Assuring Logical Consistency and Semantic Accuracy in Map Generalization

JAN-HENRIK HAUNERT & MONIKA SESTER, Hannover

Keywords: GIS, data abstraction, spatial data quality, aggregation, optimization

Summary: In recent years national mapping agencies have increasingly integrated automatic map generalization methods in their production lines. This raises the question of how to assess and assure the quality of mapping products such as digital landscape models. Generalization must not only ensure specified standards for an output scale, but also needs to keep semantics as similar as possible under these given requirements. In order to allow for objective comparisons of different generalization results we introduce a semantic distance measure. We present results that optimize this measure subject to constraints reflecting database specifications and show how this measure can be used to compare the results of different methods, including exact and heuristic approaches.

Zusammenfassung: *Gewährleistung logischer Konsistenz und semantischer Genauigkeit in der Generalisierung.* In zunehmendem Maß werden automatische Generalisierungsverfahren für die Produktion amtlicher digitaler Landschaftsmodelle eingesetzt. Dadurch entsteht ein wachsender Bedarf nach Verfahren zur Qualitätskontrolle und Qualitätssicherung. Generalisierung muss nicht nur für den Zielmaßstab definierte Standards realisieren, sondern dabei auch die Semantik repräsentierter Objekte nach Möglichkeit erhalten. Wir definieren ein semantisches Distanzmaß, um einen objektiven Vergleich unterschiedlicher Generalisierungsergebnisse zu ermöglichen, präsentieren Ergebnisse, die unter Nebenbedingungen aus existierenden Spezifikationen hinsichtlich dieses Maßes optimal sind, und zeigen Vergleichsmöglichkeiten von Ergebnissen exakter und heuristischer Verfahren auf.

1 Introduction

According to MORRISON (1995) there are seven elements of spatial data quality: Lineage, positional accuracy, attribute accuracy, completeness, logical consistency, semantic accuracy and temporal information. Most of them are affected by map generalization, for example, when applying displacement or simplification algorithms to lines, their positional accuracy is reduced; selection of objects affects completeness. The assessment and assurance of these quality criteria are important problems, especially, when heuristic generalization methods are applied. In this article we discuss how to assure semantic accuracy and logical consistency, that is, compliance with database specifications. Following this aim, we have developed a method for area aggregation by mixed-integer programming (HAUNERT & WOLFF 2006). This method has technically been presented in sufficient depth: we have proven the NP-hardness of the problem, tested multiple optimization criteria (HAUNERT 2007a), and developed specialized heuristics to obtain a better performance (HAUNERT 2007b). This method yields very good results, especially compared with commonly used iterative approaches, that locally merge too small objects with their best compatible neighbors (Haunert2007b). However, we have not sufficiently elaborated the usefulness of this method for quality assurance and quality assessment. This article focuses on these issues.

The issue of semantic accuracy is especially relevant, when map objects change their classes. In map generalization this happens in two cases: class abstraction and object aggregation. The latter case leads to class changes, when multiple objects of different classes are replaced by a single composite object. Class abstraction means, for example, to replace all churches and all post offices by objects of type public building. This only needs to be done one time by an expert on a conceptual level and thus it can easily be implemented. In contrast, object aggregation is a labor-intensive problem that needs to be automated. When masses of data are processed, also the quality assessment becomes difficult. CHENG & LI (2006) suggest to measure the semantic accuracy according to the area that changes its class in generalization. YAOLIN et al. (2002) introduce a symmetric semantic similarity matrix to compare object types of areas before and after reclassification. RODRÍGUEZ & EGENHOFER (2004)

propose an asymmetric similarity measure. Similarity values are derived from the given data model, taking class hierarchies into account and comparing attribute definitions. AHLQUIST (2005) uses a similarity measure based on fuzzy membership functions to assess land cover changes over time.

The quality of map generalization is normally defined by comparison of input and output data sets (BARD 2004; FRANK & ESTER 2006). These methods mainly depend on measures that characterize the shapes of objects and their spatial relationships. Generally, observed changes are penalized in the assessment. However, generalization naturally cannot always preserve the original situation: there are driving forces to change the data set, for example, minimal allowed sizes for the target scale. Thus, we compare the results of heuristic generalization methods with results that are optimal under given constraints. These results can be obtained with our mixed-integer program. Though this is only possible for small samples, these offer new possibilities to detect shortcomings of heuristics.

