The Cartographic Journal *Vol. 46 No. 3 pp. 214–227 August 2009*
© *The* British Cartographic Society 2009

REFEREED PAPER

Specifying Map Requirements for Automated Generalization of Topographic Data

Jantien Stoter[1], *John van Smaalen*[2], *Nico Bakker*[3] *and Paul Hardy*[4]

[1]Delft University of Technology, OTB, GIS technology, The Netherlands (research carried out at ITC, Enschede). [2]John van Smaalen, ESRI Nederland, The Netherlands. [3]Nico Bakker, Kadaster, The Netherlands. [4]Paul Hardy, ESRI Europe, Cambridge, United Kingdom
Email: j.e.stoter@tudelft.nl

This study aims at acquiring knowledge on map requirements for automated generalization. First, interactively generalized map series were visually analysed together with the specifications that cartographers use to generalize the maps. Second, these map specifications were experimentally implemented on real data in automated processes and compared to an interactively generalized map to see if the results are according to the specifications; to see if the specifications are complete and well-formalized; and to identify situations that were not addressed in the specifications. If required, the specifications were enriched and re-implemented also adding extra information from other sources. The experiments revealed the 'deep' knowledge which cartographers add to the interactive process. Based on this revealed knowledge, recommendations are formulated to specify map requirements for automated generalization of topographic data.

Keywords: automated map generalization, map requirements, knowledge acquisition

INTRODUCTION

Automated generalization of topographic data would be a significant step towards highly efficient and flexible map production at National Mapping Agencies (NMAs).

An absolute necessity for successful generalization is to have map specifications that can be unambiguously understood by the system that carries out the generalization, or at least by the person who triggers the process. In this respect, the automated process differs fundamentally from the interactive generalization process, where cartographers can add their interpretation during the whole process.

At NMAs, knowledge is available on what a generalized map should look like. This knowledge is captured in written map specifications used by cartographers who interactively generalize maps as well as in minds of cartographers and even in software code. However, NMA requirements for automated generalization are not directly available, although some automation in generalization has been introduced in many NMAs (Stoter, 2005). The reason for these missing requirements are that: first, currently no formalism has proved to be adequate for fully capturing the specifications of a map, second, not all requirements are easily to be formalized and third, much knowledge on generalization requirements and processes still needs to be revealed.

This paper will specifically address the last two issues. It aims at specifying map requirements for automated generalization using a combined approach of reverse engineering, i.e. learning from existing maps and map specifications, and machine-learning techniques, i.e. learning from experimentally implemented specifications. These techniques have been successfully applied to acquire knowledge on map requirements for automated generalization (Buttenfield *et al.*, 1991; Leitner and Buttenfield, 1995; Plazanet *et al.*, 1998; Mustière, 2005). Our contributions to these previous studies are experiments with real data and existing specifications carried out by a group of cartographic and software experts. Based on the results of the experiments, map requirements are specified for automated generalization of topographic data.

Our study starts with acknowledging Muller and Mouwes (1990) who state that generalization knowledge can be divided into two distinct layers. Superficial knowledge is written down in map specifications meant for interactive generalization, while deep knowledge is added to the process by cartographers when superficial knowledge does not suffice. Deep knowledge is much more complex to automate.

To expose this deep knowledge, our study contained two steps. First, we visually examined existing maps and studied

DOI: 10.1179/174327709X446637

their specifications. Second, we implemented available specifications using real data and compared the outputs to interactively generalized maps. The results identified whether the specifications were apparent and complete for automated processes. If required, the specifications were enriched with the obtained deep knowledge to make them suited for automated generalization and re-implemented with extra information from other sources.

This study is part of a research project funded by the Dutch research programme 'Space for Geo-information'. The aim of this overall project is to provide the most applicable base maps for portraying spatial thematic information in a Web portal. The base map should be generated from a topographic database by means of automated generalization (Foerster *et al.*, 2008). To specify requirements for this automated generalization, the subproject as presented in this paper studied how currently available specifications and data used for interactive generalization can serve automated processes.

