Generalisierung von Landnutzungsdaten

Frank Thiemann

ZUSAMMENFASSUNG:
Der Beitrag stellt einen Ansatz zur Ableitung von CORINE-Land-Cover-Daten aus dem ATKIS-Basis-DLM vor. Nach der Auswahl der relevanten topographischen Objekte werden diese mit einer Übersetzungsmatrix in CLC-Klassen überführt. Durch semantische und geometrische Generalisierung werden dann Objekte unter den Mindesterfassungsmaßen eliminiert. Die Hauptprobleme dabei sind die Identifikation und Reklassifizierung von kleingliedrigen Mischnutzungen durch lokale Analysen, die Behandlung von schmalen langgestreckten Objekten sowie die Behandlung der großen Datenmenge für ganz Deutschland. Der Ansatz bedient sich mehrerer Operatoren, unter anderem der Aggregation, der Zerlegung von Objekten, der Identifikation von Mischklassen und der Vereinfachung der Umringe. Die Ergebnisse und das Laufzeitverhalten werden detailliert diskutiert.

ABSTRACT:
The paper presents an approach for generating CORINE Land Cover (CLC) features from German Authoritative Topographic Cartographic Information System (ATKIS) data for the whole are of Germany. Based on a given transformation matrix describing the transition from topographic data to land-use data, a semantic and geometry based generalization of too small features for the target scale is performed. The challenges of the problem are the identification and reclassification of heterogeneous feature classes by local interpretation, the presence of concave, narrow or very elongated features and the processing of very large data sets. The approach is composed of several steps consisting of aggregation, feature partitioning, identification of mixed feature classes and simplification of feature outlines. The results will be discussed in detail, including runtimes.

1. Einleitung

1.1 Projekthintergrund

Mit CORINE (Coordinated Information on the European Environment) Land Cover dokumentiert die Europäische Umweltagentur (EEA) Landnutzungen und deren Änderungen innerhalb der Europäischen Union. Die Mitgliedsstaaten sind dabei verpflichtet der EEA die Daten zu liefern. Traditionell wurden diese Daten durch Fernerkundung insbesondere aus Satellitenbildern gewonnen. Durch immer kürzer werdende Bereitstellungszeiträume ist diese Art der Gewinnung jedoch sehr kostenintensiv. Das Bundesamt für Kartographie und Geodäsie verfügt mit dem ATKIS-DLM über sehr detaillierte und aktuelle Landnutzungsdaten. Durch einen automatischen Prozess sollen diese Informationen in einen CLC-Datensatz überführt werden.

1.2 CORINE Land Cover (CLC)

Der CORINE-Land-Cover-Datensatz besteht aus 2D-Polygonen in Form einer Tesselation, d.h. die Polygone bilden eine geschlossene Fläche ohne Überlappungen oder Lücken. Der Erfassungsmaßstab entspricht etwa 1:100.000. Als Mindesterfassungsgröße wurden 25 Hektar festgelegt. Benachbarte Polygone sind grundsätzlich von unterschiedlicher Landnutzung, anderenfalls wären sie zu verschmelzen.
Die Landnutzungen sind in 46 Klassen eingeteilt, die hierarchisch über drei Ebenen strukturiert sind. Der Klassencode ist eine dreistellige Zahl wobei jede Ziffer für eine Hierarchieebene steht. Die sieben Hauptebenen sind:

1. Bebaute Flächen (städtisch, industriell, Bergbau)
2. Landwirtschaft (Acker, Dauerkulturen, Grünland, heterogene Flächen)
3. Wald und naturnahe Flächen (Wald, Gestrüpp, offene Vegetation)
4. Feuchtflächen (im Festland, an der Küste)
5. Wasserflächen (Binnengewässer, maritime Gewässer)

Dabei ist CLC im Bereich der Landwirtschaft sehr detailreich. Kleinparzellige landwirtschaftliche Nutzungen und Mischnutzungen können als Mischklassen (24x-Klassen) erfasst werden. Für Deutschland sind dabei die Klassen 242 (komplexe Parzellenstruktur) und 243 (Landwirtschaftliche Nutzfläche mit signifikantem Anteil natürlicher Vegetation) relevant. Bei einer komplexen Parzellenstruktur handelt es sich um eine rein landwirtschaftliche Mischnutzung mit kleinen Parzellen unterschiedlicher Kulturen. Bei der Klasse 243 mischen sich hingegen naturnahe und natürliche Bodenbedeckung mit Landwirtschaft.

1.3 ATKIS Basis DLM

Das digitale Landschaftsmodell aus dem Amtlichen Topographisch-Kartographischen Informationssystem (ATKIS) enthält neben flächenhaften Objekten auch linien- und punktförmig erfasste topographische Objekte. Der Erfassungs-

maßstab beträgt etwa 1:10.000, die Mindesterfassungsgröße für Polygone etwa einen Hektar. Der Datensatz ist in mehrere thematische Ebenen organisiert. Objekte verschiedener Ebenen überlappen sich zum Teil, d.h. die Landnutzungsinformationen bilden keine Tesselation, sondern sind über viele Themenebenen verteilt.
Jedes ATKIS-Objekt hat einen Objektartencode. Neben der Objektart wird es durch objektartenabhängige Attribute näher spezifiziert. Die Attribute bestehen aus einem dreistelligen Schlüssel und einem vierstelligen alphanummerischen Wert. Die Objektarten sind in Objektbereiche und -gruppen eingeteilt. Die sieben Objektgruppen sind Präsentation, Siedlungen, Verkehr, Vegetation, Gewässer, Relief und Gebiete.

In der folgenden Tabelle werden ATKIS und CLC gegenübergestellt.

Datensatz	ATKIS Basis DLM	CORINE LC
Erfassungsmaßstab	1:10.000	1:100.000
Datengrundlage	Luftbilder, ALK	Satellitenbilder
Erfassungsgröße	1 ha	25 ha
Topologie	Überlappungen zwischen Ebenen. Lücken, z. B. bei komplexen Verkehrsobjekten	Tesselation
Objektarten	90 relevante (155 mit Attributen)	44 (37 relevant für Deutschland)
Objektarten Landwirtschaftlicher Nutzung	5 relevante (9 mit Attributen) Keine Mischklassen	11 (6 relevant für Deutschland) 4 (2) Mischklassen

Tabelle 1: Vergleich von ATKIS mit CLC

1.4 Automatische Ableitung von CLC aus dem ATKIS Basis DLM

Ziel des Projektes ist die automatische Ableitung eines CLC-Datensatzes aus dem ATKIS-Basis-DLM. Diese Ableitung ist ein Generalisierungsprozess. Sie erfordert neben Selektion und Reklassifizierung auch eine geometrische Generalisierung zur Vereinfachung der Geometrie entsprechend des Zielmaßstabes. Der Arbeitsablauf besteht aus zwei Teilen:

- der Modelltransformation: Auswahl, Reklassifizierung und topologische Korrektur der Daten und
- der geometrisch-semantischen Generalisierung: Aggregation und Vereinfachung für den kleineren Maßstab.