In the sequence of the article, we first explain our conceptions of logical consistency (Section 2) and semantic accuracy (Section 3) and then present our approach to assure and assess these elements of quality (Section 4).

2 Logical Consistency

KAINZ (1995) defines logical consistency as follows:

> "A spatial data set is said to be logically consistent when it complies with the structural characteristics of the selected data model and when it is compatible with the attribute constraints defined for the set."

The data model in our work is a planar subdivision, that is, an exhaustive coverage of the plane by areas that must not overlap. This representation is often used for land cover data in topographic databases. The generalization of such data sets is a well known problem (GALANDA 2003). Often additional requirements are defined, for example, the shapes of features need to be contiguous. Formally it means that for each two points in a contiguous area, there is a connecting path that is totally contained in the area. These structural requirements are independent of scale and we need to ensure their preservation during generalization. In contrast, requirements on attributes and geometries are often different for the input and output scale, thus we need to change the map. Tab. 1 compares the definitions of forest areas in three different countries. In each example, a minimal size is defined as criterion for selection, which naturally increases for smaller scales. The term "Guaranteed size", which is used in the Canadian specifications, unmistakably states that the defined thresholds must not be violated in any case. This is described accordingly in the other specifications. These strong claims are needed to reduce the influence of subjectivity in map generalization and to provide standardized map products.

Germany ATKIS (ADV 2003)		Canada National Topographic Database (NATURAL RESOURCES CANADA 1996)		Australia National Topographic Database (GEOSCIENCE AUSTRALIA 2006)	
Wald, Forst		Wooded Area		Forest or Shrub	
„Fläche, die mit Forstpflanzen (Waldbäume und Waldsträucher) bestockt ist.“		„An area of at least 35% covered by trees or shrubs having a minimum height of 2 m.“		„An area of land with woody vegetation greater than 10% foliage cover (includes trees and shrubs).“	
scale	selection criterion	scale	selection criterion	scale	selection criterion
1:25k	area ≥ 0,1 ha	1:50k	area ≥ 1ha AND width ≥ 50m	1:25k	area ≥ 0.25ha
1:50k	area ≥ 1 ha			1:100k	area ≥ 4ha
1:250k	area ≥ 40 ha	1:250k	area ≥ 25ha AND width ≥ 250m	1:250k	area ≥ 25ha
1:1000k	area ≥ 500 ha				

Tab. 1: Selection criteria for forest areas in three different national databases. The Canadian specifications use the term "Guaranteed size". In the Australian specifications this is called "Minimum size for inclusion".

Since areas below threshold in the target scale are not allowed, they need to be aggregated with others in order to keep the coverage exhaustive. As all other aims of generalization need to be subordinated, class changes need to be accepted, for example, if there is no neighbor of the same class. Formally, we define the area thresholds for different classes by $\theta: \Gamma \rightarrow R^+$, with Γ being the set of all classes. The term constraint fits well for the requirements given by database specifications. However, this must not be mixed up with constraints that allow for a gradual degree of satisfaction. Most researchers in the field of map generalization point out that constraints are often conflicting and compromises need to be found (WEIBEL & DUTTON 1998; HARRIE 1999). As constraints that ensure logical consistency do not allow any compromise, we distinguish hard constraints and soft constraints.

If the input data set is logically consistent, we normally can define simple generalization algorithms that produce logically consistent results. For instance, we can apply merge or collapse procedures to resolve size and proximity conflicts (Bader & Weibel 1997; HAUNERT & SESTER 2007). However, at this early stage of our discussion we should not commit too much to particular algorithms. Database specifications define the feasibility of solutions, but there remains much freedom in deciding for different options. Thus, we need to formalize additional aims of generalization. So far we have interpreted the selection criteria as a prohibition to keep small areas in the target scale. Additionally, we can understand the given thresholds as instruction to keep areas that have sufficient size, meaning that their classes must not be changed. Our method can simply be modified to respect this interpretation (Haunert & Wolff 2006). In this sense we also take the quality element completeness into account.

3 Semantic Accuracy

SALGÉ (1995) gives the following definition of semantic accuracy:

> "The purpose of Semantic Accuracy is to describe the semantic distance between geographical objects and the perceived reality."