Another motivation for this research lies in the EuroSDR project that evaluates automated generalization implemented in commercial software (Stoter *et al.*, 2008a). For the EuroSDR generalization tests, four NMAs specified their map requirements based on available specifications and cartographers' knowledge. The case study of this paper is one of these cases. Therefore, this research also provides insights into how map requirements of one NMA as specified for the EuroSDR project are suitable (i.e. complete, sufficiently formalized, etc.) for automated generalization of topographic data.

The section on 'METHOD' introduces the method of this research. Results are presented in the section on 'RESULTS'. Based on these results, the section on 'SPECIFYING MAP REQUIREMENTS FOR AUTOMATED GENERALIZATION' defines what and how deep cartographers' knowledge should be added to existing specifications to make them suited for automated generalization. This paper ends with conclusions in the section on 'CONCLUSIONS'.

METHOD

This section further outlines our research method (the section on 'Outlining the method'). To better understand the context of the case study, which is generalization applied for the Netherlands' Kadaster, the section on 'Current and past generalization processes of Kadaster' describes the current and past generalization processes of Kadaster. Finally, the section on 'Integrating model and cartographic generalization' explains why we consider model and cartographic generalization together in our experiments.

Outlining the method

The research method consists of two main steps. First, the current maps at all scales were visually analysed to identify generic and specific characteristics of the maps as supportive knowledge for automated generalization. Also the corresponding map specifications were studied. Results are presented in the section on 'Visual analysis of existing maps'.

Second, we conducted experiments to automatically generalize a 1 : 50k map (called TOP50 in the remainder of this paper) from TOP10NL data according to requirements of the Netherlands' Kadaster. TOP10NL is the object oriented version of TOP10vector, completed in 2007. Object oriented equivalents at the smaller scales are being created. We started our experiments by implementing current map specifications that cartographers use to generalize TOP50 from TOP10NL (Kadaster, 2005) with ESRI tools. These first results were compared to an interactively generalized map to see whether the specifications resulted in expected outputs when applying these without adding any human knowledge. Based on the intermediate results, two actions were taken to improve the automated process. First, the specifications were refined and reformulated to make them better suited for automation. Second, apart from TOP10NL data, other information was added to the process. We rerun the automated process with the enriched specifications and additional information.

The experiments were applied on three topographic classes: buildings, roads and land use, since automating generalization of these features will offer significant efficiency gains for future production lines. The reasons are several. First, they are the classes with most features and the most significant for users of the map. In addition, most critical and challenging generalization actions are required for these classes and their generalization results are most dominant for the final result. Finally, these classes are highly related in the generalization process, requiring a holistic approach to the generalization of these feature classes. The results of the experiments are reported in the section on 'Experiments of TOP10NL to TOP50 generalization'.

From the results of the experiments, we analysed which knowledge should be added to specifications and how to make them suited for automated processes. Our proposals to specify map requirements for automated processes are described in a later section.

Several aspects outline the scope of this research. First, we assume that maps generalized by cartographers according to specifications result in satisfying (but not always consistent) maps. Consequently, this research will not assess the quality of specifications for interactive generalization. Instead, the aim is to get insight into how map requirements for automated generalization can be specified starting from currently available specifications and data. Another assumption is that automated processes should result in maps that are comparable to the currently available interactively generalized maps. This assumption is based on NMA surveys that showed a continuous demand for traditional, topographic maps, which implies that NMAs still have interest to produce traditional map series. If it appears that automated processes are easier to implement when allowing minor diverging from traditional maps, the requirements for automated generalization should be reconsidered. Third, the NMA who provided the test case has only applied interactive generalization until now and therefore, no knowledge was available on the suitability of current data and specifications for automated processes.