Im ersten Teil werden zunächst die relevanten Objekte aus dem Basis-DLM ausgewählt. Topologische Probleme wie Überlappungen und Lücken werden automatisch gelöst. Mithilfe einer Übersetzungstabelle werden die ATKIS-Objektarten unter Berücksichtigung der Attribute in CLC-Klassen überführt. Nicht eindeutig überführbare Objekte werden markiert und müssen unter Zuhilfenahme von Fernerkundungsdaten nachbearbeitet werden. Das so abgeleitete Modell wird DLM-DE LC (Landcover) genannt (vergleiche auch Arnold 2009).

Im zweiten Schritt, der im Folgenden ausführlicher dargestellt werden soll, werden die detaillierten DLM-DE-LC-Daten für den Maßstab 1:100.000 generalisiert. Es werden folgende Generalisierungsoperationen verwendet:

- Dissolve – verschmilzt benachbarte Objekte gleicher Objektart
- Aggreagate– verschmilzt zu kleine Objekte mit einem semantisch kompatiblen Nachbarn
- Split – teilt Polygone an zu schmalen Stellen
- 24x-Filter – erkennt und markiert Mischnutzungen
- Simplify– geometrische Vereinfachung der Umringe

1.5 Herausforderungen

Die Datenmenge stellt bei diesem Projekt die größte Herausforderung dar. Der DLM-DE-Datensatz besteht aus zehn Millionen Polygonen. Jedes Polygon besteht im Schnitt aus 30 Stützpunkten, sodass insgesamt 300 Millionen Punkte zu verarbeiten sind. Diese Datenmenge ist derzeit nicht im Hauptspeicher von Standard-PCs zu verarbeiten. Daher werden Konzepte zum sequentiellen oder parallelen Prozessieren der Daten benötigt. Schnelle Algorithmen und effiziente Datenstrukturen sind erforderlich, um eine akzeptable Laufzeit zu erreichen.
Eine weitere Herausforderung ist die Erkennung von Mischklassen. Insbesondere die Begrenzung dieser Mischklasen ist schwierig, da der Übergang zwischen heterogenen Flächen mit nicht signifikanten Kleinflächen zu Mischklassen oft fließend ist.

2. Ansatz

2.1 Daten- und Indexstrukturen

Schnelle Algorithmen benötigen schnelle Such- und Zugriffstrukturen. Die Generalisierungsoperationen arbeiten in der Regel in einer begrenzten räumlichen und topologischen Nachbarschaft. Daher ist eine topologische Datenstruktur sowie in einigen Fällen ein räumlicher Index sinnvoll.

Im Projekt werden folgende Indexstrukturen eingesetzt:

- Doppelt verkettete Kantenliste (DCEL) erweitert um Ringe
- Suchgitter (Kacheln/grid) als räumlicher Index bzw. zweidimensionales Hashing

Doppelt verkettete Kantenliste
Die doppelt verkettete Kantenliste (doubly connected edge list = DCEL) ist eine Datenstruktur zur Speicherung von Polygon-Maschen (siehe Worboys 1995). Sie stellt eine Randbeschreibung (boundary representation) dar. Die topologischen Elemente sind Masche (face), Kante (edge) und Knoten (node) – geometrisch entsprechend Fläche, Linie und Punkt. Alle topologischen Beziehungen dieser Primitive sind explizit gespeichert (siehe Abbildung 1). Um effizient über alle Primitive einer Fläche iterieren zu können, werden die Kanten als ein Paar gerichteter Halbkanten (halfedge) abgespeichert. Jede Halbkante verweist auf ihren Startknoten, die zugehörige Masche, ihre vor- und nachfolgenden Kanten, sowie auf ihre Zwillingskante. Die Knoten tragen mit den Koordinaten die geometrischen Informationen. Zudem verweisen sie auf eine der inzidenten Halbkanten (Abbildung 3). Die Flächen beinhalten einen Verweis auf eine Halbkante ihrer äußeren Umrandung sowie, für den Fall von Löchern, Verweise auf je eine Halbkante der inneren Ringe.

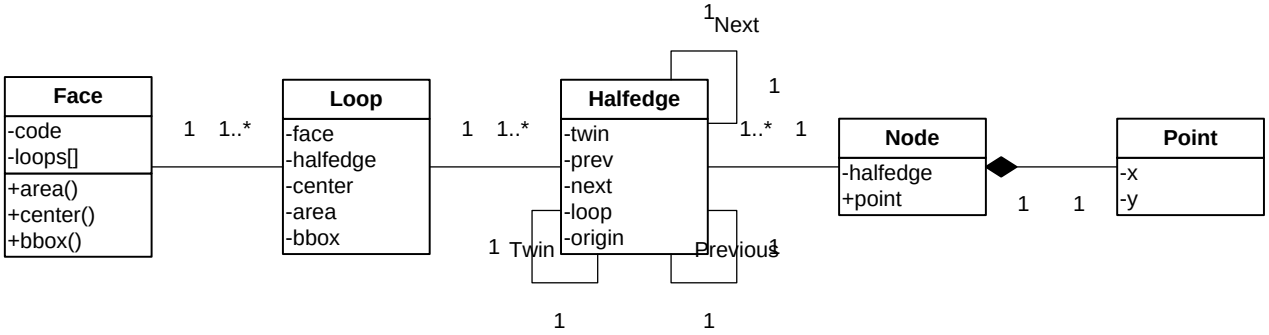

Abbildung 1: UML-Diagramm der erweiterten doppelt verketteten Kantenliste

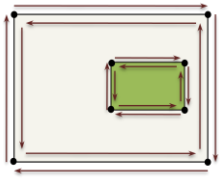

Abbildung 2: Äußere und innere Ringe

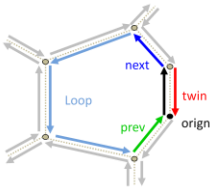

Abbildung 3: Topologie einer Halbkante

Zwischen Halbkante und Maschen wurden zusätzlich Ringe (loop) eingeführt. Diese Erweiterung ist bei 3D-Datenstrukturen gängig und wird z. B. im 3D-Modellierungskern ACIS verwendet. Ein Ring beschreibt einen geschlossenen Linienzug aus Halbkanten. Er kann eine innere oder die äußere Begrenzung einer Fläche bilden (Abbildung 2). Algorithmen zur Flächen- und Schwerpunktberechnung lassen sich nun als Methode der Klasse Ring implementieren. Zur Optimierung von wiederkehrenden Berechnungen werden die Werte in den Objekten gespeichert. Außerdem wird zur schnellen räumlichen Analyse das minimal einschließende Rechteck gespeichert. Der Code für die Landnutzung wurde an die Flächen angehängt.