Given a real value s (between 0 and 1) that measures the semantic similarity of two classes, for example, as it can be obtained with the methods proposed by YAOLIN et al. (2002), we can define the semantic distance simply as 1-s. Alternatively, a distance matrix can be defined by experts. Hierarchies defined in the data model and textual descriptions (as given in Tab. 1) need to be exploited for this task. Tab. 2 shows a distance matrix that was generated with this approach. Dividing each value with the maximum distance, we generally can normalize the distance measure, such that we obtain values between 0 and 1. High values correspond to semantically dissimilar classes like, for example, settlement and grassland. Both, farmland and grassland are classes of cultivated vegetation, thus they are semantically close. We would rather accept a change of a grassland area into farmland than into settlement. Thus, we define the semantic accuracy as the average distance between the area's classes before and after generalization. Let V be the set of all areas in the input map, $w : V \rightarrow R^+$ denote the sizes of areas, $\gamma: V \rightarrow \Gamma$ their original classes, $\gamma' : V \rightarrow \Gamma$ their new classes, and $d : \Gamma^2 \rightarrow R_0^+$ denote the semantic distance between classes, we globally measure the semantic distance by

$$\frac{\sum_{v \in V} w(v) \cdot d(\gamma(v), \gamma'(v))}{\sum_{v \in V} w(v)}.$$

In the same way, we can measure the semantic distance for a single area in the input data set, for a single area in the output data set, or for all areas of a certain class.

original class \ new class	Settlement	Farmland	Grassland	Forest
Settlement	0	1	1	1
Farmland	1	0	0.2	0.3
Grassland	1	0.2	0	0.3
Forest	1	0.3	0.3	0

Tab. 2: Semantic distance matrix. Colors and shades are used in Fig. 1 and Fig. 2.

Though the matrix that is shown in Tab. 2 is symmetric, the defined measure is not restricted to symmetric distance functions. For example, we can define $d(\gamma_1, \gamma_2) = 0.1$ and, $d(\gamma_2, \gamma_1) = 1$ meaning that a class change from γ_1 to γ_2 is more accepted than vice versa. This model is useful, as important classes like, for example, water often should not be lost. In order to take semantic changes into account, we should not only focus on their classes but also consider their shapes. An area might belong to the class forest, but does it also have the shape of a forest? This question is indeed difficult to answer. We therefore penalize shapes that have a generally untypical characteristic, that is in our case of vegetation and settlement areas, shapes that are not geometrically compact. Different measures of compactness have been discussed in an earlier work (HAUNERT 2007a). We ignore this additional criterion in the sequence of this article.

We use the result of the optimization approach as a benchmark for semantic and logical quality, as it is able to satisfy our general goal of preserving semantic accuracy in terms of semantic distance.

4 Optimization approach

The aggregation problem is often approached by iterative merging of pairs of areas, which is done until all areas satisfy the area thresholds for the target scale. It is attempted to assure quality by defining appropriate criteria for the selection of areas that are merged in each iteration (CHENG & LI 2006). Our approach is fundamentally different: we are not interested in a sequence of pairwise merge operations, but only focus on the result, thus approach the problem by optimization. For a review of the iterative algorithms and a comparison with our method we refer to HAUNERT (2007b).

4.1 *Problem formulation*

In terms of optimization, each logically consistent map is a feasible solution. The optimal feasible solution is the one that minimizes the global semantic distance measure from Section 3. We refer to this measure as cost function. The problem is to partition the set V into mutually disjoint subsets $V_1 \cup V_2 \cup ... \cup V_k = V$, where k is an unknown integer. Each of these subsets defines a composite area for the target scale. Thus, for each $i = 1 ... k$, we define the following hard constraints:

- there is a single class $\gamma_i' \in \Gamma$, such that each area $v \in V_i$ receives class γ_i', that is, $\gamma'(v) = \gamma_i'$.
- the composite area has sufficient size, that is, $\Sigma_{v \in Vi}\, w(v) \geq \theta(\gamma_i')$.
- the composite area is contiguous.
- there is an area $v \in V_i$ of unchanged class, i.e., $\gamma'(v) = \gamma(v)$. This is referred to as centre.

The last requirement simply avoids that classes appear in the generalized map, which have not been present at all. Generally, we do not assume that the set of centers is given in advance.

