen. Diese neue Vertriebsform setzt andere Mechanismen der Prozeßlenkung voraus. Die Daten und für objektorientierte Systeme auch Methoden müssen in frei zugänglichen, offenen Formaten „verpackt“ und bei Anfrage des Kunden in Bruchteilen von Sekunden „versandt“ werden. Standardisierte Schnittstellen, die diesen Zugriff auf Geodaten ermöglichen, werden derzeit vom Open GIS Consortium (OGC) entwickelt. Das OGC ist ein Zweckverband von GIS-Software-, Hardware- und Datenbankherstellern sowie von Datenproduzenten, Anwendungs-entwicklern und Forschungseinrichtungen mit dem Ziel die Interoperabilität von Geoinformation voranzutreiben (*Buehler and McKee, 1998*).

Bei diesem „online“-Zugriff auf die Daten kommen Elemente wie Schutz vor unberechtigtem Zugriff, Zahlungsmodalitäten und Zuverlässigkeit des Servers, der Verbindung (Bandbreite) und der Daten zum Tragen.

5.3.16 Qualitätsaufzeichnungen

Qualitätsaufzeichnungen sollten den direkten oder indirekten Nachweis darüber enthalten, ob die Daten die technischen Forderungen sowie die gesetzlichen und vertraglichen Forderungen erfüllen. Alle Prüfungen sind übersichtlich zu dokumentieren. Aus den Unterlagen muß hervorgehen, welcher Mitarbeiter wann mit welchen Hilfsmitteln die entsprechende Prüfung durchgeführt hat. Zu jeder Erfassungseinheit soll eine solche Liste geführt werden.

Die Qualitätsaufzeichnungen sollen erarbeitet, sicher gelagert, vor ungenehmigtem Zugang bewahrt und vom Lieferanten einen dokumentierten Mindestzeitraum aufbewahrt werden.

5.3.17 Interne Qualitätsaudits

Eine regelmäßige Überprüfung auf Realisierung der gesetzten Qualitätsziele ist notwendig. Diese sogenannten Audits sind aber auch erforderlich, um sicherzustellen, daß das Qualitätssicherungssystem nicht nur dokumentiert sondern auch praktiziert wird, damit unterschiedliche Datensätze anhand einheitlicher Qualitätskriterien und Qualitätsmaße verglichen werden können.

5.3.18 Schulung

Wichtig für die Durchführung eines Qualitätsmanagements bei der Datenerfassung ist, daß alle Mitarbeiter die Methoden verstanden haben, die Werkzeuge, also die Erfassungsfunktionen des GIS, beherrschen und die Qualitätssicherung als wichtige Aufgabe zum Erlangen des Unternehmenszieles verinnerlicht haben. Diese Bewußtseinsbildung kann nur durch systematische Schulungen zur Ausbildung von qualifiziertem Personal erreicht werden. Dabei soll nicht nur der fachliche Aspekt im Vordergrund stehen, sondern auch die Bereitschaft, in einem Team Verantwortung zu übernehmen.

5.3.19 Kundendienst

Um die Funktionsfähigkeit der Daten beim Kunden zu garantieren, soll eine geeignete Dokumentation einschließlich Gebrauchsanweisungen für den Umgang mit den Daten geliefert werden. Die Daten sind dabei systemunabhängig zu betrachten. Bei komplexen Datenmodellen soll der Produzent durch Bereitstellung von fachkundigem Kundendienstpersonals die Integration der Daten in das System des Kunden beratend unterstützen. Ein enger Kontakt zu den Kunden hilft das Produktdesign oder den Kundendienst zu verbessern.

Da die reale Welt permanenten Änderungen unterworfen ist, müssen Geodaten als verderbliche Ware betrachtet werden. Mit einer einmaligen Erfassung der Objekte der realen Welt ist es deshalb nicht getan. Der Produzent muß dem Kunden, also dem Anwender der Daten, ein dessen Bedürfnissen angemessenes Fortführungskonzept anbieten.

5.3.20 Gebrauch statistischer Methoden

Wenn statistische Methoden als Mittel zum Nachweis der Erfüllung von Qualitätsforderungen eingesetzt werden, so sind diese zu dokumentieren. Dies kann als eine Form der Qualitätsaufzeichnung genutzt werden.

Da die Bestimmung der Stützpunktkoordinaten von Geoobjekten oder quantitativer Attributwerte einen Meßvorgang mit Unsicherheiten darstellt, die allen Messungen inhärent sind, werden für die Bestimmung von Genauigkeitsmaßen statistische Methoden benötigt. Die verwendeten Methoden und Genauigkeitsmaße sind zu dokumentieren.

Wenn eine vollständige Kontrolle aller Objekte einer Datenlieferung aus Zeit- oder Kostengründen nicht möglich ist, trotzdem aber geprüft werden soll, ob die Daten den geforderten Qualitätszielen entsprechen, bietet sich das Verfahren der Stichprobenuntersuchung an. Eine Anzahl zufällig ausgewählter Objekte wird eingehend untersucht, und mit Hilfe von statistischen Methoden auf die Qualität der Daten der gesamten Lieferung geschlossen. In Abschnitt 7 wird dieses Verfahrens ausführlich diskutiert. Wenn solche Verfahren eingesetzt werden, ist dies zwischen den Vertragspartnern abzustimmen.

6 Konsistenzprüfungen

Nach der Definition von Konsistenz (siehe Abschnitt 4.3.3) müssen die Daten exakt den Regeln entsprechen, die im Datenmodell bzw. Informationsmodell festgelegt sind. Da diese Regeln eindeutig sein müssen und sich nur auf die Daten an sich beziehen, ist es im allgemeinen möglich, diese Regeln in eine strukturierte Form zu bringen, so daß sie in der Datenverwaltungskomponente des GIS zur Überwachung der Konsistenz der Daten implementiert werden können. Diese formalisierten Regeln können auch dazu verwendet werden, mit entsprechender Prüfsoftware nach Verletzungen von diesen Regeln zu suchen.

Da sich Konsistenz auf die verschiedenen Ebenen der Modellierung in physikalische, logische und konzeptionelle Konsistenz unterteilt, sind für jede Ebene entsprechende formale Regeln anzugeben. Die Regeln zur Erzielung der physikalischen Konsistenz sind im DBMS des GIS implementiert. Wenn diese Regeln verletzt sind, können die Daten nicht mehr gelesen werden, oder es kommt zu unkontrollierbaren Zuständen in Form von Verwechslungen, unsinnigen geometrischen Ausprägungen oder zu konkurrierenden Speicherzugriffen, die zu Programmabstürzen führen können.

Je nach Geometriesubschema des GIS und geometrischem Modell der Anwendung können die Bedingungen der logischen Konsistenz schon im System überwacht werden, anderenfalls muß ihre Einhaltung vom Anwender sichergestellt werden. Bei Spaghetti-Code sind weniger Regeln zu beachten als bei Geometriedaten, die als planarer Graph modelliert werden (Abschnitt 2.3.1). Die Kontrolle der logischen Konsistenz wird üblicherweise mit Prüfprogrammen durchgeführt (*USGS, 1998, Bureau of the Census, 1997*, oder *Rath und Auerbach, 1996*). *Plümer, 1996b*, schlägt eine elegantere Form vor: Regeln in deduktiven Datenbanken oder sogenannte „Trigger“ auf relationale Datenbasen sollen verwendet werden, um permanent über Regeln der logischen Konsistenz zu wachen, und somit das Konzept der konsistenzerhaltenden Transaktionen auf Geodatenbanken umzusetzen (siehe hierzu auch *Meier, 1982*).

Die Regeln der konzeptionellen Konsistenz lassen sich im allgemeinen nur schwer formalisieren, weil oft Ausnahmen existieren oder die Objekte im Kontext mit anderen Objekten betrachtet werden müssen. Es lassen sich allerdings einige Grundbedingungen aufstellen, mit denen die Regeln der konzeptionellen Konsistenz formuliert werden können.

6.1 Prüfung der logischen Konsistenz

6.1.1 Konsistenz der Attributwerte

Auf der Ebene des logischen Datenmodells müssen die Attribute bestimmte Regeln erfüllen. Die Attributwerte müssen vom richtigen Typ sein. In bestimmten Fällen ist eine implizite Typumwandlung erlaubt. So ist eine Transformation von Werten aus dem Bereich der ganzen Zahlen in reelle Zahlen ohne Probleme möglich. Die Gegenrichtung ist aber im allgemeinen mit Verlust von Information – in diesem Falle von Dezimalstellen – verbunden. Eine Zuweisung von Zeichenketten auf einfache Zahlenwerte führt zu Fehlern, genauso die Zuweisung von komplexen Datentypen wie Felder oder Mengen zu einfachen Datentypen. Über diese Regeln wacht üblicherweise das System, so daß es zu Inkonsistenzen dieses Typs erst gar nicht kommen kann, weil sie bei der Eingabe oder beim Transfer abgelehnt werden.

Die logische Bedeutung eines Attributes ist von der Anwendung und damit auch von der konzeptionellen Modellierung unabhängig. Wenn ein Attribut z.B. einen Winkel repräsentiert, so hat dies Konsequenzen auf den Wertebereich dieses Attributes. Das Attribut vom Typ der reellen Zahlen darf dann, je nach Winkelmessung, nur noch im Bereich $-\pi$ bis π oder 0 bis 2π oder –180° bis +180° oder 0gon bis 400gon liegen. Für Streckenattribute bei Bemaßungen sind nur positive Zahlen erlaubt. Diese Bedingungen können auf einer systemnahen Ebene implementiert werden. Um eine Unabhängigkeit dieser Regeln von der Anwendung zu erreichen, weil die Semantik der Attribute nicht selbsterklärend ist, muß die Menge der erlaubten Datentypen im System erweitert werden. Dies kann dadurch geschehen, daß neue Datentypen wie in den genannten Beispielen vom Typ Winkel oder

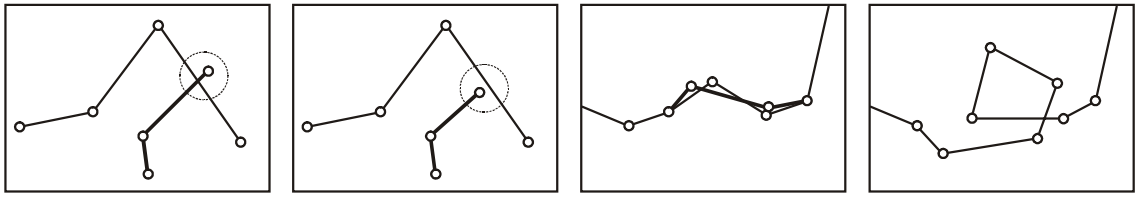
Strecke eingeführt werden. Aufgabe des Systems ist es dann, über die Einhaltung dieser Regeln zu wachen. Verletzungen dieser Bedingungen dürfen erst gar nicht zugelassen werden, so daß sich der Datenbestand bezüglich dieser Attribute immer in einem konsistenten Zustand befindet.

6.1.2 Konsistenz der Topologie

Viele Algorithmen zur Prüfung der logischen Konsistenz setzen einen planaren Graphen als logisches Datenmodell voraus (Plümer, Kainz, SDTS, ...). Planare Graphen besitzen Eigenschaften, die als Regeln formuliert werden können und sind daher für die Prüfung der logischen Konsistenz besonders geeignet. Es lassen sich allerdings nicht alle Datenmodelle auf diese Struktur abbilden. In einem topologischen Netz von Straßen dürfen an Brücken ohne Übergangsmöglichkeit zwischen den Straßen keine Knoten gebildet werden. Wenn dieses Netz trotzdem als planarer Graph modelliert werden soll, so müssen an diesen Knoten Zusatzinformationen, z.B. als Attribute oder zusätzliche Objekte der Objektart "Brücke" eingeführt und bei einer Auswertung, z.B. für die Berechnung einer optimalen Fahrroute, berücksichtigt werden.

6.1.2.1 Geometrische Konstellationen mit hoher Wahrscheinlichkeit einer topologischen Inkonsistenz

In vielen Fällen läßt sich mit einer Durchsuchung der Daten nach bestimmten geometrischen Konstellationen herausfinden, wo Fehler der Topologiebildung bei der Digitalisierung aufgetreten sind. Typische Konstellationen sind in Abbildung 35 zusammengestellt.



Overshoot *Undershoot* *Sliver Polygon* *Intersection*

Abbildung 35: Geometrische Konstellationen, bei denen mit hoher Wahrscheinlichkeit ein Digitalisierfehler vorliegt.

Das Auffinden dieser Konstellationen erfolgt über Fangradien, für die Grenzwerte angegeben werden müssen. Diese Grenzwerte sind abhängig von der geforderten geometrischen Genauigkeit. Wenn die Punkte der betroffenen Objekte außerhalb des Konfidenzbereiches liegen, so ist davon auszugehen, daß die Geometrie der Objekte von dem Bearbeiter bewußt so festgelegt wurde, oder es liegt ein grober Fehler vor, der durch die Prüfung nicht aufgedeckt werden kann. Liegen die Punkte aber innerhalb des Konfidenzbereiches, so kann es sein, daß der Bearbeiter die Datenpunkte zwar im Rahmen der erforderlichen Genauigkeit digitalisiert, aber die Funktionalitäten des Systems zur topologischen Anbindung der Objekte nicht verwendet hat. Es ist aber auch möglich, daß die Objekte des abstrakten Abbildes der realen Welt tatsächlich so zueinander in Beziehung stehen, und die Daten daher an dieser Stelle keinen Fehler aufweisen. Diese Konstellationen stellen also weder eine notwendige noch eine hinreichende Bedingung für topologische Inkonsistenzen dar. Trotzdem geben sie bei der Prüfung Hinweise, wo mögliche Fehler liegen, die aber durch den Prüfer im einzelnen verifiziert werden müssen. Je größer die Fangradien gewählt werden, um so mehr vermeintliche Fehler werden gefunden, aber um so sicherer ist es, daß keine Fehler übersehen werden.

Für den Fall eines planaren Graphen dürfen die Konstellationen „*Overshoot*“, „*Sliver polygon*“ und „*Intersection*“ nicht vorkommen, weil diese gegen die Planaritätsbedingung verstoßen. In allen drei Fällen existieren Schnittpunkte, an denen bei planaren Graphen Knoten gebildet werden müssen.

6.1.2.2 Konsistenzbedingungen für planare Graphen

Wenn das logische Datenmodell für die Geometrie als Landkarte (planarer Graph mit Knoten, Kanten und Maschen) beschrieben werden kann, dann müssen die topologischen Primitive folgende Bedingungen erfüllen (*Plümer, 1996a*).

Knoten:	1.	Für jeden (topologischen) Knoten existiert genau ein (geometrischer) Punkt. Keine zwei verschiedenen Knoten haben dieselben Koordinaten.
	2.	Jeder Knoten hat mindestens zwei inzidierende Kanten (Grad des Knoten ≥ 2).
Kanten:	3.	Jede Kante besitzt genau zwei unterschiedliche Knoten als Endpunkte.
	4.	Die geometrische Entsprechung einer Kante ist die geradlinige Strecke. Zwei Strecken haben außer an ihren Enden keine gemeinsamen Punkte.
	5.	Jede Kante besitzt genau zwei verschiedene inzidierende Maschen.
Maschen:	6.	Jede Masche hat genau einen Kreis (Folge von Kanten, in der alle Kanten verschieden sind und ebenso alle Knoten mit Ausnahme des Anfangs- und Endknoten) als Begrenzung.
	7.	Kein Punkt einer Kante liegt in der Fläche einer Masche.

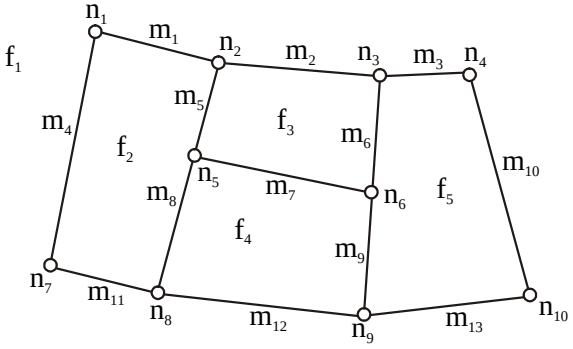
In *Plümer und Gröger, 1996*, wird eine weitere, achte Bedingung eingeführt, um für bestimmte Konstellationen bei konsistenzerhaltenden Transaktionen die Bedingung 6 nicht zu verletzen.

Zusammenhang:	8.	Der Graph ist zusammenhängend.

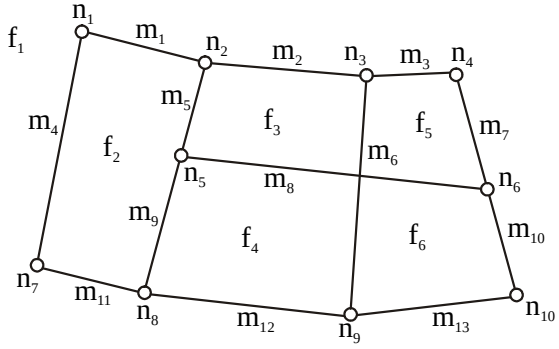
Die Autoren zeigen, daß diese Bedingungen hinreichend sind, um in Landkarten geometrisch-topologische Fehler aufzuspüren. Die Konsistenzbedingungen, wie sie bei *Kainz, 1995*, oder in *Bureau of the Census, 1997*, beschrieben werden, sind in dieser Auflistung vollständig enthalten.

Für planare Graphen muß außerdem der Eulersche Polyedersatz (*Volkmann, 1996, Wilson and Watkins, 1990*) gelten. Dieser sagt aus, daß wenn *G* ein zusammenhängender, planarer Graph ist, so gilt der Zusammenhang $n(G) + f(G) - m(G) = 2$,

wobei *n(G)* die Ordnung (Anzahl der Knoten), *f(G)* die Anzahl der Maschen und *m(G)* die Größe (Anzahl der Kanten) eines Graphen angeben. Dieser Satz, den Euler schon 1752 aufgestellt hat, stellt allerdings nur eine notwendige Bedingung für einen planaren Graphen dar. Wenn die Bedingung verletzt ist, so ist der Graph nicht planar. Wenn der Graph allerdings die Bedingung erfüllt, so kann er trotzdem nicht planar sein.



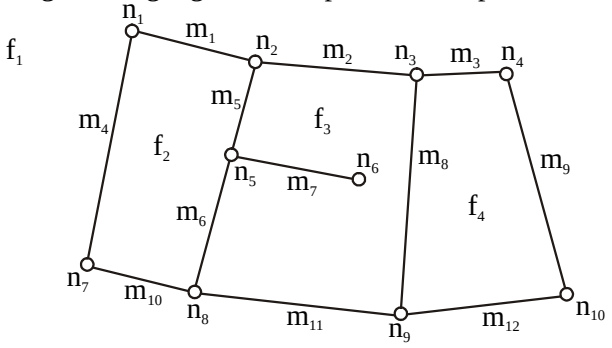
n(G) = 10, f(G) = 5, m(G) = 13 → *n+f−m = 2*



n(G) = 10, f(G) = 6, m(G) = 13 → *n+f−m ≠ 2*

Abbildung 36: Satz von Euler als notwendige Bedingung für einen planaren Graphen

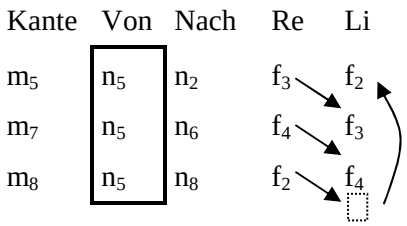
Ein weiteres Beispiel zeigt, daß der Satz von Euler nicht hinreichend ist, um topologische Inkonsistenzen zu detektieren. Obwohl die Gleichung erfüllt ist, widerspricht der Graph der Bedingung 2, nach der jeder Knoten mindestens den Grad 2 haben muß. Für den Knoten n_6 ist dies nicht erfüllt. Dieser inzidiert nur die Kante m_7 und hat damit den Grad 1.



n(G) = 10, f(G) = 4, m(G) = 12 → *n+f−m = 2*

Eine weitere Möglichkeit, um die topologische Konsistenz zu testen, ist durch den sogenannten „umbrella"-Test gegeben (*Kainz, 1995*). In einer dime-Datenstruktur (*dual independant map encoding*), wie sie in der folgenden Tabelle für das Beispiel aus Abbildung 36 gegeben ist, muß sich um jeden Knoten eine alternierende Folge von Kanten und Maschen ergeben.

Kante	Von	Nach	Rechte Masche	Linke Masche	Attribut X
m_1	n_1	n_2	f_2	f_1	...
m_2	n_2	n_3	f_3	f_1	...
m_3	n_3	n_4	f_5	f_1	...
m_4	n_1	n_7	f_1	f_2	...
m_5	n_2	n_5	f_2	f_3	...
m_6	n_3	n_6	f_3	f_5	...
m_7	n_5	n_6	f_4	f_3	...
m_8	n_5	n_8	f_2	f_4	...
m_9	n_6	n_9	f_4	f_5	...
m_{10}	n_4	n_{10}	f_5	f_1	...
m_{11}	n_7	n_8	f_1	f_2	...
m_{12}	n_8	n_9	f_1	f_4	...
m_{13}	n_9	n_{10}	f_1	f_5	...



Zyklische Folge von alternierenden rechten und linken Maschen.

Zu jedem Knoten n_1 bis n_{10} werden die Kanten selektiert, die diesen Knoten inzidieren. Die Orientierung der Kanten muß dabei so gewählt werden, daß dieser Knoten entweder einheitlich Von- oder Nach-Knoten ist. Bei einem Wechsel der Orientierung wechseln auch die linke und rechte Masche. Die inzidierenden Maschen müssen dann eine zyklische Folge von alternierenden rechten und linken Maschen ergeben (*Kainz, 1995*, *Wise, 1998*).

6.2 Prüfung der konzeptionellen Konsistenz

Im konzeptionellen Datenmodell sind implizit oder explizit Regeln vereinbart, die von allen Objekten der entsprechenden Objektklassen eingehalten werden müssen. Während die Bedingungen für die logische Konsistenz sehr allgemein formuliert werden können, ist die konzeptionelle Konsistenz von der Semantik der Daten abhängig. Es sind daher für jede Objektklasse gesondert Regeln zu vereinbaren. Die Regeln können sich auf die Sachdaten, die Geometrie oder auf eine Kombination von Sachdaten in Abhängigkeit der Geometrie beziehen. Für die Sachdaten können Regeln formuliert werden, die sich auf die Attributwerte an sich beziehen oder auf Beziehungen zwischen Attributwerten eines Objekts oder mehrerer Objekte, die in hierarchischer oder topologischer Beziehung zueinander stehen. Bezüglich der Geometrie von Objekten können Regeln aufgestellt werden, die sich nur auf ein Objekt beziehen oder auf die Beziehung mehrerer Objekte zueinander.

6.2.1 Attributregeln bezogen auf ein Objekt

6.2.1.1 Zwingend erforderliche Attributeinträge

Wenn im Datenmodell explizit vereinbart wurde, daß für bestimmte Attribute der angegebenen Objektklassen Attributwerte zwingend erforderlich sind, so müssen diese Attribute für jedes Objekt dieser Objektklassen im Datensatz mit einem Eintrag belegt sein. Diese Regel kann als Vollständigkeitsregel der Attributwerte (*attributive completeness, DIGEST, 1997*) bezeichnet werden. Die formale Beschreibung dieser Regel ist im Abschnitt 4.3.1 angegeben. Diese Definition orientiert

sich an dem Vergleich zwischen abstraktem Abbild der realen Welt und den Daten. Die Betrachtung in diesem Abschnitt bezieht sich auf Regeln, wie sie im Datenmodell festgelegt sind.

Weil die Regeln eindeutig sind, läßt sich die Vollständigkeit der Attributwerte im Gegensatz zur Vollständigkeit der Objekte durch Prüfprogramme ermitteln. Für die vorhanden Objekte kann durch eine Abfrage nach Attributen gesucht werden, die mit dem Wert *NULL* (nicht zugewiesen) belegt sind. Die gefundenen Objekte sind fehlerhaft.

Formal lautet die Regel: $\forall O \in \boldsymbol{A} \quad O.RequiredAttribute \neq NULL$ und ihre Negation d.h. die Bedingung, nach der die Regel verletzt ist: $\exists O \in \boldsymbol{A} \quad O.RequiredAttribute = NULL$. Diese Bedingung entspricht der oben beschriebenen Abfrage.

Als Beispiel für zwingend erforderliche Attributeinträge können externe Objektschlüssel (siehe Abschnitt 2.2.3.2) genannt werden. Ohne diese Attributwerte können Objekte nicht eindeutig angesprochen werden. In ATKIS sind als weitere erforderliche Sachdaten die Namen von Objekten festgelegt (*AdV, 1995*). Wenn der Name eines Objektes nicht bekannt ist, oder das Objekt keinen Eigennamen hat, so ist das entsprechende Attribut mit dem Wert „`NNNN`“ zu füllen.

6.2.1.2 Eindeutigkeit von Identifikatoren

Werden externe Objektschlüssel verwendet, wie im Beispiel des vorherigen Abschnitts beschrieben, so müssen diese auch eindeutig sein. Durch keine Schlüsselvergabe oder Manipulation der Daten darf es passieren, daß ein Schlüssel mehrere Male verwendet wird. Vor allem bei der Fortführung eines Datenbestandes muß darauf geachtet werden, daß bei Operationen wie Teilen oder Kopieren eines Objektes keine doppelten Einträge entstehen. Diese Bedingung kann als Konsistenz der Objektidentifikatoren bezeichnet werden.

Bei ATKIS werden sowohl auf Objekt- als auch auf Teilobjektebene Identifikatoren vergeben. Der Schlüssel des Teilobjekts setzt sich aus dem Schlüssel des zugehörigen Objekts und einer eindeutigen Unternummer zusammen. Abhängig von der Implementierung werden diese Schlüssel redundant geführt. In diesem Falle muß eine weitere Konsistenzbedingung eingehalten werden, nämlich daß der Objektschlüssel und der zugehörige Teil des Teilobjektschlüssels identisch sind. Werden bei Objektmanipulationen Teilobjekte zu neuen Objekten zusammengefaßt, so muß auf diese Konsistenzbedingung geachtet werden.

Mit Hilfe der Prädikatenlogik ausgedrückt lautet die Regel $\forall o_i, o_j \in \boldsymbol{a} \quad o_i.id \neq o_j.id$ mit $i \neq j$ und die Regel ist verletzt wenn $\exists o_i, o_j \in \boldsymbol{a} \quad o_i.id = o_j.id$ mit $i \neq j$ ist.

Die Prüfung dieser Regel kann erfolgen, indem jedes Objekt mit allen anderen Objekten verglichen wird, ob die Identifikatoren identisch sind. Die Anzahl der durchzuführenden Vergleiche steigt dabei quadratisch mit der Anzahl der Objekte. Daher ist es sinnvoll einen Index auf die Objektschlüssel zu setzen, weil im sortierten Zustand keine gleichen Schlüssel hintereinander auftreten dürfen. Die Anzahl der Vergleiche ist dann linear von den Objekten des Datenbestandes abhängig. In den meisten Datenbankmanagementsystemen kann die Eindeutigkeit der Indentifikatoren durch das System überwacht werden, indem das Attribut mit dem Typ „`unique`“ vereinbart wird. Diese Möglichkeit der automatischen Überwachung existiert nur wenn der gesamte Datenbestand auf einem zentralen Server verwaltet wird, oder die Daten auf dezentralen Rechnern abgelegt sind, und das DBMS als verteiltes System auf alle Daten zugreifen kann.

6.2.1.3 Wertebereiche von Attributen

Abhängig von der Semantik von Attributen, von den Objektklassen, auf die sich die Attribute beziehen, und vom Attributtyp lassen sich Wertebereiche definieren (siehe Abschnitt 2.2.3.4). Entsprechende Wertebereiche müssen für Objektklassen und für Attribute in Regeln formuliert werden. Durch Prüfprogramme können Objekte eruiert werden, deren Attribute diesen Regeln widersprechen.

Für **quantitative Attribute** läßt sich der Wertebereich als Intervall oder als Verknüpfung mehrerer Intervalle angeben. Die Regel zur Einhaltung des Wertebereiches lautet somit

$$\forall o \in \boldsymbol{a} \quad o.e \in [\min_1, \max_1] \vee o.e \in [\min_2, \max_2] \vee \ldots .$$

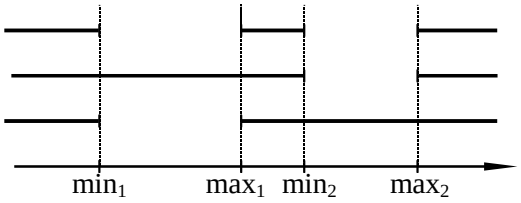
Diese Regel ist verletzt, wenn

$$\exists o \in \boldsymbol{a} \quad \neg(o.e \in [\min_1, \max_1] \vee o.e \in [\min_2, \max_2] \vee \ldots)$$
$$\Leftrightarrow \exists o \in \boldsymbol{a} \quad o.e \notin [\min_1, \max_1] \wedge o.e \notin [\min_2, \max_2] \wedge \ldots .$$

Durch Überführung der Intervallschreibweise in Kleiner- und Größerrelationen kann dieser Term auch ausgedrückt werden durch

$$\exists o \in \boldsymbol{a} \quad (\min_1 > o.e \vee \max_1 < o.e) \wedge (\min_2 > o.e \vee \max_2 < o.e) \wedge \ldots .$$

Diese Abfrage läßt sich mit SQL realisieren.



$(\min_1 > o.e \vee \max_1 < o.e) \wedge (\min_2 > o.e \vee \max_2 < o.e)$

$\min_2 > o.e \vee \max_2 < o.e$

$\min_1 > o.e \vee \max_1 < o.e$

Als Beispiel für ein quantitatives Attribut mit zwei Intervallen kann das ATKIS-Attribut „Breite der Fahrbahn (BRF)“ für die linienhaften Objektarten „Straße“, „Weg“ und „Fahrbahn“ angeführt werden. Als Attributwert wird der tatsächliche Wert gerundet auf 0,5 m angegeben. Aus Plausibilitätsgründen dürfen dieser Verkehrswege eine Breite von 0,5 m nicht unter- und von 50,0 m nicht überschreiten. Bei einer Angabe in Dezimeter ergibt sich daraus das Intervall [5; 500]. Da in ATKIS aber immer auch die Werte „9997: Attribut trifft nicht zu“, „9998: nach Quellenlage derzeit keine Zuweisung möglich“ und „9999: Sonstiges“ als gültige Einträge zugelassen sind, existiert ein zweites Intervall als Wertebereich: [9997 ; 9999]. Das Attribut „Breite der Fahrbahn (BRF)“ ist also entweder gar nicht besetzt, oder es liegt in einem dieser beiden Intervalle.

Für **qualitative Attribute** lassen sich die Wertebereiche nicht durch Intervallgrenzen angeben. Vor allem freie Texteinträge können als qualitative Attribute bezeichnet werden. Weil diese in Geoinformationssystemen eine wichtige Rolle spielen, beziehen sich die folgenden Betrachtungen ohne Einschränkung der Gültigkeit für andere qualitative Attribute vor allem auf deren Kontrolle.

Wenn die möglichen Attributwerte eine endliche Menge darstellen und die Elemente dieser Menge bekannt sind, so läßt sich die Richtigkeit der Einträge prüfen. Dabei wird nicht die Richtigkeit der Zuordnung eines Attributwertes zu einem Objekt überprüft sondern lediglich, ob der Eintrag einen gültigen Wert aus dem Wertebereich darstellt. Die Prüfung der qualitativen Attribute läßt sich automatisch durchführen, wenn die Menge der erlaubten Werte in geeigneter digitaler Form vorliegt.

Nach der Definition von Richtigkeit für Attributwerte (Abschnitt 4.3.2)

$$\forall O \in \boldsymbol{A} \quad \forall O.E_i \in O.\boldsymbol{E}_i \quad O.E_i \in O.\boldsymbol{E}_i \rightarrow o.e_i \in o.\boldsymbol{e}_i \quad \wedge \quad O.E_i \mathrel{\hat{=}} o.e_i$$

stellt die Menge der möglichen Werte für die Eigenschaft *i* eines Objekts $O.\boldsymbol{E}_i$ den Wertebereich für das qualitative Attribut dar. Da $O.\boldsymbol{E}_i$ immer eine abstrakte, wenn auch u.U. endliche Menge darstellt, ist jede digitale Auflistung der Elemente dieser Menge ein bestimmter Datensatz. Diese Referenzdaten können aber auch mit Fehlern behaftet sein. Ein Vergleich dieses unabhängig erfaßten Datensatzes mit den Attributwerten der Geodaten kann zum Auffinden von Fehlern in beiden Datenbeständen verwendet werden. Die Wahrscheinlichkeit, daß ein Wert in beiden Datensätzen auf die gleiche Weise falsch eingetragen wurde ist sehr gering. Zweifelsfälle müssen geklärt werden. Die Rechtschreibprüfung von Textverarbeitungsprogrammen arbeitet nach derselben Methode. Als Wertebereich wird ein Katalog von Wörtern eines vom Anwender erweiterbaren Wörterbuches verwendet.

Diese Methode kann auch zur Ermittlung der Vollständigkeit von Geodaten herangezogen werden, wenn sichergestellt ist, daß alle Elemente der Referenzdatei auch als Attributwerte in den Geodaten wieder auftauchen müssen. Zwischen den Elementen des Referenzdatensatzes und der Geodaten muß dann eine bijektive Abbildung möglich sein.

Das Gemeinderegister eines Bundeslandes oder das Straßenverzeichnis einer Stadt sind Beispiele für Referenzdatensätze, die zur Kontrolle der Vollständigkeit und zur richtigen Schreibweise von Attributwerten verwendet werden können. Diese Daten stehen bei statistischen Behörden im allgemeinen in digitaler Form zur Verfügung.

Wenn die Wertebereiche keine endliche Menge darstellen, oder die Werte nicht a priori bekannt sind, lassen sich für die Schreibweise trotzdem Regeln aufstellen. Diese Regeln können aber nur schwer von einer Objektklasse oder einem Attribut auf andere übertragen werden. Für Namen läßt sich z.B. die Regel aufstellen, daß Namen immer mit einem Großbuchstaben des Alphabets beginnen müssen. Für Recherchen auf den Daten ist auch eine einheitliche Schreibweise von Namen oder Abkürzungen erforderlich. So muß im Datenmodell z.B. vereinbart werden, ob bei Straßennamen das Wort „-straße“ immer ausgeschrieben wird. Bei Kurznamen von Straßen mit Widmung beginnt der Attributwert immer mit einem Großbuchstaben aus der Menge {E, A, B, L, S, K} ohne Leerzeichen gefolgt von einer Ziffernkombination von bis zu fünf Ziffern. Diese Beispiele sind sehr speziell, sollen aber aufzeigen, in welcher Art zahlreiche Fehlerquellen durch anwendungsspezifische Regeln kontrolliert werden können.

6.2.1.4 Bedingungen zwischen Attributwerten eines Objekts

Die Attribute eines Objektes beinhalten oft versteckte Redundanzen. Das bedeutet, daß aus einem Attribut oder aus der Kombination von mehreren Attributwerten der Wert oder ein Teil des Eintrages eines anderen Attributes abgeleitet werden kann. Diese Bedingungen zwischen den Attributen können herangezogen werden, um Regeln für die Einhaltung der Konsistenz zu formulieren.

Das Beispiel Widmung einer Straße kann zur Veranschaulichung herangezogen werden. Für die ATKIS-Objektart „3101 Straße“ sind unter anderen die beiden folgenden Attribute definiert:

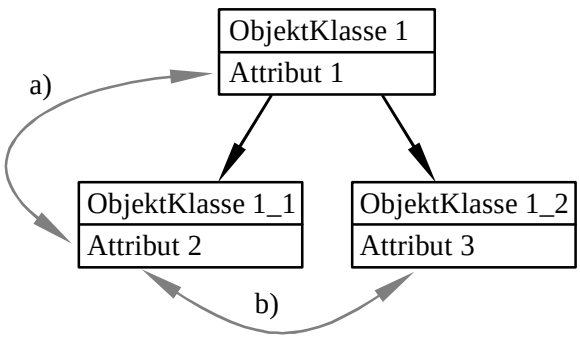
- KN Kurzbezeichnung - verkehrstechnische Bezeichnung (Kurzform, Nummer)
- WDM Widmung mit folgenden möglichen Attributwerten
  - 1301 Bundesautobahn
  - 1303 Bundesstraße
  - 1305 Landesstraße, Staatsstraße
  - 1306 Kreisstraße
  - 1307 Gemeindestraße
  - 1308 Forststraße
  - 9997 Attribut trifft nicht zu
  - 9999 sonstige

Zwischen der Kurzbezeichnung und der Widmung besteht ein Zusammenhang: wenn WDM = 1301 ist, dann muß für die Kurzform der Buchstabe „A“ eingetragen sein, bei Bundesstraße „B“, bei WDM = 1305 „L“ oder „S“ und für die Kreisstraße ein „K“ als erster Buchstabe der Kurzbezeichnung.

Für dieses Beispiel können SQL-Abfragen formuliert werden, die Objekte der Objektart Straße selektieren, bei denen ein Widerspruch zwischen der Kurzbezeichnung und der Widmung existieren. Voraussetzung ist, daß die Datenbankabfragesprache String-Operationen unterstützt. In diesem Fall muß die Möglichkeit bestehen, den ersten Buchstaben der Kurzbezeichnung zu isolieren. Diese Funktion gehört nicht zum Sprachumfang der derzeitig international normierten Versionsnummer 2 von SQL.

6.2.2 Attributregeln mit hierarchischen Beziehungen von Objekten

Besteht zwischen Objekten eine hierarchische Beziehung, können zwischen den Attributen dieser Objekte Regeln formuliert werden. Dabei lassen sich zwei Arten von Beziehungen unterscheiden.



Die erste Art, in der nebenstehenden Grafik mit a) bezeichnet, besteht aus einer Vater/Sohn-Beziehung zwischen den Objekten. Eine Regel für

diese Art könnte lauten: wenn das Attribut 1 der ObjektKlasse 1 einen bestimmten Wert hat, dann müssen alle Teilobjekte der ObjektKlasse 1_1 für das Attribut 2 einen anderen definierten Wert haben.

Die zweite Art, mit b) gekennzeichnet, gibt eine Bruder-Beziehung. Eine Regel für diese Art der Beziehung kann so formuliert werden: wenn Teilobjekte der ObjektKlasse 1_1 und 1_2 Söhne einer Instanz für die ObjektKlasse 1 sind, dann muß das Attribut 3 in Abhängigkeit von Attribut 2 bestimmte Werte annehmen.

Zum Auffinden von Objekten, die der Regel a) widersprechen, kann folgende Abfrage formuliert werden, die hier als Pseudocode wiedergegeben ist:

```
select ObjektKlasse1_1 from
  {select ObjektKlasse1 from * where ObjektKlasse1.Attribut1 = `X`}
where ObjektKlasse1_1.Attribut2 != ´Y´;
```

Diese Abfrage geht davon aus, daß alle Objekte der ObjektKlasse 1 als Ergebnismenge zwischengespeichert werden, aus dieser Teilmenge werden dann die Objekte heraus gefiltert, deren Attributwerte nicht mit denen des Vaterobjektes zusammenpassen.