Suchgitter
Um einen schnellen Zugriff auf die Knoten, Kanten und Maschen zu erhalten, werden alle Daten in Kacheln einsortiert. Dazu wird das ganze Gebiet mit einem regelmäßigen Gitter überzogen. Jede Gitterzelle beinhaltet eine Liste aller enthaltenen Punkte bzw. schneidenden Kanten oder Flächen. Diese einfache Hashingstruktur kann genutzt werden, da die Daten relativ gleichmäßig verteilt vorliegen. Optimale Ergebnisse für das DLM-DE wurden mit Gitterweiten von

100 m für Knoten und Kanten (das entspricht im Mittel weniger als zehn Objekten je Gitterzelle) und 1000 m für die Maschen (ca. 40 Maschen pro Zelle) erzielt. Experimente mit einem KD-Baum für die Punkte lieferten ein ähnliches Zeitverhalten.

2.2 Topologische Bereinigung

Vor dem Generalisierungsprozess müssen die Daten in eine topologische Struktur überführt werden. Bei diesem Schritt werden topologische und auch semantische Fehler aufgedeckt. Von jedem Polygon wird die CLC-Klasse auf Gültigkeit überprüft. Polygone unter einer Minimalgröße von z. B. 0,1 m² werden beim Import ignoriert. Punkte werden innerhalb einer vorgegebenen Distanz von z. B. einem Zentimeter gefangen. Durch einen Punkt-in-Polygon-Test und einen Segmentschnitt-Test werden überlappende Polygone erkannt und abgewiesen. Die topologische Datenstruktur erlaubt nach dem Import ein effizientes Erkennen von Lücken. Dazu werden alle äußeren Kanten verfolgt und Ringe gebildet. Ringe mit positivem Flächeninhalt sind Löcher im Datensatz. Der Ring mit dem maximalen Flächenbetrag und negativem Flächeninhalt ist der äußere Rand des Datensatzes.

2.3 Generalisierung

Das Generalisierungsmodul besteht aus verschiedenen Operatoren, die frei kombiniert werden können. Der Kern der Generalisierung ist die Aggregation der Landnutzungsflächen. Außerdem sind Methoden zur Erkennung von Mischklassen, zur Verschmelzung der Flächen und zur Vereinfachung der Umrisse notwendig. Die verwendeten Methoden werden im Folgenden vorgestellt.

Dissolve
Mit dem Dissolve-Operator werden benachbarte Flächen der gleichen Landnutzungsklasse miteinander verschmolzen. Benachbart bedeutet dabei, dass sich die Flächen mindestens eine gemeinsame Kante teilen. Der Operator entfernt die gemeinsamen Kanten und bildet neue Ringe. Dabei können wie Abbildung 4 zeigt auch zusätzliche innere Ringe entstehen.

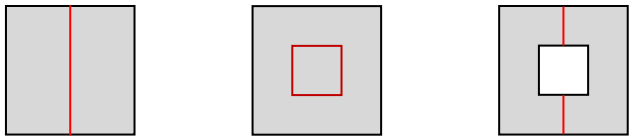

Abbildung 4: Neben den naheliegenden Fällen einer Flächenverschmelzung bei denen die Zahl der Ringe reduziert wird (links und Mitte), gibt es auch Konstellationen bei denen zusätzliche innere Ringe entstehen (rechts).

Aggregate
Der Aggregate-Operator sorgt dafür, dass die Mindestflächengröße eingehalten wird. Der Ansatz dazu wurde erstmalig von van Oosterom (1995) als GAP-tree (Generalized Area Partitioning) beschrieben. Dabei werden in einer Art Region-Growing zu kleine Flächen, beginnend bei der kleinsten, mit einem geeigneten Nachbarn verschmolzen bis alle Flächen dem Kriterium der Minimalgröße genügen. Dieser Ansatz wurde in verschiedenen Generalisierungsanwendungen verwendet, u. A. zur Ableitung vom DLM 50 aus dem Basis-DLM (Podrenek, 2002). Mit dieser Methode können die Minimalbedingungen sicher und schnell erfüllt werden. Allerdings kann dieses Vorgehen auch zu unerwarteten Ergebnissen führen (Abbildung 5), da die Entscheidungen immer nur lokal unter Beachtung der direkten Nachbarn getroffen werden. Das kann zu einer hohen Anzahl an Klassenänderungen führen. Haunert (2008) umgeht diese Nachteile durch eine globale Optimierung. Zudem kann durch Bedingungen die Kompaktheit der Objekte gesteuert werden. Der Optimierungsansatz ist jedoch NP-schwer und somit nur für sehr kleine Kartenausschnitte durchführbar. Angesichts der Datenmenge für ganz Deutschland ist der Ansatz selbst mit den zusätzlich vorgeschlagenen Heuristiken nicht performant genug.

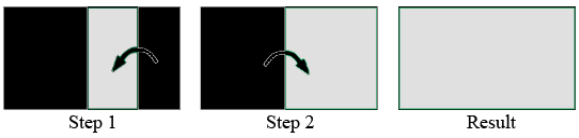

Abbildung 5: Die sequentielle Aggregation kann zu unerwarteten Ergebnissen führen: Im Ausgangsdatensatz dominieren die schwarze Flächen, im Ergebniss jedoch hat sich grau durchgesetzt. (Abb. nach Haunert (2008))

Zur Auswahl eines geeigneten Nachbarn existieren drei Optionen:

- die semantische Ähnlichkeit,
- die erzielbare geometrische Kompaktheit und
- die Kombination beider Kriterien.

Die **semantische** Ähnlichkeit kann mit einer Prioritätentabelle bestimmt werden (siehe Abbildung 6). Ein Null-Wert bedeutet dabei, dass die Klassen identisch sind. Höhere Prioritätswerte bedeuten eine größere semantische Distanz. Als semantisch geeignet wird der Nachbar mit dem kleinsten Prioriätswert gewählt. Sollten mehrere Nachbarn mit dem gleichen Prioritätswert existieren, wird zusätzlich das geometrische Kriterium herangezogen.