4.2 *Approach by mixed-integer programming*

Normally, combinatorial optimization problems in map generalization are approached by metaheuristics such as hill-climbing or simulated annealing (WARE & JONES 1998). Several theoretical achievements have been made, proving asymptotical convergence of simulated annealing under certain conditions (HAJEK 1988). However, in practice these do not allow to solve a problem with proof of optimality. An exact approach to constrained, combinatorial optimization problems is mixed-integer programming (PAPADIMITRIOU & STEIGLITZ 1998). Generally, algorithms for the solution of mixed-integer programs (MIPs) have an exponential time performance. It is unlikely that we can find a polynomial time algorithm, as the aggregation problem is NP-hard. This fact was proven in an earlier publication (HAUNERT & WOLFF 2006). We also presented and tested different MIP formulations. Due to the high complexity we were only able to optimally solve small instances (up to 50 areas) with our exact MIP, but we greatly improved the performance with three heuristics:

1. A strong definition of contiguity according to ZOLTNERS & ZINHA (1983) is applied, which excludes certain non-compact composite areas.
2. Large areas are fixed as centers, small areas are excluded from the set of potential centers.
3. Areas with a large distance in between are not merged in the same composite area.

A further heuristic has been developed that allows to decompose a dataset of arbitrary size into manageable pieces (HAUNERT 2007b). The basic idea of this method is to introduce intermediate scales. This is similar to the common iterative approach, but the generalization steps are much bigger.

4.3 *Quality assessment*

Without heuristics our optimization approach is too slow for cartographic production. However, as it yields the exact optimum for small problem instances, it can be used to test heuristic methods. For example, applying heuristics 1-3 we usually obtain results not worse than +10% from optimum. Such objective statements about the performance of generalization procedures are very rare in the literature. Often results are only visually assessed by test persons, but this approach is questionable, if the spatial data set is not only to be used for visualization, but also for statistics or other analyses. On the other hand, visualization is still the most important method to assess the quality of spatial data sets. How do +10% affect the understanding of the map content? We can only answer this question when visualizing the results.

Fig. 1 shows a sample from the German ATKIS data set at scale 1:50.000 (DLM50) with two results satisfying specifications for scale 1:250.000 (DLM250): the optimal solution and a solution, which was obtained with our heuristic approach. Some classes were changed, in order to end up with contiguous regions of sufficient size. In both cases, this was done in an apparently intelligent way: In order not to loose the settlement area in the lower left of the sample, a small forest area was sacrificed; this results in a connecting bridge to the large settlement. The first solution has an average semantic distance of 0.0519 from the input map. For the second solution this cost is 0.0564, which is approx. 9% higher. In fact, we observe some differences between both solutions, for example, on the rightmost settlement in the input: in the optimal result it changes to forest, but in the second solution it changes to farmland. If we check the distance matrix (Tab. 2), which was applied here, we see that the same distance of 1 unit is defined for both changes. It turns out that it is relatively difficult to visually detect those areas, which were not optimally aggregated by the heuristic method.

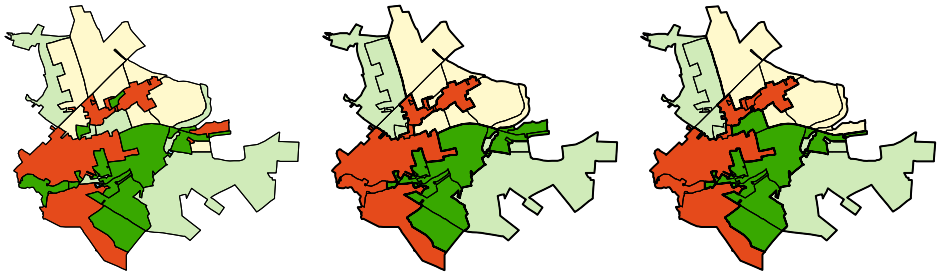

Fig. 1: A sample from the ATKIS DLM 50 (left), an optimal result satisfying specifications for ATKIS DLM 250 (center) and a result of +9% higher costs that was obtained with heuristics 1-3 (right). Boundaries of regions are bold, colors correspond to classes as shown in Tab. 2.