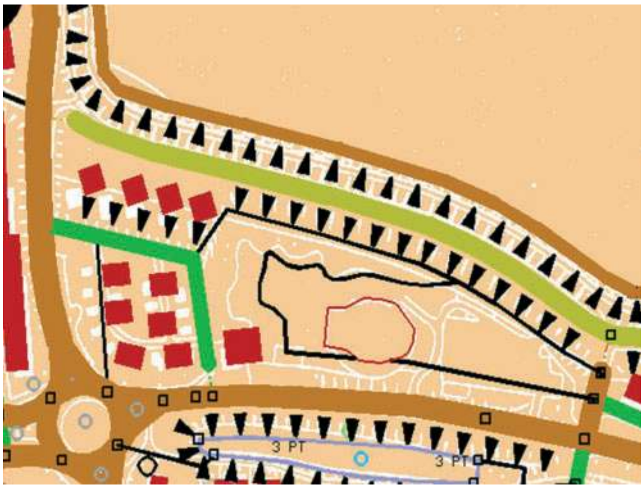


Figure 1. Interactive generalization as applied at the Kadaster: large scale map is presented as background and the cartographer generalizes the smaller scale map interactively

Consequently, this research shows how knowledge needed for the automated generalization can be made explicit for an NMA that starts automating its processes. Finally, this research does not assess the quality of the applied generalization system, which is ESRI's ArcGIS plus research prototype extensions. Instead, it will only assess the output of automated generalization with respect to the initially formulated requirements. It should be noted that the Optimizer technology used in the automated buildings generalization involved a research prototype (Monnot *et al.*, 2007a), and as such, does not indicate any commitment by ESRI to provide particular functionality in future product releases.

Current and past generalization processes of Kadaster

The Kadaster (and its predecessor organizations) has produced maps at scales from 1 : 10k to 1 : 500k since 1955. Before 1950s, the 1 : 25k map was the basis for the smaller scales (at that time, 1 : 50k, 1 : 100k and 1 : 250k). In the 1950s, the 1 : 10k map was introduced, which is currently still considered as the basis to the smaller scales.

Up to the 1990s, the entire generalization process was manual. The cartographer drew the smaller scale maps using a background of larger scale maps represented at a reduced scale (between the original and target scales). The cartographer drew the maps according to written specifications that were a result of both national and international requirements, partly originating from military map products.

From the 1990s, digital, interactive generalization was introduced following the same working process. This is still the process of today (Figure 1). Smaller scale maps are 'drawn' from the larger scale map according to specifications but now using computer tools. Apart from the larger scale map, aerial photos are used as source for cartographers' interpretation of the landscape.

The generalization workflow at the Kadaster did not change much over time. The smaller scales are generalized from the next larger scale map. The exception is the 1 : 250k dataset which was originally derived from the 1 : 50k map, until the 1 : 100k came into production (the mid-1980s).

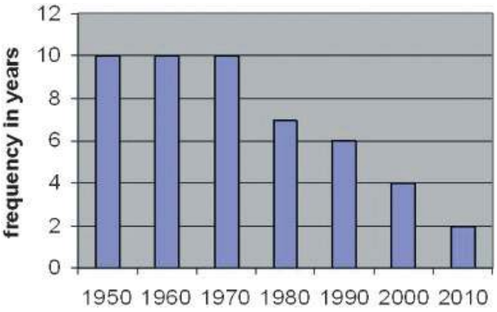


Figure 2. Development of revision cycle at Kadaster since 1950

For our experiments, it is important to understand the cartographers' use of values as indicated in the specifications. Because the interactive generalization process consists of 'drawing' rather than adjusting larger scale data, values describing size, granularity, minimal distance, etc., are not used as hard values, for instance, as criterion to eliminate certain features. Instead, they indicate a minimum threshold for optimally visualizing features at a certain scale. The cartographer may decide incidentally to display features with slightly smaller dimensions if necessary.

The revision cycles changed largely over time. The 1 : 10k and 1 : 50k maps always followed the same revision cycle which changed radically the last 50 years. Starting with a revision interval of 10 years in the 1960s that was regularly reduced, a two-year revision cycle was introduced from 2008 (Figure 2). The 1 : 100k and 1 : 250k maps had other revision cycles according to the military requirements. Often they were newer than the underlying larger scale maps and therefore, updates required new field data. From 2010, all scales should be produced in one harmonized revision cycle of two years.