CLC	111	112	121	122	123	124	131	132	133	141	142	211	212	213	221	222	223	231	241	242	243	244
111	0	1	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2	3	4
112	1	0	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2	3	4
121	3	3	0	1	1	1	2	2	2	4	4	6	6	6	6	6	6	6	5	5	6	7
122	2	2	1	0	1	1	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4
123	3	3	1	1	0	1	2	2	2	4	4	5	5	5	5	5	5	5	5	5	5	5
124	3	3	1	1	1	0	4	4	4	2	2	6	6	6	6	6	6	5	6	5	6	6
131	3	3	2	2	3	3	0	1	1	4	4	7	7	7	7	7	7	7	7	7	7	7
132	3	3	2	2	3	3	3	0	1	4	4	7	7	7	7	7	7	7	7	7	7	7
133	1	1	1	1	1	1	2	2	0	2	2	3	3	3	3	3	3	3	3	3	3	3
141	3	2	3	3	3	3	3	3	3	0	1	7	7	7	7	7	7	7	7	7	5	5
142	3	2	3	3	3	3	3	3	3	1	0	5	5	5	5	5	5	5	5	5	5	5
211	5	5	5	5	5	5	5	5	5	5	5	0	1	1	4	4	4	3	2	2	2	2

Abbildung 6: Ausschnitt aus der Prioritätentabelle von Bossard, Feranec & Otahel (2000)

Als **geometrisches** Kriterium wird die Länge der gemeinsamen Kante genutzt. Ziel des geometrischen Kriteriums ist die Erzeugung von kompakteren Formen. Die Kompaktheit kann durch das Verhältnis von Flächeninhalt zum Quadrat des Umfangs ausgedrückt werden (Formel 1). Ein Kreis erhält als die kompakteste Form den Wert 100%. Für ein Quadrat ergibt sich ein Kompaktheitswert von 78,5%. Die Kompaktheit sinkt, je länglicher und zerklüfteter die Polygone sind. Die Reduzierung von langen Kanten führt zu einer Verbesserung der Kompaktheit (Abbildung 7).

$$Kompaktheit = 4\pi \cdot \frac{A}{u^2} \cdot 100\% \qquad (1)$$

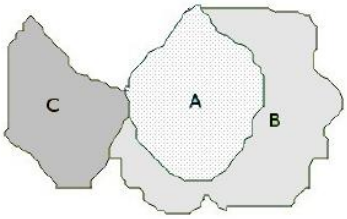

Abbildung 7: Die Vereinigung von A mit B führt zu einem kompakteren Ergebnis als die Vereinigung von A und C, da A und B eine längere gemeinsamme Kante haben.

Abbildung 8 verdeutlicht, dass weder das semantische noch das geometrische Kriterium alleine zum gewünschten Ergebnis führen. Bei der rein semantischen Aggregation entstehen sehr langestreckte oder zergliederte Formen. Bei der rein geometrischen Aggregation ändert sich die Semantik stark.

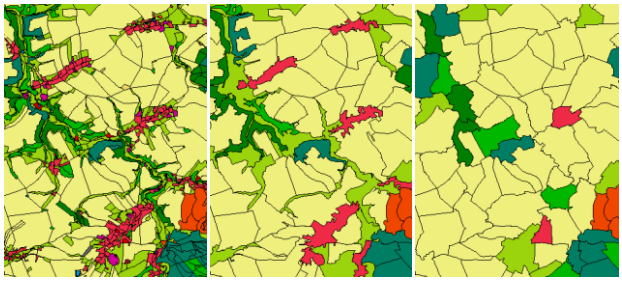

Abbildung 8: Ausgangssituation (links); Das Ergebnis der rein geometrischen Aggregation (rechts) ist kompakter jedoch semantisch ungenauer als das Ergebnis der rein semantischen Aggregation (Mitte)

Durch Kombination können beide Kriterien berücksichtigt werden. Dabei werden semantische Nähe der Flächen und Summe der Längen der gemeinsamen Kanten gewichtet. Je länger die Summe der gemeinsamen Kanten ist, umso größer darf die semantische Distanz zwischen den Flächen werden. Diese Beziehung wird durch Formel 2 ausgedrückt.

$$Distanz\ (A,B)\ =\ \frac{b^{Priorität(A,B)}}{Kantenlänge} \qquad (2)$$

Die Formel besagt, dass eine b-fach längere gemeinsame Kante die Wahl der nächst schlechteren Priorität erlaubt. Durch Veränderung der Basis b kann der Einfluss der Semantik gesteuert werden. Mit b=1 hat die Semantik keinen Einfluss, man erhält die kompaktesten Ergebnisse. Je höher b gewählt wird, umso stärker wird die Semantik berücksichtigt.

Durch leichte Abwandlung des geometrischen Aggregate-Operators ist eine Dissolve-Operation möglich, die nur bis zum Erreichen der Mindestgröße, Objekte gleicher Klasse zusammenfasst. Diese Ergänzung zum normalen Dissolve-Operator ist notwendig um zu vermeiden, dass Objekte entstehen, die sich über den gesamten Datensatz ausdehnen, wie das z. B. bei Flusssystemen zu erwarten ist.

Split
Bei der Flächengeneralisierung sind neben Minimalflächen auch Minimalbreiten zu berücksichtigen. Zu schmale Objekte sollten kollabiert oder einem geeigneten Nachbarn zugeschlagen werden. Dies trifft insbesondere auf schmale Flussläufe, Verkehrsflächen oder Begleitgrün zu. Experimente mit dem Kollapse-Algorithmus von Haunert & Sester (2008) zeigten, dass diese Implementierung aufgrund der zeitaufwendigen Puffer- und Skeletieroperationen zu langsam für die großen Datenbestände ist. Sattdessen können mit dem Split-Operator schmale Objektteile abgetrennt werden und diese über eine nachgeschaltete stark geometrisch gewichtete Aggregation anderen Nachbarn zugewiesen werden.

Abbildung 9: Der Split-Operator teilt Polygone an schmalen Stellen, sofern dort ein Knoten höherer Ordnung oder eine konkave Ecke vorliegt. Existiert auf der Gegenseite kein Knoten in der Nähe, wird sattdessen der Lotfußpunkt verwendet.