Für diese Regel bietet ATKIS einige Beispiele, weil ATKIS streng das Konzept von Objekt und Teilobjekt verfolgt, und Attribute, um Redundanzen zu vermeiden, auf einer möglichst hohen Hierarchiestufe angesiedelt werden. Bei klassifizierten Straßen beispielsweise müssen in ATKIS die Attributwerte für die Fahrbahnbreite zwischen 20 dm und 400 dm liegen und es muß mindestens ein Fahrstreifen existieren. Die Widmung von Straßen wird auf Objektebene, die beiden anderen Attribute auf Objektteilebene geführt.

Regeln nach dem Muster b) können zum Aufsuchen von fehlerhaften Objekten verwendet werden. Pseudocode dafür läßt sich folgendermaßen schreiben

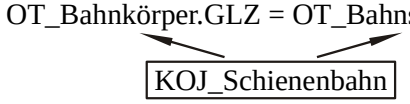
```
select ObjektKlasse1_2 from
  {select ObjektKlasse1 from * where ObjektKlasse1_1.Attribut2 = `X`}
where ObjektKlasse1_2.Attribut3 != ´A´;
```

Bei dieser Formulierung wird vorausgesetzt, daß die Abfragesprache Selektionen zuläßt, die unterschiedliche Hierarchieebenen überspannen. Wenn dies nicht zulässig ist, so muß diese Abfrage in einer Programmier- oder Makrosprache des GIS umgesetzt werden, in der einzelne Objekte angesprochen und verarbeitet werden können.

Aus dem ATKIS Datenmodell gibt es auch für diesen Zusammenhang Beispiele. Zwischen den Attributwerten der Streckengleiszahl (GLS: 1000 = eingleisig, 2000 = zweigleisig), der Breite des Verkehrsweges (BRV) und Anzahl der Gleise auf dem Bahnkörper (GLZ) bestehen die Beziehungen

$$3 \cdot \mathrm{GLZ} + 6 = \mathrm{BRV} \quad \text{und} \quad \mathrm{GLZ} = \mathrm{GLS} / 1000 \,.$$

Im Normalfall bezieht sich das Attribut Streckengleiszahl (GLS) auf das Teilobjekt Schienenbahn. Die Attribute GLZ und BRV werden bei dem Teilobjekt Bahnkörper geführt. Wenn eine komplexe Modellierung erforderlich ist, weil z.B. mehrere Bahnstrecken auf einem Bahnkörper verlaufen, dann bezieht sich das Attribut GLS auf das komplexe Objekt Schienenbahn. Das komplexe Objekt Schienenbahn umfaßt die Objekte Bahnkörper und Bahnstrecke. Die Objekte und deren Objektteile stehen also nur über das übergeordnete komplexe Objekt in Beziehung.

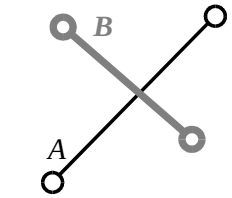


6.2.3 Attributregeln mit topologischen Beziehungen zwischen Objekten

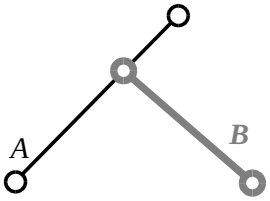
Zur Beschreibung der Topologie stellten *Egenhofer, Mark und Herring 1994,* eine 3x3-Matrix auf, mit der alle möglichen Fälle der topologischen Beziehungen zwischen Objekten abgedeckt sind. Die *ebenda* eingeführte Notation zur Beschreibung der topologischen Beziehungen zwischen Objekten ist in Appendix A zusammengefaßt.

Die Elemente finden sich in raumbezogenen Operatoren des GIS wieder. Zum Auffinden von Objekten, deren Attributwerte sich widersprechen, weil die zugehörigen Objekte in einer bestimmten topologischen Beziehung zueinander stehen, müssen Abfragen formuliert werden, die sowohl diese Operatoren verwenden als auch auf Attributwerte dieser Objekte zugreifen können.

Die Beschreibung von Referenzen zwischen Objekten durch Attribute (Abschnitt 2.2.3.3) gehört zu dieser Kategorie von Attributregeln.

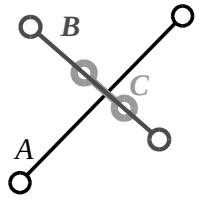


$$R(A,B) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} B^0 & \partial B & B^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix} \end{array}$$



$$R(A,B) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} B^0 & \partial B & B^- \end{array} \\ \begin{pmatrix} \emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix} \end{array}$$

Abbildung 37: Topologische Beziehung zweier linienhafter Objekte mit möglicher gegenseitiger Referenzierung.



$$R(A,B) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} B^0 & \partial B & B^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix} \end{array} \quad R(A,C) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} C^0 & \partial C & C^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix} \end{array} \quad R(B,C) = \begin{array}{c} \\ B^0 \\ \partial B \\ B^- \end{array} \begin{array}{c} \begin{array}{ccc} C^0 & \partial C & C^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix} \end{array}$$

Abbildung 38: Standardfall für die Referenzierung von drei linienhaften Objekten.

Bei einer Beteiligung von drei linienhaften Objekten ist ein Objekt zwischengeschaltet, das nach beiden Richtungen Referenzen besitzt (z.B. Objekte der Objektklasse „Brücke“). Dabei müssen neben dem Standardfall (Abbildung 38) weitere Konstellationen berücksichtigt werden. Diese unterscheiden sich alle in ihren topologischen Beziehungen (Abbildung 39). Teilweise sind vier oder fünf Objekte beteiligt.

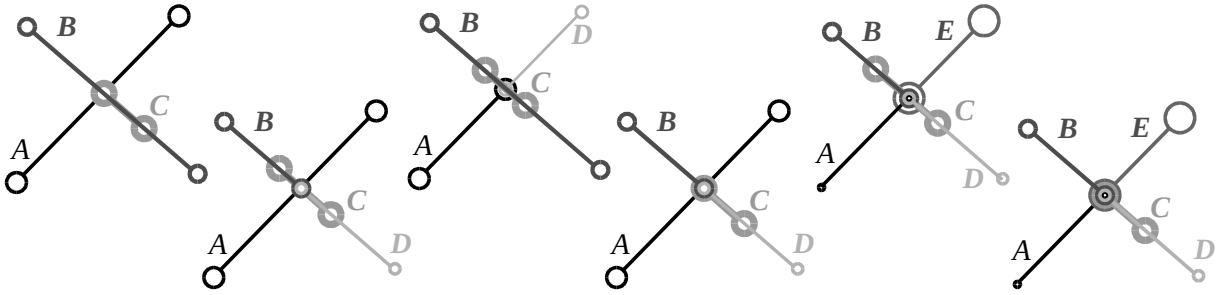


Abbildung 39: Allgemeinfälle von linienhaften Objekten mit potentieller Referenzierung.

Die gegenseitige Referenzierung von Objekten ist eine Attributbedingung, da die Referenz nach unten oder nach oben als Attribut zu den für Referenzierungen in Frage kommenden Objektklassen modelliert werden kann. Die Vergabe von Attributwerten ist nur sinnvoll, wenn die Objekte in einer der skizzierten topologischen Beziehungen stehen.

Eine weitere Gruppe von Attributbeziehungen ergibt sich durch die Nachbarschaft der Objekte. Manche Zusammenhänge zwischen Attributwerten schließen sich aus oder sind zwingend, wenn die

Objekte direkt nebeneinander liegen. Diese Konstellation kann sich sowohl auf punkt-, linien- oder flächenhafte Objekte beziehen.

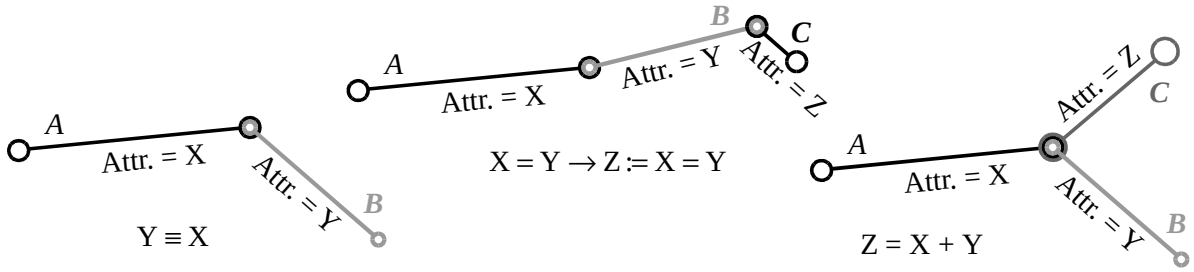


Abbildung 40: Beispiele für Beziehungen zwischen Attributen, deren linienhafte Objekte benachbart sind.

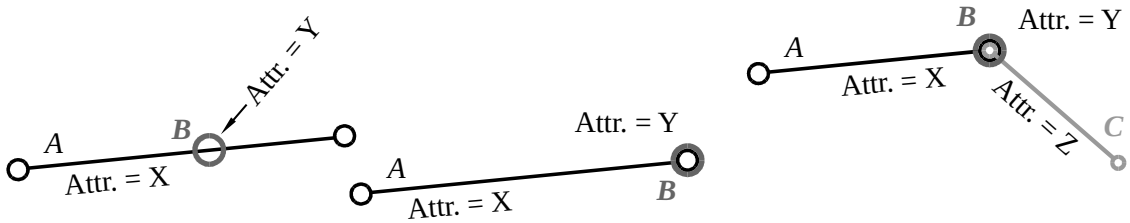
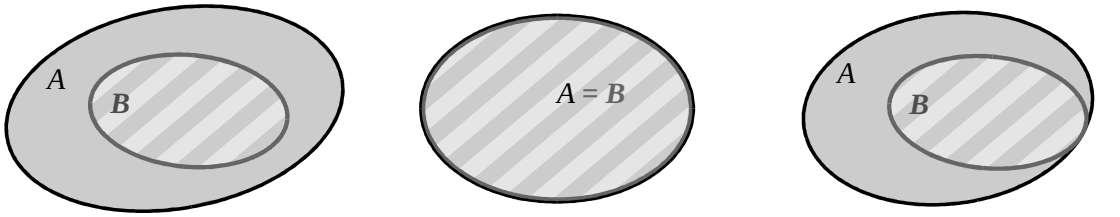


Abbildung 41: Beziehungen von Attributen, deren punkt- und linienhafte Objekte benachbart sind.

Als Beispiel kann das Attribut „Verbindungsart“ für Rohre aus dem Leitungskataster genannt werden. Zwischen zwei Leitungen unterschiedlichen Durchmessers mit dem Wert für das Attribut Verbindungsart „geschraubt“, muß ein Objekt der Objektklasse „Übergang“ existieren und dieses darf nicht den Wert „geschweißt“ für die Verbindungsart besitzen.

6.2.4 Objektregeln mit topologischen Beziehungen zwischen Objekten

Aufgrund der Semantik sind nur bestimmte topologische Beziehungen zwischen Objekten verschiedener Objektklassen zulässig. Dabei ist es möglich, daß Objekte nicht selbständig, sondern nur in Verbindung mit einem Objekt bestimmter anderer Objektklassen vorkommen dürfen. Das heißt das Objekt muß innerhalb dieses anderen Objekts liegen. Für innerhalb gibt es topologisch drei Möglichkeiten, die in Abbildung 42 dargestellt sind. Darunter befinden sich die zugehörigen Matrizen der 9-Intersection.

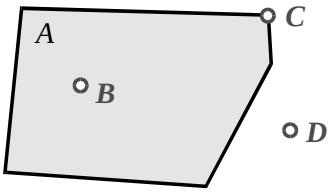


$$R(A,B) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} B^0 & \partial B & B^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix} \end{array} \qquad R(A,B) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} B^0 & \partial B & B^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \emptyset & \emptyset \\ \emptyset & \neg\emptyset & \emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix} \end{array} \qquad R(A,B) = \begin{array}{c} \\ A^0 \\ \partial A \\ A^- \end{array} \begin{array}{c} \begin{array}{ccc} B^0 & \partial B & B^- \end{array} \\ \begin{pmatrix} \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix} \end{array}$$

Abbildung 42: Objekt *B* liegt innerhalb von Objekt *A*: drei Möglichkeiten der topologischen Beziehung.

Auch der Punkt-im-Polygon-Test (*Bill, 1996*), der in GIS sehr häufig benötigt wird, z.B. bei der Auswahl von Objekten am Bildschirm, kann mit Hilfe des 9-Intersection Modell dargestellt werden. Effektive Algorithmen zur analytischen Lösung dieses Tests finden sich in *Aumann und Spitzmüller,*

1993, oder *Meier, 1986*. Für die Prüfung der konzeptionellen Konsistenz wird dieser Test auch benötigt. Zum Beispiel darf ein Punkt *B*, der innerhalb eines Flurstückes liegt, kein Grenzpunkt sein.



$$R(A,B)=\begin{pmatrix}\neg\emptyset & - & \neg\emptyset\\ \emptyset & - & \neg\emptyset\\ \emptyset & - & \neg\emptyset\end{pmatrix}\quad R(A,C)=\begin{pmatrix}\emptyset & - & \neg\emptyset\\ \neg\emptyset & - & \neg\emptyset\\ \emptyset & - & \neg\emptyset\end{pmatrix}\quad R(A,D)=\begin{pmatrix}\emptyset & - & \neg\emptyset\\ \emptyset & - & \neg\emptyset\\ \neg\emptyset & - & \neg\emptyset\end{pmatrix}$$

Abbildung 43: Punkt-im-Polygon-Test.

Die neun in Abbildung 44 dargestellten Beziehungen zwischen flächen- und linienhaften Objekten zeigen nur eine Auswahl der möglichen Fälle. Sie beziehen sich auf einfache geradlinige Objekte ohne weitere Stützpunkte außer den Anfangs- und Endknoten. Die Konstellationen zu *A*, wie sie bei den Objekten *E*, *F* und *G* gegeben sind, können auf die Objektbildung Auswirkung haben, da durch sie neue Maschen gebildet werden. Gehören z.B. die Objekte *E*, *F* und *G* im ATKIS-Datenmodell der Objektart „Straße“ an, so muß das Objekt *A* bei einer Zugehörigkeit zu „Wohnbaufläche“ in separate Objekte aufgeteilt werden. Bei „Industrie- und Gewerbefläche“ ist eine Aufteilung in Objekte und auch Objektteile zulässig. Auf diese Weise können Inkonsistenzen ausfindig und automatisch repariert werden.

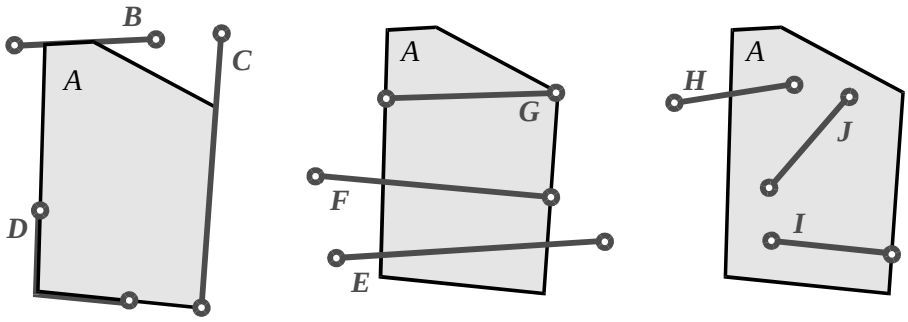


Abbildung 44: Auswahl von einfachen Konstellationen zwischen linien- und flächenhaften Objekten.

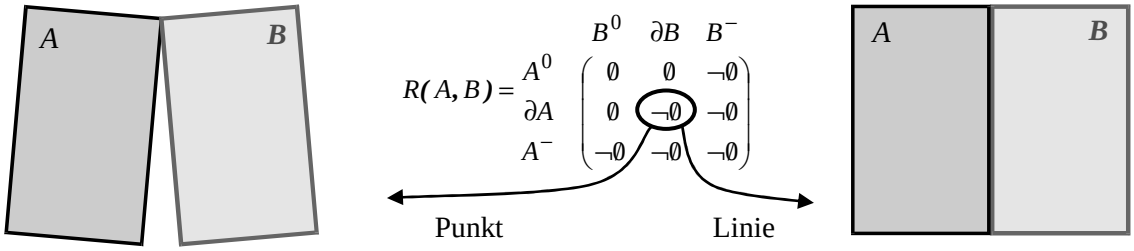


Abbildung 45: Unterschiedliche Berührungsmengen von flächenhaften Objekten führen zur gleichen 9-Intersection.

Mit dem 9-Intersection Modell kann nicht unterschieden werden, ob sich flächenhafte Objekte in einem Punkt oder in einer Kante berühren. Die Schnittmenge der Ränder ist in beiden Fällen nicht leer. Die Elemente der Matrix sind nur als Indikatoren gedacht, die eine leere oder nicht leere Schnittmenge anzeigen. Aus diesem Grund wurde von *Clementini und Di Felice, 1994*, das um die Dimension erweiterte 9-Intersection Modell eingeführt. Dabei wird nicht nur zwischen leere oder nicht leere Menge unterschieden, sondern die Dimension der Schnittmenge angegeben.

6.3 Einfluß der Datenverwaltung auf Konsistenzprüfungen

Die Art der Datenhaltung hat auf die Prüfung der Konsistenz einen großen Einfluß. Bei einer Verwaltung der Geodaten in abgeschlossenen Gebieten mit scharfer Randbegrenzung (Kacheln) treten andere Inkonsistenzen auf als bei einer blattschnittfreien Datenhaltung. Wenn die Abgabe von Geodaten sich auf Teilgebiete oder thematische Extrakte aus dem Gesamtdatenbestand beziehen, dann

können bei Erstellung dieser Auszüge Datenfehler entstehen, die im ursprünglichen Datenbestand nicht vorhanden waren. Für den Anwender oder Käufer ist aber die Konsistenz des erworbenen Datenbestandes von Interesse. Der Datenproduzent oder -vertrieb muß diese Fehlerquellen bei der Abgabe von Geodaten berücksichtigen.

6.3.1 Kachelung

Eine Verwaltung der Geodaten in scharf gegeneinander abgegrenzten Gebieten (Kacheln), zum Beispiel bezogen auf Blattschnitte von topographischen Karten oder Katasterkarten, führt implizit zu einer Verletzung der Objektbildungsregeln. Die Objekte werden dabei nicht aufgrund ihrer Semantik begrenzt, sondern zusätzlich durch willkürliche, fiktive Grenzen. Allerdings haben die Kacheln den Vorteil, daß sie abgeschlossene Einheiten darstellen, die in sich konsistent sein müssen. Wenn nur vollständige Kacheln an Anwender abgegeben werden, so bleibt die Konsistenz der Daten beim Transfer erhalten, vorausgesetzt das Schnittstellenformat oder die Ausgabe- und Einleseprogramme erzeugen keine Datenverluste.

Für den Anwender bringt die Kachel als Erfassungseinheit den Nachteil, daß ein abstraktes Objekt in mehreren digitalen Objekten wiedergegeben werden kann. Das abstrakte Objekt, aufgeteilt auf mehrere Dateien, kann dabei mehrere Objektidentifikatoren erhalten oder der Datenproduzent stellt sicher, daß für jedes abstrakte Objekt immer nur ein Objektidentifikator existiert, der allen Fragmenten dieses Objekts zugewiesen wird. Da letztere Methoden einen sehr hohen organisatorischen Aufwand vor allem bei der Fortführung von Geodaten bedeutet, wird im allgemeinen die Verletzung der Objektbildungsregeln zugunsten eines leichter konsistent zu haltenden Datenbestandes hingenommen.

Einfache Analysen wie z.B. die statistische Abfrage der durchschnittlichen Fläche aller Objekte der Objektklasse Ackerland werden aufgrund dieser Art der Datenhaltung verfälscht. Für Anwendungen, die eine bijektive Abbildung zwischen abstrakten Objekten und digitalen Objekten voraussetzen, kann das Objekt durch topologische Operatoren und attributive Abfragen rekonstruiert und durch Verschmelzung der Geometrie wieder zusammengeführt werden. Benachbarte Objekte derselben Objektklasse, die in allen Attributen außer dem Objektidentifikator übereinstimmen, und die nur durch den jeweiligen Blattschnitt voneinander getrennt sind, dürfen im Anwendungssystem vereinigt werden. Mit dieser Manipulation der Daten kann die Datenhaltung in Kacheln in eine blattschnittfreie Datenhaltung überführt werden. Da eine differentielle Fortführung nur über Objektidentifikatoren realisiert werden kann, müssen im Fortführungsfall alle Änderungen an den einzelnen Objekten zurückverfolgt werden können.

6.3.2 Abgabeeinheiten

Bei der Abgabe von Teilgebieten oder thematischen Extrakten des Gesamtdatenbestandes können Inkonsistenzen entstehen, die nur durch die Auswahl der Objekte erzeugt worden sind.

Wenn die Abgabeeinheiten im Falle der Kachelung nicht mit den Erfassungseinheiten übereinstimmen, oder bei blattschnittfreier Datenhaltung, wenn die Objekte abgegeben werden, die ganz oder teilweise innerhalb des Interessengebietes liegen, so ragen Teile der Objekte aus diesem Gebiet hinaus, und die Objekte, mit denen sie eine topologische Beziehung haben können, sind nicht im abgegebenen Datenbestand vorhanden.

Auch bei einer thematischen Auswahl von Objekten bestimmter Objektklassen fehlen unter Umständen die Objekte, zu denen diese nach dem konzeptionellen Modell in dort festgelegter Beziehung stehen müssen. Beispielsweise können Referenzen auf Objekte bestimmter Objektklassen angegeben sein, die nicht zu der Auswahl gehören. Oder es wird eine vollständige Überdeckung der gesamten Fläche des Interessengebietes verlangt, aber nicht alle Objektklassen, die zu dieser Überdeckung beitragen sollen, sind in dem Auszug aus dem Datenbestand vorhanden.

Durch die Selektion der Objekte für die Abgabe können Fehler entstehen, die in den ursprünglichen Daten nicht vorhanden waren. Diese Fehler werden bei einer Prüfung anhand des Regelwerkes aufgedeckt. Sie treten allerdings nur bei isolierter Betrachtung einer einzelnen Abgabeeinheit auf. Bezieht ein Kunde alle Abgabeeinheiten, so daß er wieder den Gesamtdatenbestand aufbaut, so müssen

bei der Prüfung einer Abgabeeinheit alle benachbarten Einheiten mit herangezogen werden, oder die Prüfung muß auf den zusammengeführten Datensatz angewandt werden.

Durch eine Prüfung von Teilgebieten können tatsächlich in den Daten vorhandene Inkonsistenzen übersehen werden. Diese Fehler treten dann in den Randbereichen auf. Diese Bereiche sind vor allem deshalb sehr sensibel, weil hier bei der Datenerfassung eine enge Absprache zwischen den erfassenden Institutionen erfolgen muß. Daher sind diese Bereiche auch besonders fehleranfällig und es ist deshalb kritisch, wenn die Methode zum Aufdecken dieser Fehler an diesen Stellen versagt. Betroffen sind die Regeln der vollständigen Überdeckung und der verbotenen Überlagerung.

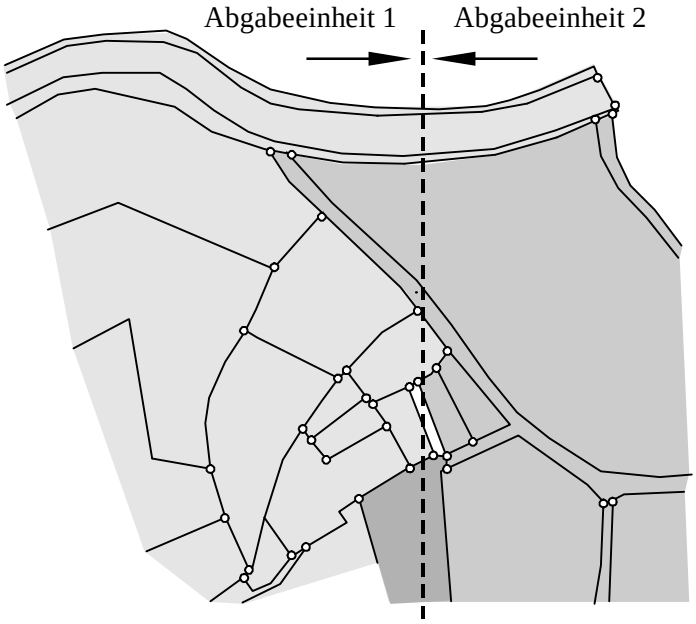


Abbildung 46: Problematik der Konsistenzprüfung am Rand von unregelmäßig begrenzten Erfassungs- oder Abgabeeinheiten.

In dem Beispiel für Flurstücke in Abbildung 46 wurde ein Flurstück im Randbereich nicht erfaßt, ein anderes Flurstück wurde sowohl in der linken als auch in der rechten Erfassungs- bzw. Abgabeeinheit digitalisiert. Diese Fehler können bei getrennter Betrachtung dieser Datensätze nicht aufgedeckt werden, sondern nur bei gleichzeitiger Verwendung aller angrenzender Objekte aus den benachbarten Abgabeeinheiten.

6.4 Formaler Regelkatalog FRACAS

Die Prüfroutinen zur Kontrolle der konzeptionellen Konsistenz können, wenn das GIS die erforderlichen Funktionalitäten bereitstellt, mit Hilfe der Analysewerkzeuge des GIS umgesetzt werden. Damit derjenige, der eine Schemaanpassung durchführt, keine Programmierkenntnisse benötigt, ist es erforderlich, daß eine Trennung zwischen der Formulierung der Regeln für eine Anwendung und der Umsetzung dieser Regeln in Prüfroutinen stattfindet. Diese Trennung kann über eine Steuerdatei erfolgen, mit welcher die Inhalte der Regeln verändert werden können. Diese Steuerdatei kann entweder tabellarisch aufgebaut sein, oder durch einen allgemeinverständlichen Syntax beschrieben werden.

Eine solche formale Beschreibung der Regeln ist durch den Regelkatalog FRACAS (***F**ormal **R**ules for **A**ssessing the **C**onsistency with respect to **A**pplication **S**chemata*) gegeben (*Joos, 1999*). Diese Art der Steuerung von Kosistenzprüfungen hat gegenüber einer festen Einbindung der Regeln in die Makrosprache des GIS entscheidende Vorteile.

- Anwender erhalten eine leicht verständliche Schnittstelle zur Steuerung der Konsistenzprüfungen.
- Der Regelkatalog läßt sich in beliebige Landessprachen umsetzen, ohne daß die Prüfroutinen verändert werden müssen.
- Der Regelkatalog kann an unterschiedliche Datenschemata angepaßt werden.

- Bei der Fortführung des Datenschemas ist keine Programmierung erforderlich.
- Die Programme sind auf unterschiedliche Anwendungen übertragbar.
- FRACAS kann in die Makrosprachen verschiedener GIS überführt werden.
- Eine Vergleichbarkeit von Konsistenzprüfungen ist gewährleistet, auch wenn sie in verschiedenen Systemen durchgeführt wurden.
- Der Regelkatalog muß bei einem Systemwechsel nicht neu erstellt werden.
- Während der automatischen Prüfung wird ein Fehlerbericht erzeugt.
- Es besteht die Möglichkeit zur interaktiven Nachbearbeitung der fehlerhaften Objekte durch die Ablage in einer Queue.

6.4.1 Struktur des Regelkatalogs

Der Regelkatalog ist aus drei Typen von Elementen aufgebaut: der Schemaanpassung, den Regelblöcken und den Statistikblöcken. Diese Blöcke geben dem Regelkatalog zur besseren Lesbarkeit eine Struktur und steuern die Ausgabe des Fehlerberichts.

In der Schemaanpassung werden die Namen der Attribute für benutzerdefinierte Objektidentifikatoren und für Referenzattribute vereinbart. Da diese Attribute in jedem Datenschema anders genannt werden können, ihre Namen aber für die Fehlerausgabe benötigt werden, müssen sie im Kopf des Regelkatalogs vereinbart werden. Der Objektidentifikator und die Referenzattribute müssen für alle Objektklassen innerhalb des Datenschemas identisch benannt sein.

Nach der Schemaanpassung können beliebig viele Regelblöcke und Statistikblöcke in beliebiger Reihenfolge folgen. Die Regeln und Statistiken beziehen sich auf Objektklassen und Attributwerte verschiedener Hierarchiestufen und deren Beziehungen zueinander. Wenn mehrere Objektklassen oder Attribute angesprochen werden sollen, können diese in einer Objektklassenliste oder Attributliste zusammengefaßt werden. Eine solche Liste wird durch geschweifte Klammern ({,}) gegrenzt und die Elemente durch Kommata getrennt.

Am Anfang eines Regelblocks wird vereinbart, ob die nachfolgenden Regeln einen Fehler oder einen Mangel ausgeben sollen. Der Anwender der Prüfsoftware kann dabei selbst entscheiden, welche Regelverletzung er als Warnung oder Hinweis und welche er als echten Fehler eingestuft haben möchte. Die Unterscheidung sollte dabei nach folgender Einteilung getroffen werden (*ISO 8402, 1991*):

- Fehler: Die Nichterfüllung festgelegter Forderungen
- Mangel: Die Nichterfüllung von Forderungen im Hinblick auf den beabsichtigten Gebrauch.

Innerhalb eines Regelblocks können beliebig viele, auch unterschiedliche Regeln aufgelistet werden. Die einzelnen Regeln werden durch ein Semikolon beendet. Zwischen und innerhalb des Regelblocks können Kommentare eingefügt werden.

Der Statistikblock dient zur Ausgabe des Anteils der mit Attributwerten belegten Objekte einer Objektklasse. Nicht alle Attributwerte sind nach dem Datenmodell zwingend erforderlich. Ein Anwender der sich speziell für ein optionales Attribut interessiert, kann dadurch den Befüllungsgrad dieses Attributes ablesen. Auch zur Planung von Nacharbeiten ist diese statistische Angabe nützlich.

6.4.2 Das Regelwerk

Der Regelkatalog FRACAS bezieht sich auf die Prüfung der konzeptionellen Konsistenz beschrieben in Abschnitt 6.2. Im Regelkatalog werden diese Regeln in einer einfach verständlichen Form definiert. Die eigentliche Prüfung und die Prüfroutinen, die hinter diesen Regeln stecken, werden vom Anwender unbemerkt durch einen Regelcompiler in die Abfrage- und Programmiersprache der GIS-Software übersetzt und dort ausgeführt. Das Regelwerk kann daher als eine Definitionssprache bezeichnet werden.

Die folgenden Bedingungen geben die Syntax und die Grammatik für die Abfrage auf Einhaltung der Regeln wieder. Die Regeln sind primär aus dem ATKIS-Datenmodell abgeleitet, lassen sich aber auf andere Datenmodelle anwenden und gegebenenfalls erweitern.

- Attributvergabe

```
Objektklasse benoetigt_Werte_fuer Attributliste | Attributliste wenn
                Attribut = Wert | Attributliste wenn Attribut != Wert;
```

Bei zwingend erforderlichen Attributwerten muß geprüft werden, ob diese für alle Objekte einer Objektklasse eingetragen sind (Abschnitt 6.1.1). Der Eintrag kann an eine Bedingung geknüpft werden, die von einem anderen Attributwert desselben Objekts abhängig ist. Der Wert kann dabei je nach Typ des Attributes ein ganzer oder rationaler Zahlenwert sein, eine Zeichenkette oder NULL (als Zeichen dafür, daß kein Wert zugeordnet ist).

- Wertebereiche

```
Objektklasse hat_Wertebereich_fuer Attribut in [Zahl1, Zahl2] |
                oder [Zahl3, Zahl4] | oder ...;
```

Die Attributwerte machen nur in einem bestimmten Bereich Sinn. Die Abfragen, die hinter dieser Regel stehen werden in Abschnitt 6.1.1 ausführlich diskutiert.

- Verbotene Überlagerung

```
{Objektklassenliste} darf_nicht_ueberlagern;
```

Alle Objekte der in der Liste angegebenen Objektklassen dürfen sich gegenseitig und sich selbst nicht überlagern.

- Notwendige Überlagerung

```
{Objektklassenliste1} innerhalb | ausserhalb {Objektklassenliste2}
                muss_ueberlagert_werden_von {Objektklassenliste3};
```

Verschiedene Objekte dürfen als Einzelobjekt nicht auftreten. Sie müssen immer vollständig mit Objekten bestimmter anderer Objektklassen überlagert sein. Beispiele solcher unselbständiger Objekte sind in ATKIS „Bergbaubetrieb“ oder „Kläranlage“. Sie dürfen nur auftreten, wenn sich „darunter“ ein „Industriegebiet“ befindet, und innerhalb von Ortslagen sind auch „Flächen gemischter Nutzung“ oder „Flächen besonderer funktionaler Prägung“. Aus diesem Grund wurde die Unterscheidung zwischen innerhalb oder außerhalb eingeführt. Die Objekte der Liste 3 bilden praktisch die dritte Schicht der Überlagerung und steuern diese Bedingung.

- Vollständigkeit der Überdeckung

```
Vollstaendige_Ueberdeckung_durch {Objektklassenliste1}
                [ausser_wenn_benachbart {{Objektklassenliste2}];
```

Bei manchen Themen soll zu jedem Punkt des Erfassungsgebietes eine Aussage gemacht werden. Das Datenmodell muß in diesem Fall so aufgebaut werden, daß alle Flächeneigenschaften, die in der realen Welt vorkommen können, durch das Modell abgedeckt sind. Die zugehörigen Objektklassen müssen dann in die Objektklassenliste1 aufgenommen werden. Es darf keine Masche existieren, der nicht genau ein Objekt der Objektklassen aus Liste1 zugeordnet ist.

Das ATKIS-Konzept sieht vor, daß komplexe Straßen (also Straßen, bei denen die beiden Richtungsfahrbahnen baulich voneinander getrennt sind, wie z.B. Autobahnen) im Regelfall durch die Straßenachse (Straßenkörper) und die Achsen der Fahrbahnen als linienhafte Objekte erfaßt werden. Zwischen diesen Linienobjekten entstehen im Datensatz Leerflächen, denen keine Nutzung zugewiesen werden kann, da sie in der realen Welt zu den Fahrbahnen gehören. Werden in die Objektklassenliste1 alle Objektklassen der Flächennutzung aufgenommen, so würde das Prüfprogramm bei allen komplexen Straßen einen Fehler finden, der aber wegen dieser Art der Modellierung keinen Fehler darstellt. Aus diesem Grund wurde für die Prüfung der vollständigen Überdeckung die Objektklassenliste2 als Ausnahme zugelassen. Für das genannte Beispiel kann der vermeintliche Fehler

ausgeschlossen werden, wenn die Objektklasse Straßenkörper als Angrenzung zu erlaubten Leerflächen in die Liste2 aufgenommen wird.

- Referenzierung

Wegen der zweidimensionalen Modellierung von ATKIS können Über- und Unterführungen nur durch Zusatzinformationen abgeleitet werden. Diese Zusatzinformationen sind in ATKIS über Attribute, den sogenannten Referenzattributen modelliert. Es wird zwischen Referenz nach oben und nach unten unterschieden. Der Attributwert ist der Objektidentifikator des Objekts, dessen Lage im Raum über oder unter dem referenzierenden Objekt liegt. Im ATKIS-Modell sind zwei Typen vorgesehen: Referenzen, bei denen zwei oder drei Objekte beteiligt sind. Bei drei beteiligten Objekten gibt es ein zwischengeschaltetes Objekt, z.B. Brücke, das sowohl eine Referenz nach oben als auch nach unten hat.

Die Referenzierung muß immer gegenseitig sein. Die Über- oder Unterführung dieser Objekte muß topologisch möglich sein, d.h. sie müssen mindestens einen gemeinsamen Punkt besitzen. Während in manchen Bundesländern nur Objekte referenziert werden, die sich echt kreuzen (*overlap*), ist in anderen eine Berührung ausreichend (siehe dazu auch Abbildung 37). Dies erschwert die Prüfung. Durch Verwendung der stärkeren Bedingung, nämlich die Kreuzung, wird verhindert, daß Konstellationen als Fehler ausgegeben werden, bei denen keine Referenzierungen erforderlich sind.

```
{Objektklassenliste1} hat_Referenz_nach_oben {Objektklassenliste2} unten
                {Objektklassenliste3};
```

Diese Regel des Regelkatalogs beschreibt die Referenzierung mit drei Objekten. Das Objekt aus der Objektklassenliste1 stellt dabei das zwischengeschaltete Objekt dar. Jedes Objekt der Objektklassen aus Liste1 muß nach unten referenziert sein, wenn es ein Objekt der Objektklassen aus Liste2 kreuzt. Und es darf zu keinem anderen Objekt nach unten referenziert sein, dessen Objektklasse nicht in Liste2 aufgeführt ist, oder dessen Objektklasse zwar in Liste2 enthalten ist, das aber das zwischengeschaltete Objekt nicht kreuzt.

```
Objektklasse referenziert_oben {Objektklassenliste1} ohne
                {Objektklassenliste2} [ausser_wenn Attribut=Wert];
```

Diese Regel der direkten Referenzierung bezieht sich auf Referenzen zwischen genau zwei Objekten. Zum Ausschluß, daß kein drittes Objekt an der Konstellation beteiligt ist, können die Klassen, die Einfluß haben könnten, in Liste2 explizit ausgeschlossen werden. Da Ausnahmen bestehen können, die vom Attributwert der zu referenzierenden Objektklasse abhängen, kann optional eine Zusatzbedingung angegeben werden.

6.4.3 Beispiele

Die folgenden Beispiele zeigen Datenfehler auf, wie sie in einer ATKIS-Abgabeeinheit typischerweise mehrere Male vorkommen. Die Fehler wurden mit FRACAS ermittelt und in einer Queue abgelegt.

Jedes Beispiel wird der Regel zugeordnet, gegen die bei der Erfassung verstoßen wurde, und es wird versucht, eine mögliche Fehlerursache anzugeben.

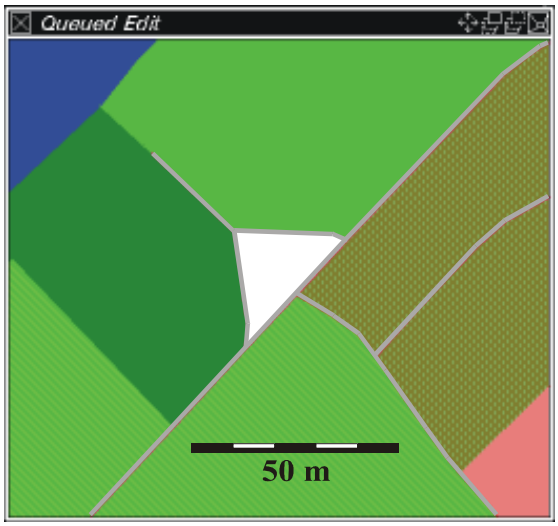


Abbildung 47: Fläche in einer Wegegabelung ohne Zuweisung einer Nutzungsart.

Abbildung 48: Abgetrennte Fläche ohne Nutzungsart.

Nach dem ATKIS-Datenmodell müssen allen Maschen, die sich aus dem Netz der linienhaften Objekte ergeben, Nutzungsarten zugewiesen werden. Bei kleinen Maschen, wie zum Beispiel auf Verkehrsinseln oder zwischen Wegegabelungen wird diese Zuweisung häufig vergessen (Abbildung 47). Wenn die Erfassungsquelle keine Aussage macht oder schwer zu interpretieren ist, kommt es vor, daß der Erfasser unsicher wird und diese Fläche ausläßt, um den Sachverhalt zu einem späteren Zeitpunkt oder gemeinsam mit einem Kollegen zu klären. Danach wird vergessen, diese Fläche zu erfassen. Es liegt ein Verstoß gegen das Gebot der vollständigen Überdeckung des Erfassungsgebietes vor.