The degree of generalization across the scales was never harmonized (Figure 3). Consequently, differences in the level of detail are partly caused by the original purposes of the different scales. The most detailed scale is scale 1 : 10k (formerly TOP10vector, nowadays TOP10NL). It is used for GIS analyses, orientation, as background for thematic information and displayed as paper and digital raster map at scale 1 : 25k. From its start, the 1 : 50k map must give as 'much' detail as possible according to international military standards, because it was the main map for military use. On the other hand, the 1 : 100k served as an overview map for militaries and less detail was needed and even desired. Consequently, the 1 : 25k and 1 : 50k maps are very detailed, while the 1 : 100k is very open and much more generalized than expected (see also the section on 'Visual analysis of existing maps').

The legends of the 1 : 25k, 1 : 50k and 1 : 100k maps are rather well harmonized. However, the 1 : 250k and 1 : 500k maps are visualized with other colours and symbols (Figure 3). The underlying vector databases at all scales (i.e. TOPxxvector databases) do have the same structure and coding system.

Integrating model and cartographic generalization

In our study, we specified and implemented requirements that do not distinguish between model generalization,

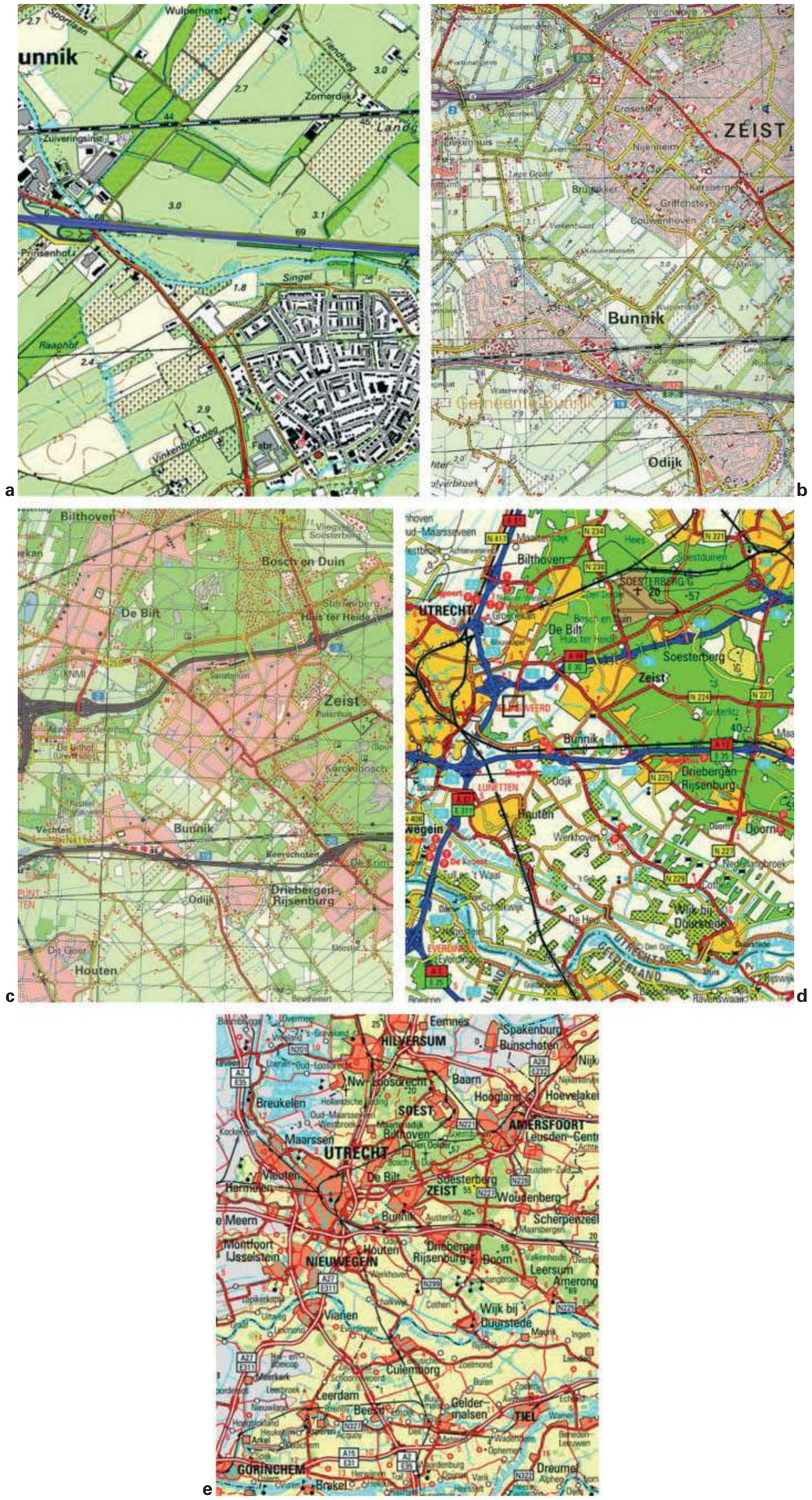


Figure 3. Topographic maps at different scales, as produced by Kadaster: (a) 1 : 25k, (b) 1 : 50k, (c) 1 : 100k; (d) 1 : 250k, (e) 1 : 500k

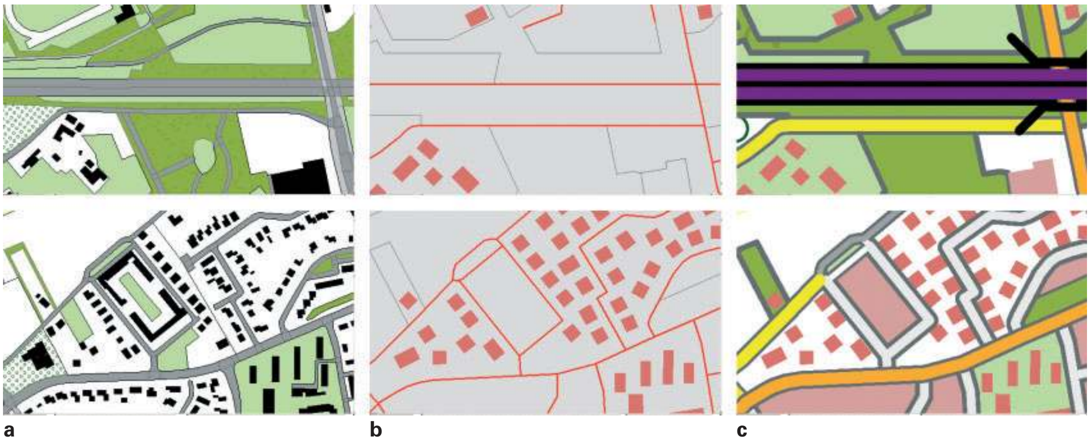