Der Split-Operator teilt Polygone an schmalen Stellen. Dazu wird der Knoten mit dem geringsten Abstand zu einer nicht-adjazenten Kante bestimmt. Es werden dabei nur konkave Knoten (dort sind die kürzesten Abstände zu erwarten) und Knoten höherer Ordnung (dort geschnitten entstehen kompaktere Formen) berücksichtigt. Das Lot auf diese Kante muss innerhalb der Fläche verlaufen und eine signifikante Abkürzung gegenüber dem Weg auf dem Umring darstellen. Auf der Gegenseite wird der Lotfußpunkt als Knoten eingefügt, sofern nicht bereits ein Knoten in der Nähe besteht.

24x-Filter
Anders als in ATKIS gibt es in CORINE-Landcover eine Gruppe von Klassen für landwirtschaftliche Mischnutzungen. Da in ATKIS jede Nutzung für sich klassifiziert wird, ist ein spezieller Operator notwendig, der Mischnutzungen erkennt um sie nach CLC zu überführen. Die Mischnutzung ist dabei durch kleine Flächen (unter der Mindesterfassungsgröße) mit unterschiedlicher, vorwiegend landwirtschaftlicher Nutzung gekennzeichnet. Für die Erkennung der Klasse 242 sind ausschließlich landwirtschaftliche Nutzflächen (2xx) relevant. Bei der Klasse 243 müssen zusätzlich Wald, naturnahe und natürliche Flächen berücksichtigt werden.
Der Operator berechnet für jede Masche eine Nachbarschaftsstatistik. Dazu werden alle Maschen mit dem Schwerpunkt innerhalb eines vorgegebenen Radius und dem Flächeninhalt kleiner als der gesuchten Mindesterfassungsgröße über eine Tiefensuche in der topologischen Datenstruktur zusammengesammelt. Innerhalb dieser Daten werden die Flächenanteile der landwirtschaftlichen Nutzung (2xx) und der naturnahen und natürlichen Flächen (3xx, 4xx, 512) berechnet. Falls eine Klasse dominiert (ihr Anteil ist größer als 75%), wird diese Klasse für die untersuchte Masche übernommen. Um heterogene Gebiete von Rändern größerer homogener Flächen zu unterscheiden, wird ein Maß für die Heterogenität benötigt. Eine Fläche ist heterogen genutzt, wenn viele Nutzungsartenwechsel auf kleiner Fläche stattfinden. Die Nutzungsartenwechsel drücken sich als Kanten zwischen den Flächen aus. Als Maß kann die Kantenlänge bezogen auf den betrachteten Flächeninhalt zwischen den zu berücksichtigenden Nutzungen verwendet werden. Um zwischen den Klassen 242 und 243 zu unterscheiden, wird der Anteil der naturnahen Flächen herangezogen. Haben die naturnahen Flächen (3xx, 4xx, 512) bezogen auf die relevanten Klassen einen Anteil von mehr als 25% wird die Fläche als 243 klassifiziert.

Simplify
Der Simplify-Operator dient der Vereinfachung der Polygonumringe, indem er nicht benötigte Stützpunkte entfernt. Ein Punkt wird nicht benötigt, wenn sich die Geometrie ohne den Punkt nur unwesentlich (innerhalb einer definierten Schranke) verändert und die Topologie beibehalten wird. Die Implementierung verwendet den Algorithmus von Douglas & Peucker (1973). Er wurde um die Behandlung von geschlossenen Polygonen und einen Topologietest erweitert.

Der Operator arbeitet über alle Ringe jeweils stückweise zwischen zwei topologischen Knoten (Grad > 2). Bei Ringen ohne topologischen Knoten wird ein beliebiger Punkt als Startknoten ausgewählt. Der Algorithmus versucht wie der von Douglas-Peuker zunächst die direkte Verbindung von Anfangs- und Endknoten zu verwenden und prüft den Abstand aller ausgelassenen Punkte. Sollten Anfangs- und Endpunkt identisch sein, wird anstelle des Punkt-Geraden-Abstandes der Punkt-Punkt-Abstand verwendet. Der Punkt-Segment-Abstand wurde nicht verwendet, da auch schmale Spitzen entfernt werden sollen. Ist die Distanz größer als der vorgegebene Schrankenwert, der sinnvollerweise im Bereich der Kartiergenauigkeit gewählt wird, wird der Punkt mit der maximalen Distanz wieder hinzugenommen, der Algorithmus für beide Teilsegmente erneut durchgeführt. Anderenfalls können die Zwischenpunkte nicht sofort gelöscht werden, sondern es erfolgt zunächst der Topologietest. Dazu werden alle Punkte innerhalb der Boundingbox um das aktuelle Liniensegment auf Seitenwechsel bezüglich der Linie getestet. Sollte ein Punkt die Seite wechseln, wird wiederum der Punkt mit der größten Distanz beibehalten.

2.4 Verkettung der Operatoren

Im Folgenden wird dargestellt, wie die Operatoren im Programmablauf angeordnet werden. Die Parameter sind zur Ableitung der CLC-Daten im Maßstab 1:100.000 mit 25 Hektar Mindesterfassungsgröße gewählt.

1. Import und Datenbereinigung
2. Lücken füllen
3. Dissolve: Flächen kleiner als 25 ha
4. Split: Flächen schmaler als 50 m
5. Aggregate: Flächen kleiner als 1 ha (stark geometrisch gewichtet, Basis=1,2)
6. 24x-Filter (Radius=282 m entspricht 25 ha Kreis)
7. Aggregate: Flächen kleiner als 5 ha (gewichtet, Basis=2)
8. Aggregate: restliche Flächen kleiner als 25 ha rein semantisch
9. Simplify (Toleranz 20 m)
10. Dissolve: ohne Größenbeschränkung