Im Fall der Abbildung 48 ist ein Flächenobjekt ohne ersichtlichen Grund geradlinig abgeschnitten. Die Datenerfassung erfolgt in diesem Fall zwar blattschnittfrei, aber die analogen Datenquellen beziehen sich im allgemeinen auf nach Koordinatenlinien zugeschnittene Gebiete. Der Bearbeiter mußte daher das flächenhafte Objekt am Blattrand abschließen. Beim Wechsel zur benachbarten Erfassungsquelle, hätte der restliche Teil des Objektes erfaßt und mit dem schon vorhandenen Objekt verschmolzen werden müssen. Bei der Bearbeitung des oberen Bereiches wurde das Objekt vergessen zu digitalisieren. Dieses „Loch“ widerspricht der geforderten vollständigen Überdeckung mit Objekten der Flächennutzung.

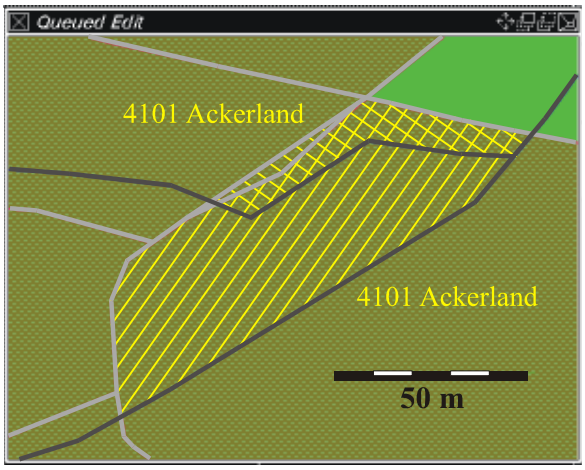


Abbildung 50: Fläche, der zweimal die Nutzung als Ackerland zugewiesen wurde.

Abbildung 49: Überschneidung von See und Wald.

Nach dem ATKIS-Objektartenkatalog, gibt es Objektklassen, die sich gegenseitig ausschließen. Insbesondere verbietet es die Semantik der Objekte, daß einer Fläche zweimal ein Objekt derselben Objektklasse zugewiesen wird. In Abbildung 50 ist ein Fall dargestellt, bei dem sich zwei Objekte der Objektklasse Ackerland überlagern. Die Flächen der Objekte sind mit gegeneinander gedrehten Schraffuren dargestellt. Die Fläche, bei der zwei Zuweisungen erfolgten, ist doppelt schraffiert. Der Flächeninhalt beträgt 700 m^2. Der Erfasser hat vermutlich beim Digitalisieren den Überblick verloren, welche Flächen er schon erfaßt hatte, und hat dadurch ein Objekt gebildet, das ein vorhandenes Objekt überlagert. Durch eine Darstellung erfaßter Objekte mit einer auffälligen Flächensignatur kann diese Fehlerquelle reduziert werden.

Die verbotene Überlagerung in Abbildung 49 ist erst bei starker Vergrößerung sichtbar. Auch hier ist der Überdeckungsbereich mit einer Schraffur verdeutlicht. Die Objekte der Objektklassen See und Wald überlagern sich bis zu einem Betrag von 10 cm. Dieser Wert liegt zwar innerhalb der zulässigen Digitalisierunsicherheit, aber aus topologischer Sicht sind die Daten fehlerhaft. Der Fehler kann entstehen, wenn der Rand zu benachbarten Objekten freihändig digitalisiert wird, ohne die Hilfsmittel des Erfassungssystems zur topologisch korrekten Anbindung an vorhandene Objekte zu benutzen.

Die nächsten Beispiele stellen keine Fehler dar, sondern verdeutlichen die Notwendigkeit, die Regeln von FRACAS für bestimmte Konstellationen einzuschränken.

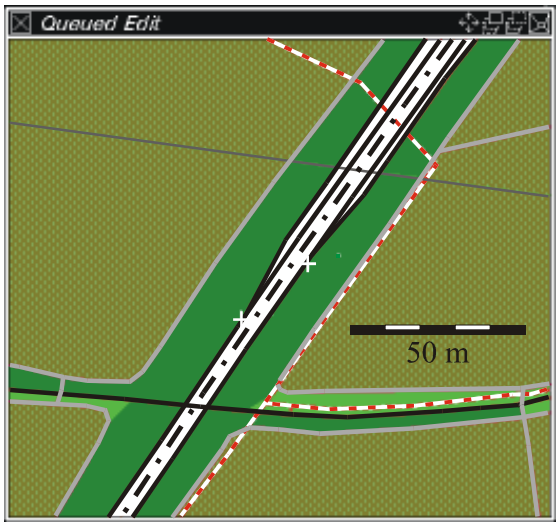


Abbildung 51: Freie Flächen im Bereich von komplexen Straßen.

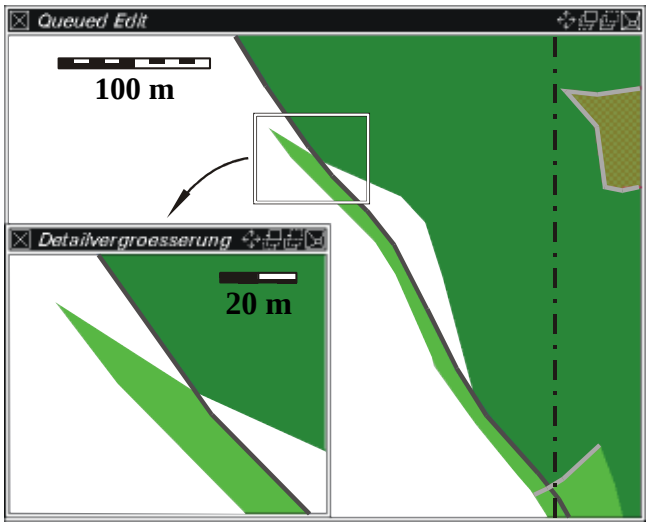


Abbildung 52: Konsistenzbedingungen am Rand von Abgabeeinheiten.

Indem die Straßenfläche nur durch ihre Achse modelliert und erfaßt wird, ergeben sich bei komplexen Straßen zwischen den Fahrbahnen und dem Straßenkörper Flächen, denen keine Flächennutzung zugewiesen werden kann. Aus diesem Grund wurde bei der Definition der vollständigen Überdeckung eine Einschränkung bezüglich benachbarter Objekte eingeführt. Abbildung 51 verdeutlicht den Zusammenhang anhand des Beispieles einer Bundesstraße mit baulich getrennten Fahrstreifen und mit separaten Fahrbahnen zur Beschleunigung.

Abbildung 52 zeigt, wie am Rande von nicht blattschnittbezogenen Abgabeeinheiten vermeintliche Inkonsistenzen entstehen können. Der linienhaft erfaßte Fluß gehört noch zu dieser Einheit, weil er teilweise innerhalb der durch die strichpunktierte Linie angezeigten Gebietsgrenze liegt. Wenn die Objekte der benachbarten Einheiten nicht bei der Prüfung berücksichtigt werden, so findet das Programm an dieser Stelle einen Fehler, weil eine Innenfläche ohne Zuweisung einer Flächennutzung existiert.

7 Stichprobenprüfung

Während sich für die Konsistenzbedingungen eindeutige Regeln aufstellen lassen, deren Einhaltung automatisch mit Hilfe der GIS-Software geprüft werden kann, müssen für die Prüfung der Qualitätskriterien Vollständigkeit, Richtigkeit und Genauigkeit weitere Daten als Referenz herangezogen werden, die ein höheres Qualitätsniveau besitzen. Inwieweit die zu prüfenden Daten mit den Referenzdaten im Rahmen der geforderten Qualitätsziele übereinstimmen, kann in der Regel nicht automatisch untersucht werden, da die Referenzdaten nicht notwendigerweise als Vektordaten digital vorliegen.

Wenn digitale Referenzdaten vorliegen, müssen diese nach dem gleichen Modell erfaßt worden sein wie die zu prüfenden Daten, oder die Modelle müssen zumindest eine nicht leere Schnittmenge besitzen, nämlich die Menge der prüfbaren Objektklassen. Bei einer digitalen Prüfung sind Algorithmen erforderlich, die identische Objekte einander eindeutig zuordnen und bezüglich der Qualitätskriterien miteinander vergleichen. Liegen keine digitalen Referenzdaten vor, dann können diese durch eine Zweitdigitalisierung des Datenbestandes anhand von unabhängigen oder u.U. sogar derselben Datenquellen erzeugt werden.

Der Vergleich zwischen den zu prüfenden Geodaten und den analogen Referenzdaten kann auch visuell erfolgen. Beispielsweise können die Daten in einer für die Prüfung angepaßten Signatur auf transparenten Träger geplottet und mit analogen Datenquellen im selben Maßstab überlagert werden. Oder die Daten werden auf feldtaugliche Rechner überspielt und direkt mit der realen Welt verglichen. Streng genommen findet natürlich kein Vergleich mit der realen Welt sondern auch in diesem Fall mit dem abstrakten Abbild der realen Welt statt, da der Prüfer vor Ort die Welt durch den Filter der Modellierung betrachtet. Wenn der Prüfer ein anderes Objektverständnis hat, so kommt es zu Verfälschungen des Prüfergebnisses. Die Identifizierung der Objekte wird dann über Koordinaten erfolgen, die im Feld mit Hilfe von GPS oder terrestrischer Vermessungsverfahren realisiert werden. Bei sehr hohen Genauigkeitsanforderungen wird sich die Feldprüfung nicht vermeiden lassen.

Bei jeder Prüfung anhand von Referenzdaten ist nachzuprüfen, ob sich das abstrakte Abbild der realen Welt zwischen der Erfassung der Referenzdaten und der zu prüfenden Daten geändert hat. In diesem Fall wird die Prüfung der Daten mit einer Aktualisierung überlagert, was die Beurteilung der Qualität der Daten verzerrt. Wird mit der Prüfung der Daten auch gleichzeitig eine Korrektur durchgeführt, kann diese Methode negative Auswirkungen haben, falls die Referenzdaten älteren Datums sind als die zu prüfenden Daten.

Wenn es digitale Referenzdaten gibt, die ein höheres Qualitätsniveau besitzen, und nach exakt dem selben Modell erfaßt worden sind, ist zu überlegen, ob ein Austausch der Daten gegenüber einer Prüfung nicht wirtschaftlicher ist, auch im Hinblick darauf, daß dann nur noch ein Datensatz fortgeführt werden muß. Wenn allerdings die Referenzdaten mit einer höheren Auflösung erfaßt wurden und das höhere Qualitätsniveau sich daraus ergibt, dann kann diese Auflösung dem Modell der zu prüfenden Daten widersprechen. Diese Referenzdaten werden gegenüber den zu prüfenden Daten teurer in der Anschaffung sein.

Unabhängig, ob eine visuelle Prüfung durch einen Bearbeiter erfolgt, oder ob teuere Referenzdaten zur Kontrolle mit Algorithmen herangezogen werden, die u.U. noch zu entwickeln sind, der zeitliche, personelle und finanzielle Aufwand ist im Vergleich zu der Konsistenzprüfung sehr hoch. Aus diesem Grund wird eine Prüfung des gesamten Datenbestandes ausscheiden. Statt dessen kann für einen repräsentativen Teil des Datenbestandes eine sorgfältige Qualitätsuntersuchung durchgeführt und daraus auf die Qualität des gesamten Datenbestandes geschlossen werden.

Diese Vorgehensweise wird als statistische Qualitätskontrolle oder Stichprobenkontrolle bezeichnet (*Uhlmann, 1982*). Zwei Hauptaufgaben lassen sich unterscheiden: 1. die Kontrolle der Annahme und ebenso der Auslieferung von Waren (Eingangs- und Endkontrolle) und 2. die laufende Kontrolle einer Produktion. Im ersten Fall ist das Ziel, für den Empfänger und den Lieferanten sicherzustellen, daß die Qualität der Geodaten den vereinbarten Bedingungen entspricht. Die laufende Kontrolle soll prüfen, ob die Qualitätsmaße der Objekte innerhalb der zulässigen Grenzen liegen, oder ob in den Produktions-

prozeß eingegriffen werden muß, z.B. durch Nachschulung der erfassenden Personen, durch Kalibrierung der Erfassungshardware, durch Veränderung der Methode oder durch Modifizierung der Software. Die laufenden Kontrollen sind bei dem Qualitätsmanagementsystem der Datenerfassung zu berücksichtigen (Abschnitt 5.3).

Auch Totalkontrollen können nicht immer fehlerlos durchgeführt werden. Ein qualitativ hochwertiger Datensatz kann auch fehlerhafte Objekte enthalten. Daraus folgt, daß ein Vergleich zweier Datenbestände nur die relative Qualität dieser Daten ergeben wird.

Bei einer visuellen Kontrolle kann der Kontrolleur ermüden. Wenn ein Interpretationsspielraum wie z.B. bei unscharfer geometrischer oder thematischer Abgrenzung von Objekten besteht, kann er einseitige Entscheidungen treffen (vermutlich eher zu Ungunsten eines zweifelhaften Objektes um seine Aufgabe als Kontrolleur zu rechtfertigen). Bei verhältnismäßig kleinen Stichproben dagegen ist eine sorgfältige und durch qualifiziertes Fachpersonal ausgeführte Kontrolle der Stichprobenelemente weit eher durchführbar, so daß man auf diese Weise nicht nur Kosten einspart, sondern auch zuverlässigere Ergebnisse erhält.

7.1 Ziel der Stichprobenkontrolle

Bevor über Details wie Stichprobenumfang, Auswahlverfahren oder Signifikanzlevel nachgedacht werden kann, muß geklärt sein, was bei der Stichprobenerhebung geprüft werden soll. Prinzipiell können zwei Ziele formuliert werden, die sich entscheidend auf das Prüfverfahren auswirken.

- Soll die Erfassung als solche kontrolliert werden?
- Sollen die Daten gegenüber der abstrakten Realität kontrolliert werden?

Der entscheidende Unterschied liegt in der Bezugsgröße. Wenn die Erfassung kontrolliert werden soll, so muß der Vergleich zwischen den zu kontrollierenden Daten und den zugehörigen Originalerfassungsquellen erfolgen. Dabei wird davon ausgegangen, daß die Datenquellen die reale Welt hinreichend gut repräsentieren, oder daß nur der Prozeß der Digitalisierung kontrolliert werden soll. Eine Kontrolle der Datenqualität entsprechend der Definition dieser Arbeit (Abschnitt 4.2) liegt dann nicht vor, wird aber in Fällen Anwendung finden, bei denen die Qualität der Produktion eines Werkvertragnehmers beurteilt werden soll oder keine unabhängigen Datenquellen existieren und ein Feldvergleich nicht möglich ist.

Da mit einem Geoinformationssystem Aussagen über die reale Welt gemacht werden sollen, ist die Bezugsgröße zum Vergleich der Geodaten das abstrakte Abbild der realen Welt. In diesem Fall sind für die Kontrolle unabhängige Datenquellen heranzuziehen, da Fehler in den Datenquellen sich unmittelbar auf die digitalen Geodaten auswirken werden und bei der Erfassung durch Fehlinterpretationen oder Fahrlässigkeit weitere Fehler hinzukommen. Die beste Möglichkeit eine unabhängige Kontrolle unter Ausschluß von Vorinterpretationen durchzuführen ergibt sich durch eine Aufnahme oder den direkten Vergleich im Feld (*ground truth*). Je höher der Abstraktionsgrad einer Datenquelle, um so mehr Interpretationen und Klassifizierungen mit allen Möglichkeiten, Fehler zu begehen, haben im voraus stattgefunden. Eine Karte hat z.B. einen höheren Abstraktionsgrad als ein Orthophoto oder ein Luftbild.

Das Problem, das sich ergibt, wenn die Erfassungsquellen unterschiedlichen Aktualitätsstand besitzen wurde im Abschnitt 3.2.6 angesprochen. Vermeiden läßt sich die Fehleinschätzung der Datenqualität aufgrund von Veränderungen in der realen Welt nur dadurch, daß unabhängige Datenquellen mit vergleichbarem Aufnahmedatum, am besten so neu wie möglich verwendet werden, und die Daten permanent aus diesen aktuellen Datenquellen fortgeführt werden. Wann das Datum der Aufnahme verschiedener Erfassungsquellen vergleichbar ist, hängt von der Dynamik der zu erfassenden Objekte ab. Das Aufnahmedatum ist streng genommen immer ein Zeitintervall, da die Aufnahme einen Prozeß mit einer bestimmten Dauer darstellt. Die Aufnahmedauer kann sich von Bruchteilen einer Sekunde bei Luftbildaufnahmen über Minuten bei Fernerkundungsdaten bis hin zu Tagen und Wochen bei Feldvermessungen erstrecken.

7.2 Der Begriff einer Stichprobe

In der Statistik nennt man die Gesamtmenge der Elemente, über die eine Aussage gemacht werden soll, die Grundgesamtheit oder einfach Gesamtheit. Eine endliche Teilmenge, die aus Elementen der Grundgesamtheit besteht und in der alle Elemente mit dem Ziel auf die Grundgesamtheit zu schließen untersucht werden, heißt Stichprobe. Die Anzahl *n* der Elemente der Stichprobe wird als Umfang bezeichnet *(Fisz, 1976)*.

Die Elemente der Gesamtheit kann man hinsichtlich verschiedener Merkmale untersuchen. Ohne Einschränkung der Allgemeinheit soll zunächst das Merkmal *Y* der Elemente dieser Gesamtheit behandelt werden. Dieses Merkmal hat in der Gesamtheit eine Verteilung, die durch die Verteilungsfunktion *F(y)* charakterisiert ist.

Die Eigenschaft *Y* wird für jedes Elements der Grundgesamtheit durch eine reelle Zahl *y* beschrieben. Da bei jedem Geoobjekt bezüglich der Qualitätskriterien Vollständigkeit, Konsistenz, Richtigkeit und Genauigkeit genau festgestellt werden kann, ob jedes einzelne dieser Kriterien erfüllt ist, lassen sich die Objekte der Grundgesamtheit in fehlerhafte und korrekte Objekte bezüglich eines oder beliebiger Kombinationen dieser Kriterien unterscheiden. Diese qualitative Unterscheidung wird durch die Eigenschaft *Y* beschrieben, indem z.B. fehlerhaften Objekten der Wert $y = 1$ und korrekten Objekten $y = 0$ zugewiesen wird.

Die interessierende Eigenschaft der Elemente der Stichprobe wird durch einen *n*-dimensionalen Zufallsvektor $(Y_1, Y_2, ...,Y_n)$ beschrieben, d.h. Y_k $(k=1,2, ..., n)$ ist eine Zufallsvariable, deren Realisierungen y_k, aus den Beobachtungen am *k*-ten Element jeder möglichen *n*-elementigen Stichprobe besteht, die nach einer Zufallsmethode aus der Grundgesamtheit ausgewählt wurde.

Die Auswahlmethode von Elementen in die Stichprobe ist zufällig, wenn

- alle Elemente mit der gleichen Wahrscheinlichkeit gezogen werden.
- das zum Ziehen benutzte Auswahlverfahren im anschaulichen Sinne unabhängig vom interessierenden Merkmal ist.

7.3 Voraussetzungen zur Durchführung einer Stichprobenkontrolle

Da bei einer Stichprobenkontrolle von einer Teilmenge der zu beurteilenden Geoobjekte auf die Gesamtheit geschlossen wird, muß die Gesamtheit der Daten bestimmte Voraussetzungen erfüllen. Wenn diese nicht gegeben sind und bei sehr großen Grundgesamtheiten wird eine Unterteilung in Lose durchgeführt, die für die statistischen Untersuchungen als Grundgesamtheit betrachtet werden.

Ein Los kann aus einem oder mehreren Datensätzen, aus Daten eines definierten Gebietes oder Objekten selektierter Objektklassen als Teilmenge von einem oder mehreren Datensätzen bestehen. Die Größe eines Loses ist wohl abzuwägen (*DIN ISO 2859-0, 1991*). Sie sollte in Einklang mit dem Produktionsprozeß und dem Produzenten stehen, damit eine Größe bestimmt wird, die einvernehmlich als geeignet gilt. Aus dem Blickwinkel der Kontrolle betrachtet sind große Lose von Vorteil, da größere Stichproben gezogen werden können, die zu schärferen Aussagen führen. Der für signifikante Aussagen erforderliche Stichprobenumfang ist in erster Näherung von der Losgröße unabhängig, daher ist das Verhältnis zwischen Stichprobenumfang und Losgröße bei großen Losen günstiger. Allerdings sollten die Lose nicht zu groß gewählt werden, vor allem wenn sie aus verschiedenen Datensätzen zusammengestellt sind, da eine Stichprobenkontrolle über die Annahme oder Ablehnung des gesamten Loses entscheidet.

Eine Stichprobe von Geoobjekten kann als zufällig bezeichnet werden, wenn die Wahrscheinlichkeit, ein fehlerhaftes Objekt anzutreffen, innerhalb der Grundgesamtheit unabhängig von der Lage der gezogenen Objekte gleich ist, also außer in besonders gekennzeichneten Teilgebieten keine Anhäufungen existieren. Mit anderen Worten: als Voraussetzung für die Durchführung von Stichprobenuntersuchungen muß eine Homogenität der Daten bezüglich ihres Fehlerverhaltens gefordert werden, beziehungsweise die Fehlerrate muß für beliebige Teilgebiete identisch sein.

Man hat berechtigte Grund anzunehmen, daß die Daten homogen sind, wenn sie

- auf Grundlage desselben Typs von Datenquellen
- nach derselben Erfassungsmethode
- von derselben Institution (strenggenommen sogar von derselben Person)

erfaßt wurden. Da anhand der Metadaten Gebiete identifiziert werden können, die diese drei Voraussetzungen für Homogenität erfüllen, lassen sich dadurch Lose zur Stichprobenkontrolle zusammenstellen.

Eine Untersuchung auf Homogenität kann mit statistischen Methoden erfolgen. In Abschnitt 7.9 wird dargestellt, wie die Annahme über die Homogenität mit Hilfe eines Hypothesentests anhand der Stichproben verifiziert oder widerlegt werden kann. Da Homogenität für das Verfahren der statistischen Qualitätskontrolle vorausgesetzt wird, gibt dieser Test einen guten Indikator für die Anwendbarkeit der statistischen Qualitätsuntersuchung bei einem bestimmten Datensatz.

Die genannten drei Kriterien stellen allerdings lediglich notwendige Bedingungen für die Homogenität der Fehlerverteilung von Gebieten dar. Eine weitere Bedingung kommt aus den Daten selbst: die u.U. mehrfach zusammenhängenden Gebiete einer Grundgesamtheit müssen gemeinsame Charakteristika aufweisen. Das heißt die Gebiete müssen eine vergleichbare räumliche Verteilung der Objekte und damit eine vergleichbare Informationsstruktur aufweisen. So sollten z.B. nicht Daten eines urbanen Gebietes mit Daten über eine ländliche Struktur zu einem Los zusammengefaßt werden, da in diesen Gebieten unterschiedliche Objektklassen und verschiedene Objektbildungen dominieren werden.

Aus rechtlichen Gründen dürfen statistischen Methoden nicht angewendet werden bei der Prüfung von Daten, auf deren Basis Entscheidungen getroffen werden, die eine Gefahr für Leib und Leben darstellen können. Statistische Methoden sind nämlich nur Stichprobenprüfungen, bei denen nur einige, aber nicht alle Teile eines Loses wirklich geprüft werden (*Brauer u. Kühme, 1996*).

7.4 Auswahl von Stichproben

Die Entscheidungsregeln bei den vorgestellten Verfahren der Stichprobenkontrolle setzen voraus, daß die Auswahl von Untersuchungsobjekten rein zufällig erfolgt. Stichprobenuntersuchungen nach diesem Ideal der **Urnenauswahl** (englisch: *lottery sampling*) können für Geodaten zwar herangezogen werden, sind aber meist sehr teuer. Zur Einsparung von Kosten sollte das Auswahlverfahren bei gleichem Erhebungsnutzen modifiziert werden. Dazu gibt es zwei Wege. Einmal eine bloß technische Modifikation, mit der das Urnenmodell im Prinzip beibehalten, jedoch technisch modifiziert wird, zum anderen eine modellmäßige („peristatische“) Änderung, bei der das Urnenmodell durch kompliziertere Modelle abgelöst wird. Zu den technischen Modifikationen zählen die systematische Auswahl und das Landkartenverfahren. Flächenstichproben, wenn Objekte mehrerer Gebiete zu einer Stichprobe zusammengefaßt werden, sind eine Sonderform der Landkartenverfahren und zählen zu den modellmäßigen Änderungen, deren Effizienz noch eingehend zu untersuchen ist.

Die Entscheidung, welches Verfahren Anwendung finden soll, hängt vor allem vom Prüfvorgang ab. Wenn die Objekte mit der realen Welt vor Ort verglichen werden müssen, so würden bei Einzelziehungen gleichmäßig verteilt auf das Gebiet der Grundgesamtheit erhebliche Fahrtkosten entstehen (z.B. bei der Prüfung von Navigationsdaten für Europa, siehe *Claussen, 1995a, b* und *1996*). Werden die Daten mit anderen Quellen als den Erfassungsquellen verglichen, so können erhebliche Kosten für die Beschaffung dieser Quellen entstehen. Schon das Zusammenstellen der ursprünglichen Erfassungsquellen, um jeweils nur ein oder je nach Stichprobenumfang zufälligerweise auch mehr Objekte zu prüfen, kann zu einem großen Aufwand führen. Manche Fehler sind auch nur dann erkenntbar, wenn die zu prüfenden Objekte im Kontext mit benachbarten Objekten gesehen werden. Eine isolierte Betrachtung des Einzelobjekts ist dabei nicht zulässig. Eine Umsetzung des Urnenverfahrens für die Kontrolle von Geoobjekten ist aber denkbar, wenn Erfassunngsquellen und digitale Daten hybrid dargestellt werden können und die Kontrolle am Bildschirm durchgeführt wird. Das Verfahren setzt einen schnellen Zugriff auf alle Daten voraus.

Das Pendant zum Durchmischen und zufälligen Ziehen aus einer Urne ist in technischen Systemen das Erzeugen von Zufallszahlen. Diese zufälligen Zahlen[6] können verwendet werden, um aus den Ordnungsstrukturen der Geodaten eine Zufallsauswahl zu treffen. Verschiedene Möglichkeiten zur Ziehung von Objekten in Geodaten werden in den Kapiteln 7.4.1 bis 7.4.3 vorgestellt.

Als Stichprobenelement wird dabei immer ein Geoobjekt betrachtet. Dabei müssen nicht alle Objekteigenschaften von Interesse sein. Die Zielsetzung der Prüfung bedingt, ob die Klassenzugehörigkeit, die Geometrie, spezielle oder alle Attributwerte oder Relationen zu anderen Objekten von Interesse sind und nach welchen Qualitätskriterien untersucht werden muß.

7.4.1 Ziehung von Objektidentifikatoren

Eine Möglichkeit beruht auf der Verwendung eines eindeutigen internen oder externen Objektidentifikators. Besteht er aus fortlaufenden Zahlen, wird eine gleichverteilte Zahl zwischen der kleinsten und größten Objektnummer erzeugt, und das zugehörige Objekt zur Prüfung ausgewählt. Der Vorgang wird so lange wiederholt, bis der festgelegte Stichprobenumfang *n* erreicht ist. Bei nicht fortlaufenden Zahlen, wird geprüft, ob zu der gezogenen Nummer ein Objekt mit dieser Objektnummer existiert. Wenn das Objekt existiert, wird es zur Stichprobe hinzu genommen, andernfalls wird diese Zufallszahl verworfen und die nächste erzeugt.

7.4.2 Landkartenverfahren

Zu den Landkartenverfahren zählen Punkt-, Linien-, Routen- und Flächenstichproben. Bei der Realisierung eines Landkartenverfahrens kann das Ordnungsprinzip der 2D-Lagekoordinaten verwendet werden. Das Gebiet des zu prüfenden Loses besitzt eine minimale und maximale Ausdehnung entlang der beiden Koordinatenachsen. Zur Ziehung eines Untersuchungsobjektes werden zwei gleichverteilte Zufallszahlen zwischen den jeweiligen Extremwerten erzeugt. Verlaufen die Begrenzungen des Loses nicht parallel zu den Koordinatenachsen, so muß geprüft werden, ob das Koordinatenpaar zum Gebiet der Grundgesamtheit gehört. Falls der Punkt außerhalb liegt, wird das nächste Paar von Zufallszahlen erzeugt. Wenn der Punkt zum Interessengebiet gehört, wird das Objekt ausgewählt, das am nächsten zu diesem Punkt liegt. Es ist zwar unwahrscheinlich, daß es mehrere Objekte mit gleichem Abstand zu diesem Punkt gibt, aber sollte dieser Fall eintreten, so kann entweder das erste Objekt, das gefunden wurde, oder eine weitere Zufallszahl zwischen Eins und der Anzahl der Objekte mit gleichem Abstand das zu untersuchende Objekt selektieren. Die Abstandsfunktion zwischen linienhaften, flächenhaften und zusammengesetzten, komplexen Objekten zu einem Punkt muß definiert werden.

Zur Linienstichprobe werden zufällige Linien durch das Erhebungsgebiet gezogen und die von den Linien geschnittenen oder tangierten Objekte in die Stichprobe aufgenommen. Für die Auswahl von Geoobjekten kann dieses Verfahren dadurch Vorteile bringen, daß nur Erfassungsquellen entlang dieser Linien zur Kontrolle benötigt werden. Diese Bündelung kann aber zu ungewollten Verzerrungen führen, da keine flächendeckende Kontrolle erfolgt.

Eine Modifikation der Linienstichprobe erhält man, wenn die zufälligen Linien entlang einer Route (z.B. Straßenverlauf oder anderer netzartiger Strukturen) herausgegriffen werden. Für eine Vor-Ort-Prüfung bringt dieses Verfahren natürlich den Vorteil, daß die Abfolge der Prüfungsdurchführung optimiert ist. Allerdings können durch dieses Auswahlverfahren systematische Fehler übersehen werden, insbesondere dann, wenn die Routenplanung auf Basis der Daten erfolgt, die durch Abfahren der Route geprüft werden sollen.

---

[6] Streng betrachtet handelt es sich bei den nach einem Algorithmus erzeugten Zufallszahlen um Pseudozufallszahlen, weil die Algorithmen i.A. periodisch sind, allerdings mit einer so großen Periode, daß sie hier als zufällig betrachtet werden können.

7.4.3 Flächenstichprobe

Bei der Flächenstichprobe wird ein Raster über das Interessengebiet gelegt, und es werden die Objekte einzelner, zufällig ausgewählter Felder als Stichprobe verwendet. Da aber die Objektdichte nicht als konstant betrachtet werden kann, ergibt sich bei vorgegebener Anzahl von Feldern und konstanter Rasterweite ein sich ändernder Stichprobenumfang. Dies kann vermieden werden, indem bei konstanter Rasterweite so lange Felder gezogen werden, bis der erforderliche Stichprobenumfang gerade überschritten wurde. Eine zweite Möglichkeit besteht darin, die Anzahl der zu untersuchenden Gebiete vorab festzulegen, und die Größe der Gebiete so zu variieren, daß gerade die anteilsmäßig erforderlichen Objekte (Stichprobenumfang dividiert durch Anzahl der Untersuchungsgebiete) enthalten sind.

Die Auswahl von mehreren zusammenhängenden Elementen in Gruppen oder Bündeln wird in der Statistik mit **Klumpenauswahl** (englisch: *cluster sampling*) bezeichnet (*Menges und Skala, 1973*). In der Frühzeit der Repräsentativstatistik hielt man die Klumpenasuwahl für prinzipiell schlechter als die reine Zufallsauswahl von individuellen Einheiten. Bald bemerkte man jedoch, daß die Klumpenauswahl oft einen kleineren Stichprobenfehler aufweist. Der Einfluß, den die Klumpung, d.h. die Zusammenlegung von Untersuchungseinheiten zu Klumpen, auf den Stichprobenfehler ausübt, wird als Klumpungseffekt bezeichnet.

Der Klumpungseffekt ist Null, wenn die Klumpung streng zufällig erfolgt. Erfolgt die Zusammenlegung systematisch, wie am Anfang des Abschnittes beschrieben, so kann ein positiver oder ein negativer Klumpungseffekt auftreten. Wenn durch die Klumpung eine Verminderung der Zufallsfehler auftritt, spricht man von einem positiven Klumpungseffekt. Dieser positive Effekt tritt auf, wenn die Klumpen in sich sehr heterogen sind. Negativ wirkt sich die Klumpung aus, wenn die Klumpen aus ähnlichen Einheiten bestehen.

7.5 Verteilungsfunktionen

Die Eigenschaft eines digitalen Objektes, fehlerhaft oder nicht fehlerhaft zu sein, wird durch die Zufallsvariable Y mit den Werten {0;1} beschrieben (siehe Abschnitt 7.2). Bei einer Grundgesamtheit von N Objekten, von denen M bezüglich eines oder einer Kombination von Qualitätskriterien fehlerhaft sind, ergibt sich die Wahrscheinlichkeit, daß ein zufällig gezogenes Objekt fehlerhaft ist, zu

$$\Phi := P(Y = 1) = \frac{M}{N}.$$

Die Anzahl N der Elemente einer Grundgesamtheit von digitalen Objekten kann im GIS über eine einfache Abfrage mit der SQL-Aggregatfunktion `COUNT` ermittelt werden. Die Anzahl M der fehlerhaften Objekte der Grundgesamtheit ist im allgemeinen unbekannt, denn sonst müßte keine Stichprobenuntersuchung durchgeführt werden.

y_i stellt die Realisierung der Zufallsvaribale Y am i-ten Objekt dar. Bei der Prüfung einer Stichprobe vom Umfang n wird die Anzahl der fehlerhaften Objekte gezählt. Diese Zahl m ist die Realisierung der neuen, diskreten Zufallsvariable X_n, deren Zusammenhang mit Y durch eine einfache Summation hergestellt werden kann

$$m = \sum_{i=1}^{n} y_i \ .$$

7.5.1 Binomialverteilung

Werden die zu prüfenden Objekte zufällig aus der Grundgesamtheit ausgewählt und nach erfolgter Prüfung nicht gekennzeichnet, besteht die Möglichkeit, daß ein Objekt mehrere Male in die Untersuchung eingeht. In diesem Fall folgt die Wahrscheinlichkeit, daß genau m fehlerhafte Objekte aus einer n-elementigen Stichprobe gezogen werden der Binomialverteilung

$$P(X_n = m) = \binom{n}{m} \Phi^m (1-\Phi)^{n-m} \qquad (m = 0,1,2,\ldots,n).$$

Kurz: $X_n \sim Bi(n,\Phi)$

Die Wahrscheinlichkeit, bei den ersten *m* Ziehungen fehlerhafte Objekte zu ziehen und bei den darauffolgenden *n-m* nicht fehlerhafte, ist $\Phi^m(1-\Phi)^{n-m}$. Da die Reihenfolge der Ziehung fehlerhafter und fehlerfreier Objekte ohne Bedeutung, hat jede der

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

möglichen Reihenfolgen dieselbe Wahrscheinlichkeit.

Zur Berechnung des Erwartungswertes und der Varianz der Zufallsvariable X_n gilt (*Fisz, 1976*)

$$E(X_n) = n \cdot \Phi = n\frac{M}{N} \qquad D^2(X_n) = n \cdot \Phi(1-\Phi) = n \cdot \frac{M(N-M)}{N^2}.$$

7.5.2 Hypergeometrische Verteilung

Bei der Stichprobenuntersuchung von Geoobjekten wird die Auswahl der zu prüfenden Objekte ohne Zurücklegen stattfinden. Damit kann jedes Objekt der Grundgesamtheit nur einmal untersucht werden. Die Wahrscheinlichkeit innerhalb einer Stichprobe ein weiteres fehlerhaftes Objekt zu finden, ist deshalb davon abhängig, wie viele Objekte unter den zuvor untersuchten Objekten schon fehlerhaft waren. Die Zufallsvariable *X* (Anzahl der fehlerhaften Objekte in einer Stichprobe vom Umfang *n)* gehorcht der hypergeometrischen Verteilung (*Uhlmann, 1982*).

Die Anzahl der Möglichkeiten, aus einer Grundgesamtheit von *N* Elementen zufällig *n* Elemente herauszugreifen, wird durch die Binomialkoeffizienten gegeben. Sie lautet

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N(N-1)(N-2)\cdot\ldots\cdot(N-n+1)}{1\cdot 2\cdot 3\cdot\ldots\cdot(n-1)\cdot n}.$$

Wenn die Grundgesamtheit genau *M* fehlerhafte Objekte enthält, so beträgt die Anzahl der Möglichkeiten in *n* zufälligen Zügen *m* fehlerhafte und *(n-m)* fehlerfreie Objekte zu erhalten

$$\binom{M}{m} \cdot \binom{N-M}{n-m}.$$

Wird diese Anzahl der für das Ereignis $X_n = m$ günstigen Fälle durch die Gesamtzahl der möglichen Fälle dividiert, so erhält man die die Wahrscheinlichkeitsfunktion

$$P(X_n = m) = \frac{\binom{M}{m} \cdot \binom{N-M}{n-m}}{\binom{N}{n}} \quad \text{mit} \quad 0 \le m \le n, \quad m \le M, \quad n\text{-}m \le N\text{-}M.$$

Man sagt daher: die Zufallsvariable *X* besitzt eine hypergeometrischen Verteilung mit den expliziten Parametern *N*, *n* und *M* (oder *N*, *n* und Φ), oder kurz:

$$X_n \sim H(N,n,M) \text{ oder } X_n \sim H(N,n,\Phi).$$

Zur Berechnung des Erwartungswertes und der Varianz der Zufallsvariable *X* gilt (*Uhlmann, 1982*)

$$E(X_n) = \sum_{m=0}^{n} m \cdot \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} = n\frac{M}{N}\ , \qquad D^2(X_n) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right).$$

Zur Vermeidung von großen Zahlen als Zwischenergebnisse und damit für eine höhere numerische Stabilität kann die Wahrscheinlichkeit für das Ereignis $X_n=m$ mit folgender Produktformel berechnet werden

$$P(X_n = m) = \prod_{i=0}^{m-1}\frac{n-i}{m-i} \cdot \prod_{i=0}^{m-1}\frac{M-i}{N-n+m-i} \cdot \prod_{i=0}^{n-m-1}\frac{N-M-i}{N-i}\ .$$

Außerdem gilt die Rekursionsformel

$$P(X_n = m+1) = \frac{(n-m)(M-m)}{(m+1)(N-M-n+m+1)} \cdot P(X_n = m)\ .$$

7.5.3 Approximation der hypergeometrischen Verteilung durch andere Verteilungen

Wenn für beliebige N, der Quotient M/N konstant bleibt, die Fehlerdichte also über das gesamte Gebiet als homogen betrachtet werden kann, so nähert sich für große Grundgesamtheiten die hypergeometrische Verteilung der Binomialverteilung.