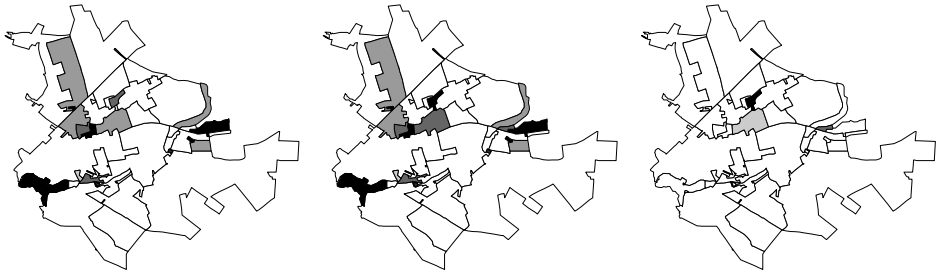

Fig. 2: Semantic distances for classes of single areas before and after generalization. Left: Optimum. Center: Result obtained with heuristics. Right: Difference of both (center-left). Grey shades correspond to values as given in Tab. 2.

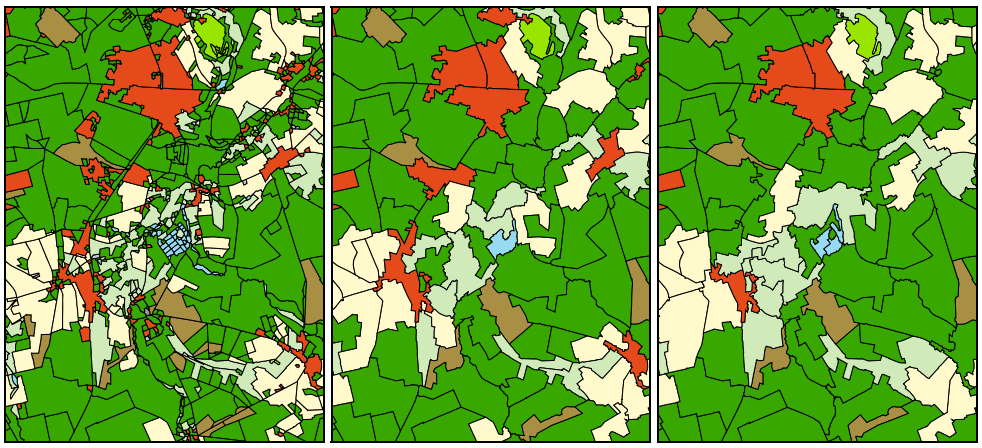

Fig. 3: A sample from the ATKIS DLM 50 (left), a result of the heuristic developed by Haunert (2007b), and a result of a simple iterative merging procedure (right). Both results satisfy the specification for the ATKIS DLM 250.

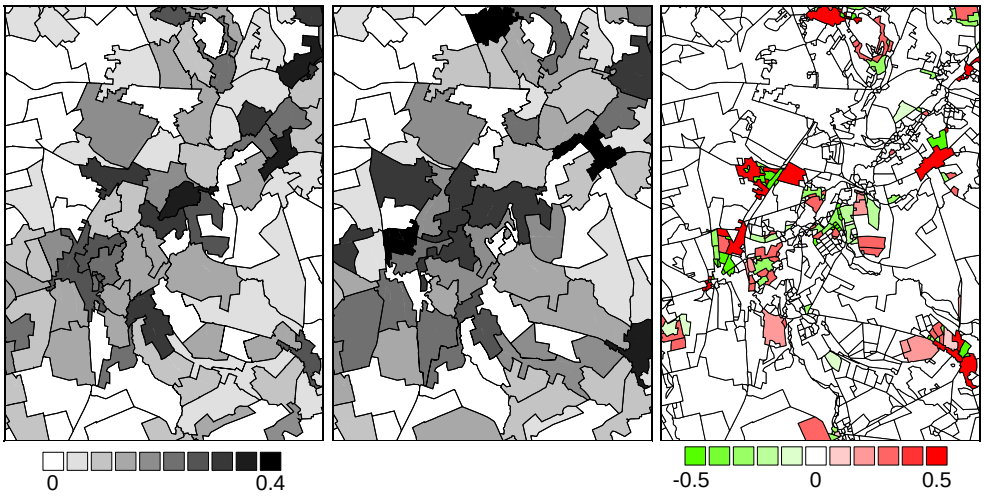

Fig. 4: Semantic distances for regions in Fig. 3. Left: Result from Fig. 3 (center). Center: Result from Fig. 3 (right). Right: Difference of both for areas of input scale (center-left).