Während des Imports (1) werden Semantik und Topologie überprüft. Kleine topologische Fehler werden durch einen Punktfang bereinigt. Objekte die einen topologischen Fehler verursachen werden abgewiesen. Objekte mit ungültigem Klassencode erhalten einen Platzhaltercode, ebenso wie die im zweiten Schritt (2) gefundenen Lücken. Diese Platzhalterobjekte werden in folgenden Schritten bevorzugt Nachbarobjekten zugeschlagen.
Der erste Dissolve-Schritt (3) dient der Bildung von größeren Objekten. Insbesondere Siedlungsflächen sind oft sehr kleinteilig erfasst. Die Schranke verhindert die Bildung von sehr großen Polygonen, die sich negativ auf Laufzeitverhalten und Speicherbedarf auswirken würden. Hierbei entstehende unkompakte Flächen werden durch den Split-Operator (4) an Stellen, die schmaler als 50 Meter sind, zerlegt. Anschließend werden Flächen, die kleiner als ein Hektar sind, mit geometrisch geeigneten Nachbarn verschmolzen (5).
Mit dem 24x-Filter (6) werden durch Analyse des 25-Hektar-Umkreises Kandidaten für heterogene Objekte gesucht. Anschließend werden alle Flächen in zwei Schritten bis auf die Mindesterfassungsgröße aggregiert. Zunächst wird bis zu einer Größe von fünf Hektar unter Berücksichtigung des geometrischen Kriteriums zusammengefasst (7). Im Größenbereich über fünf Hektar wird danach ausschließlich nach semantischen Kriterien zusammengefasst (8) um großflächige Klassenänderungen mit größeren semantischen Distanzen zu vermeiden. Die Knotenreduzierung (9) erfolgt vor der abschließenden Zusammenführung (10). Als geometrische Toleranz für die Reduzierung werden 20 Meter (entsprechend 0,2 mm in der Karte) gewählt.

3. Ergebnisse

3.1 Laufzeitverhalten und Speicherbedarf

Die implementierten Algorithmen sind schnell, benötigen jedoch relativ viel Arbeitsspeicher. Die Daten- und Indexstrukturen belegen bis zu 160 Bytes pro Punkt auf einem 32-bit-Rechner. Auf einem 64-bit-Rechner mit 6 GB freien Arbeitsspeicher gelang es ohne Partitionierung, Datensätze mit bis zu 30 Millionen Punkten zu prozessieren – das entspricht etwa einem Zehntel von Deutschland.
Laufzeittests wurden auf einem 32-bit Rechner mit einem 2,66 GHz Intel Core 2 Prozessor durchgeführt. Das System ist nach dem Windows-Performance Test ausgewogen konfiguriert - RAM, Festplatte und Prozessor haben Windows Performance Index 5,5.
Den größten Zeitanteil der Prozessierung benötigen die Ein-/Ausgabe-Operationen. Dabei können bis zu 100.000 Punkte pro Sekunde aus Shapefiles gelesen und in die topologische Datenstruktur eingefügt werden. Die gemessenen Schreibgeschwindigkeiten waren vom Festplattenschreibcache beeinflusst. Ohne Cache entsprechen sie etwa der Lesegeschwindigkeit.

Die Laufzeit der Generalisierungsoperationen ist stark von den Daten beeinflusst. Die Reihenfolge der Generalisierungsoperationen beeinflusst damit auch signifikant die Laufzeit der gesamten Generalisierung. Im vorgestellten Prozessablauf benötigt die Split-Operation die meiste Zeit. Sie hat ein quadratisches Laufzeitverhalten bezüglich der Stützpunkte je Polygon. Beim ersten Dissolve-Schritt werden etwa 75% der Polygone beseitigt. Dabei reduziert sich die Anzahl der Stützpunkte jedoch nur um 25%, wodurch die Zahl der Stützpunkte je Polygon auf das Dreifache steigt. Damit benötigt der Split-Algorithmus an dieser Stelle trotz Datenreduktion mehr Zeit als er vor dem Dissolve benötigen würde. Der Zeitbedarf liegt hierbei in der Größenordnung der Lesezeit.
Alle anderen Generalisierungsoperationen sind mindestens 10-mal so schnell wie das erstmalige Lesen. Der Aggregate-Operator prozessiert eine Millionen Punkte pro Sekunde. Die Liniengeneralisierung ist mit 0,7 Millionen Punkten pro Sekunde etwas langsamer, wird jedoch auch erst auf dem stark reduzierten Datensatz am Ende der Prozesskette ausgeführt.

Operation	Anzahl Polygone	Lesen	Operation	Schreiben (gecached)
1+2) Import+Lücken	91717 → 91781	20 s	>1 s	2 s
3) Dissolve 25 ha	91781 → 26431	21 s	2,6 s	1 s
4) Split 50 m	26431 → 68451	10 s	20 s	2 s
5) Aggregate 1 ha	68451 → 16565	11 s	1,5 s	0 s
6) 24x-Filter	16565	6 s	0,3 s	0 s
7) Aggregate 5 ha	16565 → 7397	6 s	0,4 s	0 s
8) Aggregate 25 ha	7397 → 3734	4 s	0,2 s	0 s
9) Simplify 20 m	3734	3 s	0,7 s	0 s
10) Dissolve	3734 → 1240	1 s	0,2 s	0 s
Summe (ca.)		82 s	26 s	5 s

Tabelle 2: Zeitbedarf der Generalisierungsoperationen am Beispiel Dresden auf einem 2,66 GHz PC

In Tabelle 2 ist beispielhaft der Zeitbedarf für die Prozessierung des Dresden-Datensatzes aufgeführt. Die Zeiten für die Lese-/Schreibzugriffe beziehen sich auf den Zugriff auf Shapedateien.

3.2 Semantische und geometrische Genauigkeit

In Abbildung 10 werden Eingangsdaten (DLM-DE), unser Ergebnis und der Referenzdatensatz (CLC 2006) der Region Dresden gegenübergestellt. Erkennbar ist eine deutliche Vereinfachung des Ausgangsdatensatzes. Der Fluss Elbe ist in seiner Form erhalten. Die Struktur der Nutzung ist weiterhin erkennbar. Unter den weggefallen Objekten sind auch einige langgestreckte zergliederte Straßendörfer. An einigen Stellen wurden Mischklassen gebildet. Im Vergleich zum CLC 2006 ist festzustellen, dass dort die Flächen etwas kompakter sind. Ferner sind im CLC 2006 sind mehr Straßendörfer erhalten. Einige im CLC 2006 erfassten Wald- und Industrieflächen finden sich jedoch mit entsprechender Klassifizierung nicht im DLM-DE wieder. Mischklassen wurden in unserem Ergebnis in etwas geringerem Umfang gebildet. Die Lage der Mischklassen ist augenscheinlich plausibel, stimmt jedoch oft nicht mit den Mischklassen im Referenzdatensatz überein.
Diese optischen Eindrücke werden auch durch die Statistiken belegt. Zur Beurteilung der semantischen und geometrischen Genauigkeit wurden Vergleiche zwischen dem Ausgangsdatensatz (DLM-DE), unserem Ergebnis und dem Referenzdatensatz (CLC 2006) gerechnet.