Die Analogie kann anhand der verwendeten Produktformel gezeigt werden, ein Beweis findet sich in *Fisz, 1976*.

$$P(X_n = m) = \prod_{i=0}^{m-1}\frac{n-i}{m-i} \cdot \prod_{i=0}^{m-1}\frac{M-i}{N-n+m-i} \cdot \prod_{i=0}^{n-m-1}\frac{N-M-i}{N-i} \qquad i,n,m << N,M$$

$$\lim_{N\to\infty} P(X_n = m) = \binom{n}{m} \cdot \left(\frac{M}{N}\right)^m \cdot \left(\frac{N-M}{N}\right)^{n-m} = \binom{n}{m}\Phi^m(1-\Phi)^{n-m}$$

Weiterhin kann gezeigt werden, daß die Binomialverteilung sich der Poissonschen Verteilung annähert, wenn der Stichprobenumfang *n* gegen unendlich strebt und zugleich $n\frac{M}{N}$ konstant bleibt, ($\Phi \to 0$).

$$\lim_{n\to\infty} P(X_n = m) = \frac{\lambda^m}{m!}e^{-\lambda} \qquad \text{mit } \lambda = n\frac{M}{N} = n\cdot\Phi$$

Der Beweis kann in *Fisz, 1976*, nachvollzogen werden.

Die Approximation ist auch schon für nicht ganz so große Stichproben relativ gut. Abbildung 53 stellt für verschiedene feste Werte von *N*, *M* und *n* die drei Verteilungsfunktionen in Abhängigkeit von der Fehleranzahl in der Stichprobe *m* gegenüber.

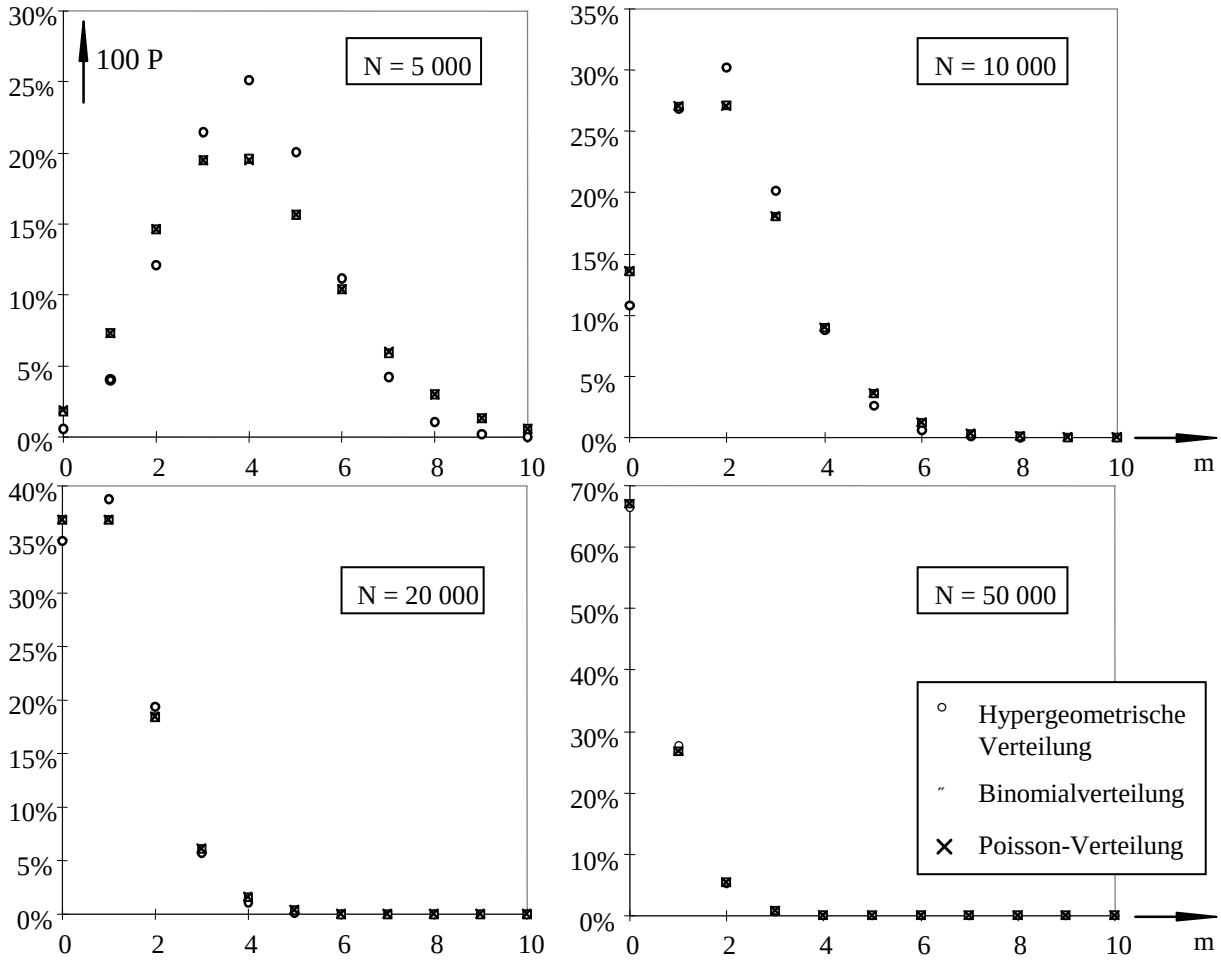


Abbildung 53: Die Werte der Wahrscheinlichkeitsfunktion 100 P der hypergeometrischen Verteilung, der Binomialverteilung und Poisson-Verteilung mit den Parametern *M = 10* und *n* = 2 000 für das Antreffen von *m* fehlerhaften Objekten.

Um die Unterschiede anschaulicher zu machen, wurden in Abbildung 54 für dieselben Parameter die Differenzen der Wahrscheinlichkeiten der Binomial- und der Poisson-Verteilung gegenüber der hypergeometrischen Verteilung graphisch aufgetragen.

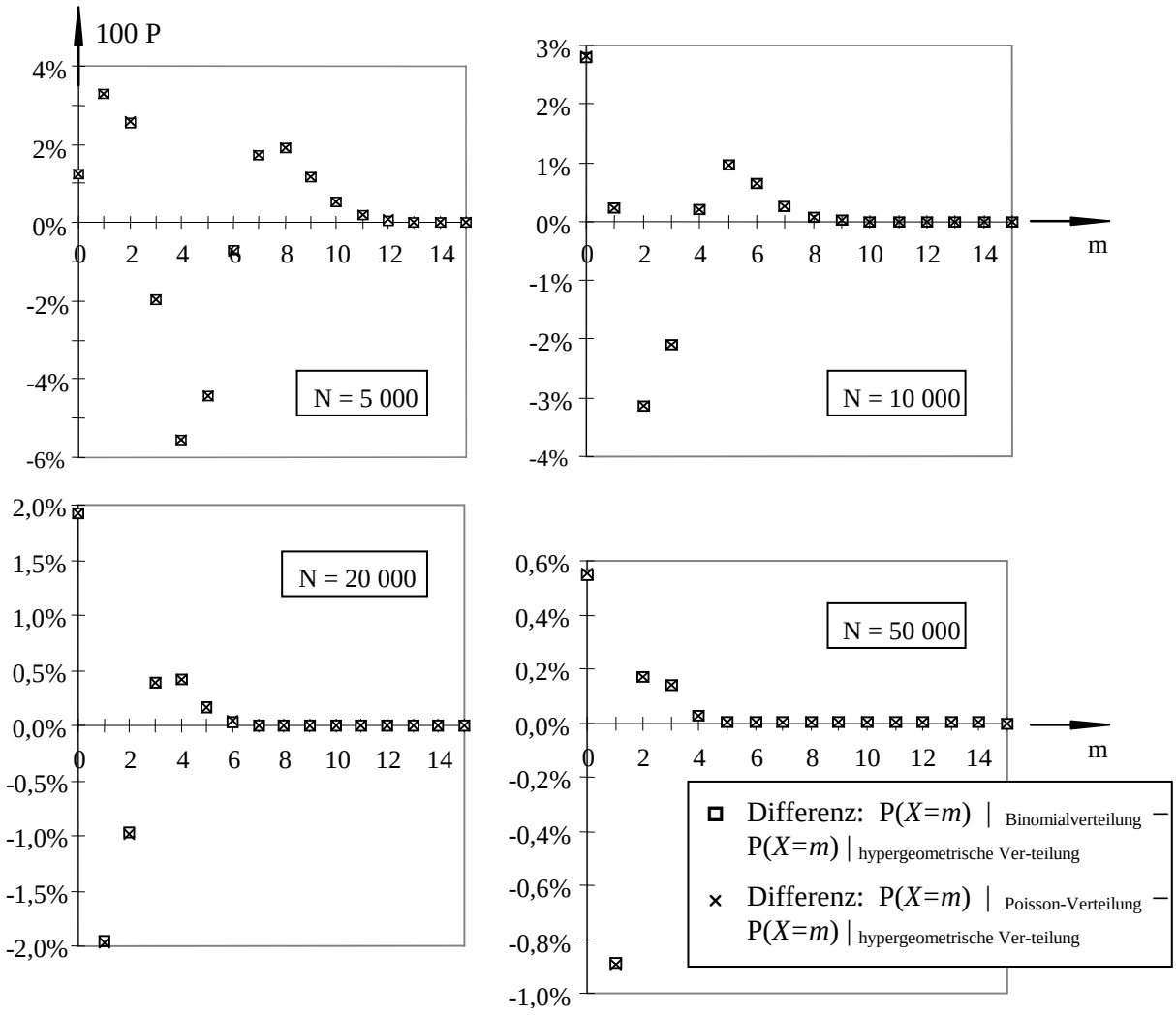


Abbildung 54: Differenzen der Wahrscheinlichkeiten unter Binomialverteilung / Poisson-Verteilung und hypergeometrischen Verteilung *m* fehlerhafte Objekte anzutreffen.

Die beiden Kurven fallen für alle Werte fast zusammen, da die Binomialverteilung für kleine Werte Φ sehr gut durch die Poisson-Verteilung approximiert wird. Allerdings ist in Abbildung 54 zu sehen, daß die Differenzen zur hypergeometrischen Verteilung beachtliche Größenordnungen annehmen. Für die Berechnungen im Zusammenhang mit der Stichprobenkontrolle wurde daher soweit möglich die hypergeometrische Verteilung herangezogen. Die Fälle, bei denen die hypergeometrische Verteilung durch eine der beiden Verteilungen approximiert wurde, sind besonders gekennzeichnet.

Die Binomialverteilung kann für große *n* auch durch eine Normalverteilung angenähert werden.

$$\lim_{n\to\infty} P\left(\frac{X_n - E(X_n)}{D(X_n)} \le \lambda\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} \exp\left(\frac{-t^2}{2}\right) dt$$

Nach *Bronstein et al, 1995,* ist dies mit allgemein ausreichender Genauigkeit möglich, wenn $n \cdot \Phi > 4$ und $n \cdot (1-\Phi) > 4$ ist.

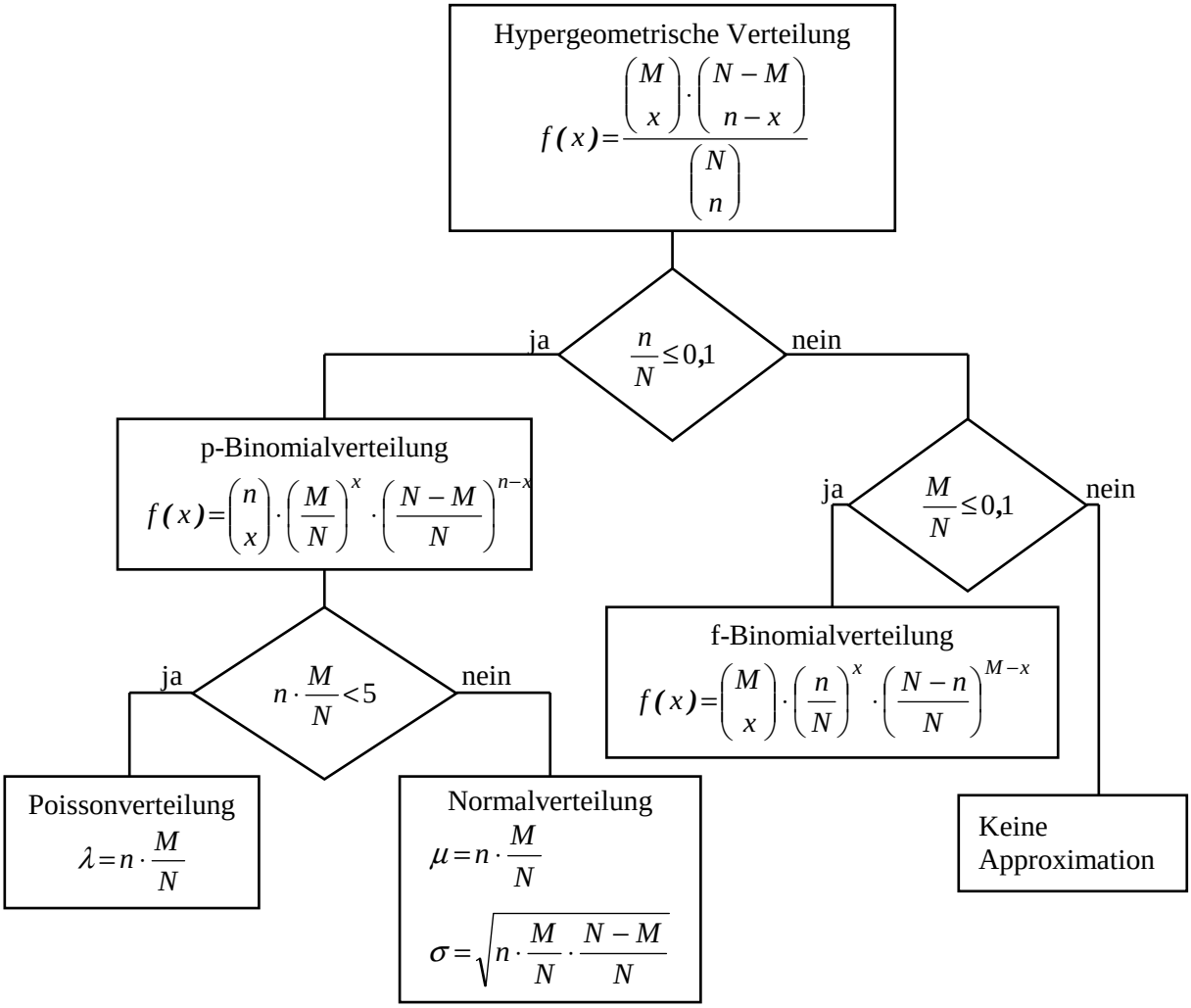


Abbildung 55: Flußdiagramm zu möglichen Approximationen der Dichtefunktion der hypergeometrischen Verteilung (*Schilling, 1982*).

Für eine bessere Approximation gibt *Mace, 1964*, eine Transformation an, die binomialverteilte Zufallsvariablen in normalverteilte Zufallsvariablen überführt. Die Transformationsgleichung lautet

$$U = (2\arcsin\sqrt{P} - 2\arcsin\sqrt{\Phi})\sqrt{n}\ .$$

Für einen Stichprobenumfang *n* ergibt sich ein Schätzwert *P* für die Fehlerrate Φ der Grundgesamtheit durch den Quotienten

$$P=\frac{X_n}{n} \quad \text{mit} \quad E(P)=\Phi \quad \text{und} \quad D^2(P)=\frac{\Phi(1-\Phi)}{n}$$

Die diskrete Zufallsvariable X_n wird durch die transzendente Transformation in eine stetige Zufallsvariable *u* überführt, deren Verteilung näherungsweise einer Normalverteilung mit Erwartungswert 0 und Varianz 1 entspricht.

$$U=\left(2\arcsin\sqrt{\frac{X_n}{n}}-2\arcsin\sqrt{\frac{M}{N}}\right)\sqrt{n} \quad \Rightarrow \quad U \sim N(0;1)$$

Die Normalverteilung hat gegenüber der Binomialverteilung und der hypergeometrischen Verteilung den Vorteil, daß ihre Quantile in Tafelwerken aufgelistet und in vielen Softwareprodukten als Standardfunktionen implementiert sind.

7.6 Verschiedene Begriffe der Testtheorie

Zur Untersuchung, ob die ermittelte Fehlerrate einer zu bewertenden Grundgesamtheit eine mittlere akzeptable Fehlerrate überschreitet, kann ein Hypothesentest durchgeführt werden. Der Hypothesentest wird so formuliert, daß bei der Durchführung des Tests zwischen zwei Hypothesen, der Null- und der Alternativhypothese, entschieden werden kann. Die Hypothesen beruhen auf Annahmen über die Grundgesamtheit, welche durch den Test verifiziert oder verworfen werden müssen.

Die **Nullhypothese H_0** macht eine Aussage darüber, wie viele Objekte der Grundgesamtheit höchstens fehlerhaft sind.

$$H_0 : M \le M_0 \quad \text{bzw.} \quad \Phi \le \Phi_0$$

Die **Alternativhypothese H_A** wird akzeptiert, wenn die Nullhypothese verworfen wird.

$$H_A : M > M_0 \quad \text{bzw.} \quad \Phi > \Phi_0$$

Zur Verifikation oder Falsifikation der Nullhypothese zugunsten der Alternativhypothese wird eine Stichprobenprüfung durchgeführt. Da nicht alle Objekte der Grundgesamtheit untersucht werden, kann die Entscheidung zwischen den Hypothesen nur mit einer bestimmten statistichen Sicherheit getroffen werden.

Auf der Grundlage eines konkreten Stichprobenergebnisses $(x_1, x_2, ..., x_n)$ soll zu einer Entscheidung über H_0 gelangt werden. Dies geschieht mit Hilfe einer Testgröße (auch Teststatistik oder Prüfgröße), die aus den Beobachtungen errechnet wird und deren Verteilung unter der Voraussetzung der Richtigkeit der Hypothesen H_0 oder H_A vollständig bekannt ist (*Caspary und Wichmann, 1994*).

Die Hypothesen über die Fehlerraten bzw. die Geamtzahl von zulässigen fehlerhaften Objekten muß sich aus der Anwendung ergeben. Dabei sind zwei ausgezeichnete Werte zu beachten (*DIN ISO 2859-Teil 0, 1991*):

$\Phi_0 = \frac{M_0}{N}$ ... zu erwartende bzw. unter den gegebenen Randbedingungen mindestens zu erreichende Fehlerrate

$\Phi_A = \frac{M_A}{N}$ ... im Einzelfall gerade noch akzeptable Fehlerrate

Mit der statistischen Qualitätskontrolle soll gewährleistet werden, daß die Qualität mit vertretbarem Aufwand zwischen diesen beiden Eckwerten liegt, und nur in Ausnahmefällen schlechter wird. Dabei darf die Qualität natürlich besser als die vorgegebenen Eckwerte sein. Wird sie allerdings zu schlecht, so muß das Los zurückgewiesen und überarbeitet werden.

Aus diesem Grund werden die Hypothesen für die weiteren Untersuchungen konkreter formuliert:

$$\begin{aligned} H_0 &: M = M_0 \quad \text{bzw.} \quad \Phi = \Phi_0 \\ H_A &: M = M_A \quad \text{bzw.} \quad \Phi = \Phi_A \end{aligned}.$$

Eine Irrtumswahrscheinlichkeit α, auch Test- oder Signifikanzniveau genannt, wird vorgegeben. Das Komplement zu 100%, 1-α, wird auch als Sicherheitswahrscheinlichkeit oder Sicherheit bezeichnet. Zu einem vorgegebenen α kann eine kritische Region K als Teilmenge des Stichprobenraumes bestimmt werden, in die die Prüfgröße bei richtiger H_0 mit der Irrtumswahrscheinlichkeit α fällt, so daß in α% der Fälle eine falsche Entscheidung getroffen wird.

$$M \le M_0 \rightarrow P\left((x_1, x_2, \ldots, x_n) \in K\right) \le \alpha$$

Dieser Zusammenhang kann im allgemeinen nur durch eine Ungleichung formuliert werden, weil die x_i diskret verteilt sind.

Die Testvorschrift lautet dann:

I. Wenn das Stichprobenergebnis $(x_1, x_2, ..., x_n)$ in der kritischen Region K liegt, so lehnt man die Nullhypothese ab, nimmt also die Alternative H_A an. Im Hinblick auf H_0 nennt man K daher auch die Ablehnungsregion.

II. Liegt dagegen $(x_1, x_2, ..., x_n)$ nicht in K, so kann man H_0 nicht unter Zugrundelegung dieser Irrtumswahrscheinlichkeit α ablehnen. Also nimmt man mangels besseren Wissens die Nullhypothese H_0 an. Man nennt das Komplement von K die Annahmeregion.

Bei der Anwendung des Tests gibt es genau zwei Arten, falsche Entscheidungen zu treffen, die unvermeidbar sind, weil die Stichprobe vom Zufall abhängt.

Erstens kann es vorkommen, daß $(x_1, x_2, ..., x_n) \in K$ ist, obwohl $M \leq M_0$ ist. Laut Testvorschrift wird dann die Nullhypothese H_0 abgelehnt, obwohl H_0 richtig ist; man nennt dies den **Fehler 1. Art**. Die Wahrscheinlichkeit, diesen Fehler 1. Art zu begehen, ist nach der obigen Ungleichung gleich der vorgegebenen Irrtumswahrscheinlichkeit α. Weil in der Literatur für statistische Qualitätskontrolle gerne auf Warenlieferungen von dem Produzenten an den Konsumenten eingegangen wird, hat sich für den Fehler 1. Art auch der Begriff **Produzentenrisiko** etabliert.

Zweitens kann es vorkommen, daß $(x_1, x_2, ..., x_n) \notin K$ ist, obwohl $M > M_A$ ist. Hier wird die Nullhypothese H_0 nicht abgelehnt, obwohl H_0 falsch ist; man nennt dies den **Fehler 2. Art**. Bei der Durchführung des Tests handelt es sich dabei ebenfalls um eine Fehlentscheidung, weil die falsche Nullhypothese angenommen wird. In Analogie zu dem Begriff Produzentenrisiko wird für den Fehler 2. Art oft der Begriff **Konsumentenrisiko** verwendet.

Die Wahrscheinlichkeit, einen Fehler 2. Art zu begehen, hängt im allgemeinen von der gewählten Irrtumswahrscheinlichkeit α, von dem vorliegenden M_A aus H_A und dem Stichprobenumfang n ab. Bei einem „guten“ Test wird die Wahrscheinlichkeit für den Fehler 2. Art mit wachsendem n immer kleiner. Es ist anschaulich klar, daß die Wahrscheinlichkeit für den Fehler 2. Art um so größer ist, je kleiner die Irrtumswahrscheinlichkeit α gewählt wird, und umgekehrt.

Ein nützliches Hilfsmittel, um die Eigenschaften eines Tests im Hinblick auf die Fehler 1. und 2. Art darzustellen, ist die **Gütefunktion** $g(M)$ oder $g(\Phi)$, die definiert wird durch

$$g(M) = P_M\left((x_1, x_2, \ldots, x_x) \in K\right).$$

Die Gütefunktion gibt also in Abhängigkeit der Anzahl M der fehlerhaften Objekte der Grundgesamtheit bzw. der Fehlerrate Φ die Wahrscheinlichkeit für das Ablehnen der Nullhypothese H_0 an. Nach der Definition ist $g(M) \leq \alpha$ für alle $M \leq M_0$. Für $M = M_A$ dagegen ist $1- g(M)$ die Wahrscheinlichkeit für den Fehler 2. Art.

In der Qualitätskontrolle ist es üblich, statt der Gütefunktion die **Operationscharakteristik** L (*DIN ISO 2859-0, 1991*) zu benutzen:

Operationscharakteristik = 1 - Gütefunktion

Die Operationscharakteristik gibt die Wahrscheinlichkeit für die Annahme eines Loses in Abhängigkeit vom Anteil Φ der gelieferten fehlerhaften Objekte an. Die Operationscharakteristik $L(\Phi)$ wird auch Annahmekennlinie oder Arbeitscharakteristik genannt und kurz mit OC-Kurve (englisch: *operating characteristic curve*) bezeichnet (Siehe dazu auch Abbildung 57).

7.7 Stichprobenplan

Bei der qualitativen Beurteilung der Elemente der Grundgesamtheit beziehungsweise der Stichprobe, wenn also nur zwischen fehlerhaften und korrekten Objekten unterschieden wird, ist durch die Testdurchführung determiniert, welcher Verteilungsfunktion das Stichprobenergebnis folgt. Bei der Untersuchung von Geoobjekten wird ein Objekt bezüglich des relevanten Qualitätsmerkmales nur einmal untersucht, daraus folgt, daß eine hypergeometrische Verteilung $H(N, n; \Phi)$ vorliegt. Es soll die Nullhypothese H_0 gegen die Alternative H_A getestet werden.

Als Prüfgröße wird die Anzahl *m* der fehlerhaften Objekte der Stichprobe verwendet. Sie kann als Summe der Realisierungen x_i der Zufallsvariablen X_n, die nur die Werte 0 oder 1 annimmt, dargestellt werden:

$$m = \sum_{i=1}^{n} x_i$$

Die Nullhypothese wird abgelehnt, wenn *m* zu groß ausfällt, nämlich wenn *m* einen von der Irrtumswahrscheinlichkeit α abhängigen Wert *Ac(α)* übersteigt. *Ac(α)* kann aus den Werten der hypergeometrischen Verteilung abgeleitet werden.

Der kritische Bereichung *K* lautet damit für einen einstufigen Stichprobenplan $K: m > Ac$.

7.7.1 Einstufiger Stichprobenplan

Ein einstufiger Stichprobenplan wird durch zwei Zahlen beschrieben: den Stichprobenumfang *n* und die Anzahl der fehlerhaften Objekte in der Stichprobe, mit denen das Los noch akzeptiert werden kann, *Ac*.

Fisz, 1976, führt für die Festlegung eines einstufigen Stichprobenplanes folgende Kurzbezeichnung ein: Einstufiger Stichprobenplan : $Ac \,\|\, n$

In der *(DIN ISO 2859-0, 1991)* wird ein Beispiel (*Example 32*) für einen einstufigen Stichprobenplan gegeben:

- *Acceptable quality level* : AQL = 1,5%
- *Sample size* : *n = 315 items*
- *Acceptance number* : *Ac = 10 nonconforming items*
- *Rejection number* : *Re = 11 nonconforming items*

Damit aus einer einzigen Stichprobe eine eindeutige Entscheidung über Annahme oder Zurückweisung des Loses getroffen werden kann, ist beim einstufigen Stichprobenplan die Rückweisezahl *Re* immer um eins höher als die Annahmezahl *Ac* (*Re= Ac + 1*).

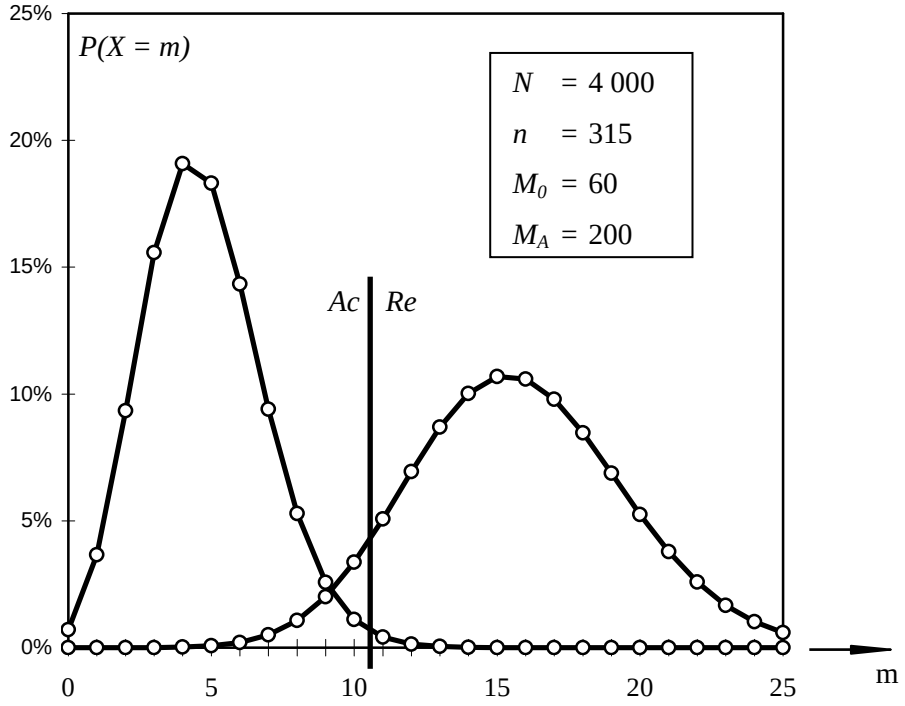


Abbildung 56: Wahrscheinlichkeiten, bestimmte Anzahlen von fehlerhaften Objekten zu ziehen in Abhängigkeit von den hypothetischen Fehleranzahlen der Grundgesamtheit M_0 und M_A nach der hypergeometrischen Verteilung.

Die Erwartungswerte und Varianzen der Dichtefunktionen für das skizzierte Beispiel lassen sich nach den im Abschnitt 7.5.2 gegebenen Gleichungen berechnen.

$$E\big|_{M=M_0}(X) = n\frac{M_0}{N} = 4{,}725 \approx 5 \qquad D^2\big|_{M=M_0}(X) = n\frac{M_0}{N}\left(1-\frac{M_0}{N}\right)\left(\frac{N-n}{N-1}\right) = 4{,}3 \qquad \sqrt{D^2\big|_{M=M_0}(X)} = 2{,}1$$

$$E\big|_{M=M_A}(X) = n\frac{M_A}{N} = 15{,}75 \approx 16 \qquad D^2\big|_{M=M_A}(X) = n\frac{M_A}{N}\left(1-\frac{M_A}{N}\right)\left(\frac{N-n}{N-1}\right) = 13{,}8 \qquad \sqrt{D^2\big|_{M=M_A}(X)} = 3{,}7$$

Mit dem im Beispiel gegebenen Anahme- und Rückweisezahlen *Ac = 10* und *Re = 11* und mit den Hypothesen über die wahre Fehleranzahl, können auch die Irrtumswahrscheinlichkeiten errechnet werden.

Fehler 1. Art:
$$\alpha = P\big|_{M=M_0}(m \geq Re) = 1 - P(m < Re) = 1 - \sum_{i=0}^{Re-1} P(m=i) = 1 - \sum_{i=0}^{Re-1} \frac{\binom{M_0}{i}\binom{N-M_0}{n-i}}{\binom{N}{n}}$$

$$\alpha = 0{,}6\% \qquad \text{mit } M_0 = 60$$

Fehler 2. Art:
$$\beta = P\big|_{M=M_A}(m \leq Ac) = \sum_{i=0}^{Ac} P(m=i) = \sum_{i=0}^{Ac} \frac{\binom{M_A}{i}\binom{N-M_A}{n-i}}{\binom{N}{n}}$$

$$\beta = 7{,}3\% \qquad \text{mit } M_A = 200$$

Zum Vergleich dazu die Werte für dasselbe Beispiel bei verschärfter Kontrolle (siehe Abschnitt 7.8). Der Stichprobenplan lautet in diesem Fall $8 \,\|\, 315$, und $\alpha = 4{,}3\%, \beta = 1{,}9\%$.

7.7.2 Annehmbare (AQL) und Rückzuweisende Qualitätsgrenzlage (LQ)

Der Zusammenhang zwischen dem Ausschußprozentsatz Φ_a % des Loses und der Annahmewahrscheinlichkeit a % ist, wie in Abschnitt 7.6 beschrieben, durch die Operationscharakteristik gegeben (siehe dazu Abbildung 57):

$$L(\Phi_a) = a\,.$$

Der 90%-Punkt heißt in der deutschen Literatur Annahmegrenze, und man bezeichnet 100% - 90% = 10% das zugehörige Produzentenrisiko. Es entspricht der Wahrscheinlichkeit für den Fehler 1. Art, weil bei einem tatsächlichen Ausschußanteil von $\Phi_{90\%}$ im Durchschnitt noch 10% der Lieferungen wegen zufällig schlechten Ausfalls der Stichprobe an den Produzenten oder Lieferanten zurückgehen, obwohl die Partie ausreichend gut ist. In der anglo-amerikanischen Literatur bevorzugt man statt dessen $\Phi_{95\%}$ und nennt diesen Ausschußanteil „*acceptable quality level*“ und auf deutsch **annehmbare Qualitätsgrenzlage** oder abgekürzt AQL (*Uhlmann, 1982*).

Im Sinne eines Hypothesentests entspricht $\Phi_{1-\alpha}$ der Nullhypothese H_0, bei der ein Ausschußanteil von $M_0 = N\,\Phi_{1-\alpha}$ angenommen wird. AQL beschreibt die Fehlerrate, die zu der 95% Annahmewahrscheinlichkeit eines Loses für einen bestimmten Stichprobenumfang gehört. Φ_β korrespondiert analog dazu mit der Fehlerrate, die zu der Alternativhypothese H_A gehört.

Der Punkt $\Phi_{10\%}$ heißt Ablehnungsgrenze („*lot tolerance per cent defective*“, kurz: LTPD), und die zugehörige Annahmewahrscheinlichkeit von 10% für die Partie nennt man Konsumentenrisiko. Dies entspricht der Wahrscheinlichkeit für den Fehler 2. Art, weil hier der Verbraucher noch mit 10% Wahrscheinlichkeit eine nicht mehr ausreichend gutes Los annimmt.

DIN ISO 2859 Teil 2, 1993, enthält Stichprobenanweisungen, die nach der **rückzuweisenden Qualitätsgrenzlage** (LQ) geordnet sind. Bei den dort aufgelisteten Stichprobenplänen haben Lose, deren Qualitätslage gleich der LQ ist, abhängig von der Stichprobenanweisung eine Annahmewahrscheinlichkeit, die gewöhnlich kleiner als 10%, in jedem Fall jedoch kleiner als 13% ist.

Für einen Anwender der Daten bedeutet dies aber, daß in 10% der Fälle noch Lose akzeptiert werden, obwohl sie nur auf dem Qualitätsniveau der rückzuweisenden Grenzlage sind. Für viele Anwendungen im Bereich der Planung, Energieversorgung oder Einsatzleitung ist dieser Anteil noch zu hoch. Für die weiteren Untersuchungen wird daher für den Fehler 2. Art ein Prozentsatz von 5% gewählt. Damit steigt zwar der erforderliche Stichprobenumfang *n*, da eine geringere Irrtumswahrscheinlichkeit angesetzt wird, aber das Konsumentenrisiko wird dadurch geringer und die Risiken falsche Entscheidungen zu treffen sind auf beide Vertragspartner gleich verteilt.

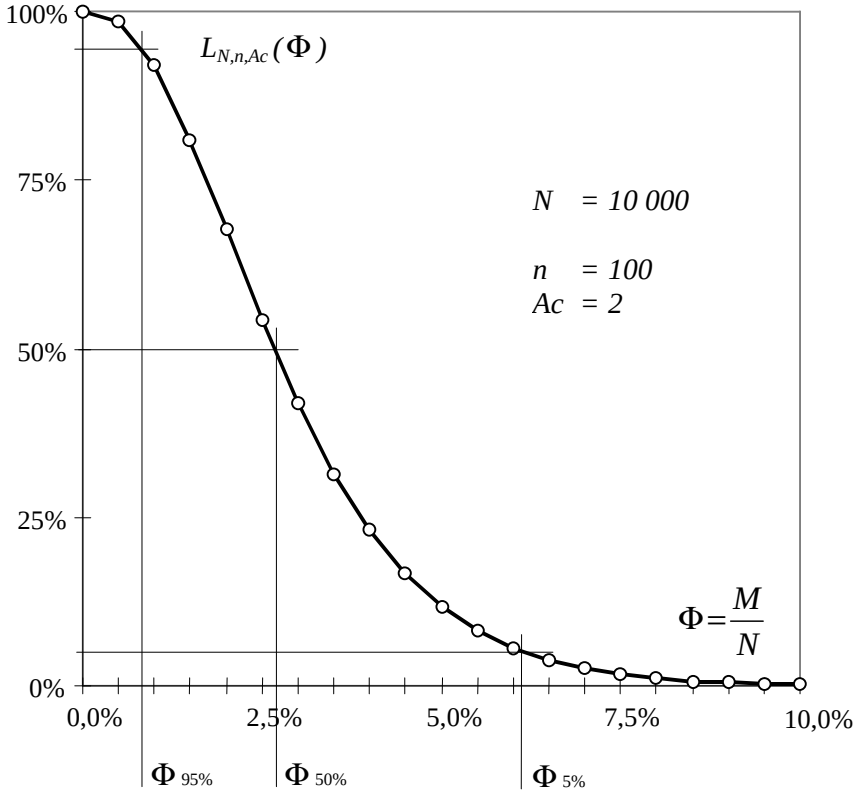


Abbildung 57: Annahmegrenze $\Phi_{95\%} = 0{,}83\%$, Indifferenzpunkt $\Phi_{50\%} = 2{,}66\%$ und rückzuweisende Qualitätsgrenzlage $\Phi_{5\%} = 6{,}15\%$ beim Stichprobenumfang $n = 100$ und Annahmezahl $Ac = 2$ berechnet mit der hypergeometrischen Verteilung.

Anders als die AQL, welche dem Hersteller einen Anhaltspunkt für diejenige Qualitätslage gibt, die er liefern muß, damit er in den meisten Fällen die Annahmekriterien erfüllen kann, gibt die rückzuweisende Qualitätsgrenzlage dem Abnehmer keinen zuverlässigen Anhaltspunkt hinsichtlich der wahren Qualität der angenommenen Lose. Aus diesem Grunde sollte für die rückzuweisende Qualitätsgrenzlage realistischerweise eine mindestens dreimal höhere Fehlerrate gewählt werden als die gewünschte Qualitätslage. Das ermöglicht dem Lieferer mit der gewünschten Qualität zu liefern und dabei eine angemessene Annahmewahrscheinlichkeit der vorgestellten Lose zu haben.

Die AQL ist ein Parameter des Stichprobenplans und sollte nicht mit der mittleren Qualitätslage verwechselt werden, welche die Qualitätslage des Prozesses beschreibt. Es wird davon ausgegangen, daß die mittlere Qualitätslage kleiner oder gleich der AQL ist, um zu häufige Rückweisungen zu vermeiden (*DIN ISO 2859-1, 1993*). Besonders wird in der Norm folgende Aussage hervorgehoben: „Die Festlegung eines AQL-Wertes bedeutet nicht, daß der Lieferant das Recht hat, wissentlich auch nur ein einziges fehlerhaftes Objekt zu liefern."

Die anzuwendende AQL muß in einem Vertrag oder nach Maßgabe der zuständigen Stelle festgelegt werden. Es können verschiedene AQL´s für einzelne Qualitätskriterien oder für einzelne Objektklassen festgelegt werden. Die Zuordnung zu Gruppen sollte für die Qualitätsforderung in der jeweiligen Situation angemessen sein.

7.7.3 Stichprobenumfang

Um die beiden Parameter *Ac* und *n* des Stichprobenplanes festzulegen, müssen Hypothesen über die Häufigkeiten von Fehlern der Grundgesamtheit getroffen und statistische Irrtumswahrscheinlichkeiten für den Fehler 1. und 2. Art gewählt werden.