Figure 4. Displacement of features in TOP50vector triggered by graphical conflicts of symbolized features in 1 : 50k map: (a) TOP10NL; (b) TOP50vector data; (c) TOP50 map

aiming at lower resolution geographic databases, and cartographic generalization, aiming at readable maps at smaller scales. Brassel and Weibel (1988) and Gruenreich (1992) introduced this separation and identify a digital landscape model (DLM) that contains the basic primary model of reality. They explain 'model generalization' as the derivation of primary models at lower semantic and geometric resolution from the basic DLM. The digital cartographic model (DCM) is the result of applying cartographic generalization, i.e. reduction, enlargement and modification of graphic symbols to the DLMs.

Several reasons justify our approach of integrating requirements for model and cartographic generalization. First, the Kadaster does not distinguish between a database and a cartographic representation in its current production line, which is the starting point of our study. Consequently, the TOPxxvector datasets integrate aspects of the DLM and the DCM: the geometries in the vector datasets take into account the way they appear on the map. For example, a motorway at 1 : 50k is portrayed with a line-symbol of width 1.5 mm, which is 75 m in reality. To avoid overlap of the motorway symbol with other features such as buildings, features are displaced and simplified in current TOPxxvector products. Creating the map is a simple operation, which adds symbols to the geometries in the vector datasets (Figure 4).