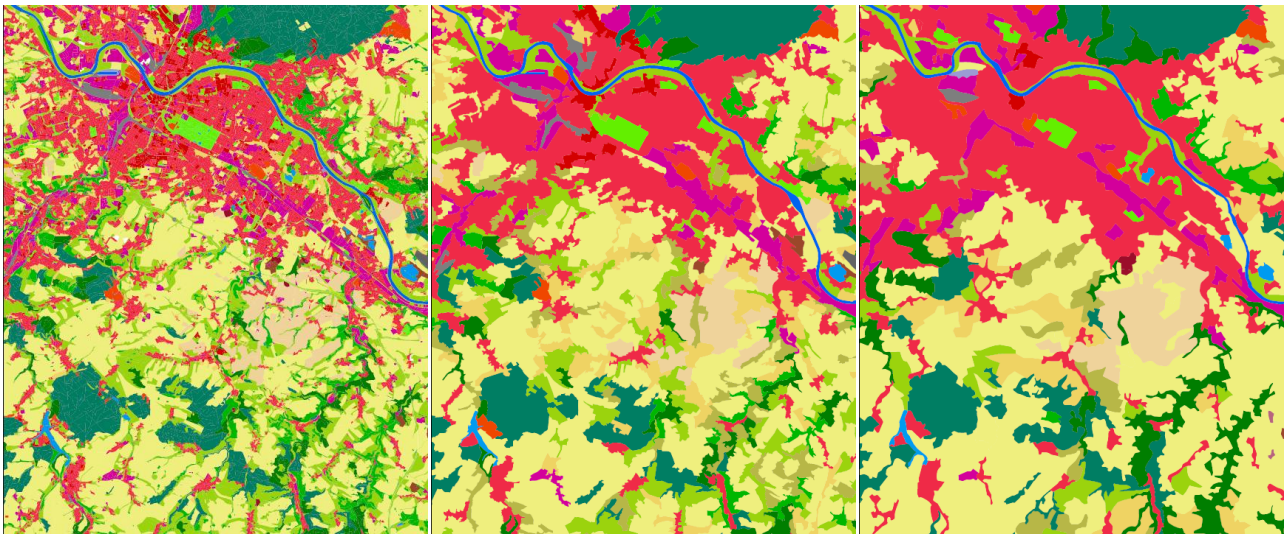

Abbildung 10: Testgebiet Dresden (von links nach rechts): DLM-DE, Ergebnis und CLC 2006 als Referenz

Zunächst wurden die Übereinstimmungen zwischen den Datensätzen überprüft (siehe Abbildung 11). Die größte Übereinstimmung (75%) weisen das DLM-DE und der daraus abgeleitete CLC-Datensatz auf. Referenz und DLM-DE stimmen nur zu 60% überein. Unser aggregierter Datensatz ähnelt dem CLC-2006 mit 66% Übereinstimmung mehr als die Ausgangsdaten.
Die jeweiligen Flächenanteile am Datensatz sind in Tabelle 3 aufgeführt und in Abbildung 12 dargestellt. Besonders deutliche Unterschiede zwischen DLM-DE und CLC 2006 zeigen sich dabei beim Grünland (231), beim Laubwald (311) sowie bei den durchgängig städtisch geprägten Flächen (111). Die Flächenanteile im generalisierten Datensatz liegen in der Regel zwischen denen im Ausgangs- und Referenzdatensatz. Die Mischklassen 242 und 243 kommen im DLM-DE kaum bzw. nicht vor. Durch die Generalisierung wurden sie bezogen auf den Referenzdatensatz in etwas geringerem Umfang erzeugt.

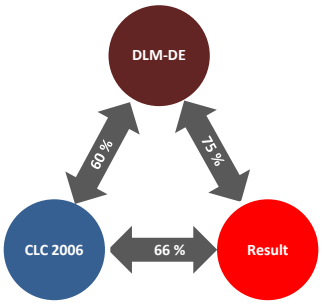

Abbildung 11: Anteil der übereinstimmenden Flächennutzung

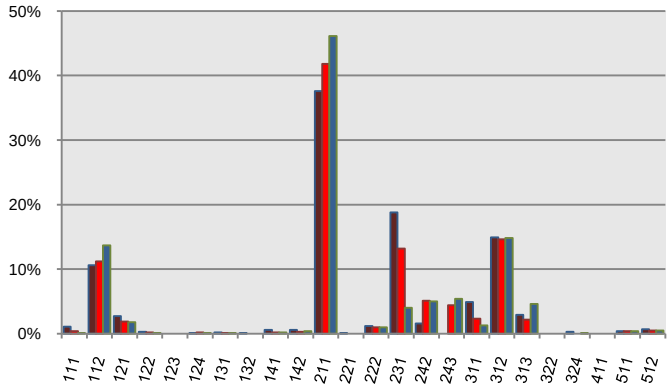

Abbildung 12: Flächenanteile der Klassen

Tabelle 3 stellt außerdem Vollständigkeit und Korrektheit der Klassen von den möglichen Paarungen (Ziel/Referenz) dar. Die Vollständigkeit gibt den Flächenanteil der Referenzobjekte an, die im Zieldatensatz an der gleichen Stelle vorhanden sind. Die Korrektheit gibt den Flächenanteil der Zielobjekte an, die im Referenzdatensatz an der gleichen Stelle vorhanden sind. An den geringen Werten für die Mischklasen (24x) wird deutlich, dass diese Klassen nicht an den gleichen Stellen gebildet werden. Es ist zu prüfen, inwiefern eine eindeutige Erkennung dieser Klassen möglich ist oder ob auch bei manueller Interpretation keine eindeutigen Ergebnisse erzielt werden.