Die Werte müssen die beiden Gleichungen für das Risiko, einen Fehler 1. oder 2. Art zu begehen, gleichzeitig erfüllen.

$$\alpha = 1 - \sum_{i=0}^{Ac} \frac{\binom{M_0}{i}\binom{N-M_0}{n-i}}{\binom{N}{n}} \qquad \beta = \sum_{i=0}^{Ac} \frac{\binom{M_A}{i}\binom{N-M_A}{n-i}}{\binom{N}{n}}$$

Da nur ganzzahlige Lösungen in Frage kommen, werden die Gleichungen nicht exakt erfüllt sein können, sondern nur mit einer hinreichend guten Näherung. Die Lösungen können in diesem Fall nur durch systematisches Probieren ermittelt werden.

Ein analytisches Lösungsverfahren ergibt sich durch bestimmte Vereinfachungen und Transformationen. In einem ersten Schritt wird die hypergeometrische Verteilung durch die Binomialverteilung approximiert und in einem zweiten Schritt durch die in Abschnitt 7.5.3 gegebene Transformation in eine neue Zufallsvariable überführt, die näherungsweise standardnormalverteilt ist.

Für einen Stichprobenumfang *n* ergibt sich ein Schätzwert für die Fehlerrate Φ der Grundgesamtheit durch den Quotienten

$$p = \frac{m}{n} \qquad \text{mit } E(p) = \Phi \text{ und } D^2(p) = \frac{\Phi(1-\Phi)}{n}.$$

Die Realisierung *m* hypergeometrisch verteilte Größe *M* kann hinreichend genau durch die binomial verteilte Variable *K* approximiert werden, deren Wahrscheinlichkeitsfunktion gegeben ist durch

$$P(K = i) = \binom{n}{i} \Phi^i \, \Phi^{n-i}.$$

Die diskrete Zufallsgröße *K* wird durch die transzendente Transformation

$$U = \left(2 \arcsin \sqrt{\frac{K}{n}} - 2 \arcsin \sqrt{\frac{M}{N}} \right) \sqrt{n}$$

in eine stetige Zufallsvariable *U* überführt mit $U \sim N(0;1)$.

Unter der Annahme des schlechtesten, gerade noch akzeptablen Falles, daß die Fehlerhäufigkeit in der Grundgesamtheit der annehmbaren Qualitätsgrenzlage entspricht, kann die Beziehung zwischen der Wahrscheinlichkeit für einen Fehler 1. Art, α, mit Hilfe der Verteilungsfunktion der standardisierten Normalverteilung ermittelt werden.

$$P(u < u_{1-\alpha}) = 1 - \alpha \quad \text{mit } M = M_0$$

Der kritische Wert $u_{1-\alpha}$ entspricht dabei dem Quantil der Normalverteilung

$$\Phi(u_{1-\alpha}) = 1 - \alpha = \int_{-\infty}^{u_{1-\alpha}} \varphi(t)\,dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u_{1-\alpha}} e^{-t^2/2}\,dt.$$

Wenn die Fehlerhäufigkeit der Grundgesamtheit dem Wert M_A der Alternativhypothese entspricht, so wird die Nullhypothese mit der Wahrscheinlichkeit β fälschlicherweise akzeptiert.

$$P(u < u_\beta) = \beta \quad \text{mit } M = M_A$$

Da die beiden Gleichungen für die Wahrscheinlichkeiten gleichzeitig erfüllt sein müssen, ergibt sich das folgende System von Ungleichungen:

$$\left(2\arcsin\sqrt{\frac{k}{n}} - 2\arcsin\sqrt{\frac{M_0}{N}}\right)\sqrt{n} < u_{1-\alpha}$$

$$\left(2\arcsin\sqrt{\frac{k}{n}} - 2\arcsin\sqrt{\frac{M_A}{N}}\right)\sqrt{n} < u_{\beta}$$

Zur Berechnung des mindestens erforderlichen Stichprobenumfanges interessiert der Grenzwert, daher können die „Kleiner"-Relationen durch „Gleichheits"-Relationen ersetzt und das Gleichungssystem nach *n* aufgelöst werden.

$$n = \left(\frac{u_\alpha + u_\beta}{2\arcsin\sqrt{\frac{M_0}{N}} - 2\arcsin\sqrt{\frac{M_A}{N}}}\right)^2 \quad \text{mit} \quad u_{1-\alpha} = -u_\alpha$$

Mit dieser Formel läßt sich unter den genannten Vereinfachungen der Mindeststichprobenumfang bestimmen, damit die Risiken 1. oder 2. Art auf die vorgegebenen Wahrscheinlichkeiten begrenzt werden.

Die Quantile der Normalverteilung sind in Formelsammlungen vertafelt oder können als Standardfunktionen von Statistikbibliotheken oder Tabellenkalkulationsprogrammen aufgerufen werden. In EXCEL z.B. lautet die Funktion zur Berechnung der Quantile der Normalverteilung „NORMINV()" mit den Parametern Wahrscheinlichkeit, Erwartungswert und Standardabweichung. Für die Quantile der standardisierten Normalverteilung steht die Funktion „STANDNORMINV()" zur Verfügung.

Uhlmann, 1982, gibt zur Bestimmung der Parameter *Ac* und *n* des Stichprobenplanes ein anderes Verfahren, das auf *Peach und Littauer, 1946*, zurückgeht.

Eine stetige Verteilungsfunktion mit der Dichte

$$g_f(y) = \begin{cases} 0 & \text{für } y \le 0 \\ \dfrac{1}{2^{f/2}\Gamma\left(\dfrac{f}{2}\right)} y^{\frac{f}{2}-1} e^{-\frac{y}{2}} & \text{für } y > 0 \end{cases}$$

heißt χ^2-Verteilung mit Freiheitsgrad $f \ge 1$. Die dabei auftretende Gamma-Funktion ist für $y > 0$ definiert durch

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$$

Für jede reelle Zahl $z > 0$ gilt $\Gamma(z+1) = z \cdot \Gamma(z)$, für jede natürliche Zahl $m \neq 1$ ist $\Gamma(m) = (m\text{-}1)!$, und außerdem gilt

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\ .$$

Wenn für große *N* und *n* und für kleine Φ die Verteilung der fehlerhaften Objekte der Stichprobe durch die Poisson-Verteilung genähert wird (Abschnitt 7.5.3), dann ergibt sich für die Operationscharakteristik mit $\Phi = M / N$ und $\lambda = n \cdot \Phi$ die Gleichung

$$L^*_{n,Ac}(\Phi) = \sum_{m=0}^{Ac} \frac{n^m \Phi^m}{m!} e^{-np}$$

Für Ac = 0, 1, 2, ... und für jedes reelle $\lambda \neq 0$ gilt außerdem

$$\sum_{m=0}^{Ac} \frac{\lambda^m}{m!} e^{-\lambda} = 1 - \int_0^{2\lambda} \frac{1}{2^{Ac+1}\, Ac!}\, y^{Ac} e^{-y/2} dy .$$

Der Integrand entspricht der Dichte $g_k(y)$ der χ^2-Verteilung mit Freiheitsgrad $k = 2\,(Ac+1)$. *Uhlmann* skizziert einen Beweis, indem er die Identität der Terme bei $\lambda = 0$ und in ihrer ersten Ableitung nach λ nachweist.

Die Funktion $G^*(\gamma;k)$ wird als Umkehrfunktion der Verteilungsfunktion der χ^2-Verteilung mit Freiheitsgrad k eingeführt, mit

$$\int_0^{G^*(\gamma;k)} g_k(y)\,dy = \gamma$$

Dadurch ist die Gleichung $L^*_{n,Ac}(\Phi) = \gamma$ gleichbedeutend mit

$$1 - \int_0^{2\lambda} g_{2(Ac+1)}(y)\,dy = \gamma$$

und damit

$$2n\Phi = G^*\left(1-\gamma;2(Ac+1)\right).$$

Für einen Stichprobenplan müssen die beiden Bedingungen korrespondierend mit dem Produzenten- und Konsumentenrisiko eingehalten werden:

$$L_{n,Ac}(\Phi_{1-\alpha}) \geq 1-\alpha \quad \text{und} \quad L_{n,Ac}(\Phi_{\beta}) \leq \beta .$$

Nach der Umkehrfunktion der χ^2-Verteilung läßt sich der Zusammenhang auch ausdrücken mit

$$\frac{1}{2n} G^*\left(\alpha;2(Ac+1)\right) \geq \Phi_{1-\alpha} \quad \text{und} \quad \frac{1}{2n} G^*\left(1-\beta;2(Ac+1)\right) \leq \Phi_{\beta} .$$

Da $G^*(\gamma;k)$ monoton wachsend vom Freiheitsgrad k abhängt, kann das folgende Verfahren zur Bestimmung der Parameter für einen Stichprobenplan abgeleitet werden.

Der Stichprobenumfang n und die Annahmezahl Ac erfüllen genau dann die Bedingungen, wenn

$$\frac{G^*\left(\beta;2(Ac+1)\right)}{2\Phi_{\beta}} \leq n \leq \frac{G^*\left(1-\alpha;2(Ac+1)\right)}{2\Phi_{1-\alpha}}$$

gilt. Man erhält ein möglichst kleines n, wenn man ein möglichst kleines Ac so wählt, daß sich diese Bedingung gerade noch mit einer passenden natürlichen Zahl n erfüllen läßt.

Wenn der Quotient $\Phi_\beta / \Phi_{1-\alpha} = M_A / M_0 = q$ gegeben ist, so kann durch Einsetzen der natürlichen Zahlen 0, 1, 2, ... für die Annahmezahl Ac die kleinste Zahl n ermittelt werden, für die die folgende Gleichung gilt $G^*\left(\beta;2(Ac+1)\right) < q \cdot G^*\left(1-\alpha;2(Ac+1)\right)$.

Für die Werte $\alpha = \beta = 5\%$ und für die Annahmezahlen 0 bis 25 sind die Quantile der χ^2-Verteilung in Verbindung mit drei Quotienten q = 2, 3 und 5 aufgelistet.

Ac	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
G(5%;2(Ac+1))*	5,99	9,49	12,59	15,51	18,31	21,03	23,68	26,30	28,87	31,41	33,92	36,42
2 G(95%;2(Ac+1))*	0,21	1,42	3,27	5,47	7,88	10,45	13,14	15,92	18,78	21,70	24,68	27,70
3 G(95%;2(Ac+1))*	0,31	2,13	4,91	8,20	11,82	15,68	19,71	23,88	28,17	32,55	37,01	41,55
5 G(95%;2(Ac+1))*	0,51	3,55	8,18	13,66	19,70	26,13	32,85	39,81	46,95	54,25	61,69	69,24

12	13	14	15	16	17	18	19	20	21	22	23	24	25
38,89	41,34	43,77	46,19	48,60	51,00	53,38	55,76	58,12	60,48	62,83	65,17	67,50	69,83
30,76	33,86	36,99	40,14	43,33	46,54	49,77	53,02	56,29	59,57	62,88	66,20	69,53	72,87
46,14	50,78	55,48	60,22	64,99	69,81	74,65	79,53	84,43	89,36	94,32	99,29	104,29	109,31
76,90	84,64	92,46	100,36	108,32	116,34	124,42	132,55	140,72	148,94	157,19	165,49	173,82	182,19

Die hervorgehobenen Felder der Tabelle geben die kleinsten Zahlen an, bei denen das mit *q* multiplizierte (1-α)-Quantil das β-Quantil übersteigt. Durch Division mit der doppelten Fehlerrate $2\Phi_\beta$ ergibt sich ein Bereich, in dem der erforderliche Stichprobenumfang *n* liegt. Indem die untere Schranke zur nächsten ganzen Zahl aufgerundet wird, wird der optimale Stichprobenumfang ermittelt, die zugehörige Annahmezahl kann direkt aus der Tabelle abgelesen werden.

Auch dieses Verfahren beruht darauf, die unbekannten Größen auf vertafelte Verteilungsfunktionen zurückzuführen, deren Quantile entweder aus Tafelwerken oder mit Hilfe von Statistikprogrammen ermittelt werden können. Die Umkehrfunktion der χ^2-Verteilung wird z.B. in EXCEL mit dem Befehl „CHIINV()“ aufgerufen.

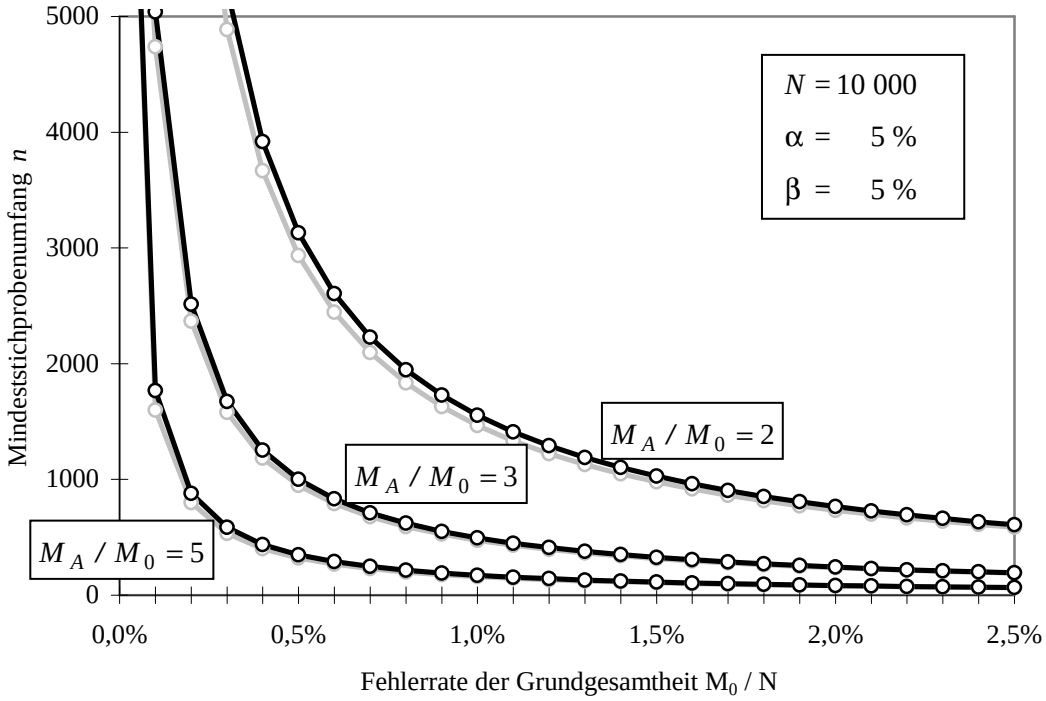


Abbildung 58: Stichprobenumfang in Abhängigkeit von der Fehlerrate der Grundgesamtheit für verschiedene Trennschärfen (M_A / M_0) bei α = β = 5%. Die Verfahren nach *Mace* (in dunkler Signatur) und nach *Peach und Littauer* (grau) werden gegenübergestellt.

Aus der Abbildung 58 kann abgelesen werden, daß

- je weniger fehlerhafte Objekte in der Grundgesamtheit vorhanden sind, um so mehr Objekte müssen zur statistischen Qualitätskontrolle untersucht werden.
- und je kleiner der Unterschied zwischen der Fehleranzahl M_0 unter der Hypothese H_0 und der Fehleranzahl M_A, gegen die bei der Kontrolle getestet werden soll, um so größer muß der Stichprobenumfang gewählt werden.

Außerdem gilt (siehe Abbildung 59):

- je kleiner die Wahrscheinlichkeit, einen Fehler 1. oder 2. Art zu begehen, gewählt wird, um so größer ist der erforderliche Stichprobenumfang, wobei der Zusammenhang nicht linear ist, sondern näherungsweise exponentiell mit negativem Exponenten.

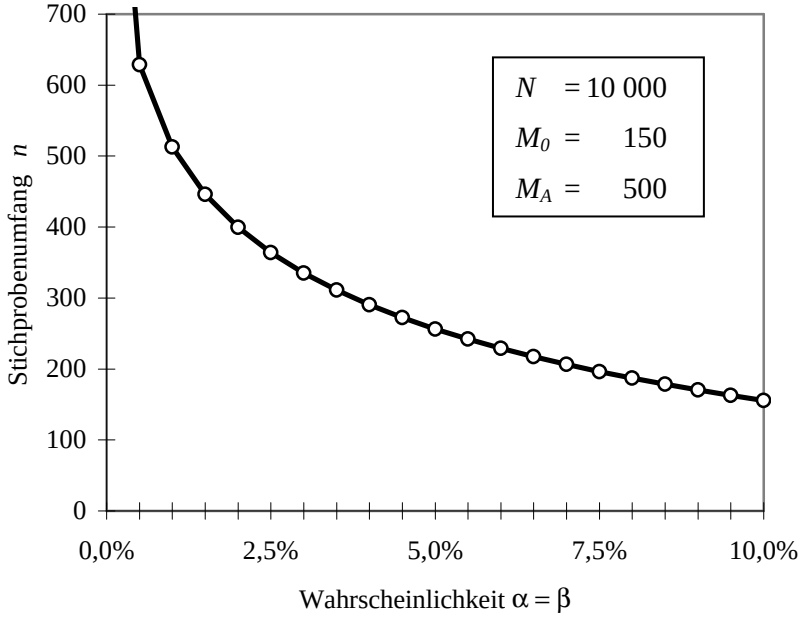


Abbildung 59: Zusammenhang zwischen den Wahrscheinlichkeiten, einen Fehler 1. und 2. Art zu begehen, und dem erforderlichen Stichprobenumfang bei konstanter Fehlerrate M_0 und konstantem Verhältnis zwischen M_0 und M_A .

7.7.4 Abgebrochene Kontrolle

Eine Modifizierung des Verfahrens für einfache Stichprobenpläne mit dem Ziel, bei Losen mit schlechter Qualität den Stichprobenumfang zu reduzieren, ist durch die **abgebrochene Kontrolle** (englisch: *curtailed sampling* oder *truncated sampling*) gegeben.

Wenn in einer Stichprobe mehr als *Ac* fehlerhafte Objekte festgestellt werden, wird das Los abgelehnt. Es ist daher naheliegend, die Kontrolle bei Erreichen dieser Zahl sofort abzubrechen. Allerdings müssen die Objekte der Stichprobe schon vorab festgelegt sein, da sonst beim Prüfen die Gefahr besteht, daß ein Prüfer aus dem gesamten Los die Fehler zusammensucht, um möglichst schnell auf (*Ac*+1) fehlerhafte Objekte zu kommen.

Zur Berechnung des durchschnittlichen Stichprobenumfanges bei der abgebrochenen Kontrolle werden die folgenden Größen definiert.

$$m_i = \text{Anzahl der fehlerhaften Objekte unter den ersten } i \text{ Objekten der Stichprobe}$$

Bei der abgebrochenen Kontrolle ist der tatsächliche Prüfumfang

$$n' = \begin{cases} i, \text{falls } \; m_{i-1} \le Ac \wedge m_i > Ac \quad \text{für } \; i = Ac+1, Ac+2, \ldots, n \\ n, \text{falls } \; m_n \le Ac \end{cases}$$

Das Ergebnis „$m_{i-1} \le Ac$ und zugleich $m_i > Ac$“ tritt genau dann ein, wenn $m_{i-1}=Ac$ und $m_i-m_{i-1}=1$ ist. Für den Stichprobenumfang ergibt sich daher ein Erwartungswert von

$$E(n') = \sum_{i=Ac+1}^{n} i \cdot P(m_{i-1} = Ac \wedge m_i > Ac) + n \cdot P(m_n \le Ac) .$$

Die Wahrscheinlichkeit für das Ereignis, das zum Abbruch der Kontrolle führt, läßt sich mit Hilfe der Kombinatorik berechnen. Da die Objekte ohne Zurücklegen untersucht werden, gibt es $\binom{N}{i-1}$ Möglichkeiten, $(i-1)$ aus den vorhandenen N Objekten auszuwählen. Für jede gibt es $(N-(i-1))$ Möglichkeiten, ein weiteres Objekt zu ziehen. Die Anzahl der möglichen, gleichwahrscheinlichen Elementarereignisse ist somit

$$\binom{N}{i-1}\cdot(N-i+1).$$

Für die günstigen Elementarereignisse gibt es $\binom{M}{Ac}$ Möglichkeiten, Ac fehlerhafte Objekte aus der Gesamtheit der Fehler auszuwählen. Dann müssen gleichzeitig aus den $N-M$ tadellosen Objekten $i-(Ac+1)$ fehlerfreie Objekte gezogen werden. Daß das darauffolgende i-te Objekt fehlerhaft ist, dafür existieren dann nur noch $M-Ac$ Möglichkeiten. Die Anzahl der für das Ereignis günstigen Elementarereignisse beträgt also

$$\binom{M}{Ac}\binom{N-M}{i-Ac-1}(M-Ac).$$

Die Wahrscheinlichkeit ergibt sich aus dem Quotienten von Anzahl der günstigen durch Anzahl der möglichen Elementarereignisse

$$P(m_{i-1}=Ac\wedge m_i-m_{i-1}=1)=\frac{\binom{M}{Ac}\binom{N-M}{i-Ac-1}(M-Ac)}{\binom{N}{i-1}(N-i+1)}.$$

Da in den Binomialkoeffizienten sehr große Fakultäten vorkommen und diese numerisch nicht mehr handhabbar sind, wird der Term in eine Produktform überführt:

$$P(m_{i-1}=Ac\wedge m_i-m_{i-1}=1)=\frac{M}{Ac!\,(N-M-i+Ac+1)}\cdot\prod_{k=1}^{Ac}\frac{M-k}{N-i+k}\cdot\prod_{k=0}^{i-Ac-1}\frac{N-M-k}{N-k}\cdot\prod_{k=1}^{Ac}(i-k).$$

Da die Wahrscheinlichkeit $P(m_n \le Ac)$ direkt der Operationscharakteristik $L_{N,n,Ac}(M/N)$ entspricht, kann der Erwartungswert für den mittleren Stichprobenumfang n' ermittelt werden. In Abbildung 60 werden der mittlere Stichprobenumfang der vollständigen und der abgebrochenen Prüfung anhand des Beispiels aus Abschnitt 7.7.1 gegenübergestellt.

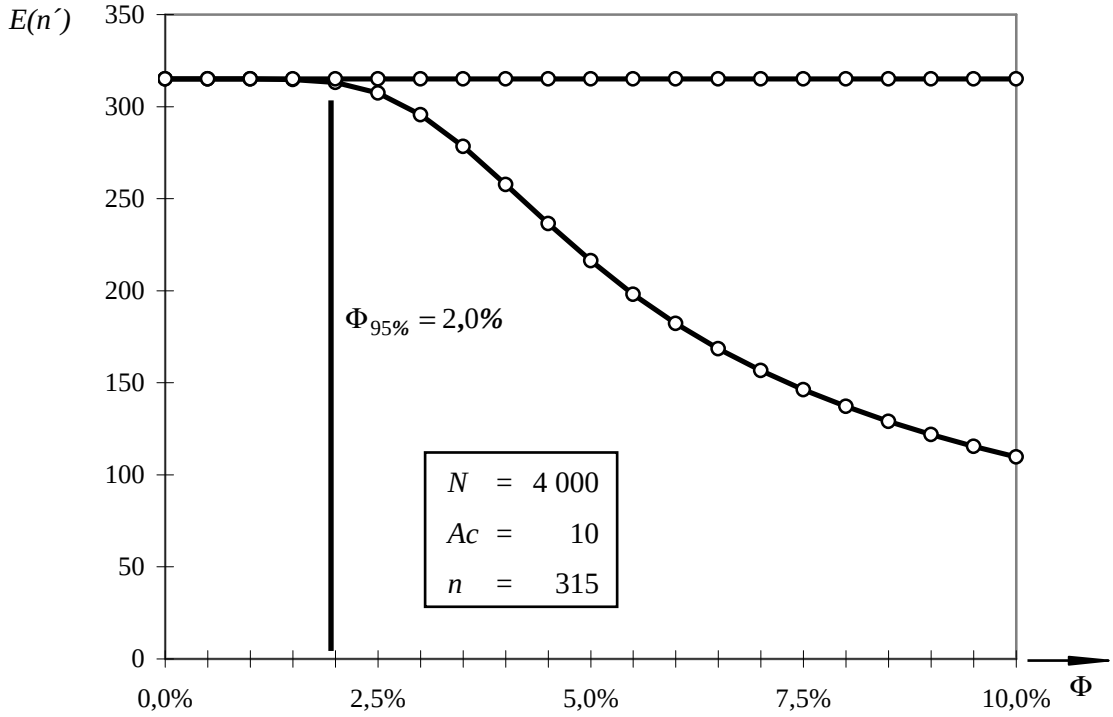


Abbildung 60: Erwartungswert für den Stichprobenumfang der abgebrochenen Kontrolle in Abhängigkeit von der Fehlerrate der Grundgesamtheit.

Da für $\Phi = 0$ keine fehlerhaften Objekte in der Stichprobe gefunden werden können, und da mindestens $Ac + 1$ Objekte untersucht werden müssen, auch wenn alle Objekte der Grundgesamtheit fehlerhaft sein sollten, so ergeben sich für den Erwartungswert die beiden Extreme

$$E(n') = \begin{cases} n & \text{für } \Phi = 0 \quad (M = 0) \\ Ac + 1 & \text{für } \Phi = 1 \quad (M = N) \end{cases}.$$

7.7.5 Mehrstufige und sequentielle Stichprobenpläne

Eine für die statistische Qualitätskontrolle sehr nützliche Verallgemeinerung des einstufigen Alternativtests stellen die mehrstufigen Tests dar (*Uhlmann, 1982*). Bei einem *m*-stufigen Test für eine Nullhypothese H_0 gegen die Alternative H_A zieht man in einem ersten Schritt eine Stichprobe vom Umfang n_1. Die Testvorschrift teilt nun den Stichprobenraum in einen Annahme- und einen Ablehnungsbereich und einen dritten Bereich, der in einstufigen Verfahren nicht vorkommt. Fällt das Stichprobenergebnis in diesen dritten Bereich, wird zunächst keine Entscheidung über H_0 und H_A getroffen, sondern eine weitere Stichprobe vom Umfang n_2 zufällig ausgewählt. Man trifft also in der ersten Stufe eine der folgenden drei Entscheidungen:

$d_0 = H_0$ wird angenommen,

$d_1 = H_A$ wird angenommen,

d_2 = weitere Stichprobe ziehen.

Mit der Entscheidung d_0 oder d_1 ist der Test beendet. Lautet die Entscheidung, eine weitere Stichprobe zu ziehen, dann beläuft sich der Stichprobenumfang auf n_1+n_2. Auf Basis dieser erweiterten Stichprobe ist nun wieder eine Entscheidung zwischen den drei möglichen Ausgängen zu treffen. Dieses Verfahren

kann sich, falls es nicht durch eine der Entscheidungen d_0 oder d_1 vorher abgebrochen wird, bis zu *s* mal wiederholt werden. Der Indifferenzbereich wird bei jeder Vergrößerung der Stichprobe enger. Die Akzeptanz- und Ablehnungszahl der *s*-ten Stufe ist so festzulegen, daß eine Entscheidung d_2 nicht getroffen werden kann. Damit ist der erforderliche Stichprobenumfang mit $n_1+n_2+...+n_s$ nach oben beschränkt.

Ein Spezialfall der mehrstufigen Stichprobenpläne ist gegeben, wenn der Stichprobenumfang n_i für jede Stufe *i* mit $n_i=1$ festgelegt wird und die Anzahl der Stufen a priori nicht vorgeschrieben ist. Diese speziellen Tests werden als sequentielle Stichprobenpläne bezeichnet. Vor jedem Ziehen eines weiteren Objektes ist zu berechnen, ob schon eine Entscheidung gefällt werden kann, oder ob weitere Untersuchungen erforderlich sind. Diese Berechnung ist gegenüber der Prüfung mit einem hohen Aufwand verbunden. Zum Beispiel steht von vornherein fest, daß nach der Untersuchung des ersten Objektes noch keine Entscheidung d_0 oder d_1 getroffen werden darf. Außerdem ist das Verfahren zur Prüfung von Geodaten so festzulegen, daß mehrere Objekte aus unterschiedlichen Gebieten des Loses geprüft werden müssen. Aus diesen Gründen sind sequentielle Verfahren bei der Kontrolle von Geodaten nicht geeignet.

Mehrstufige Stichprobenverfahren haben den Vorteil, daß in klaren Fällen, bei denen die Lose sehr wenige fehlerhafte Objekte enthalten, oder wenn die Fehlerrate die akzeptable Grenze weit übersteigt, sehr schnell eine Entscheidung getroffen werden kann. Im Grenzbereich steigt die Anzahl der zu untersuchenden Objekte, ehe eine statistisch signifikante Aussagen gemacht werden kann.

Bei sehr langwierigen Untersuchungen am Einzelobjekt, die auch parallel ausgeführt werden können, sind mehrstufige Stichprobenpläne ungeeignet. Solche Untersuchungen kommen jedoch bei Geodaten in der Regel nicht vor.

Bei der Qualitätskontrolle von Geodaten steigt der Kontrollaufwand näherungsweise linear mit der Anzahl der zu prüfenden Objekte. Aus wirtschaftlichen Überlegungen ist es daher sinnvoll, den erforderlichen Stichprobenumfang gering zu halten. Daher bietet sich für die statistische Qualitätskontrolle von Geodaten an, mehrstufige Untersuchungen durchzuführen, deren Umfang der Einzelstufe n_i so gewählt wird, daß Objekte aus mehreren Gebieten des Loses untersucht werden können, um weitestgehend auszuschließen, daß lokale Inhomogenitäten die Entscheidung über Akzeptanz oder Ablehnung verfälschen.

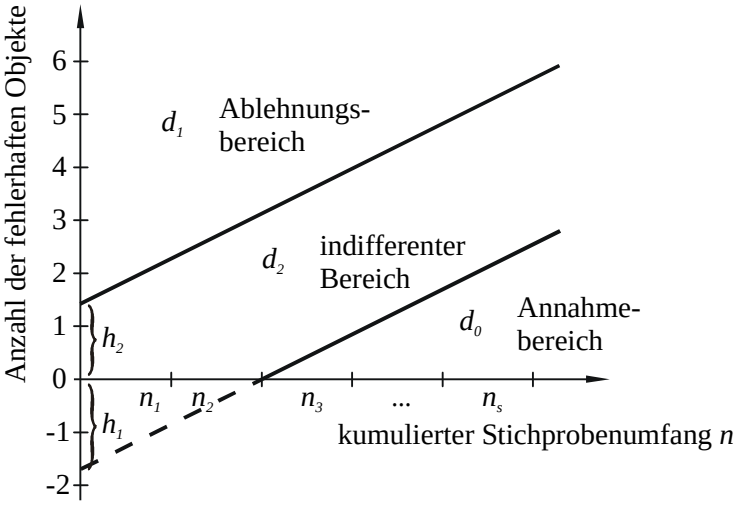
Der Stichprobenplan für ein mehrstufiges Prüfverfahren wird üblicherweise in Tabellenstruktur dargestellt.

Stufe	Stichproben-umfang	Umfang insgesamt	Annahmezahl	Rückweisezahl
1	n_1	n_1	Ac_1	Re_1
2	n_2	$n_1 + n_2$	Ac_2	Re_2
3	n_3	$n_1 + n_2 + n_3$	Ac_3	Re_3
:	:	:	:	:
s	n_s	$\sum_{i=1}^{s} n_i$	Ac_s	$Re_s = Ac_s + 1$

Die Annahme- und Rückweisezahlen beziehen sich jeweils auf den gesamten Stichprobenumfang bis zu dieser Stufe. Solange die Fehlerzahl im indifferenten Bereich d_2 liegt, müssen weitere Stichproben entnommen werden.

Der mathematische Zusammenhang zwischen den Annahme- und Rückweisezahlen und den Irrtumswahrscheinlichkeiten für den Fehler 1. (α) und 2. Art (β) in Abhängigkeit von den Fehlerraten $\Phi_0 = M_0 / N$ beziehungsweise $\Phi_A = M_A / N$ ist in *Montgomery, 1991*, und *Menges u. Skala, 1973*, gegeben.

Die Ableitung der Formeln geht auf die Sequentialanalyse (*Wald, 1945, Fisz, 1976)* zurück und gilt streng nur für sequentielle Stichprobenverfahren. Mehrstufige Stichprobenpläne können als Spezialfall von Sequentialtests betrachtet werden, wobei pro Stufe nicht einzelne Elemente, sondern ein kleinerer Stichprobenumfang ausgewählt werden.



$$h_1 = \left(\log\frac{1-\alpha}{\beta}\right)/k \quad h_2 = \left(\log\frac{1-\beta}{\alpha}\right)/k$$

$$k = \log\frac{\Phi_A(1-\Phi_0)}{\Phi_0(1-\Phi_A)} \quad s = \left(\log\frac{1-\Phi_0}{1-\Phi_A}\right)/k$$

$$X_{Ac} = -h_1 + s \cdot n$$

$$X_{Re} = h_2 + s \cdot n$$

Abbildung 61: Indifferenter, Ablehnungs- und Annahmebereich eines mehrstufigen Stichprobenplanes.

Der indifferente Bereich d_2 wird dabei als Fläche zwischen zwei parallelen Geraden beschrieben. Die Steigung der Geraden und die y-Achsenabschnitte sind Funktionen dieser statistischen Größen.

7.7.6 Mittlerer Stichprobenumfang bei mehrstufigen Stichprobenplänen

Aus der Wahrscheinlichkeit, mit der eine bestimmte Stufe des mehrstufigen Stichprobenplanes erreicht wird, kann der Erwartungswert für den mittleren Stichprobenumfang zur Kontrolle eines Loses bis zur Entscheidung über Akzeptanz oder Ablehnung errechnet werden.

$$E[\bar{n}] = n_1 + \sum_{i=2}^{s} n_i \cdot P\left(Ac_{i-1} < X_{i-1} < Re_{i-1}\right)$$

Die Wahrscheinlichkeit dafür, daß die Anzahl der fehlerhaften Objekte im indifferenten Bereich d_2 liegt, ist eine Funktion des tatsächlichen Qualitätsniveaus des Loses. Die in *Uhlmann, 1982*, gegebene Formel zur Berechnung des Erwartungswertes für den mittleren Stichprobenumfang

$$E[\bar{n}] = n_1 + n_2 \sum_{m=Ac_1+1}^{Re_1-1} \binom{M}{m}\binom{N-M}{n_1-m} \Big/ \binom{N}{n_1}$$

gilt nur für einen zweifachen Stichprobenplan. Bei mehrstufigen Verfahren hängt die Wahrscheinlichkeit über den Ausgang der Untersuchung von den vorhergehenden Stufen ab, weil zur Entscheidung über ein Los immer die kumulierte Stichprobe betrachtet wird. Darum werden zur Bestimmung dieser Wahrscheinlichkeit nur die hinzukommenden fehlerhaften Objekte betrachtet.

$$P\left(Ac_{i-1} < X_{i-1} < Re_{i-1}\right) = \sum_{i=1}^{s} \sum_{k_i=Ac_i+1}^{Re_i-1} \prod_{j=1}^{s} P(X = k_j - k_{j-1})$$

$$\text{mit} \quad X \sim H\left(N - \sum_{l=1}^{j-1} n_l, n_j, M - \sum_{l=1}^{j-1} k_l\right); \; N > \sum_{l=1}^{j-1} n_l; \; M > \sum_{l=1}^{j-1} k_l$$

Dieser Zusammenhang trägt der Tatsache Rechnung, daß sich mit jeder Stufe die Anzahl der Objekte der Grundgesamtheit um n_i vermindert, und daß sich mit jedem gefundenen fehlerhaften Objekt die Fehler der Grundgesamtheit um Eins reduzieren.

In Abbildung 62 sind die möglichen Elementarereignisse als Baumdiagramm dargestellt. Jeder Pfeil steht für das Ereignis, daß eine bestimmte Anzahl (teilweise im Oval angegeben) an fehlerhaften Objekten bei der i-ten Ziehung hinzugekommen sind.

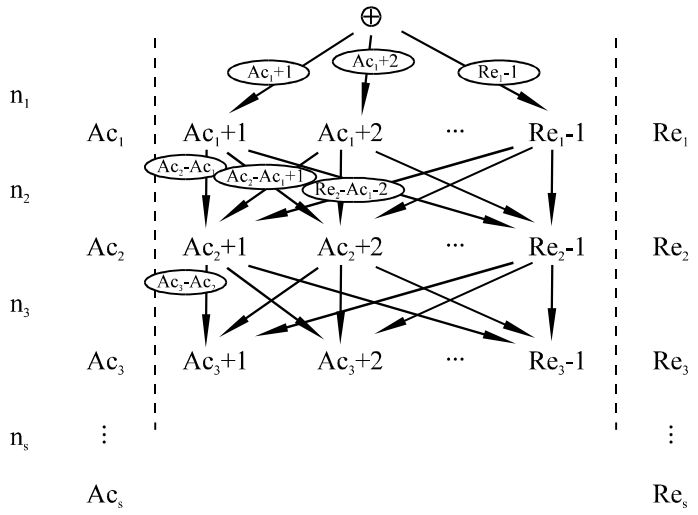


Abbildung 62: Baumdiagramm der Elementarereignisse für die Berechnung der Wahrscheinlichkeit, daß zwischen einer Stufe und der vorhergehenden eine Anzahl von Fehlern entdeckt werden.

Aus den berechneten Wahrscheinlichkeiten läßt sich der Erwartungswert für den mittleren Stichprobenumfang ermitteln. Für das folgende Beispiel wird der Zusammenhang in Abbildung 63 graphisch dargestellt.

Stufe i	n_i	Σn_i	Ac_i	Re_i
1	50	50	#[7]	3
2	50	100	0	3
3	50	150	1	4
4	50	200	2	5
5	50	250	3	6
6	50	300	4	6
7	50	350	6	7

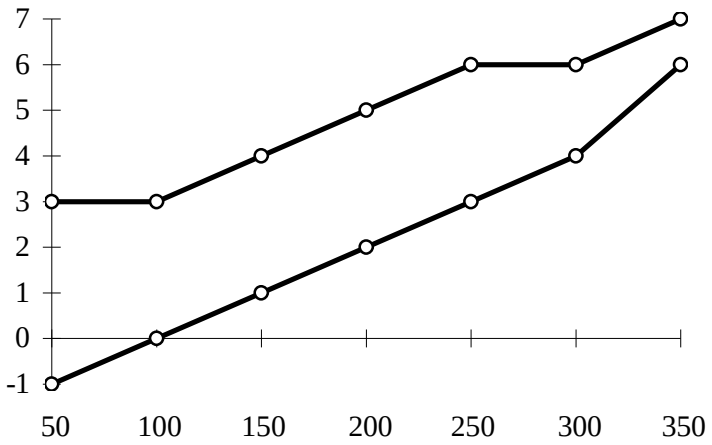


Abbildung 63: Beispiel für einen mehrstufigen Stichprobenplan nach ISO 2859-0.

Für die einzelnen Stufen lassen sich nach der hypergeometrischen Verteilung die Wahrscheinlichkeiten für die Fehler 1. Art (α) - nämlich das Los abzulehnen, ob wohl es den Qualitätsanforderungen entspricht - und 2. Art (β) - nämlich das Los zu akzeptieren, obwohl das Qualitätsniveau zu niedrig ist - berechnen. Zur Berechnung wurden folgende Annahmen getroffen:

Die Anzahl der Objekte der Grundgesamtheit beträgt N = 10 000. Die Anzahl der fehlerhaften Objekte nach der Nullhypothese liegt bei M_0 = 65, die nach der Alternativhypothese ist um den Faktor 5 größer bei M_A = 325.