In order to identify the reason for the difference in costs, we need to visualize the differences of semantic distances for single areas. Fig. 2 displays distances of class changes as grey shades of areas, dark grey corresponds to expensive changes. In Fig. 2 (right) we see the difference of both solutions: Only for four small areas the reclassification done by the heuristic approach is suboptimal.

Normally, we do not have the optimal solution to compute the differences as shown in Fig. 2 (right). There would not be any reason to apply heuristics, if we generally could exactly solve the problem. However, similar to comparisons with optimal solutions for small samples, we can visually compare results of different heuristics. Fig. 3 shows a sample that was processed with our heuristic approach based on intermediate scales (HAUNERT 2007b) and with the common iterative merging procedure,

which, for example, is explaind by CHENG & LI (2006). Both objectives were considered: class similarity and compactness. In each iteration of the merging procedure, the smallest area was assigned to one of its neighbours. Each time, this neighbour was selected to minimize the cost function; of course, this does not lead to the global optimum. Considering compactness, both procedures performed similarly. Using our optimization method, we only obtained an improvement by 2% of costs for non-compact shapes. However, we obtained 20% less costs for class changes. In Fig. 3 we observe that several settlement areas are lost with the simple iterative procedure. This is due to the fact that the algorithm does not foresee consequences of merge actions for further processing steps. Thus it is not able to sacrifice small areas in order to safe bigger ones.

In Fig. 2 we investigated the quality measure for the original areas; these are minimal mapping units in the aggregation problem. We can do the same analysis on a less detailed level, that is, for each area of the target scale; this is shown in Fig. 4 (left) and (center). A comparison of both results is only possible on the highest level of detail, as the regions in both solutions are different. Fig. 4 (right) reveals those areas whose semantics were kept more similar with the optimization approach (red) and those that were kept more similar with the iterative approach (green). We clearly observe the dominance of red areas. Also we can see that red and green areas are often in vicinity. This confirms our assumption that, in contrast to the simple iterative procedure, the optimization approach can sacrifice unimportant areas, in order to safe more important ones.

5 Conclusion

The definition of semantic similarity or distance measures is considered as the key to quality assessment in map generalization. We have shown that, with given semantic distance values for classes, we can optimally solve the area aggregation problem in map generalization for small instances. With such theoretically proven optima, we have found the “absolute zero” for the degree of badness. This allows to make objective, quantitative statements about the performance of heuristic methods. Additionally, we can compare the performance of heuristics relative to each other. In both cases we have seen that shortcomings of heuristic methods can be detected by visualization. In particular, we have seen that our heuristic based on intermediate scales results in 20% less cost for class change than the simple iterative method. We observed that its relatively good performance is due to its capabilty of sacrificing smaller areas, such that bigger ones can be safed. Future research should focus on better semantic distance measure, not only considering the class memberships of objects. Semantics can also be carried by shapes and patterns of objects. This becomes relevant for other generalization operators, such as typification. In fact, pattern recognition techniques are often applied in map generalization. However, metrics are missing that measure the semantic distance between patterns.

6 Acknowledgements

The authors acknowledge support from grant SE 645/2-1 of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

References

ADV, 2003: ATKIS-Objektartenkatalog. – http://www.atkis.de (accessed 9 November 2007).

AHLQUIST, O., 2005: Using semantic similarity metrics to uncover category and land cover change. – GeoSpatial Semantics: Proceedings of first international conference, GeoS 2005, Mexico City, Mexico. – Volume 3799 of Lecture Notes in Computer Science. – Springer, Berlin, Germany.

BADER, M. & WEIBEL, R., 1997: Detecting and resolving size and proximity conflicts in the generalization of polygonal maps. – Proceedings of 18th International Cartographic Conference, Stockholm, Sweden, pp. 1525–1532.

BARD, S., 2004: Quality Assessment of Cartographic Generalization. – Transactions in GIS, **8**(1):63–81.

CHENG, T. & LI, Z., 2006: Toward Quantitative Measures for the Semantic Quality of Polygon Generalization. – Cartographica, **41**(2):487–499.