	DLM-DE	Ergebnis	CLC 2006	CLC 2006 / DLM-DE		Ergebnis / DLM-DE		Ergebnis / CLC 2006	
Code	Anteil	Anteil	Anteil	Korrekt	Vollständig	Korrekt	Vollständig	Korrekt	Vollständig
111	1,1%	0,3%	0,1%	50%	5%	71%	22%	16%	53%
112	10,6%	11,1%	13,8%	53%	70%	69%	73%	79%	64%
121	2,7%	1,9%	1,8%	53%	36%	67%	48%	54%	57%
122	0,3%	0,2%	0,1%	44%	13%	59%	45%	17%	44%
123	0,0%	-	0,0%	53%	73%	-	0%	-	0%
124	0,1%	0,2%	0,1%	71%	64%	90%	91%	67%	76%
131	0,2%	0,1%	0,1%	41%	14%	64%	29%	7%	10%
132	0,1%	0,0%	0,0%	18%	5%	40%	13%	0%	0%
133	-	-	0,0%	0%	-	-	-	-	0%
141	0,6%	0,2%	0,2%	39%	12%	69%	24%	30%	35%
142	0,6%	0,2%	0,4%	16%	10%	74%	32%	22%	14%
211	37,6%	42,4%	46,3%	72%	88%	84%	95%	87%	79%
221	0,1%	-	-	-	0%	-	0%	-	-
222	1,2%	1,0%	1,0%	65%	56%	85%	72%	70%	67%
231	18,8%	15,0%	4,0%	71%	15%	78%	62%	20%	76%
242	1,6%	2,8%	5,1%	1%	3%	17%	30%	6%	4%
243	-	3,7%	5,3%	0%	-	0%	-	30%	21%
311	4,9%	2,5%	1,3%	47%	13%	82%	42%	24%	46%
312	14,9%	15,1%	14,8%	81%	80%	92%	93%	82%	84%
313	2,9%	2,2%	4,5%	18%	27%	78%	58%	33%	16%
322	0,0%	0,0%	-	-	0%	69%	37%	0%	-
324	0,3%	0,0%	0,1%	7%	3%	86%	4%	0%	0%
411	0,0%	-	-	-	0%	-	0%	-	-
511	0,4%	0,4%	0,4%	85%	73%	90%	89%	74%	85%
512	0,7%	0,5%	0,5%	85%	59%	84%	63%	75%	80%
997	0,1%	-	-	-	0%	-	0%	-	-
998	0,1%	-	-	-	0%	-	0%	-	-

Tabelle 3: Flächenanteile der CLC-Klassen sowie Korrektheit und Vollständigkeit bezogen auf Ziel/Referenz am Beispiel Dresden. Hervorgehoben sind die Klassen mit deutlichen Unterschieden zischen DLM-DE und CLC 2006

In Tabelle 4 sind Komplexität und Kompaktheit der Polygone der einzelnen Datensätze am Beispiel von Dresden gegenübergestellt. Die Komplexität drückt sich in der Anzahl der Punkte je Polygon sowie in der Anzahl und Größe der Polygone aus. Die Kompaktheit setzt den Flächeninhalt in Beziehung zum Umfang (Formel 1).
In der Gegenüberstellung ist erkennbar, dass der generalisierte Datensatz deutlich reduziert wurde. Er ist geringfügig komplexer und etwas weniger kompakt als der CLC-Datensatz von 2006.

Datensatz	DLM-DE	Ergebnis	CLC 2006
Polygone	91324	1341	878
∅ Punkte pro Polygon	24	104	77
∅ Polygonfläche	2,3 ha	155 ha	238 ha
∅ Polygonumfang	0,6 km	9,4 km	10,1 km
∅ Kompaktheit	50%	24%	33%

Tabelle 4: Komplexität und Kompaktheit der Dresden-Datensätze

4. Zusammenfassung und Ausblick

Es wurde eine Sammlung von Operationen entwickelt, mit dem eine Generalisierung von tesselationsartig vorliegenden Flächennutzungsdaten möglich ist. Die Bearbeitung des DLM-DE für ganz Deutschland am Stück ist aufgrund der hohen Datenmenge mit Standard-PCs nicht möglich. Eine sequentielle Bearbeitung ist theoretisch möglich, benötigt aber viel Zeit, da in der Regel die räumliche und topologische Umgebung des zu bearbeitetenden Objekts berücksichtigt werden muss und somit Objekte wiederholt gesucht und gelesen werden müssen. Um Schreib-/Lese-Zugriffe zu sparen wird derzeit eine Partiontierungslösung entwickelt, die den Datensatz weitgehend kachelweise abarbeitet. Diese Partitionierung ist auch Voraussetzung für eine eventuelle parallele Prozessierung. Randeffekte an den Partitionsgrenzen sollen dabei möglichst gering bleiben.
Es bleibt außerdem zu untersuchen, wie empfindlich die Generalisierungsergebnisse bezüglich der Parameter und der Reihenfolge der Generalisierungsoperationen sind. Dies ist insbesondere deshalb von Interesse, da aus dem Datensatz auch Änderungen der Landnutzung abgeleitet werden sollen. Dabei müssen die tatsächlichen Änderungen der Nutzung von Pseudo-Änderungen, die durch die Generalisierung auftreten, getrennt werden. Es ist zu prüfen, ob zunächst die Änderungen aus dem hochauflösenden Datensatz abgeleitet und generalisiert werden sollten und anschließend der neue Datensatz durch Verschneidung mit den generalisierten Änderungen entsteht.

5. Literatur

Arnold, S., 2009. Digital Landscape Model DLM-DE – Deriving Land Cover Information by Intergration of Topographic Reference Data with Remote Sensing Data. *Proceedings of the ISPRS Workshop on High-Resolution Earth Imaging for Geospatial Information*, Hannover.

Bossard, M., Feranec, J. & Otahel, J., 2000. EEA CORINE Land Cover Technical Guide – Addendum 2000. – Technical Report No. 40, Kopenhagen.

Büttner, G., Feranec, G. & Jaffrain, G., 2006. EEA CORINE Land Cover Nomenclature Illustrated Guide – Addendum 2006. – European Environment Agency.

Douglas, D. & Peucker, T., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, *The Canadian Cartographer* 10 (1973) S. 112-122.

Geoff, B. et al. 2007:. UK Land Cover Map Production Through the Generalisation of OS MasterMap®. *The Cartographic Journal*, 44 (3). S. 276-283.

Haunert, J.-H., 2008. Aggregation in Map Generalization by Combinatorial Optimization, Vol. Heft 626 of *Reihe C*, Deutschen Geodätische Kommission, München.

Haunert, J.-H. & Sester, M., 2008. Area collapse and road centerlines based on straight skeletons, *GeoInformatica*, vol. 12, no. 2, S. 169-191.

Pondrenk. M, 2002. Aufbau des DLM50 aus dem Basis-DLM und Ableitung der DTK50 – Lösungsansatz in Niedersachsen. In: *Kartographische Schriften, Band 6, Kartographie als Baustein moderner Kommunikation*, S.126-130, Bonn.

van Oosterom, P., 1995. The GAP-tree, an approach to 'on-the-fly' map generalization of an area partitioning, in: J.-C. Müller, J.-P. Lagrange & R. Weibel, eds, *GIS and Generalization - Methodology and Practice*, Taylor & Francis, S. 120-132.

Worboys, M. 1995. *GIS – A Computing Perspective*. Taylor & Francis, S. 196-202.