[7] Bei der ersten Stufe darf ein Los nicht angenommen werden.

$$\alpha_i = P(m \geq Re_i) = 1 - P(m \leq Re_i - 1) = 1 - \sum_{k=0}^{Re_i} P(m = k) \Big|_{m \sim H(N, \sum_{j=1}^{i} n_j; M_0 / N)}$$

$$\beta_i = P(m \leq Ac_i) = \sum_{k=0}^{Ac_i} P(m = k) \Big|_{m \sim H(N, \sum_{j=1}^{i} n_j; M_A / N)}$$

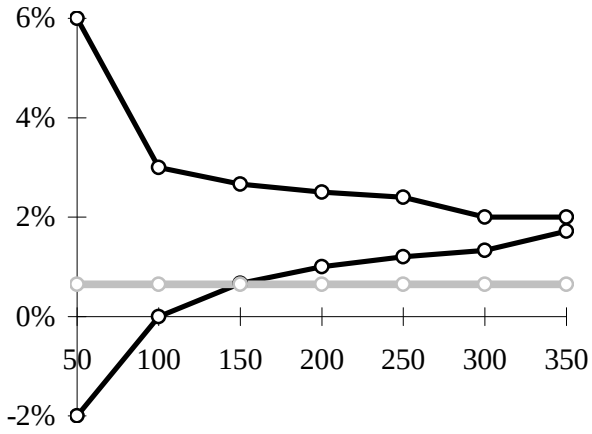
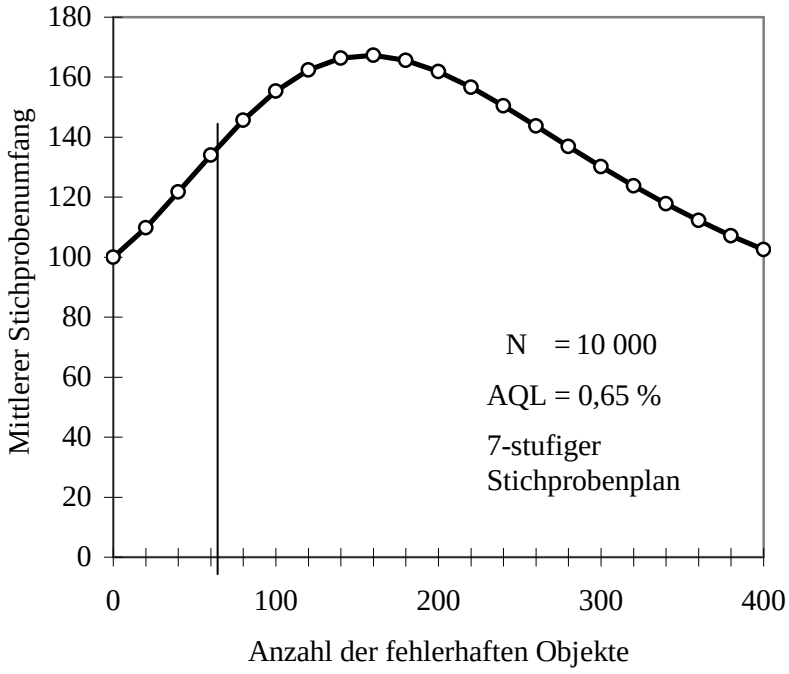


Abbildung 64: Fehlerrate bei der Annahme- und der Rückweisezahl und in grau: Fehlerrate, wenn sich die Grundgesamtheit auf AQL-Niveau befindet.

Stufe i	α_i	β_i
1	0,4 %	0,0 %
2	2,7 %	3,6 %
3	1,6 %	4,1 %
4	1,0 %	3,9 %
5	0,6 %	3,5 %
6	1,3 %	3,0 %
7	0,7 %	5,8 %

Die Wahrscheinlichkeiten für die Fehler 1. und 2. Art sind in dieser Tabelle zusammengestellt.

Der Erwartungswert für den mittleren Stichprobenumfang eines mehrstufigen Stichprobenplans hängt von der wahren Fehlerrate der Grundgesamtheit eines Loses ab. Wenn die Fehlerrate sehr gering ist, so kann das Los schon nach wenigen Untersuchungen akzeptiert werden. Bei einer sehr hohen Fehlerrate



Der Stichprobenumfang einer einfachen Stichprobe mit annähernd gleicher Operations-charakteristik beträgt $n = 268$.

Abbildung 65: Erwartungswert für den mittleren Stichprobenumfang anhand des Beispiels.

übersteigen nach wenigen Schritten die gefundenen Fehler die Akzeptanzgrenze und das Los kann verworfen werden, ohne daß weitere Untersuchungen angestellt werden müssen. Im mittleren Bereich ist es wahrscheinlicher, daß die Fehleranzahl im indifferenten Bereich liegt, und somit weitere Stichproben untersucht werden müssen.

7.7.7 Mittlerer Durchschlupf

Auch nach der Fehlerbehandlung bei Ablehnung eines Loses kann nicht davon ausgegangen werden, daß die Fehlerrate komplett auf null Fehler reduziert ist. Allerdings muß erwartet werden, daß durch die Revision der Daten die Fehlerrate erheblich vermindert wird. Diese Verminderung kann entweder als Faktor[8] $f << 1$ betrachtet werden, oder als eine Reduzierung auf eine unvermeidbare Fehlerrate[9] $\Phi_u = M_u / N$. Wenn die unvermeidbare Fehleranzahl nicht bekannt ist, wird oft ein Wert von null fehlerhaften Objekten angenommen. Für den Erfassungsprozeß von Geodaten wäre diese Annahme allerdings sehr optimistisch.

Die Fehlerrate nach der Stichprobenkontrolle beträgt

$$\hat{\Phi} = \begin{cases} M / N & \text{für } m \le Ac \\ \hat{\Phi}_u = M_u' / N \quad \text{bzw. } \Phi_u = M_u / N & \text{für } m > Ac \text{ mit } M_u' = f \cdot M \end{cases}$$

und der mittlere Ausschußanteil

$$E(\hat{\Phi}) = \frac{M}{N} \cdot L_{N,n,Ac}(M / N) + \frac{M_u}{N} \cdot \left(1 - L_{N,n,Ac}(M / N)\right) = \frac{M_u}{N} + \frac{M - M_u}{N} \cdot L_{N,n,Ac}(M / N).$$

Dieser Anteil wird mit **mittlerer Durchschlupf** oder auch mittlere Auslieferungsqualität (englisch: *average outgoing quality* = AOQ) bezeichnet. Er gibt an, mit welcher Fehlerrate ein Anwender auch nach Durchführung einer statistischen Qualitätskontrolle im Mittel zu rechnen hat.

Die Funktion des mittleren Durchschlupfes in Abhängigkeit von der wahren Fehlerrate hat einen Maximalwert, der Höchstwert des mittleren Durchschlupfes (englisch: *average outgoing quality limit* = AOQL) genannt wird. Der Höchstwert kann durch einfaches Einsetzen von Werten für *M* ermittelt werden. Setzt man für die Fehlerrate der Stichprobe eine Binomialverteilung voraus, so kann der Höchstwert durch Lösen einer Extremalaufgabe ermittelt werden (*Uhlmann, 1982*).

---

[8] Dieser Ansatz geht davon aus, daß die Anzahl der verbleibenden Restfehler um so größer ist, je größer die Fehlerrate

[9] Dieser Ansatz beruht auf der Annahme, daß unabhängig von der ursprünglichen Fehleranzahl eine bestimmte Anzahl unvermeidbarer Restfehler übrig bleiben.

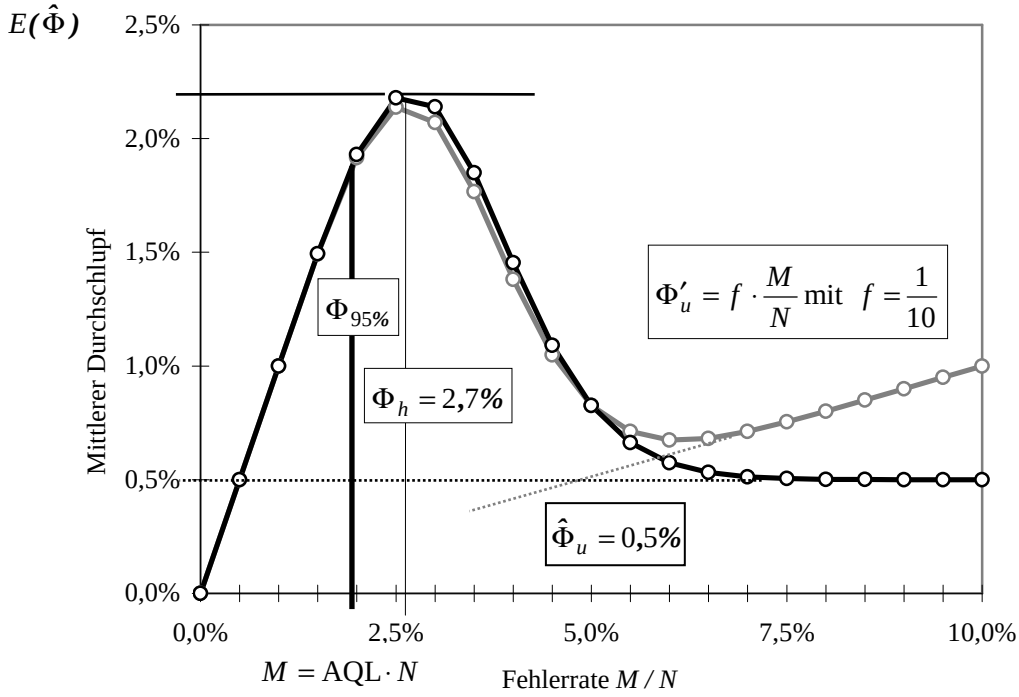


Abbildung 66: Mittlerer Durchschlupf in % in Abhängigkeit von der Fehlerrate der Grundgesamtheit für den Stichprobenplan aus Beispiel in Abschnitt bei konstantem Anteil von Restfehlern (dunkel) und bei proportionaler Restfehlerrate (grau).

7.7.8 Kostenoptimale Prüfpläne

Wenn beim Einsatz eines Geoinformationssystemes mit fehlerhaften Objekten gearbeitet wird, so ziehen diese Fehler im allgemeinen Kosten nach sich, die von verschiedenen Faktoren abhängig sind (z.B. Art des Fehlers: Welches Qualitätskriterium wurde verletzt? Welcher Objektklasse gehört das fehlerhafte Objekt an?). Andererseits kann der Aufwand zur Vermeidung oder Eliminierung von Fehlern so weit getrieben werden, daß die damit verbundenen Kosten die Vermeidung von Folgekosten durch Datenfehler nicht mehr rechtfertigen.

Unter bestimmten Annahmen über die Kosten, die von Erfahrungswerten oder aus Aufzeichnungen über längere Zeiträume abgeleitet werden können, lassen sich die Gesamtkosten eines Systems ermitteln. Diese Kosten werden den Kosten gegenübergestellt, die entstanden wären, hätte keine Qualitätsprüfung stattgefunden.

Für die Berechnungen wurden folgende Kosten berücksichtigt:

- K_1: Kosten, die durch fehlerhafte Objekte entstehen, unter Berücksichtigung der Qualitätskontrolle
- K_2: Kosten der Qualitätskontrolle
- K_3: Kosten zur Korrektur der Fehler

Dem wird gegenübergestellt

- K_1^*: Kosten durch fehlerhafte Daten ohne Qualitätskontrolle

Die durch fehlerhafte Objekte verursachten Kosten ergeben sich aus den mittleren Kosten, die durch das Arbeiten mit einem fehlerhaften Objekt entstehen (z.B. durch Mehraufwand, Planungsfehler, Sachschaden), aus der Häufigkeit, mit der auf die Geodaten zugegriffen wird, und aus der Wahrscheinlichkeit, ein fehlerhaftes Objekt anzutreffen. Diese Wahrscheinlichkeit ist unmittelbar von der Qualität der Daten abhängig, denn sie ist identisch mit dem mittleren Durchschlupf.

Die Kosten einer Stichprobenuntersuchung setzen sich aus fixen Kosten (z.B. durch Einrichtung eines GIS-Arbeitsplatzes für die Prüfung mit Hardware, Software und Peripherie oder Einarbeitung von Prüfpersonal) und den Kosten zur Untersuchung der einzelnen Objekte zusammen.

Eine Nachbearbeitung aller fehlerhaften Objekte eines Loses ist nur dann erforderlich, wenn das Los bei der Ein- oder Ausgangskontrolle zurückgewiesen wurde. Dann müssen zuerst alle verbleibenden Objekte, die bei der Stichprobenauswahl nicht enthalten waren, untersucht werden, d.h. es muß eine Totalkontrolle erfolgen. Alle Fehler, die bei den Untersuchungen gefundenen wurden, müssen eliminiert werden. Für beide Arbeitsgänge fallen Kosten an, zu denen noch Fixkosten hinzukommen.

Für die einzelnen Posten werden Variablen eingeführt:

- b_1: Mittlere Kosten, die aus der Arbeit mit einem fehlerhaften Geoobjekt resultieren
- a_2: Fixe Kosten für die Qualitätskontrolle
- b_2: Kosten zur Kontrolle eines einzelnen Objekts (z.B. Stundenlohn eines Prüfers / Prüfleistung pro Stunde)
- a_3: Fixe Kosten für die Fehlerkorrektur
- c_3: Kosten zur Korrektur eines Objektes im Datenbestand
- N_{GIS}: Anzahl der Zugriffe auf Geoobjekte bei der Anwendung

Mit Hilfe dieser Variablen können Erwartungswerte für die aufgeführten Kosten berechnet werden.

$$E(K_1) = b_1 \cdot \frac{N_{GIS}}{N} \cdot \left(M \cdot L_{N,n,Ac}(M/N) + M_u \cdot \left(1 - L_{N,n,Ac}(M/N)\right)\right) = b_1 \cdot N_{GIS} \cdot E(\hat{\Phi})$$

$$E(K_2) = a_2 + b_2 \cdot n$$

$$E(K_3) = \left(a_3 + b_2 \cdot (N-n) + c_3 \cdot (M - M_u)\right) \cdot \left(1 - L_{N,n,Ac}(M/N)\right)$$

$$E(K_1^*) = b_1 \cdot \frac{N_{GIS}}{N} \cdot M$$

Damit der Aufwand zur Durchführung der Qualitätsprüfung gerechtfertigt ist, müssen die Folgekosten aufgrund von Datenfehlern samt der Kosten für die Kontrolle und Verbesserung stets unter den Folgekosten durch ein ungeprüftes Annehmen der Lose liegen. Auch Imageverluste und schwindende Akzeptanz des Geoinformationssystems wegen einer zu großen Fehlerhäufigkeit müssen dabei pekuniär ausgedrückt und bei der Bestimmung von b_1 berücksichtigt werden.

$$K_1 + K_2 + K_3 < K_1^*$$

Da sich die Investitionen zur Durchführung der Qualitätskontrolle durch Vermeidung von Datenfehlern und dadurch auch von Folgekosten im Laufe der Zeit ausgleichen, wird für den Punkt, bei dem sich die Kosten die Waage halten, der Begriff **Amortisationspunkt** (englisch: *break-even point*) eingeführt. Da die zur Bestimmung des Amortisationspunktes angesetzte Anzahl von Zugriffen auf Geoobjekte in einer bestimmten Zeitspanne erfolgen, kann dieser Zeitraum auch als Amortisationsdauer bezeichnet werden.

Für einen bestehenden Prüfplan *(n, Ac)* können die dadurch zu erwartenden Kosten ermittelt werden. Im folgenden Beispiel werden für die Posten Annahmen getroffen, die in beliebigen Währungseinheiten ausgedrückt sind. Tatsächlich ist die Einheit, in der die pekuniären Posten ausgedrückt sind, unerheblich, da zur Bestimmung des Amortisationspunktes nur relative Angaben benötigt werden. Der Prüfplan bezieht sich auf das Beispiel in Abschnitt 7.7.1.

Bei einem Losumfang von N = 4 000 Objekten wird erwartet, daß in einem bestimmten Zeitraum N_{GIS} = 1000 Objekte abgefragt werden. Beim Arbeiten mit einem fehlerhaften Objekt wird geschätzt, daß der dadurch entstehende Schaden im Mittel b_1 = 500 WE (Währungseinheiten) beträgt. Für die Prüfung und die Nachbearbeitung der Objekte sind fixe Kosten von jeweils $a_2 = a_3$ = 5 000 WE veranschlagt. Die Prüfung eines einzelnen Objektes soll b_2 = 2,50 WE und die Korrektur eines einzelnen Fehlers $c_3 = 10 \cdot b_2$ = 25 WE betragen. In Abbildung 67 werden in Abhängigkeit von der tatsächlichen Fehlerrate des Loses die einzelnen Kosten graphisch dargestellt. In Abbildung 68 sind die

Gesamtkosten und die Kosten, die ohne Qualitätsprüfung entstünden, gegeneinander aufgetragen. Bei den Kosten K_1 und K_2 sind die beiden Modelle der unvermeidbaren Fehler M_u berücksichtigt. *K1´* und *K3´* gehen von einem fehlerproportionalen Anteil von unvermeidbaren Fehlern nach der Kontrolle aus (M_u=M/10). Bei *K1* und *K3* (ohne Strich) ist diese Fehlerzahl konstant (M_u = 20).

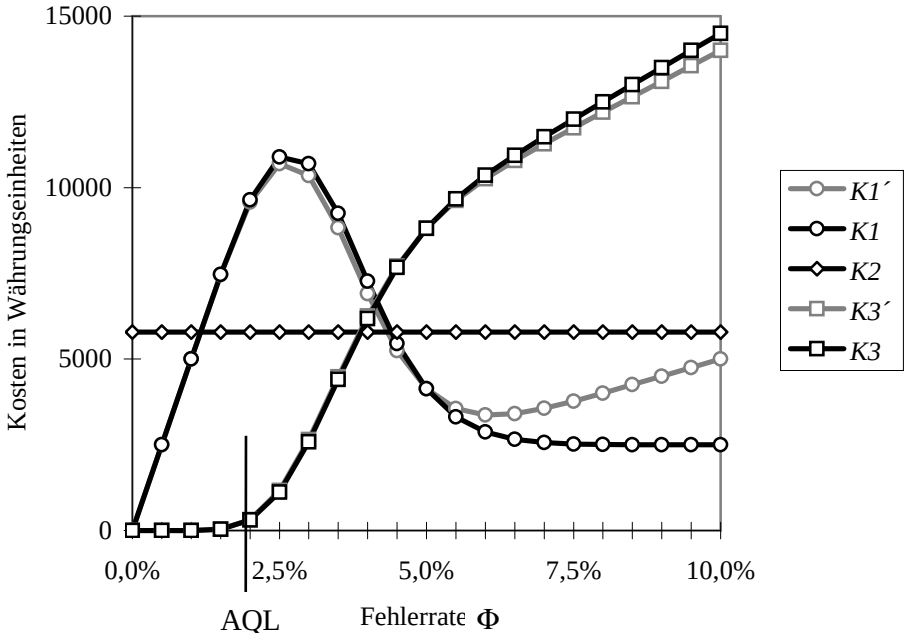


Abbildung 67: Darstellung der einzelnen Kostenanteile *K1, K2* und *K3* in Abhängigkeit von der Fehlerrate Φ.

Bei der Kostenbetrachtung werden immer die Kosten berücksichtigt, die dem Anwender entstehen. Abhängig von der vertraglichen Vereinbarung zwischen Datenproduzent und Anwender, muß der Anwender die Kosten für die Fehlerkorrektur selbst tragen, oder er lehnt das Los ab und gibt es dem Produzent zurück. Im letzteren Fall entstehen die Kosten für die Korrektur der Daten (*K3* oder *K3´*) nicht dem Anwender sondern dem Produzenten. Das ist das unternehmerische Risiko bei der Datenproduktion, das allerdings im Bereich der annehmbaren Qualitätsgrenzlage sehr gering ist, wie aus Abbildung 67 ersichtlich ist.

Der Schnittpunkt der Kostenentwicklungen ohne und mit Qualitätsprüfung gibt den Punkt an, ab dem die Kosten des Prüfverfahrens geringer sind als die Kosten der Fehler ohne Qualitätsprüfung. Dieser Schnittpunkt entspricht dem eingeführten **Amortisationspunkt**. Wenn die Kosten K1 und K2 berücksichtigt werden, so liegt der Amortisationspunkt im Beispiel bei Φ_{A12} = 3,2 % und wenn der Anwender für K1, K2 und K3 aufkommen muß, so verschiebt sich der Amortisationspunkt auf Φ_{A123} = 3,8 %. Ob die unvermeidbaren Restfehler konstant oder proportional zur Fehlerrate sind, wirkt sich dabei kaum aus.

Mit diesem Verfahren kann die Qualitätsgrenzlage nach objektiven Gesichtspunkten ermittelt werden. Die Kosten für die Erfassung, Kontrolle und Prüfung sind im allgemeinen bekannt. Die Folgekosten, die im Mittel durch das Arbeiten mit fehlerhaften Objekten entstehen, können von Anwendern aus Erfahrungswerten abgeschätzt werden. Unter der Bedingung, daß die Kosten für die Qualitätskontrolle die zu erwartenden Folgekosten nicht übersteigen sollten, läßt sich die akzeptable Qualitätsgrenzlage schon vor der Digitalisierung bestimmen. Damit kann der Aufwand bei der Erfassung den Anforderungen aus der Anwendung angepaßt werden. Wenn das GIS auch nach wirtschaftlichen Gesichtspunkten betrieben werden soll, ist das hier entwickelte Verfahren auch für die praktische Anwendung relevant.

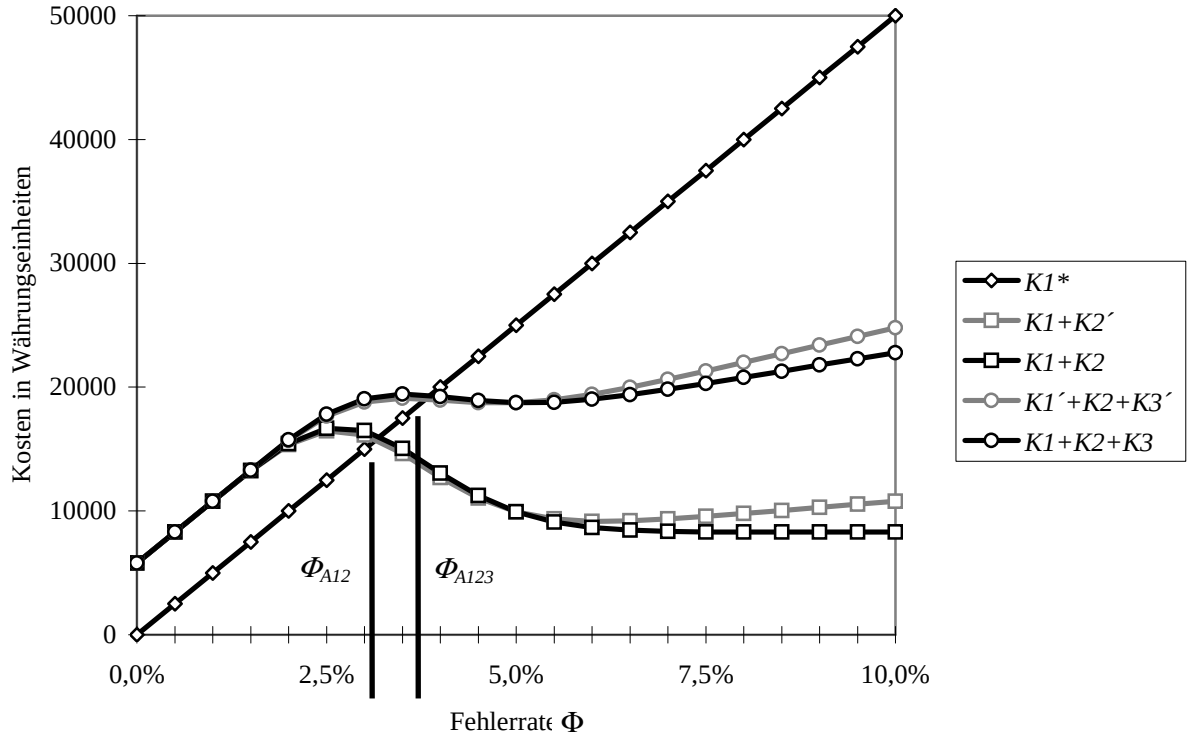


Abbildung 68: Gegenüberstellung der Gesamtkosten ohne Qualitätsprüfung K_1^* und mit Qualitätsprüfung *K1+K2+K3* bzw. *K1+K2* in Abhängigkeit von der Fehlerrate Φ.

7.8 Normale, verschärfte und reduzierte Prüfung

Damit der Stichprobenumfang immer möglichst gut der Qualitätslage der Geodaten angepaßt ist, kann nach *DIN ISO 2859-1, 1993*, einen Verfahrenswechsel erfolgen, falls die Objekte über längere Zeit als besonders gut oder besonders schlecht eingestuft wurden. Dadurch soll gewährleistet werden, daß eine Verschlechterung der Qualitätslage frühzeitig entdeckt wird. Dies bietet einen automatischen Schutz für den Abnehmer. Außerdem wird dadurch der Prüfaufwand möglichst gering gehalten, da jede Prüfung auch immer mit nicht unerheblichen Kosten verbunden ist. Der Wechsel zwischen normaler, verschärfter und reduzierter Prüfung ist dann gerechtfertigt, wenn eine konstante Qualität der Datenerfassung auch über mehrere Lose hinweg mit gutem Grund angenommen werden kann.

Sofern nichts anderes vereinbart ist, muß nach *DIN ISO 2859-1, 1993*, mit normaler Prüfung begonnen werden. In besonderen Regeln wird festgelegt, wann auf ein anderes Prüfverfahren übergegangen werden soll. Die Regeln resultieren aus statistischen Überlegungen.

Prinzipiell sieht *DIN ISO 2859-0, 1991*, vor, daß bei der verschärften Prüfung der Stichprobenumfang dem der normalen Prüfung entspricht, aber die Annahmezahl heruntergesetzt wird. Wenn die Annahmezahl allerdings schon bei 1 liegt und eine Reduzierung auf 0 eine ungerechtfertigte Verschärfung darstellte, oder wenn bei einer Annahmezahl von 0 keine Reduzierung mehr möglich ist, dann wird vorgeschlagen, die Verschärfung zu erreichen, indem die Annahmezahl der normalen Prüfung beibehalten, aber der Stichprobenumfang erhöht wird.

Wenn für die reduzierte Prüfung der Stichprobenumfang verringert wird, so ist dabei die Rückweisezahl so zu wählen, daß die Wahrscheinlichkeit für einen Fehler 1. Art, also das Produzentenrisiko, verringert wird, weil ansonsten der Produzent für das anhaltend gute Qualitätsniveau bestraft würde, indem im Schnitt mehr gute Lose zurückgewiesen würden.

DIN ISO 2859-1, 1993, beinhaltet folgende Regeln zum Verfahrenswechsel. Abbildung 71 verdeutlicht die Abläufe zum Verfahrenswechsel in einem Flußdiagramm.

Wechsel von normal nach verschärft: Wenn normal geprüft wird, muß auf verschärfte Prüfung übergegangen werden, wenn von fünf oder weniger aufeinanderfolgenden Losen zwei in der Erstprüfung als unannehmbar beurteilt werden. Wiedervorgestellte Lose sollen bei diesem Verfahren nicht mitgezählt werden.

Unter der Annahme, daß die Nullhypothese richtig ist, ergibt sich für die Wahrscheinlichkeit, daß zwei von fünf aufeinanderfolgenden Losen abzulehnen sind, folgender mathematischer Zusammenhang:

$$P(\text{2 aus 5 Losen unannehmbar}) = \sum_{i=2}^{5} (i-1)\left(P(\text{Los unannehmbar})\right)^2 \left(P(\text{Los annehmbar})\right)^{i-2}$$

$$= \sum_{i=2}^{5} (i-1)\left(1 - P(m \le Ac)\right)^2 \left(P(m \le Ac)\right)^{i-2}$$

$$= \sum_{i=2}^{5} (i-1)\left(1 - \sum_{k=0}^{Ac} P(m=k)\right)^2 \left(\sum_{k=0}^{Ac} P(m=k)\right)^{i-2} \quad \text{mit} \quad m \sim H(N, n, M_0)$$

Die Annahmezahlen sind so festgelegt, daß mit der Vermutung der Richtigkeit der Nullhypothese, die Wahrscheinlichkeit, ein Los abzulehnen, gerade der Irrtumswahrscheinlichkeit α entspricht. Damit ergibt sich die Wahrscheinlichkeit für einen Wechsel von normaler zu verschärfter Kontrolle zu

$$P(\text{Wechsel von normal nach verschärft}) = \sum_{i=2}^{5} (i-1) \cdot \alpha^2 \cdot (1-\alpha)^{i-2}$$

und mit einer Irrtumswahrscheinlichkeit α = 5% ein Wert von 2,3%. Weitere Werte sind in der folgenden Tabelle aufgelistet.

Irrtumswahr-scheinlichkeit α	Wahrscheinlichkeit eines Verfahrenswechsel
0,5%	0,02%
1,0%	0,1%
1,5%	0,2%
2,0%	0,4%
2,5%	0,6%
3,0%	0,8%
3,5%	1,1%
4,0%	1,5%
4,5%	1,8%
5,0%	2,3%
7,5%	4,8%
10%	8,1%

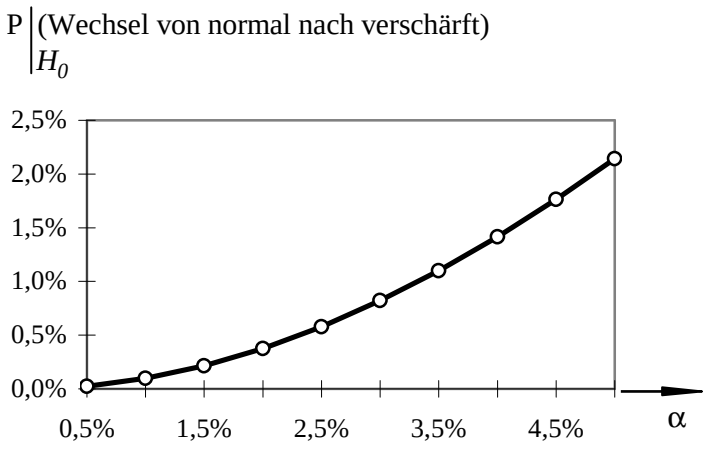


Abbildung 69: Zusammenhang zwischen der festgelegten Irrtumswahrscheinlichkeit und der Wahrscheinlichkeit für einen Verfahrenswechsel von normal zu verschärfter Kontrolle unter der Annahme, daß H_0 richtig ist.

Die Wahrscheinlichkeit, daß ein Verfahrenswechsel von normal zu verschärft entsprechend ISO 2859-1 durchgeführt werden muß, hängt nicht nur von der Irrtumswahrscheinlichkeit, sondern auch von der tatsächlichen Fehlerrate der Objekte der Grundgesamtheit ab. Abbildung 70 veranschaulicht dies für ein Beispiel.

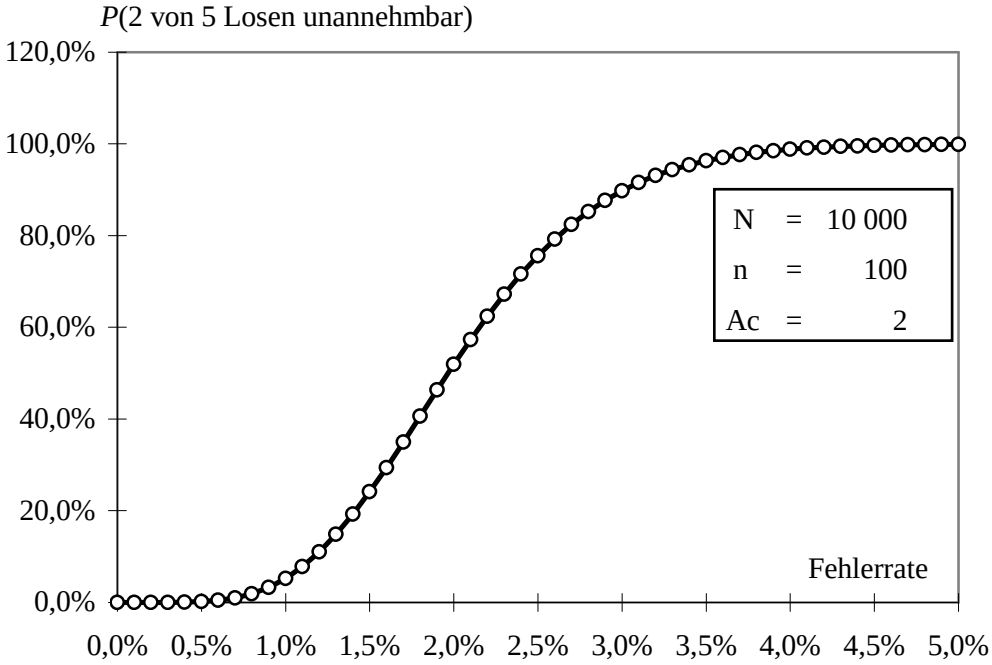


Abbildung 70: Wahrscheinlichkeit für einen Verfahrenswechsel von normal zu verschärft in Abhängigkeit der Fehlerrate für N = 10 000, n = 100 und Ac = 2.

Wechsel von verschärft nach normal: Wenn 5 aufeinanderfolgende Lose in der Erstprüfung als annehmbar beurteilt wurden, muß wieder auf normale Prüfung übergegangen werden,.

Die Wahrscheinlichkeit, daß 5 aufeinanderfolgende Lose als annehmbar eingestuft werden, wird analog zum umgekehrten Verfahrenswechsel errechnet:

$$\begin{aligned} P(\text{5 aufeinanderfolgende Lose annehmbar}) &= \left(P(\text{1 Los annehmbar})\right)^5 \\ &= \left(P(m \le Ac)\right)^5 \\ &= \left(\sum_{i=0}^{Ac} P(m=k)\right)^5 \quad \text{mit} \quad m \sim H(N,n;M) \end{aligned}$$

Je geringer die Fehlerrate, um so größer ist die Wahrscheinlichkeit, daß die verschärfte Prüfung zugunsten der normalen Prüfung ausgesetzt wird.

Wechsel von normal nach reduziert: Wenn normal geprüft wird, muß auf reduzierte Prüfung übergegangen werden, wenn alle folgenden Bedingungen erfüllt sind:

a) Die vorhergehenden 10 Lose sind zur normalen Prüfung vorgestellt worden und alle wurden in der Erstprüfung als annehmbar beurteilt.

b) Die Gesamtzahl fehlerhafter Objekte in den Stichproben der 10 vorhergehenden Lose ist kleiner oder gleich dem betreffenden Grenzwert, der sich aus der Nullhypothese und den statistischen Sicherheitsniveaus errechnen läßt.

c) Die Produktion (Datenerfassung) läuft gleichmäßig.

d) Die zuständige Stelle sieht den Wechsel zu reduzierter Prüfung als erwünscht an.

Die Bedingung a) ist mit einer Wahrscheinlichkeit von

$$P(\text{10 aufeinanderfolgende Lose annehmbar}) = \left(\sum_{i=0}^{Ac} P(m = k) \right)^{10} \quad \text{mit} \quad m \sim H(N, n, M)$$

erfüllt.

Zur Berechnung der Wahrscheinlichkeit für Bedingung b) werden alle untersuchten Objekte der 10 Lose zu einer Stichprobe zusammengefaßt. Der Umfang n^* dieser zusammengefaßten Stichprobe und die Gesamtzahl der gefundenen fehlerhaften Objekte m^* ergibt sich aus der Summe der Einzelstichproben, die Gesamtzahl aller Objekte N^* und aller fehlerhaften Objekte M^* als Summe aus allen Losen.

$$n^* = \sum_{i=1}^{10} n_i, \; m^* = \sum_{i=1}^{10} m_i, \; N^* = \sum_{i=1}^{10} N_i \quad \text{und} \quad M_0^* = \sum_{i=1}^{10} M_{0_i}$$

$$P(m^* \le Ac^*) = 1 - \alpha \quad \text{mit} \quad m^* \sim H(N^*, n^*, M_0^*)$$

$$P(m^* \le Ac^*) = \beta \quad \text{mit} \quad m^* \sim H(N^*, n^*, M_A^*)$$

$$\text{Aus } n^*, m^*, N^*, M_0^*, M_A^* \text{ und } \beta \text{ folgt } Ac^* \text{ und } \alpha.$$

Je größer der Stichprobenumfang, um so schärfere Aussagen können getroffen werden. Durch die Verwendung der kumulierten Stichprobe kann mit einer höheren statistischen Sicherheit entschieden werden, ob die gefundene Fehlerrate durch zufälliges Auswählen von Prüfobjekten entstanden ist, oder ob sie die wahre Fehlerrate der Grundgesamtheit aller Lose hinreichend gut approximiert. Diese Untersuchung setzt wiederum die Bedingung c) voraus, nämlich, daß die Geodaten auch über mehrere Lose hinweg homogen sind.

Wechsel von reduziert nach normal: Wenn reduziert geprüft wird, muß zur normalen Prüfung übergegangen werden, wenn mindestens einer der folgenden Fälle bei der Erstprüfung eintritt:

a) Ein Los ist unannehmbar.
b) Die Anzahl der fehlerhaften Objekte eines Loses zwischen der Annahmezahl und der Rückweisezahl liegt. In diesem Fall wird das Los zwar akzeptiert, aber es muß ein Verfahrenswechsel durchgeführt werden.[10]
c) Die Produktion (Datenerfassung) wird unregelmäßig oder verzögert.
d) Andere Umstände rechtfertigen den Übergang zu normaler Prüfung.

Die Wahrscheinlichkeit für einen Wechsel von einer reduzierten zu einer normalen Prüfung nach Bedingung a) entspricht genau der Wahrscheinlichkeit α für einen Fehler 1. Art.

[10] Der Stichprobenplan muß so konzipiert sein, daß $Ac - Re > 1$ ist.