Second, generalization leads inevitably to accuracy loss, whether this is for the database or the map. The implicit inaccuracies of current vector products are no problem for users that perform GIS analyses at small scales. If more accurate data are needed, one can use TOP10NL data that does not contain inaccuracies because of symbolization.

The last reason to not distinguish between model and cartographic generalization is that often it is not easy to determine whether an operation is purely cartographic or that it belongs to the domain of model generalization. Since we wanted to include both model and cartographic aspects of generalization in the project and we did not want to lose time by addressing the highly complex issue of model versus cartographic generalization, it was decided not to make the distinction.

RESULTS

This section presents the results of our study. The section on 'Visual analysis of existing maps' presents results from the visual analysis of the interactively generalized maps. The section on 'Experiments of TOP10NL to TOP50 generalization' presents the results of implementing and enriching map specifications for TOP50 generalization.

Visual analysis of existing maps

When overviewing the current maps, the first observation of interest for the automated generalization process is that the content of TOP50 is much more similar to a 1 : 10k map than to a 1 : 100k map. This is not in line with the scale reduction factors: TOP50 is five times smaller than a 1 : 10k map, whereas a 1 : 100k map is only twice smaller than TOP50. Also when we compare the sizes of the datasets (Table 1), we see a similar trend in data reduction across the scales. As explained in the section on 'Outlining the method', this is due to differences in original purpose of the different scales which were never harmonized.

A second observation from the specifications relevant for automated generalization is the strict hierarchy of features which may be displaced:

- important dikes and dams (least likely to be displaced);
- railways;
- highways;
- main roads;
- important canals and rivers;
- other roads;

Table 1. Size of topographical datasets at different scales covering the complete area of The Netherlands (size of dataset for 1 : 100k map is not available)

Scale	Size of datasets (MB)
1 : 10k	3000
1 : 50k	1100
1 : 250k	50
1 : 500k	15

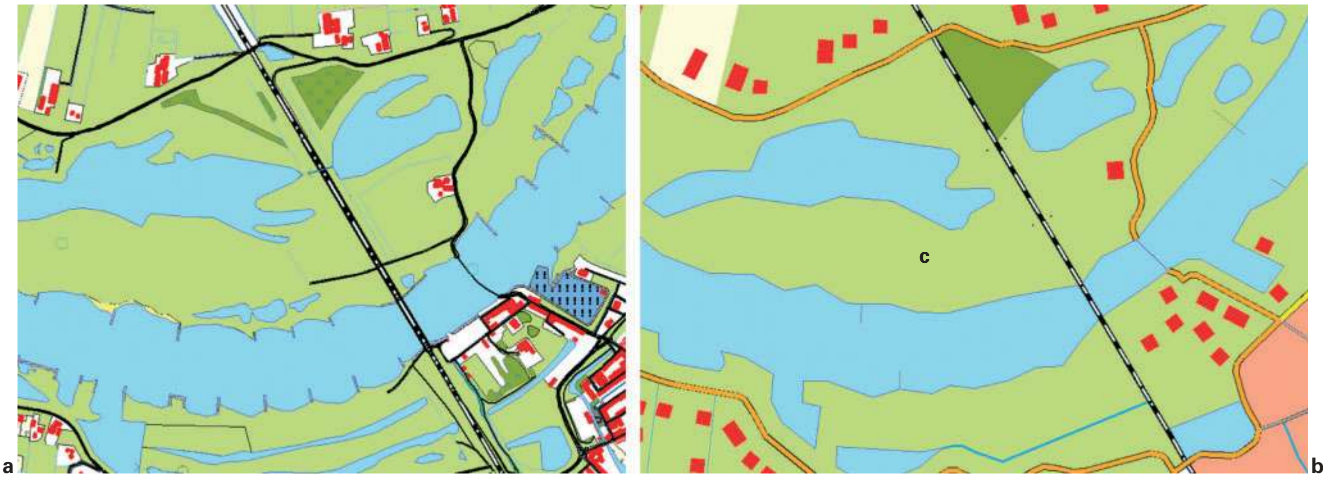


Figure 5. Water in TOP10NL and in TOP50 map: (a) TOP10NL – original water features, (b) current TOP50 map – water is considerably simplified, but not consistent as to whether aggregated, eliminated or collapsed

- smaller waterways;
- land use (most likely to be displaced).