FRANK, R. & ESTER, M., 2006: A Quantitative Similarity Measure for Maps. – Riedl, A., Kainz, W. & Elmes, G. A. (Editors): Progress in Spatial Data Handling. – Springer, Berlin, Germany.

GALANDA, M., 2003: Automated Polygon Generalization in a Multi Agent System. PhD thesis, Department of Geography, University of Zurich, Switzerland, 2003.

GEOSCIENCE AUSTRALIA, 2006: Topographic data and map specifications. http://www.ga.gov.au/mapspecs/250k100k/ (accessed 9 November 2007).

HAJEK, B., 1988: Cooling schedules for optimal annealing. – Mathematics of operations research, 13(2):311–329.

HARRIE, L. E., 1999: The Constraint Method for Solving Spatial Conflicts in Cartographic Generalization. – Cartography and Geographic Information Science, **26**(1):55–69.

HAUNERT, J.-H. & WOLFF, A., 2006: Generalization of land cover maps by mixed integer programming. – GIS '06: Proceedings of the 14th annual ACM international symposium on advances in geographic information systems, Arlington, Virginia, USA, pp. 75–82.

HAUNERT, J.-H., 2007a: A formal model and mixed-integer program for area aggregation in map generalization. – PIA07: Photogrammetric Image Analysis, Munich, Germany. – IAPRS Volume XXXVI, part 3/W49A, pp. 161-166.

HAUNERT, J.-H., 2007b: Efficient area aggregation by combination of different techniques. – Proceedings of 10th ICA Workshop on Generalisation and Multiple Representation, Moscow, Russia.

HAUNERT, J.-H. & SESTER, M., 2007: Area collapse and road centerlines based on straight skeletons. – GeoInformatica, doi:10.1007/s10707-007-0028-x (published online 17 August 2007).

KAINZ, W., 1995: Logical consistency. – GUPTILL, S. T. & MORRISON (editors): Elements of spatial data quality, chapter 6, pp. 109–137 – Elsevier Science, Oxford, UK.

MORRISON, J. L., 1995: Spatial data quality. – GUPTILL, S. T. & MORRISON (editors): Elements of spatial data quality, chapter 1, pp. 1–12 – Elsevier Science, Oxford, UK.

NATURAL RESOURCES CANADA, 1996: National topographic data base: data dictionary. – http://ftp2.cits.rncan.gc.ca/pub/bndt/doc/dictntd3_en.pdf (accessed 9 November 2007).

PAPADIMITRIOU, C. H. & STEIGLITZ K., 1998: Combinatorial Optimization. – Dover Publications, Inc., Mineola, NY, USA.

RODRÍGUEZ, M. A. & EGENHOFER, M. J., 2004: Comparing geospatial entity classes: an asymmetric and context dependent similarity measure. – International Journal of Geographical Information Science, **18**(3): 229–256.

SALGÉ, F., 1995: Semantic accuracy. – GUPTILL, S. T. & MORRISON (editors): Elements of spatial data quality, chapter 7, pp. 139–151 – Elsevier Science, Oxford, UK.

WARE, J. M. & JONES, C. B., 1998: Conflict reduction in map generalization using iterative improvement. – GeoInformatica, **2**(4):383–407.

WEIBEL R. & DUTTON. G., 1998: Constraint-based automated map generalization. – Proceedings of the 8th International Symposium on Spatial Data Handling, Vancouver, Canada, pp. 214–224.

YAOLIN, L., MOLENAAR, M. & KRAAK, M.-J., 2002: Semantic similarity evaluation model in categorical database generalization. – Proceedings of ISPRS Commission IV Symposium on Geospatial Theory, Processing and Applications, Ottawa, Canada. – IAPRS, Volume 34, part 4, pp. 279–285.

ZOLTERNS, A. A. & SINHA, P. , 1983: Sales territory alignment: A review and model. – Management Science, **29**(11):1237–1256.

Anschriften der Autoren:

Dipl.-Ing. JAN-HENRIK HAUNERT, Prof. Dr.-Ing. habil. MONIKA SESTER, Leibniz Universität Hannover, Institut für Kartographie und Geoinformatik, Appelstraße 9a, D-30167 Hannover, Tel.: +49-511-762-3588, Fax: +49-511-762-2780, e-mail: {Jan.Haunert, Monika.Sester}@ikg.uni-hannover.de