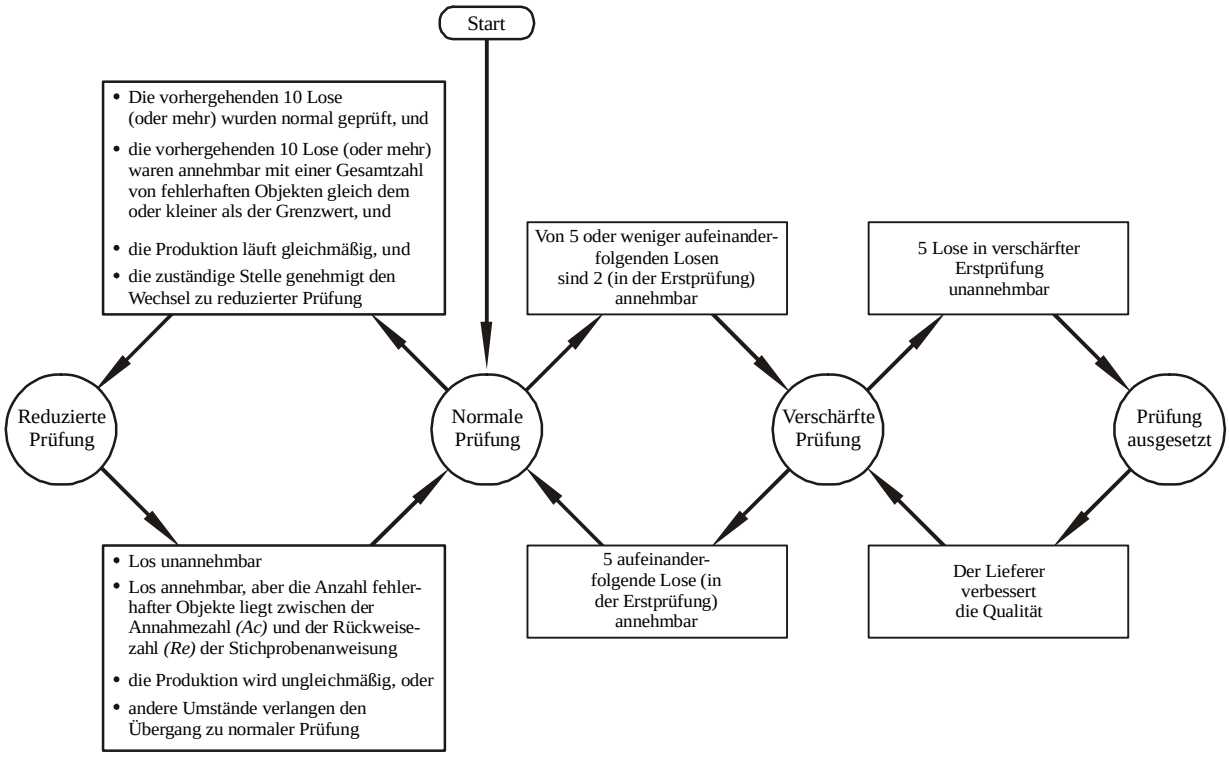


Abbildung 71: Flußdiagramm der Regeln für den Verfahrenswechsel (nach *DIN ISO 2859-1, 1993*).

7.9 Test auf Homogenität

Eine starke Voraussetzung für die Zulässigkeit der Durchführung von repräsentativen Stichprobenuntersuchungen ist die Homogenität der zu prüfenden Lose. Zur Vermeidung von falschen Schlüssen soll nicht durch nur eine Stichprobe aus einem Teilgebiet auf das gesamte Gebiet der Grundgesamtheit geschlossen, sondern die Stichprobe auf mehrere Gebiete verteilt werden. Auch dann ist die Homogenität zu fordern. Allerdings kann wegen der Verteilung auf mehrere Gebiete mit statistischen Methoden untersucht werden, ob signifikante Unterschiede zwischen den Fehlerraten der einzelnen Kontrollgebiete bestehen. Wenn durch diese Untersuchung Häufungsbereiche von Fehlern festgestellt werden, so darf die Stichprobenkontrolle nicht zu einer Annahme oder Ablehnung des Loses führen. In diesem Fall ist eine eingehende Ermittlung der Ursache für diese Häufung anzustellen, die gegebenenfalls doch zu einer Ablehnung des Loses, oder zu einer neuen Festlegung von homogenen Losen führt.

Der Test auf Gleichheit der Fehlerraten von Stichprobengebieten wird als Signifikanztest formuliert. Dabei werden ohne Einschränkung der Allgemeingültigkeit Stichprobengebiete immer paarweise untersucht. Zur Bezeichnung der Testgröße wird eine Nullhypothese in der Weise formuliert, daß die Fehlerraten in beiden Gebieten identisch sind. Die Hypothese ist im Rahmen einer bestimmten statistischen Sicherheit zu verifizieren bzw. zu falsifizieren.

$$H_0 : \frac{M_1}{N_1} = \frac{M_2}{N_2}$$

Es wird weiter vorausgesetzt, daß die Teilgebiete - zumindest annähernd - die selbe Anzahl von Objekten beinhalten. Abhängig von der Art der Objekte und von dem Einfluß, den Fehler bei bestimmten Objektklassen haben, ist sogar eine Forderung nach der gleichen Anzahl von Objekten bestimmter Objektklassen nötig.

Aus der Nullhypothese ergibt sich durch Äquivalenzumformungen

$$\frac{M_1}{N_1} = \frac{M_2}{N_2} \Leftrightarrow 2\arcsin\sqrt{\frac{M_1}{N_1}} = 2\arcsin\sqrt{\frac{M_2}{N_2}} \Leftrightarrow \delta_0 = 0$$

unter Einführung einer neuen Variablen

$$\delta_0 := 2\arcsin\sqrt{\frac{M_1}{N_1}} - 2\arcsin\sqrt{\frac{M_2}{N_2}}\,.$$

Unter Verwendung der Transformation (siehe Abschnitt 7.5.3)

$$u_1 = \left(2\arcsin\sqrt{\frac{m_1}{n_1}} - 2\arcsin\sqrt{\frac{M_1}{N_1}}\right)\sqrt{n_1} \qquad \sim N(0;1)$$

$$u_2 = \left(2\arcsin\sqrt{\frac{m_2}{n_2}} - 2\arcsin\sqrt{\frac{M_2}{N_2}}\right)\sqrt{n_2} \qquad \sim N(0;1)$$

kann eine neue Zufallsvariable gebildet werden mit

$$u' = \frac{u_1}{\sqrt{n_1}} - \frac{u_2}{\sqrt{n_2}} = 2\arcsin\sqrt{\frac{m_1}{n_1}} - 2\arcsin\sqrt{\frac{M_1}{N_1}} - 2\arcsin\sqrt{\frac{m_2}{n_2}} + 2\arcsin\sqrt{\frac{M_2}{N_2}}\,,$$

deren Varianz nach dem Varianzenfortpflanzungsgesetz ermittelt werden kann

$$D^2(u') = \frac{1}{n_1}D^2(u_1) + \frac{1}{n_2}D^2(u_2) = \frac{1}{n_1} + \frac{1}{n_2}\,.$$

Unter der Annahme, daß H_0 mit $\delta_0 = 0$ erfüllt ist, vereinfacht sich die Prüfgröße u', welche durch eine Normierung in eine standardnormalverteilte Zufallsvariable u überführt wird.

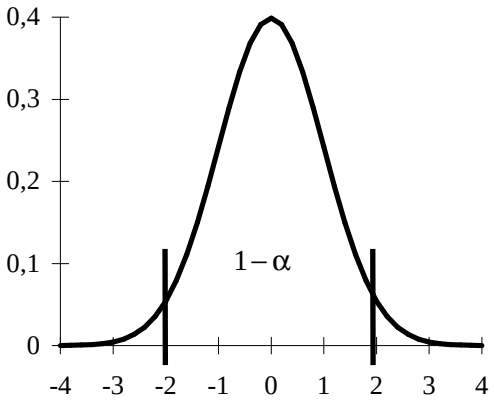
$$u = \frac{u'}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2\arcsin\sqrt{\frac{m_1}{n_1}} - 2\arcsin\sqrt{\frac{m_2}{n_2}}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \sim N(0;1)$$

Die Wahrscheinlichkeit dafür, die Hypothese H_0 zu verwerfen, obwohl sie richtig ist, kann aus der Verteilungsfunktion der Normalverteilung errechnet werden. Es wird ein zweiseitiger Test angewandt, da nur der Betrag von u von Bedeutung ist.

$$P\left(|u| \le u_{1-\alpha/2}\right) = 1 - \alpha$$

Die Mächtigkeit des Tests kann unter Vorgabe einer Alternativhypothese H_A ermittelt werden. Die Wahrscheinlichkeit, die Nullhypothese anzunehmen, obwohl die Alternativhypothese zutrifft, wird mit β bezeichnet.

$$P\left(\frac{u'_A}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le u_\beta\right) = \beta \Leftrightarrow P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,u'_A \le u_\beta\right) = \beta$$



Bei einer Gesamtheit von $n = n_1 + n_2$ untersuchten Objekten ergibt sich die Frage nach der optimalen Aufteilung auf die beiden Stichproben.

Da u'_A nur von vorgegebenen Werten der Alternativhypothese abhängt, wird die Wahrscheinlichkeit β dann einen minimalen Wert annehmen, wenn die Wurzel einen Maximalwert erreicht.

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \sqrt{\frac{n_1 (n - n_1)}{n}}$$

Notwendige Bedingung für einen Extremwert der Wurzel in Abhängigkeit vom Stichprobenumfang n_1 lautet

$$\frac{\partial}{\partial n_1}\sqrt{\frac{n_1 (n - n_1)}{n}} = 0 \qquad \frac{n - 2n_1}{2\sqrt{n \cdot n_1 (n - n_1)}} = 0 \quad \Rightarrow \quad n_1 = n_2 = \frac{n}{2}$$

Um der Stichprobenprüfung zum Test auf Homogenität eine möglichst große Macht zu geben, muß also die Anzahl der zu untersuchenden Objekte in beiden Teilgebieten gleich groß sein.

Mit optimaler Stichprobenverteilung lautet die Gleichung für den Fehler 2. Art

$$P\left(\sqrt{n}\left[\arcsin\sqrt{\frac{2m_1}{n}} - \arcsin\sqrt{\frac{2m_2}{n}}\right] \le u_\beta + \sqrt{n}\left[\arcsin\sqrt{\frac{M_1}{N_1}} - \arcsin\sqrt{\frac{M_2}{N_2}}\right]\right) = \beta$$

Zur Berechnung des erforderlichen Stichprobenumfanges müssen die Gleichungen zur Berechnung der Fehler 1. und 2. Art gleichzeitig erfüllt sein.

$$\left|\sqrt{n}\left[\arcsin\sqrt{\frac{2m_1}{n}} - \arcsin\sqrt{\frac{2m_2}{n}}\right]\right| \le u_{1-\alpha/2}$$

$$\sqrt{n}\left[\arcsin\sqrt{\frac{2m_1}{n}} - \arcsin\sqrt{\frac{2m_2}{n}}\right] \le u_\beta + \sqrt{n}\left[\arcsin\sqrt{\frac{M_1}{N_1}} - \arcsin\sqrt{\frac{M_2}{N_2}}\right]$$

Wenn das Teilgebiet mit den meisten gefundenen fehlerhaften Objekten als Stichprobe mit dem Index 1 deklariert wird, können die Betragsstriche weggelassen werden. Da der mindestens erforderliche Stichprobenumfang von Interesse ist, können die Kleiner-oder-gleich-Zeichen durch Gleichheitszeichen ersetzt und das Gleichungssystem nach *n* aufgelöst werden.

$$u_{1-\alpha/2} = u_\beta + \sqrt{n}\,\frac{\delta_A}{2} \;\Leftrightarrow\; n_1 = n_2 = \frac{n}{2} = 2\left[\frac{u_{1-\alpha/2} - u_\beta}{\delta_A}\right]^2 \quad \text{mit } \delta_A = 2\arcsin\sqrt{\frac{M_{1A}}{N_1}} - 2\arcsin\sqrt{\frac{M_{2A}}{N_2}}$$

So groß muß der Stichprobenumfang mindestens sein, damit das vorgegebene Signifikanzniveau und die Mächtigkeit des Testes erreicht werden können.

7.10 Stichprobenuntersuchungen bei besonderen Objektklassen

Wenn bekannt ist, daß bestimmte Objektklassen ein besonderes Fehlerverhalten aufweisen, oder wenn Fehler bei Objekten dieser Klassen besonders schwerwiegend sind, muß dies bei der Stichprobenauswahl berücksichtigt werden. In manchen Fällen genügt es sicherzustellen, daß ausreichend Objekte dieser Klasse in der zufällig ausgewählten Stichprobe vorhanden sind. Wenn die Objektklasse selten auftritt oder wenn das Auftreten dieser Objektklasse Häufungspunkte aufweist, kann es erforderlich sein, Stichproben nur aus der Menge der Objekte dieser Klasse zu entnehmen.

Für sehr kritische Fälle, wenn also die Anforderungen an die Qualität der Daten sehr hoch sind, oder wenn Datenfehler Menschenleben gefährden können, muß eine 100%-Prüfung durchgeführt werden.

8 Normentwürfe zu Datenqualität und Metadaten

Eine Reihe von nationalen und internationalen Gremien haben Qualitätsmodelle entwickelt, die in unterschiedliche Normen zur Beschreibung und zum Austausch von Geodaten eingeflossen sind. Die Normen unterscheiden sich in der Nomenklatur, der Stringenz und im Inhalt.

Das erste Qualitätsmodell wurde 1987 unter der Schirmherrschaft des amerikanischen Kongresses für Vermessung und Kartographie (ACSM: *American Congress of Surveying and Mapping*) veröffentlicht. Die dort definierten fünf Qualitätsaspekte wurden von der Internationalen Assoziation für Kartographie (ICA) als Bestandteil des Austauschformates SDTS (Spatial Data Transfer Standard) vom USGS (*U.S. Geological Survey) (USGS, 1998*) übernommen. Nach geringfügigen Änderungen führte das amerikanische Institut für Standardisierung und Technologie (NIST: *National Institute of Standards and Technology*) 1994 diesen Standard in die Norm für Informationsverarbeitung FIPS PUB 173 über (*Morrison, 1995*). Die fünf Qualitätsaspekte des SDTS sind: Abstammung (*lineage*), Lagegenauigkeit (*positional accuracy*), Attributgenauigkeit (*attribute accuracy*), Vollständigkeit (*completeness*) und logische Konsistenz (*logical consistency*).

Nachfolgende Qualitätsnormen wie z.B. in DIGEST, Geoinformation (CEN TC 287), Geoinformation / Geomatics (ISO TC 211) und die Norm zur Beschreibung von Fahrzeugnavigationsdaten, GDF, (CEN TC 278 und ISO TC 204: *pr ENV ISO 14825, 1996*) bauten mehr oder weniger auf diesem Qualitätsmodell auf. Weitere Aspekte wurden hinzugefügt, andere weggelassen oder modifiziert. Die verwendeten Begriffe sind nicht einheitlich definiert, daher ist es schwer die Normen miteinander zu vergleichen *(Caspary und Joos, 1998*). Wenn nur Qualitätskriterien genannt werden, ohne daß auch Qualitätsmaße eingeführt sind, ist die Vergleichbarkeit von Qualitätsangaben in Frage gestellt.

DIGEST (*DIgital Geographic Information Exchange STandard*) als Austauschformat für Geodaten wurde von einer Arbeitsgruppe für digitale Geodatenverarbeitung der militärischen Geodienste einiger NATO-Länder, DGIWG (*Digital Geographic Information Working Group*), entwickelt und herausgegeben. Die derzeit aktuelle Version 2.0 enthält zusätzlich zu den SDTS-Qualitätselementen spezielle Elemente zur Sicherheit von Geodaten. Es sind Abgabebeschränkungen und Geheimhaltungsstufen eingeführt. Sobald ein einziges Objekt oder Attribut als geheim gilt, wird der gesamte Datenbestand als geheim eingestuft. Zusätzlich gibt es in DIGEST einen Indikator, wie oft ein Objekt abgeschnitten wurde. Der Indikator bezieht sich auf das verbleibende Objekt und wird als ganze Zahl angegeben. Er kann auf durch Kachelung zerteilte Objekte hinweisen. Er gibt damit einen Hinweis auf Objekte, die von der in Abschnitt 6.3.1 angesprochenen Problematik der Konsistenzprüfung bei blattschnittbezogenen Erfassungseinheiten betroffen sind. Der Nutzen einer solchen Zahlenangabe, z.B. *clipping indicator = 4*, ist aus der Sicht des Autors nicht nachzuvollziehen.

In der Monographie der internationalen kartographischen Assoziation (ICA) zur Qualität von Geodaten, herausgegeben von *Guptill und Morrison, 1995*, werden gegenüber dem SDTS Qualitätsmodell zwei weitere Qualitätskriterien eingeführt. Auf Betreiben Frankreichs wurde das Element semantische Genauigkeit aufgenommen (*Morrison, 1995*). *Salgé, 1995*, definiert semantische Genauigkeit als Übereinstimmung der Semantik von Geoobjekten mit den Festlegungen im Datenmodell. Das Konzept der semantischen Genauigkeit stimmt mit dem Konzept der Richtigkeit von Geoobjekten, wie es in dieser Arbeit eingeführt wurde, überein. Sprachlich ist der Begriff der semantischen Genauigkeit verwirrend, da die Genauigkeit im Zusammenhang mit Geometrie und quantitativen Attributen nach der *DIN ISO 5725, 1991*, als Nähe zwischen gemessenem Wert und anerkanntem Referenzwert definiert ist. Der Begriff Richtigkeit ist im Zusammenhang mit einer falschen Klassifizierung aus diesem Grund zutreffender als der Begriff Genauigkeit. Als zweites, zusätzliches Element führte die ICA die zeitliche Information ein. Die zeitliche Information wird zum Zeitpunkt der Erfassung oder Fortführung als Stempel aufgedrückt (*Guptill, 1995*). Obwohl diese Zeitinformation eine zwar wichtige aber, wie der Name auch schon angibt, eben nur eine weitere Information zum Entstehungs- oder Änderungsdatum beisteuert, wird dieses Attribut als siebtes Qualitätselement bezeichnet. Die Aktualität, wie dieses Element in deutschsprachiger Literatur üblicherweise genannt wird, wird in der vorliegenden Arbeit unter den Metadaten geführt.

In der Vornorm des europäischen Komitees für Normung (CEN) zur Datenqualität *pr ENV 12656, 1998*, beziehen sich die Qualitätselemente auf einen Datensatz und damit auf alle Objekte, Attribute

und Beziehungen in diesem Datensatz. Für jeden Datensatz muß nach CEN mindestens eine der folgenden Angaben gemacht werden:

- Abstammung des Datensatzes
- Referenzanwendungen
- Satz von Qualitätsparametern, der durch beliebige Parameter ergänzt werden kann
- Angaben zur Homogenität

Es ist vorgesehen, aber nicht erforderlich, daß zu jeder Qualitätsangabe zusätzliche Angaben über die Zuverlässigkeit dieser Informationen gemacht werden. Dies soll nach pr ENV 12656 mit Werten für die Standardabweichung, durch ein Vertrauensintervall oder durch eine textliche Beschreibung erfolgen. Optional kann beschrieben werden, wie die Qualitätsinformationen ermittelt wurden und welchen Effekt die Abstraktion beim Übergang von realen Welt zu ihrem abstrakten Abbild ausmacht. Dieser Effekt wird in der vorliegenden Arbeit unter dem Begriff Modellqualität behandelt.

Im Rahmen der Normierung von Navigationsdaten für den Straßenverkehr wurde im CEN TC 278 ein Qualitätsmodell entwickelt und von ISO TC 204 als internationale Norm übernommen *(prENV ISO 14825, 1996*). Hier wird die Übereinstimmung der digitalen Daten mit der Situation in der realen Welt abhängig vom Zweck innerhalb definierter Schranken gefordert. Durch die Schranken lassen sich anhand dieses Qualitätsmodells unmittelbar Aussagen darüber machen, ob die Qualität der Daten ausreichend ist. Diese Norm enthält nicht nur Definitionen von Qualitätskriterien, sondern beschreibt auch Methoden zur Ermittlung von Qualitätsmaßen. Dabei wird zwischen einer Formatkontrolle und einer Kontrolle der Semantik unterschieden. Bei der Formatkontrolle sollen die folgenden Fehlergruppen berücksichtigt werden:

- Syntaxfehler, wie z.B. Verwendung von falschen Zeichensätzen, oder falscher Feldtypen
- Fehler von Werten, z.B. falsche Feldbezeichner, Verwendung von falschen Attributkodierungen oder Werte außerhalb des zulässigen Wertebereiches
- Integritätsfehler der Datenbank, z.B. fehlerhafte Zeiger oder Schlüssel
- Topologische Fehler, z.B. fehlende Verknüpfungen
- Integritätsfehler der Werte, z.B. fehlerhafte Relationen zwischen Objekten und Attributen

Die Kontrolle der Semantik ist eingeteilt in indirekte und direkte Qualitätsprüfungen. Bei der indirekten Qualitätsprüfung wird untersucht, ob bei der Datenerfassung ein ISO 9000 konformes Qualitätsmanagement angewendet wurde. Bei der direkten Qualitätsprüfung werden die Daten oder repräsentative Stichproben mit der realen Situation oder einer anderen gültigen Referenz verglichen. Bei der Auswertung der Stichprobenuntersuchungen verweist diese Norm auf ISO 2859.

Von den 20 Arbeitspaketen des ISO TC 211, Geographic information/Geomatics, befassen sich zwei mit der Qualität von Geodaten. In *ISO 19113, 1999*, wird ein Modell zur Beschreibung der Datenqualität eingeführt. Neben den allgemein anerkannten Elementen Vollständigkeit, logische Konsistenz, Positionsgenauigkeit, zeitliche und thematische Genauigkeit, steht es dem Anwender dieser Norm frei, noch weitere Elemente zu definieren. Die genannten Elemente werden in Unterelemente gegliedert. Die logische Konsistenz hat z.B. als Unterelemente Konsistenz des Wertebereiches, des Formates und der Topologie. Zur Beschreibung der nicht quantifizierbaren Qualität eines Datensatzes werden Überblickelemente eingeführt. Diese sind Zweck, Verwendung, Herkunft und vom Benutzer definierte Elemente. Es wird eine Schnittstelle zu den Metadaten beschrieben, die in einem separaten Teil dieser Norm behandelt werden (*ISO 19115, 1999*).

ISO 19114, 1999, behandelt Methoden, wie die Qualität von Geodaten ermittelt werden kann. Auch in dieser Norm wird zwischen direkten und indirekten Verfahren unterschieden. Die direkten Verfahren können entweder auf den gesamten Datensatz oder auf Stichproben angewandt werden. Bei den indirekten Methoden wird die Qualitätsinformation aus anderen Quellen als den Daten abgeleitet. Als solche Quellen kommen in Betracht: Metadaten, die Information über den Zweck der Datenerfassung, die Beschreibung der Herkunft der Daten, andere Anwendungen, die mit den Daten durchgeführt wurden, oder Prüfprotokolle der Datenerfassung. Jede Information über den Erfassungsprozeß gibt Hinweise zur Beurteilung der Eignung der Daten, wenn keine vollständige Beschreibung des Qualitätsmodells der Geodaten vorliegt.

Die in den Normen eingeführten Elemente zur Beschreibung der Qualität von Geodaten sind als Übersicht in der folgenden Tabelle dargestellt.

		FIPS 173 SDTS 1988	***DIGEST 2nd edition 1997***	***ICA 1995***	***CEN TC 278 1996***	***CEN TC 287 1996***	***ISO TC 211 1999***
Metadaten bezogen auf die Quellen	Herkunft	✓	✓	✓		✓	✓
	Zweck						✓
	Verwendung					✓	✓
	benutzerdefiniert						✓
Metadaten bezogen auf das Modell	Auflösung				✓		
	Präzision				✓		
	Indikator über Abschneidungen		✓				
	Veränderung durch die Abstraktion					✓	
Metadaten bezogen auf die Verfügbarkeit	Abgabebeschränkung		✓				✓
	Geheimhaltung		✓				✓
Übereinstimmung mit der Datenspezifikation	Format		✓		✓		
Qualitätselemente bezogen auf die Genauigkeit	Position	✓	✓	✓	✓	✓	✓
	Attribut	✓	✓	✓	✓	✓	✓
	Semantik			✓		✓	✓
	Zeit		✓	✓	✓	✓	✓
Qualitätselemente bezogen auf das konzeptionelle Modell	Richtigkeit				✓		✓
	Vollständigkeit	✓	✓	✓	✓	✓	✓
Qualitätselemente bezogen auf das logische Modell	logische Konsistenz	✓	✓	✓	✓	✓	✓
Qualitätselemente bezogen auf die Qualitätsangaben	Zuverlässigkeit					✓	

9 Zusammenfassung und Ausblick

Der Einsatz eines GIS für die Planung, für Verwaltung von rechtsverbindlichen Objekten, z.B. Flurstücke oder Schutzgebiete, für Fahrzeugnavigation und Flottenmanagement, für Geomarketing und für Leitungsdokumentation bringt nicht nur Vorteile in Form von Kosteneinsparung, schnellere Verfügbarkeit, höhere Aktualität und vor allem Analysemöglichkeiten, sondern birgt auch Gefahren in sich. Auf der Basis der Geodaten werden Entscheidungen gefällt, die sowohl Einfluß auf Investitionen haben als auch auf unsere Umwelt, und somit auch auf unsere Lebensqualität. Die Qualität der Daten ist deshalb besonders wichtig. Um qualitativ hochwertige Daten zu erhalten, muß ein der Anwendung angemessener Aufwand getrieben werden. Dieser Aufwand darf nicht nur mit den Kosten der Datenerfassung in Relation gestellt werden, sondern auch mit den Auswirkungen, die fehlerhafte Objekte in einem Geoinformationssystem bewirken können.

Um die Qualität von Geodaten ermitteln oder beurteilen zu können, muß zuerst klar definiert sein, welche Objekte der realen Welt wie in einem Geoinformationssystem repräsentiert werden sollen. Dazu muß ein Datenmodell aufgestellt werden, das aus den Teilen konzeptionelles, logisches und physikalisches Datenmodell besteht. Das Datenmodell legt die Objektauswahl, deren Eigenschaften, die Struktur und Regeln fest.

Die Beschreibung der Daten erfolgt durch Metadaten. Sie beinhalten nicht nur das Datenmodell, sondern auch alle Informationen über den Entstehungsprozeß der Daten und ein Qualitätsmodell, das einem potentiellen Anwender von Geodaten ein Urteil ermöglicht, ob die Geodaten für eine beabsichtigte Anwendung geeignet sind. Zur Beschreibung des Qualitätsmodells müssen Qualitätskriterien und Qualitätsmaße eingeführt werden. Die Kriterien sind erforderlich, um Datenfehler taxieren zu können, da nicht alle Fehlerarten gleiche Auswirkungen für die Anwendungen haben. Es wurde gezeigt, daß die vier Kriterien „Vollständigkeit“, „Richtigkeit“, „Genauigkeit“ und „Konsistenz“ zur Einordnung von Datenfehlern ausreichend, aber nicht immer eindeutig sind. Wenn keine Mehrdeutigkeiten zugelassen sind, müssen Zusatzregeln als Entscheidungshilfen bei der Behandlung aller anwendungsspezifischer Spezialfälle angegeben werden.

Auf Basis dieser vier Kriterien werden die Qualitätsmaße eingeführt, die entweder objektbezogen oder bezogen auf Gebiete definiert werden können. Die Qualitätsmaße können durch Festlegung von Grenzwerten zur Formulierung von Qualitätszielen verwendet werden. Qualitätsziele sind insbesondere im Rahmen eines Qualitätsmanagements erforderlich. Das Qualitätsmanagement dient zur Überwachung und zur Dokumentation, daß die Ziele eingehalten und mögliche Fehlerursachen früh erkannt werden, damit die Erfassung und Fortführung von Geodaten auf einem hohen Qualitätsniveau erfolgen und durch ständige Rückkopplungen weiter verbessert werden. Das Qualitätsmanagement kann durch 20 QS-Elemente in Anlehnung an ISO 9000 ff beschrieben werden. Kernstück des Qualitätsmanagement sind regelmäßige Prüfungen. Zur Prüfung von Geodaten können zwei Arten der Prüfung unterschieden werden. Prüfungen können automatisch ablaufen oder interaktiv durch einen menschlichen Prüfer und sie können auf den gesamten Datenbestand angewandt werden oder nur auf Stichproben, deren Prüfergebnisse auf den Gesamtdatenbestand hochgerechnet werden.

Automatische Prüfungen können nur zur Kontrolle der Konsistenz, d.h. der Einhaltung von Regeln des Datenmodells, herangezogen werden. Die Regeln beziehen sich auf die konzeptionelle, logische oder physikalische Ebene der Datenmodellierung. Zur Überprüfung der konzeptionellen Konsistenz wurde ein Regelwerk aufgestellt, mit dem Konsistenzbedingungen für topographische Daten formuliert werden können. Eine Prüfsoftware, die diejenigen Objekte und Konstellationen aufdeckt, die diesen Regeln widersprechen, wurde entwickelt. Die Prüfung wird von einem implementierungsunabhängigen Regelkatalog gesteuert. Zur Anpassung an Änderungen und Erweiterungen des Datenschemas braucht nur der Regelkatalog fortgeführt zu werden. Die Bedingungen zur Suche nach Inkonsistenzen werden mit Hilfe der Prädikatenlogik und des 9-Intersection-Modells zur Beschreibung von topologischen Beziehungen formuliert. Automatische Prüfungen müssen auf dem gesamten Datenbestand durchgeführt und entdeckte Fehler korrigiert werden, bevor die Daten einer Anwendung zugeführt werden dürfen.

Wenn eine vollständige Prüfung des gesamten Datenbestandes mit zu großem Aufwand verbunden ist, weil diese z.B. zu lange dauert und die Daten schnell benötigt werden oder die Kosten für die Prüfung das Budget oder den Nutzen übersteigen, so kann unter bestimmten Voraussetzungen eine Stichprobenkontrolle durchgeführt werden. Die Wahrscheinlichkeit, eine bestimmte Anzahl von fehlerhaften Objekten in einer Stichprobe zu finden, errechnet sich nach der hypergeometrischen Verteilung. Zur Aufstellung eines Stichprobenplanes, bestehend aus dem erforderlichen Stichprobenumfang und einer zugehörigen Annahmezahl, wird eine annehmbare Qualitätsgrenzlage und rückzuweisende Qualitätsgrenzlage benötigt. Diese kann entweder durch Absprache zwischen Produzent und Anwender der Daten festgelegt oder nach wirtschaftlichen Gesichtspunkten ermittelt werden. Verschiedene Strategien zur Reduzierung des durchschnittlichen Stichprobenumfanges werden diskutiert.

Die Dokumentation der durchgeführten Kontrollen und deren Ergebnisse ist für die Einschätzung der Zuverlässigkeit der Geodaten wichtig. Verschiedene Ansätze zur Verwaltung dieser Informationen zusammen mit weiteren Metadaten wurden diskutiert. Abhängig davon, ob das einzelne Objekt oder eine Gruppe von Objekten die Bezugsgröße für Qualitätsinformationen darstellt, sind unterschiedliche Konzepte zur Verwaltung der qualitätsbezogenen Metadaten möglich.

In dieser Arbeit wurden die Aspekte der Datenqualität behandelt. An verschiedenen Stellen wurde deutlich, wie wichtig das Modell für die Verwendbarkeit von Geodaten ist. Die Aspekte der Modellqualität wurden nur gestreift. Der Autor ist sich bewußt, daß in diesem Bereich noch weiterer Forschungsbedarf besteht. Die Crux besteht darin, daß die Bewertung von Modellen sehr eng an die jeweilige Anwendung gebunden ist, und somit ein allgemein gültiger Formalismus zur Bewertung der Modellqualität kaum angegeben werden kann.

Wenn die Qualität von Geoinformationssystemen oder von aus GIS abgeleiteten Informationen beurteilt werden soll, müssen alle Komponenten eines GIS betrachtet werden (*Joos, 1994*). Dabei kommen Aspekte wie Ausfallsicherheit von Hardware, Antwortzeitverhalten bei Datenzugriffen, Effizienz und Richtigkeit von Algorithmen zur Analyse von Geodaten sowie Auswirkungen oder Sensibilitätsuntersuchungen von Datenfehlern zum Tragen. Dazu ist ein Qualitätsmodell für Arbeitsergebnisse und Entscheidungen, die mit einem GIS erzeugt wurden wünschenswert. Da ein solches Modell nicht verfügbar ist, besteht auch hier Forschungsbedarf. Mit den in dieser Arbeit vorgeschlagenen Kriterien erscheint eine Berechnung oder Abschätzung der Auswirkung von der Datenqualität auf die Ergebnisqualität möglich. Unter dieses Thema fällt auch die Beurteilung der Qualität von Diensten mit Geodaten, wie sie durch die Interoperabilität von GIS möglich werden.

Die Aktualität von Geodaten wurde nicht als Qualitätskriterium behandelt, sondern als Element der Metadaten betrachtet. Sie ist ein entscheidendes Indiz für die Verwendbarkeit der Geodaten in den meisten Anwendungen. Durch die Dokumentation des Entstehungs- und Änderungsdatums, welches die Daten tragen, ist noch kein aktueller Datenbestand gewährleistet. Im Bereich der Fortführung durch Einrichtung von Meldediensten oder durch automatisierte Erkennung von Objekten aus Fernerkundungsdaten, sowie durch Fortführung von Datenbeständen einer Maßstabsebene und Propagierung der Änderungen in die anderen Maßstabsebenen durch Generalisierung sind nur Teilaspekte gelöst.

Um Informationen über die Qualität von Geodaten einem Anwender zugänglich zu machen, insbesondere, wenn die Informationen nicht nur allgemeiner, beschreibender Natur sind, sondern in enger Verknüpfung mit den Daten stehen oder selbst einen Raumbezug haben, wie in dieser Arbeit vorgeschlagen, werden standardisierte Zugriffsmechanismen oder Austauschformate erforderlich. Verschiedene Organisationen und Gremien arbeiten teils konkurrierend teils kooperierend zum Zeitpunkt der Fertigstellung dieser Arbeit an Normen für Geoinformation. Zum Wohle der Nutzer von Geoinformationssystemen ist es wichtig, daß erstens diese Normen vollständig und anwendbar sind, zweitens die Systemhersteller Mechanismen bereitstellen, damit diese Normen angewandt werden können und drittens die Datenproduzenten die ermittelten Qualitätsmaße den Endnutzern zugänglich machen.

10 Summary

In order to assess and to judge the quality of digital geodata, it is necessary to define the objects of the real world which are to be represented in a geographical information system and how this can be done. This can be accomplished by establishing a data model consisting of a conceptual, a logical, and a physical level. The data model defines objects to be selected, their properties, structure, and rules.

Data is described by metadata. Metadata not only includes the data model, but also all the information about the production process and a quality model, that allows a potential user to judge whether the data is suitable for the intended use. Quality criteria and measures have to be identified in order to establish the quality model. The criteria are needed to rank the errors within the data, since not all types of errors have the same influence on applications. It has been shown that "completeness", "correctness", "accuracy", and "consistency" are sufficient to classify the errors but that they are not always unique. If ambiguities are not allowed, additional rules have to be specified. These then serve as decision aids when dealing with all the user-specific cases.

Quality measures are introduced based on these four criteria. They can be defined so that they are object-specific or area-specific. By stipulating certain thresholds these measures can serve to formulate quality goals. Quality goals are essential for quality management systems. QM is used to monitor the production process, to ensure that targets are adhered to and that possible causes of error are identified at an early stage. QM helps to keep the quality of the data produced and its maintenance at a consistently high level and even assists in the process of improving quality by providing feedback to the data capture process. QM can be described by 20 quality assurance elements following the ISO 9000 ff standards. Inspections, carried out at regular intervals, are at the heart of the QMS. The inspection of geodata can be performed in two different ways. It can be carried out interactively by a human inspector or automatically. It can be applied to the entire data set or to a representative sample, the results of which can then be applied to the entire data set.

Automatic checks can only be used to verify consistency, i.e. to check whether the data complies with the rules of the data model. These rules refer to the conceptual, logical and physical level of modelling. For verification of conceptual consistency a set of rules has been established, where consistency conditions of topographic data can be formulated. Software for identifying objects that contradict these conditions was implemented. The testing is based on a software-independent catalogue of rules. Therefore the rules can easily be adapted to changes or expansions of the data schema. The search conditions are formulated using predicate logic and the 9-intersection formalism for topologically related objects. Automatic checking is supposed to be performed on the entire data set, before the data is used in any application.

If the effort of checking the entire data set is too much, e.g. because controlling the entire data would be too time consuming or just too costly, a sample quality control can be performed under certain conditions. The probability of finding a certain amount of erroneous objects within a sample is given by the hypergeometric distribution. A sample plan consists of the required sample size and an acceptance number. In order to set up a sample plan, a value for the acceptable quality level and for the least quality is required. They can either be stated by agreement between data producer and data user, or by economic considerations. Various strategies for reducing sample size are discussed.

The documentation of the performed controls and their results are important for a user to judge the reliability of geodata. Different approaches on how to store this information together with other metadata are discussed. There are different proposals for managing metadata depending on whether the information refers to single objects or entire datasets.

Implementing a GIS for planning and administering legally binding objects such as parcels or national parks, for traffic guidance and fleet management, for geo marketing and for facility management not only brings advantages in the form of cost reduction, better availability, more up-to-date data, and analysis, but it can also carry risks. Decisions are made based on the data. These decisions can affect both investments and our environment and consequently the quality of our life. The quality of the data is therefore very important. More effort is needed to obtain higher quality data. This effort must not only be compared with the costs of data capture, but also with the consequences arising from erroneous data in a GIS.

11 Glossar

Die zentralen Begriffe dieser Arbeit werden im folgenden zusammengestellt und definiert. Für einige Begriffe wurde ein Bezug zu CEN und ISO hergestellt, insbesondere wenn für dieselbe Definition oder dasselbe Konzept andere Begriffe eingeführt worden sind. Als Referenz für die Terminologie wurden die Dokumente CEN/TC 287 N542: 287003 – Definitions, 1997, CEN TC 287 N580: 287008 – Quality, 1998, und ISO/TC 211 N478: 15046-4 – Terminology, 1998, herangezogen. Teilweise gibt es auch Widersprüche bei der Verwendung und den Definitionen der Begriffe innerhalb einer Normenfamilie.

Objekt der realen Welt	Phänomen der realen Welt mit bestimmten Eigenschaften, das als abgeschlossene Einheit betrachtet werden kann. Insbesondere sind für GIS die Objekte der realen Welt interessant, die eine Beziehung zur Erde haben (ISO: real world phenomenon, CEN: object).
Abstraktes Objekt	Element des abstrakten Abbildes der realen Welt, das durch die Modellierung aus einem Objekt der realen Welt hervorgegangen ist.
Digitales Objekt	Repräsentation eines abstrakten Objektes in einem GIS (ISO: feature, CEN: entity).
Objekt	Synonym für digitales Objekt.
Objektklasse	Klasse von Objekten, die durch gemeinsame Eigenschaften bestimmt ist (ISO: entity, CEN: entity class).
Modell	Abstraktionsvorschrift, wie Objekte der realen Welt abzubilden sind.
Konzeptionelles Modell	Modell zur inhaltlichen und semantischen Festlegung von Geodaten.
Logisches Modell	Modell, das logischen und strukturellen Konzepte von Geodaten festlegt.
Physikalisches Modell	Modell, das die Speicherung von Geodaten festgelegt.
Schema	Formale Beschreibung der Inhalte eines Modells.

12 Abkürzungen

ACSM	American Congress of Surveying and Mapping
AdV	Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland
AGILE	Association of Geographic Information Laboratories in Europe
CEN	Comité Européen de Normalisation
CERCO	Comité Européen des Responsables de la Cartographie Officielle
DDGI	Deutscher Dachverband für Geoinformation
DGIWG	Digital Geographic Information Working Group
DIGEST	DIgital Geographic Information Exchange STandard
DIN	Deutsches Institut für Normung e.V.
EUROGI	European Ambrella Organisation for Geoinformation
ETH	Eidgenössische Technische Hochschule
FIPS PUB	Federal Information Processing Standards Publication
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
ITC	International Institute for Aerospace Survey and Earth Science
NIST	National *(USA)* Institute of Standards and Technology
OGC	Open GIS Consortium
UML	Unified Modelling Language
URL	Uniform Ressource Locator: eindeutige Internet Adresse
ZfV	Zeitschrift für Vermessungswesen

13 Literaturverzeichnis

AdV, 1986: Einheitliche Datenbankschnittstelle (EDBS). Druck und Vertrieb: Landesvermessung und Geobasisinformation Niedersachsen, Landesbetrieb, Podbielskistraße 331, D-30659 Hannover.