A final relevant observation from studying the map series is related to generalization of specific classes:

- the road classifications at 1 : 10k, 1 : 50k and 1 : 100k are very similar, although the road classes are merged in 1 : 100k to fewer classes. The 1 : 250k has a different road classification, based on other (military) specifications;
- roads that are kept on smaller scales are the most important roads. Since this importance information is not encoded in the data, cartographers interpreted which roads are important;
- maps at scale 1 : 250k and smaller do not show individual buildings;
- generalization of land use areas (specifically woodland and water areas) is not consistent, i.e. when is aggregation applied; when deletion, when simplification? (Figure 5);
- enlargement of woodlands and water areas in the case of minimal width can cause displacement of roads, which is not in line with the generalization hierarchy as listed above.

Experiments of TOP10NL to TOP50 generalization

In this section, the results of our experiments, in which map specifications were studied, enriched, implemented and compared to interactively generalized maps, are presented for buildings, roads and land use separately.

Generalizing buildings

In TOP10NL, road polygons, water polygons and land use polygons form a full partition of space. Consequently, TOP10NL buildings (including aggregated buildings) are a separate layer on top of these features. Land use polygons that are located below buildings are predominantly classified as 'other' ('*overig*' in Dutch). TOP50 contains much more built-up area than TOP10NL, which is one of the possible values of land use. These areas replace TOP10NL buildings.

The following specifications describe how cartographers convert single buildings to built-up area:

If density of buildings in urban areas is sufficiently high, buildings can be replaced by built-up area, with the exception of detached houses and buildings in industrial areas. In rural areas built-up areas are never created. Important buildings are never aggregated to built-up area.

This verbal rule needs human interpretation, as the specification contains no hard value for 'sufficiently high building density'. For implementation, we refined the specification to say that land use polygons of type 'other' that are covered for at least 10% by buildings, should be portrayed as 'built-up' area, with exception of detached houses, important buildings, buildings in industrial areas and buildings in rural areas. The buildings on top of these reclassified polygons should be eliminated.

This (refined) specification seems straightforward to implement. However, information on urban/rural areas and detached houses is not encoded in the data. In addition, industrial areas are represented with point symbols on the maps and therefore, the extent of these areas is not available from the data. Cartographers interpret this extra information from aerial photographs.

To add this missing information, we enriched the process with information on industrial polygons and on building types ('detached house') from other sources. In addition, the urban areas were computed based on global building densities.

After applying the reformulated specifications to the enriched data and comparing the result with the interactively generalized map, another issue surfaced. The implementation analysed each complete land use polygon of type '*overig*' on its building coverage. It did not take into account the possible uneven distribution of buildings over a polygon, as cartographers can do. In those cases, cartographers only create built-up area in that part of the polygon where the buildings are located (Figure 6, 'A' locations). Although this approach better reflects reality, the map specifications do not say what area should be analysed on building density. In addition, the solution of only generating built-up area at locations of buildings is harder to automate, since new land use boundaries need to