AdV, 1995: ATKIS-Objektartenkatalog. Druck und Vertrieb: Landesvermessungsamt Nordrhein-Westfalen Muffendorfer Straße 19-21, D-53177 Bonn.

Aigner, M., 1984: Graphentheorie. Teubner, Stuttgart.

Aumann, G., K. Spitzmüller, 1993: Computerorientierte Geometrie. BI-Wissenschaftsverlag, Mannheim, Leipzig, Wien, Zürich.

Aronoff, S., 1989: Geographic information systems: a management perspective. WDL Publications, Ottawa, Canada.

Bartelme, N., 1995: Geoinformatik. Springer-Verlag Berlin Heidelberg New York.

Bauer, H., 1978: Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie. Walther de Gruyter, Berlin, New York.

Bethge, F., 1997: Genauigkeit geometrischer Größen aus Vektordaten. DGK, Reihe C, Nr. 473, Beck, München.

Bill, R., 1996: Grundlagen der Geo-Informationssysteme, Band 2: Analysen, Anwendungen und neue Entwicklungen. Wichmann, Karlsruhe.

Bill, R., D. Fritsch, 1994: Grundlagen der Geo-Informationssysteme, Band 1: Hardware, Software und Daten. Wichmann, Karlsruhe.

Birkin, M., G. Clarke, M. Clarke und A. Wilson, 1996: Intelligent GIS. Pearson Professional, Cambridge, UK.

Booch, G., 1996: Objektorientierte Analyse und Design: Mit praktischen Anwendungsbeispielen. Addison-Wesley.

Brassel, K., F. Bucher, E.-M. Stephan und A. Vckovski, 1995: Completeness. In: *Guptill and Morrison, 1995*, pp. 81-108.

Brauer, J.-P., E. U. Kühme, 1996: DIN EN ISO 9000 - 9004 umsetzen. Hanser, München.

Bronstein,I. N., K. A. Semendjajew, G. Musiol und H. Mühlig, 1995: Taschenbuch der Mathematik. Harri Deutsch, Thun, Frankfurt a. M.

Buehler, K., L. McKee (Eds.), 1998: The OpenGIS Guide, Introduction to Interoperable Geoprocessing. Open GIS Consortium, Inc., Wayland, MA, USA.

Bureau of the Census, 1997: TIGER/Line ® Files Technical Documentation. Washington, DC: (URL: http://www.census.gov/geo/tiger/TIGER97.pdf).

Burkhardt, R., 1997: UML – Unified Modelling Language. Addison-Wesley-Longman, Bonn.

Burrough, P. A., R. A. MacDonnell, 1998: Principles of Geographical Information Systems. Oxford University Press, Oxford.

Burrough, P. A., A.V. Frank, 1996: Geographic objects with undeterminate boundaries. Taylor & Francis, London [u.a.].

Buttenfield, B., K. Beard, 1994: Graphical and Geographical Components of Data Quality. In: *Hearnshaw, H. M. and D. J. Unwin, 1994*, pp. 150-157.

Buziek, G. (Hrsg.), 1995: GIS in Forschung und Praxis. Wittwer, Stuttgart.

Caspary, W., 1992: Qualitätsmerkmale von Geodaten. ZfV Heft 7, 1992, S. 360-367.

Caspary, W., 1993: Qualitätsaspekte bei Geo-Informationssystemen. ZfV, Heft 8/9, 1993, S. 444-450.

Caspary, W., 1994: Dealing with fuzzy sets in Geo-Information Systems. Proceedings of the FIG Commission 5 Congress, March 5 to 12 1994, Melbourne, Australia, pp. 505.3/1-8.

Caspary, W., K. Wichmann, 1994: Lineare Modelle: algebraische Grundlagen und statistische Anwendungen. Oldenbourg, München, Wien.

Caspary, W., 1995: Towards Fuzzy Geometry. GISDATA Specialist Meeting on Data Quality, July 1995, Lisbon.

Caspary, W., G. Joos, 1996: Ein Qualitätsmanagement für Geobasisdaten. In: Das Geoinformationssystem ATKIS und seine Nutzung in Wirtschaft und Verwaltung, Hrsg.: LVA Rheinland-Pfalz, Koblenz.

Caspary, W., G. Joos, 1998: Quality Criteria and Control for GIS Databases. In: *H. Kahmen, E. Brückl and Th. Wunderlich (Eds.), 1998*, pp. 436-441.

Chrisman, N., 1982: A theory of cartographic error and its measurement in digital data bases. Proceedings of Auto-Carto 5, Crystal City, pp. 159-168.

Chrisman, N., 1991: The error component in spatial data. In: *Maguire, Goodchild and Rhind, 1991*.

Claussen, H., 1995a: Digitale Karten für Fahrzeugnavigationssysteme. Kongreßdokumentation zum 79. Deutschen Geodätentag in Dortmund.

Claussen, H., 1995b: Qualitätsanforderungen an die digitale Karte aus Anwendersicht. Grazer Geoinformatiktage ´95, Mitteilungen der geodätischen Institute der TU Graz, Folge 80.

Claussen, H., 1996: Qualitätsbeurteilung Digitaler Karten für Fahrzeugnavigationssysteme. Geo-Informations-Systeme, Jg. 9, Heft 5, S. 23-29.

Claussen, H., G. Vickus, 1998: Present State of Road Databasis for Driver Information Systems and Telematics. In: *H. Kahmen, E. Brückl and Th. Wunderlich (Eds.), 1998*, pp. 407-411.

Clementini E., P. Di Felice, 1994: A Comparison of Methods for Representing Topological Relationships. Information Sciences 80, pp. 1-34.

Date, C. J., 1981: An introduction to database systems. Third edition, Addison-Wesley Publishing Company.

Date, C. J., 1995: An introduction to database systems. Sixth edition, Addison-Wesley Publishing Company.

Detreköi, Á, 1994: Data quality management in GIS systems. In: Computers, Environment and Urban Systems, Vol. 18, No. 2, pp. 81-85.

Deutsche Gesellschaft für Qualität (Hrsg.), 1988: Stichprobenprüfung für kontinuierliche Fertigung anhand qualitativer Merkmale. Beuth, Berlin.

Dogac, A., M. T. Özsu, A. Biliris, T. Sellis (Eds.), 1994: Advances in Object Oriented Database Systems. Springer, Berlin [u.a.].

Ehlschlaeger, C. R., M. F. Goodchild, 1994: Dealing with Uncertainty in Categorical Coverage Maps: Defining, Visualizing, and Managing Errors. From: Proceedings, Workshop on Geographic Information Systems at the Conference on Information and Knowledge Management, Gaithersburg MD, December 1 1994, pp. 86-91. (URL: http://geo.swf.uc.edu/~chuck/acm/paper.html).

Egenhofer, M., D. M. Mark und J. Herring, 1994: The 9-Intersection: Formalism and Ist Use for Natural-Language Spatial Predicates. NCGIA Technical Report 94-1, Santa Barbara, CA, Buffalo, NY, Orono, ME, USA.

Ferrari, G., 1996: Boundaries, Concepts, Language. In: Burrough, P. A. [Hrsg.], 1996: Geographic objects with indeterminate boundaries. Taylor & Francis, London.

Fischer, M. M., P. Nijkamp, 1993: Geographic Information Systems, Spatial Modelling, and Policy Evaluation. Springer-Verlag, Berlin, Heidelberg.

Fisher, P., 1995: Innovations in GIS 2. Taylor & Francis, London [u.a.].

Fisz, M., 1976: Wahrscheinlichkeitsrechnung und mathematische Statistik. VEB Deutscher Verlag der Wissenschaften, Berlin.

Fotheringham, A. S., P. Rogerson, 1994: Spatial Analysis and GIS. Taylor & Francis, London [u.a.].

Frank, A., 1983: Datenstrukturen für Landinformationssysteme - semantische, topologische und räumliche Beziehungen in Daten der Geo-Wissenschaften. Institut für Geodäsie und Photogrammetrie an der Eidgenössischen Technischen Hochschule Zürich, Mitteilungen Nr. 34.

Glemser, M., 1996: Inegration geometrischer Datenqualität in GIS-Funktionen. In: Interner Bericht Nr. 5, Institut für Geodäsie und Geoinformatik, Universität Rostock, S. 69-84.

Göpfert, W., 1991: Raumbezogene Informationssysteme. Wichmann, Karlsruhe.

Goodchild, M. F., S. Gopal, 1989 (Hrsg.): The Accuracy of Spatial Databases. Taylor & Francis, London [u.a.].

Goodchild, M., B. Buttenfield und J. Wood, 1994: Introduction to Visualizing Data Validity. In: *Hearnshaw und Unwin, 1994*, pp. 141-149.

Goodchild, M., L. Chih-Chang und Y. Leung, 1994: Visualizing Fuzzy Maps. In: *Hearnshaw und Unwin, 1994*, pp. 158-167.

Guptill, S. C., J. L. Morrison, 1995 (Hrsg.): Elements of Spatial Data Quality. Elsevier Science, Kidlington, Tarrytown, Tokyo.

Guptill, S. C., 1995: Temporal information. In: *Guptill und Morrison, 1995*, pp. 153-165.

Hake, G., D. Grünreich, 1994: Kartographie. Walter De Gruyter, Berlin.

Hearnshaw, H. M., D. J. Unwin, 1994: Visualization in Geographical Information Systems. John Wiley & Sons, Chichester, UK.

Herberger, M., M. Junker, 1999: Wem gehören digitale Geodaten? Vortrag auf der Fachtagung "Digitale geographische Daten im Saar-Lor-Lux-Raum", 16. April 1999, in Sandweiler/Luxemburg. http://www.jura.uni-sb.de/urheberrecht/vortraege/1999-clear-luxemburg.html

Joos, G., 1994: Quality aspects of geo-informations. In: Proceedings of the fifth European conference on geographical information systems, EGIS in Paris, March 29 – April 1, 1994, Vol. II, pp. 1147-1153.

Joos, G., 1996a: Konsistenz- und Plausibilitätsprüfungen von Geodaten. In: Nachrichten aus dem Karten- und Vermessungswesen, Reihe I, Heft Nr. 115, Verlag des Instituts für Angewandte Geodäsie, Frankfurt a.M, S. 93-100.

Joos, G., 1996b: Konsistenzprüfung für das ATKIS-Datenmodell. In: Interner Bericht Nr. 5, Institut für Geodäsie und Geoinformatik, Universität Rostock, S. 103-109.

Joos, G., 1998: Statistical quality control of geodata. Proceedings of the AGILE 1998 meeting, ITC publications.

Joos, G., 1999: Assessing the quality of geodata by testing consistency with respect to the conceptual data schema. In: M. Craglia and H. Onsrud (Eds.), Geographic Information Research: Trans-Atlantic Perspectives, Taylor & Francis, London [u.a.], pp. 509-519.

Joos, G., U. Baltzer und K.-H. Kullmann, 1997: Qualitätsmanagement beim Aufbau einer topographischen Grunddatenbank am Beispiel von ATKIS in Hessen. ZfV, Heft 4, 122. Jg., S. 149-159.

Kahmen, H., E. Brückl und Th. Wunderlich (Eds.), 1998: Proceedings of the IAG SC4 Symposium in Eisenstadt/Austria, April 20-22, 1998.

Kainz, W., 1995: Logical consistency. In: *Guptill and Morrison, 1995*, pp. 109-137.

Klemmer, W., R. Spranz, 1997: GIS-Planung und Projektmanagement. Wilfried Klemmer Roland Spranz GbR, Bonn. ISBN 3-00-001532-9.

Kraus, K., H. Kager, 1994: Accuracy of Derived Data in a Geographic Information System. In: Computers, Environment and Urban Systems, Vol. 18, No. 2, pp. 87-94.

Kreyszig, E., 1973: Statistische Methoden und ihre Anwendungen. Vandenhoeck & Ruprecht, Göttingen.

Langran, G., 1992: Time in Geographic Information Systems. Taylor & Francis, London [u.a.].

Larman, C., 1998: Applying UML and patterns: an introduction to object-oriented analysis and design. Prentice-Hall, Upper Saddle River, USA.

Lawford, G. J., 1995: Geodata quality validation. Cartography, Vol. 24, No. 2.

Leung, Y., J. P. Yan, 1998: A locational error model for spatial features. International Journal of Geographical Information Science, Vol. 12, No. 6, pp.607-620.

Longley, P., M. Batty (Eds.), 1996: Spatial Analysis: Modelling in a GIS Environment. Pearson Professional, Cambridge.

Loomis, M. E. S., 1995: Object databases. Addison-Wesley, Reading, USA.

Mace, A. E., 1964: Samle-Size Determination. Reinhold, New York.

Maguire, D. J., J. Dangermond, 1991: The functionality of GIS. In: *Maguire, Goodchild und Rhind, 1991*, pp. 319-335.

Maguire, D. J., M. F. Goodchild und D. W. Rhind (Eds.), 1991: Geographical Information Systems: Principles and Applications. Longman Scientific and Technical, Harlow, UK.

Maassen, R., 1996: Navigierbare Straßendatenbanken. Geo-Informations-Systeme, Jg. 9, Heft 5, S. 20-23.

Masser, I., H. Campbell and M. Craglia, 1996: GIS diffusion. Taylor & Francis, London [u.a.].

Mayer, K. H., 1989: Algebraische Topologie. Birkhäuser, Basel, Boston, Berlin.

Meier, A., 1982: Semantisches Datenmodell für flächenbezogene Daten. Dissertation ETH Zürich 7043.

Meier, A., 1986: Methoden der grafischen und geometrischen Datenverarbeitung. Teubner Verlag, Stuttgart.

Menges, G., H. J. Skala, 1973: Grundriß der Statistik 2: Daten. Westdeutscher Verlag, Opladen.

Möhlenbrink, W., 1998: Traffic-Guidance and Information Systems – New Technologies for the Geoinformation Market. In: *Kahmen, Brückl und Wunderlich, 1998*, pp. 395-400.

Molenaar, M., 1995: The Relationship Between the Extensional and Geometric Uncertainty of Spatial Objects. GISDATA Specialist Meeting on Data Quality, July 1995, Lisbon.

Montgomery, D. C., 1991: Introduction to statistical quality control. John Wiley & Sons, New York [u.a.].

Morrison, J., 1995: Spatial data quality. In: *Guptill and Morrison, 1995*, pp. 1-12.

Müller, K., 1998: Zweifache Stichprobenpläne für qualitative und quantitative Merkmale mit minimaler maximaler ASN. Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschafts- und Sozialwissenschaften des Fachbereiches Wirtschafts- und Organisationswissenschaften der Universität der Bundeswehr Hamburg.

Muller, J.-P., J.-Ph. Lagrange und R. Weibel, 1996: GIS and Generalisation. Taylor & Francis, London [u.a.].

Oosterom, P. van, 1993: Reactive Data Structures for Geographic Information Systems. Oxford University Press.

Peach, P., S. B. Littauer, 1946: A Note on Sampling Inspection. The Annals of Mathematical Statistics 17: pp. 81-84.

Peterson, K., O. Maus, 1996: Geo-Daten und Analyseinstrumente in Geo-Marketing. Geo-Informations-Systeme, Jg. 9, Heft 5, S. 2-9.

Peuquet, D. J., 1984: A conceptual framework and comparison of spatial data models. Cartographica 21: pp. 66-113.

Plümer, L., 1996a: Zur Prüfung der Konsistenz von Geometrie und Topologie in Landkarten. In: Nachrichten aus dem Karten- und Vermessungswesen, Reihe I, Heft Nr. 115, S. 131-140, Verlag des Instituts für Angewandte Geodäsie, Frankfurt a.M.

Plümer, L., 1996b: Achieving Integrity of Geometry and Topology in Geographical Information Systems. Proceedings of the „SAMOS“ International Conference on Geographic Information Systems in Urban, Environmental and Regional Planning, Isalnd of Samos, Greece, April 19-21, 1996, pp. 45-60.

Plümer, L., G. Gröger, 1996: Nested Maps - a Formal, Provably Correct Object Model for Spatial Aggregates. Proceedings of the 4th ACM Workshop on Advances in GIS, Rockville, Maryland, November 15-15, 1996, pp. 77-84.

Pütz, D., U. Kemp und R. Troch, 1996: Geo-Route - ein raumbezogenes Planungsinstrument für die Zustellungs- und Transportnetze der Deutschen Post AG. Geo-Informations-Systeme, Jg. 9, Heft 5, S. 9-15.

Raper, J., 1989: Three dimensional applications in GIS. Taylor & Francis, London [u.a.].

Raper, J., 1997: Multi Dimensional GIS. Taylor & Francis, London [u.a.].

Rath, Ch., C. Auerbach, 1996: Prüfung digitaler Grundrißdaten der ALK. Forum 1/1996, S. 306-321.

Roschlaub, R., 1996: Geometrische Datenqualität und Klassifikation von Geodaten. In: Interner Bericht Nr. 5, Institut für Geodäsie und Geoinformatik, Universität Rostock, S. 85-94.

Salgé, F., 1995: Semantic accuracy. In: *Guptill und Morrison, 1995*, pp. 139-151.

Sbresny, J., 1997: Fehlerquellen in Raumbezogenen Informationssystemen. Geologisches Jahrbuch, Reihe F, Heft 33, Schweizerbart´sche Verlagsbuchhandlung, Stuttgart.

Scheuring, R., 1995: Zur Qualität der Basisdaten von Landinformationssystem. Schriftenreihe des Sudiengangs Vermessungswesen, Heft 49, Neubiberg.

Schilcher, M. (Hrsg.), 1990: Kartographie: Anwendungen in der Praxis. Wichmann, Karlsruhe.

Schilling, E. G., 1982: Acceptance Sampling in Quality Control. Marcel Dekker, New York, Basel.

Schmidt, D., D. Fritsch, 1996: In transition from 2.5D GIS to a 3D-GIS. In: International Archives of Photogrammetry and Remote Sensing, Vol. 31, pp. 748-752.

Schmidt, H., 1994: Meßunsicherheit und Messungstoleranz bei Ingenieurvermessungen. Veröffentlichung des Geodätischen Instituts der Rheinisch-Westfälischen Technischen Hochschule Aachen.

Schmidt, H., 1997: Was ist Genauigkeit? – Zum Einfluß systematischer Abweichungen auf Meß- und Ausgleichungsergebnisse. Vermessungswesen und Raumordnung (VR), 59. Jg., Heft 4, S. 212-228.

Scholl, M., A. Voisard (Eds.), 1997: Advances in Spatial Databases. Springer, Berlin [u.a.].

Shi Wenzhong, 1994: Modelling Positional and Thematic Uncertainties in Integration of Remote Sensing and Geographic Information Systems. ITC Publication No. 22, Enschede, The Netherlands.

Stanek, H., A. U. Frank, 1993: Data Quality Requirements for GIS defined by Law: A Case Study. UDMS`93 Proceedings, Wien.

Stanek, H., N. Smith, A. Giordano, 1995: Modellierung und Normierung von Datenqualität im GIS. Online Papers AGIT 95 (URL: http://www.sbg.ac.at/geo/agit/papers95/hstanek.htm).

Stöcker, R., H. Zieschang, 1994: Algebraische Topologie. Teubner, Stuttgart.

Stonebraker, M., 1996: Object-relational DBMSs – the next great wave. Morgan Kaufmann, San Francisco, USA.

Tayi, G. K. (ed.), 1995: Integrating Production Lotsizing, Inspection and Rework Decisions: Useful models and applications. European Journal of operational research, Vol. 80, No. 2.

Thaer, C. (Hrsg.), 1975: Euklid - Die Elemente. Wissenschaftliche Buchgesellschaft, Darmstadt.

Tomlin, C. D., 1990: Geographic information systems and cartographic modeling. Prentice Hall, Englewood Cliffs, NJ.

Tomlinson, R., 1997: GIS planning and implementation life cycle. Seminarunterlagen zu GIS Management Seminar für GIS System- und Projektmanager, 8.-9.3.1997, Veranstalter: ESRI, Kranzberg.

Turner, A. K. (ed.), 1992: Three-Dimensional Modeling with Geoscientific Information Systems. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Uhlmann, W., 1982: Statistische Qualitätskontrolle. Teubner, Stuttgart.

USGS, 1998: American National Standard for Information Systems - Spatial Data Transfer Standard (SDTS) - Part 1, Logical Specifications. (URL: http://mcmcweb.er.usgs.gov/sdts/index.html).

Vauglin, F., 1999: Probability Assessment for the Use of Geometrical Metadata. In: M. Craglia and H. Onsrud (Eds.), Geographic Information Research: Trans-Atlantic Perspectives, Taylor & Francis, London [u.a.], pp. 497-508.

Veregin, H., 1989: Error Modelling for the map overlay Operation. In: *Goodchild and Gopal, 1989*, S. 3-18.

Volkmann, L., 1996: Fundamente der Graphentheorie. Springer, Wien, New York.

Wald, A., 1945: Sequential Tests of Statistical Hypothesis. The Annals of Mathematical Statistics 16: pp. 117-186.

Walter, V., D. Fritsch, 1998: Integration von ATKIS- und GDF-Daten mit Hilfe eines relationalen Zuordnungsansatzes. ZfV Heft 1, 1998, S. 2-10.

Wilson, R. J., J. J. Watkins, 1990: Graphs. John Wiley & Sons, New York [u.a.].

Wise, S., 1998: In for a dime, in for a dozen. GIS Europe, Vol. 7, Issue 1, pp. 16-17.

Wittig, K.-J., 1993: Qualitätsmanagement in der Praxis. Teubner, Stuttgart.

Wood, J., 1994: Visualizing Contour Interpolation Accuracy in Digital Elevation Models. In: *Hearnshaw and Unwin, 1994*, pp. 168-180.

Worboys, M. F., 1994: Innovations in GIS 1. Taylor & Francis, London [u.a.].

Worboys, M. F., 1995: GIS - A Computing Perspective. Taylor & Francis, London [u.a.].

14 Verwendete Normen

ISO/IEC 646: 1991, Information technology - 7-bit coded character set for information interchange

ISO/IEC 6937:1994, Information technology - coded graphic character set for text communication - Latin alphabet

ISO/IEC 9075:1992, Information Technology - Database Languages - SQL

ISO/IEC 10646:1993, Information technology - Universal Multiple-Octet Coded Character Set (UCS)

ISO/IEC 14772-1:1997, Information technology -- Computer graphics and image processing -- TheVirtual Reality Modeling Language (VRML)

ISO 8601: 1988, Data elements and interchange formats - Information interchange - Representation of dates and times

ISO 8859: 1988, Information processing - 8-bit single-byte coded graphic character sets

ISO 10303 Part 11, DIS: 1992, Idustrial automation systems – Product data representation and exchange – Description Methods: The EXPRESS language reference manual

ISO 19104, CD : 1999, Geographic information - Terminology

ISO 19111, CD: 1999, Geographic information – Spatial referencing by coordinates

ISO 19112, CD: 1999, Geographic information – Spatial referencing by geographic identifiers

ISO 19113, CD: 1999, Geographic Information – Quality Principles

ISO 19114, CD: 1999, Geographic Information – Quality Evaluation Procedures

ISO 19115, CD: 1999, Geographic Information – Metadata

ISO/TC 211 N365: 1997, Guidelines for terminology and collections of terms from ISO/TC 211 Geographic information/Geomatics: Short guide for writing definitions

DIN 1319 Teil 1: 1995, Grundlagen der Meßtechnik - Grundbegriffe

DIN 1319 Teil 2: 1996, Grundlagen der Meßtechnik - Begriffe für die Anwendung von Meßgeräten

DIN 1319 Teil 3: 1996, Grundlagen der Meßtechnik - Auswertung von Messungen einer einzelnen Meßgröße, Meßunsicherheit

DIN 1319 Teil 4: 1985, Grundlagen der Meßtechnik – Behandlung von Unsicherheiten bei der Auswertung von Messungen

DIN 18201: 1984, Toleranzen im Bauwesen – Begriffe, Grundsätze, Anwendung, Prüfung

Entwurf DIN 55350 Teil 11: 1992, Begriffe zu Qualitätsmanagement und Statistik

DIN 55350: 1982 – 1989, Begriffe der Qualitätssicherung und Statistik, Teile 11, 12, 13, 15, 21, 23, 24 und 31

DIN ISO 2859 Teil 0: 1991, Teil 1: 1993, Teil 2: 1993, Teil 3: 1995, Annahmestichprobenprüfung anhand der Anzahl fehlerhafter Einheiten oder Fehler (Attributprüfung)

DIS ISO 3951: 1992, Verfahren und Tabellen für Stichprobenprüfung auf den Anteil fehlerhafter Einheiten in Prozent anhand quantitativer Merkmale (Variablenprüfung)

DIN ISO 5725 Teil 1: 1991, Genauigkeit (Richtigkeit und Präzision) von Meßverfahren und Meßergebnissen

DIN ISO 8402: 1991, Qualität – Begriffe

DIN ISO 8402 A1: 1989, Qualität – Begriffe, Änderung 1

DIN ISO 9000 Teil 2: 1992, Qualitätsmanagement- und Qualitätssicherungsnormen

DIN V ENV 12009: 1997, Geoinformation – Referenzmodell

DIN V ENV 12160: 1997, Geoinformation – Datenbeschreibung - Raumbezugsschema

DIN V ENV 12656: 1999, Geographic information – Data description – Quality

DIN V ENV 12657: 1999, Geographic information – Data description – Metadata

prENV 12661: 1998, Geographic information – Referencing – Geographic identifiers

prENV ISO 14825: 1996, Geographic Data Files

A Topologische Beziehungen von Objekten im $\mathbb{R}^2$

Zur Beschreibung der topologischen Beziehung von Objekten müssen zuerst einige Begriffe der algebraischen Topologie eingeführt werden (*Mayer, 1989, Stöcker und Zieschang, 1994, Bronstein et al., 1995, Egenhofer und Herring, 1994*).

Auf einer Menge $\boldsymbol{X}$ sei jedem Paar von Elementen $x, y \in \boldsymbol{X}$ eine reelle Zahl d zugeordnet, so daß für beliebige Elemente $x, y, z \in \boldsymbol{X}$ die folgenden Eigenschaften, die Axiome des metrischen Raumes, erfüllt sind:

(M1)	$d(x,y) \geq 0$ und $d(x,y) = 0$ genau dann, wenn $x = y$	(Nichtnegativität)
(M2)	$d(x,y) = d(y,x)$	(Symmetrie) .
(M3)	$d(x,y) \leq d(x,z) + d(z,y)$	(Dreiecksungleichung)

Eine Funktion $d : \boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}^1_+$ mit den Eigenschaften (M1) bis (M3) heißt Metrik, und das Paar $\boldsymbol{X} = (\boldsymbol{X}, d)$ heißt metrischer Raum.

Eine zentrale Rolle spielt in der Topologie der Umgebungsbegriff. Mit seiner Hilfe läßt sich die räumliche Beziehung eines Punktes in einem topologischen Raum zu den übrigen Punkten und den Teilmengen dieses Raumes beschreiben. Die Topologie eines metrischen Raumes wird durch die Vorgabe von ε-Umgebungen jedes Punktes definiert.

$\boldsymbol{X}$ sei eine beliebige Menge. Ist $(\boldsymbol{X}, d)$ ein metrischer Raum mit der Metrik $d : \boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}^1_+$, so wird für jede positive reelle Zahl ε und jedes $x \in \boldsymbol{X}$ die ε-Umgebung $B(x, \varepsilon) = \{ y \in \boldsymbol{X} \mid d(x,y) < \varepsilon \}$ definiert. Mit Hilfe des Umgebungsbegriffes lassen sich die Begriffe innerer Punkt, offener Kern, Berührungspunkt und Rand für allgemeine topologische Räume definieren.

$\boldsymbol{X}$ sei ein topologischer Raum, $\boldsymbol{A}$ Teilmenge von $\boldsymbol{X}$ und $x \in \boldsymbol{X}$. x heißt **innerer Punkt** von $\boldsymbol{A}$, wenn $\boldsymbol{A}$ Umgebung von x ist. Die Menge der inneren Punkte von $\boldsymbol{A}$ heißt **offener Kern** von $\boldsymbol{A}$ und wird mit $\boldsymbol{A}^o$ bezeichnet. $\boldsymbol{A}$ ist genau dann offen, wenn $\boldsymbol{A} = \boldsymbol{A}^o$.

x heißt **Berührungspunkt** von $\boldsymbol{A}$, wenn für alle Umgebungen $\boldsymbol{U}$ von x gilt, daß $\boldsymbol{U} \cap \boldsymbol{A} \neq \emptyset$. Die Menge aller Berührungspunkte von $\boldsymbol{A}$ heißt **abgeschlossene Hülle** von $\boldsymbol{A}$ und wird mit $\overline{\boldsymbol{A}}$ bezeichnet.

x heißt **äußerer Punkt** von $\boldsymbol{A}$, wenn x innerer Punkt von dem Komplement von $\overline{\boldsymbol{A}}$, $\boldsymbol{X} \setminus \overline{\boldsymbol{A}}$, ist. Die Menge aller äußeren Punkte von $\boldsymbol{A}$ wird mit $\boldsymbol{A}^-$ bezeichnet.

x heißt Randpunkt von $\boldsymbol{A}$, wenn x Berührungspunkt von $\boldsymbol{A}$ und von $\boldsymbol{X} \setminus \boldsymbol{A}$ ist. Die Menge der Randpunkte von $\boldsymbol{A}$ heißt Rand von $\boldsymbol{A}$ und wird mit $\partial \boldsymbol{A}$ bezeichnet. Für den Rand gilt:

$$\partial \boldsymbol{A} = \overline{\boldsymbol{A}} \cap \overline{\boldsymbol{X} \setminus \boldsymbol{A}} = \overline{\boldsymbol{A}} \cap \boldsymbol{X} \setminus \boldsymbol{A}^o = \overline{\boldsymbol{A}} \setminus \boldsymbol{A}^o .$$

Außerdem gilt: $\boldsymbol{A}^o \cup \partial \boldsymbol{A} \cup \boldsymbol{A}^- = \boldsymbol{X}$ und $\boldsymbol{A}^o \cap \partial \boldsymbol{A} = \partial \boldsymbol{A} \cap \boldsymbol{A}^- = \boldsymbol{A}^o \cap \boldsymbol{A}^- = \emptyset$.

Es wird ausschließlich die euklidische Metrik für $x = (x_o, \ldots, x_{n-1})$ und $y = (y_o, \ldots, y_{n-1}) \in \mathbb{R}^n$ mit

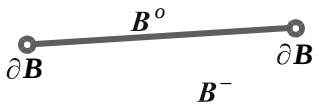
$$d(x,y) = \left(\sum_{i=0}^{n-1} (x_i - y_i)^2 \right)^{1/2}$$

betrachtet.

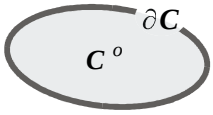
Die Menge $\boldsymbol{A}$ kann dann entweder 0- (punkthaft), 1- (linienhaft) oder 2-dimensionaler (flächenhaft) sein. Ein Punkt ist eine einfache 0-Zelle. Eine Linie ist eine einfache 1-Zelle und eine Fläche ist eine einfache 2-Zelle. Die Zellen werden als einfach bezeichnet, weil es auch zusammengesetzte, bzw. komplexe Zellen gibt.

Punkt $\boldsymbol{A}$ 

$\boldsymbol{A}^o = \neg\emptyset,\ \partial\boldsymbol{A} = \emptyset,\ \boldsymbol{A}^- = \neg\emptyset$

Linie $\boldsymbol{B}$ 

$\boldsymbol{B}^o = \neg\emptyset,\ \partial\boldsymbol{B} = \neg\emptyset,\ \boldsymbol{B}^- = \neg\emptyset$

Fläche $\boldsymbol{C}$ 

$\boldsymbol{C}^-$ $\boldsymbol{C}^o = \neg\emptyset,\ \partial\boldsymbol{C} = \neg\emptyset,\ \boldsymbol{C}^- = \neg\emptyset$

Die Beziehung zwischen zwei Mengen $\boldsymbol{A}$ und $\boldsymbol{B}$ kann durch die paarweise Bildung von Schnittmengen zwischen den inneren Kernen, den Rändern und dem Äußeren dieser Mengen beschrieben werden. Die Schnittmengen können als Matrix dargestellt werden. Sie sind entweder leer ($\emptyset$) oder nicht leer ($\neg\emptyset$), einen dritten Fall gibt es nicht.

$$R(\boldsymbol{A},\boldsymbol{B}) = \begin{pmatrix} \boldsymbol{A}^o \cap \boldsymbol{B}^o & \boldsymbol{A}^o \cap \partial\boldsymbol{B} & \boldsymbol{A}^o \cap \boldsymbol{B}^- \\ \partial\boldsymbol{A} \cap \boldsymbol{B}^o & \partial\boldsymbol{A} \cap \partial\boldsymbol{B} & \partial\boldsymbol{A} \cap \boldsymbol{B}^- \\ \boldsymbol{A}^- \cap \boldsymbol{B}^o & \boldsymbol{A}^- \cap \partial\boldsymbol{B} & \boldsymbol{A}^- \cap \boldsymbol{B}^- \end{pmatrix}$$

Von den theoretisch 2^9 möglichen Kombinationen sind nicht alle sinnvoll und manche widersprüchlich. Eine Ausarbeitung der sinnvollen Fälle mit Beispielen findet sich bei *Egenhofer und Herring, 1994*.

Weil mit dieser zweiwertigen Beschreibung nicht alle Fälle von topologischen Beziehungen eindeutig beschrieben werden können (siehe Kapitel 6.2.4), kann das Modell erweitert werden, indem die Dimension der Schnittmenge angegeben wird (*Clementini and Di Felice, 1994*). Dieses Modell wird als *Dimensionally Extended Nine-Intersection Model (DE-9IM)* bezeichnet. Die dazugehörige Matrix bekommt dann die Form:

$$R(\boldsymbol{A},\boldsymbol{B}) = \begin{pmatrix} \boldsymbol{dim}(\boldsymbol{A}^o \cap \boldsymbol{B}^o) & \boldsymbol{dim}(\boldsymbol{A}^o \cap \partial\boldsymbol{B}) & \boldsymbol{dim}(\boldsymbol{A}^o \cap \boldsymbol{B}^-) \\ \boldsymbol{dim}(\partial\boldsymbol{A} \cap \boldsymbol{B}^o) & \boldsymbol{dim}(\partial\boldsymbol{A} \cap \partial\boldsymbol{B}) & \boldsymbol{dim}(\partial\boldsymbol{A} \cap \boldsymbol{B}^-) \\ \boldsymbol{dim}(\boldsymbol{A}^- \cap \boldsymbol{B}^o) & \boldsymbol{dim}(\boldsymbol{A}^- \cap \partial\boldsymbol{B}) & \boldsymbol{dim}(\boldsymbol{A}^- \cap \boldsymbol{B}^-) \end{pmatrix}$$

Der Operator ***dim( )*** gibt die Dimension der Schnittmenge an. Der Wertebereich liegt im Falle einer 2D-Modellierung bei: -1, 0, 1, 2.

-1	leere Schnittmenge
0	Schnittmenge ist ein einziger Punkt
1	Schnittmenge ist eine Linie
2	Schnittmenge ist eine Fläche

Lebenslauf

	geboren am 7. Dezember 1963 in Ludwigsburg
1970 – 1974	Grundschule in Ludwigsburg
1974 – 1976	Friedrich-Schiller-Gymnasium Ludwigsburg
1976 – 1978	Friedrich-List-Gymnasium Asperg
1978 – 1983	Otto-Hahn-Gymnasium Ludwigsburg
1983 – 1984	Grundwehrsdienst
1984 – 1987	Studium des Vermessungswesen an der Universität Stuttgart
1987 – 1988	Studium des Vermessungswesen an der University of Calgary, Kanada im Rahmen eines DAAD Stipendiums
1988 – 1990	Studium des Vermessungswesen an der Universität Stuttgart Abschluß: Diplom-Ingenieur für Vermessungswesen
Oktober 1990 – August 1992	Wissenschaftlicher Mitarbeiter am Geodätischen Institut der Universität Stuttgart
Seit August 1992	Wissenschaftlicher Mitarbeiter am Institut für Geodäsie der Universität der Bundeswehr München

Dank

Die Anregung zu dieser Arbeit gab Herr Prof. Dr.-Ing. W. Caspary. Mit zahlreichen Veröffentlichungen ist er ein Vorreiter in dem Bereich „Datenqualität in Geoinformationssystemen“. Für die Unterstützung zur Durchführung dieser Arbeit, für viele aufschlußreiche fachliche Diskussionen und für seine Ermunterungen zum Weiterarbeiten möchte ich mich sehr herzlich bei ihm bedanken.

Herrn Prof. Dr.-Ing. W. Möhlenbrink danke ich für das Interesse an dieser Arbeit, seine hilfreichen Anregungen und für die Übernahme des Korreferates. Herr Prof. Dr.-Ing. W. Reinhardt hat mit dem Lehrstuhl für Geoinformatik neue Impulse eingebracht und mit der Übernahme des Vorsitzes im Promotionsverfahren für einen reibungslosen Ablauf gesorgt.

Durch zahlreiche Projekte innerhalb der Arbeitsgemeinschaft Geoinformationssysteme (AGIS) der Universität der Bundeswehr München konnten weitere Anregungen und interessante Fragestellungen in diese Arbeit einfließen und die Theorie wurde mit Aufgabenstellungen aus der Praxis untermauert. Bei meinen Kolleginnen und Kollegen der AGIS möchte ich mich für die konstruktive Zusammenarbeit und die hilfreichen Diskussionen bedanken.

Mit den Problemen und Fehlerquellen bei der Erfassung von Geodaten wurde ich zum ersten Mal bei der Erfassung von Testdaten für das topographische Informationssystem TOPIS für das Amt für Militärisches Geowesen (AmilGeo) konfrontiert. Die Deutsche Forschungsgemeinschaft (DFG) hat mit der Förderung des Projekts „Qualitätsmanagement für Geodaten“ den Grundstein für die theoretische Auseinandersetzung mit dem Thema gelegt. Dieses Qualitätsmanagement konnte für die Erfassung von ATKIS DLM 25/1 und DLM 25/2 und deren Fortführung erstmals mit dem Hessischen Landesvermessungsamt (HLVA) erprobt werden. Der Bedarf des Amtes für Militärisches Geowesen (AMilGeo) an einem Werkzeug zur Prüfung von ATKIS-Daten bei der Übernahme führte zu einem Regelwerk, das so allgemein gehalten wurde, daß es auch für andere Datenbestände anwendbar ist. Bei der Kontrolle der Betriebsmitteldaten des Netzinformationssystems bei den Stadtwerken München (SW/M) konnte das Verfahren der statistischen Qualitätskontrolle in die Praxis umgesetzt werden.

Bei den genannten Institutionen möchte ich mich für die interessanten Fragestellungen und für das entgegengebrachte Vertrauen in eine wissenschaftlich fundierte und praktisch verwertbare Lösung ihrer Probleme bedanken.

Verschiedene GIS-Softwarehersteller haben durch eine Bereitstellung von Hochschullizenzen diese Arbeit erst möglich gemacht. Bedanken möchte ich mich bei den Firmen Intergraph, ESRI, Bentley, AutoDesk und Smallworld (in der Reihenfolge, wie ich mit ihren Systemen vertraut wurde).

Zu guter Letzt möchte ich mich bei meiner Familie und bei allen Freunden für die Geduld und die herzliche und vertraute Atmosphäre bedanken, die zur Fertigstellung dieser Arbeit erforderlich war.