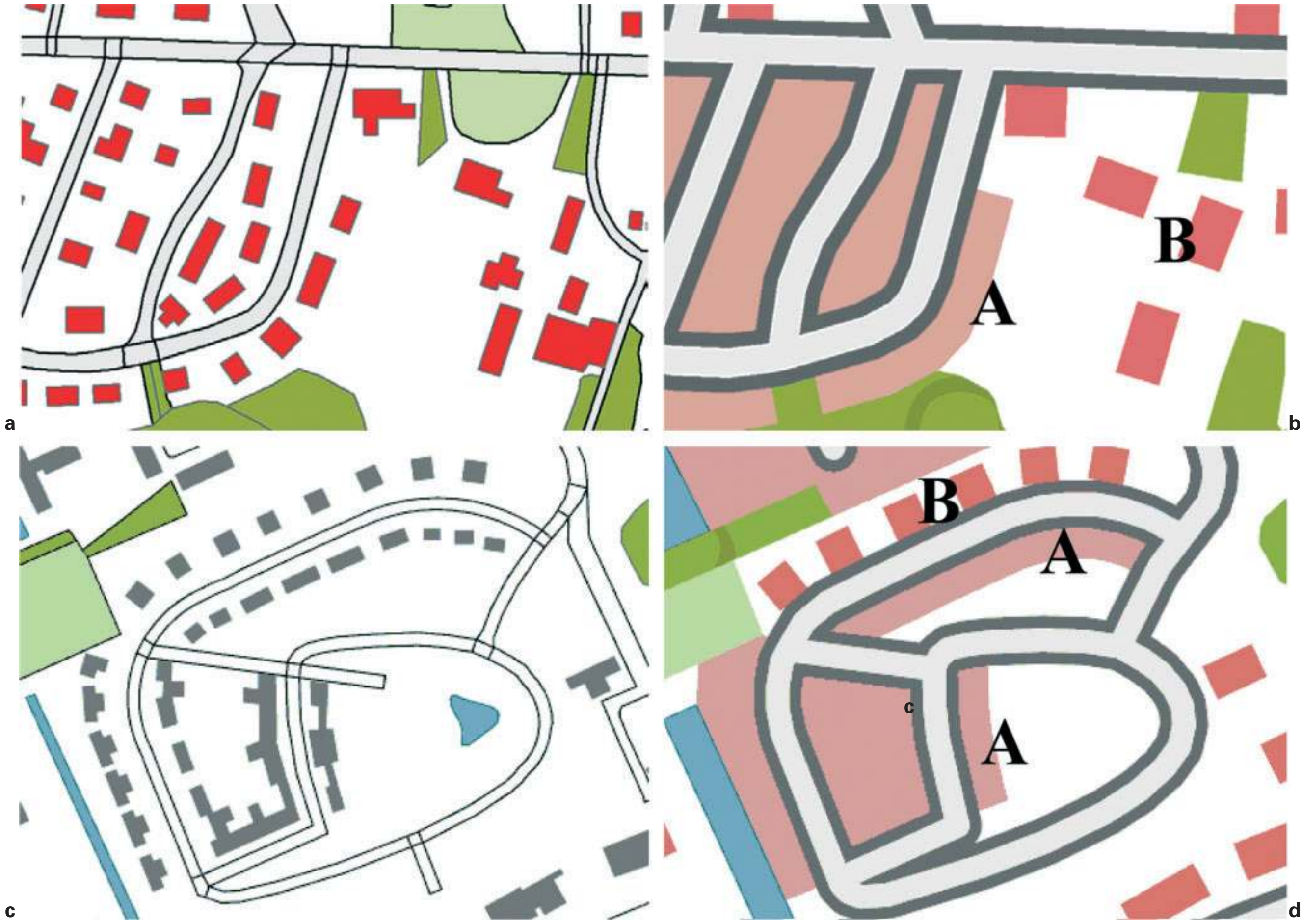


Figure 6. Conversion of TOP10NL buildings to TOP50 built-up areas, as a result of interactive generalization: TOP10NL (left) and interactive generalized TOP50 (right). 'A' shows locations where the uneven distribution of buildings across a polygon was taken into account. 'B' shows locations where 'typification' was applied and a set of buildings was replaced by a smaller number of larger examples

be generated. In any case, the specifications should be refined to avoid that land use polygons that contain large open areas are converted to built-up area.

Still some TOP10NL buildings remain in TOP50. This is because the density was too low, or because they are located in rural or industrial areas, or because they are important buildings or detached houses.

Several specifications describe how to treat such buildings. One specification describes that these remaining buildings should never be aggregated. Other specifications describe what to do with buildings that are too small to be depicted at scale 1 : 50k. For example, buildings should have a minimum size (20 × 20 m; in exceptional cases 15 × 15 m; protrusions should at least be 15 × 15 m), as well as a minimum distance to other buildings (10 m). Elimination is recommended to meet the distance requirement, where both important buildings and the last buildings in a line should be kept. Other operations such as displacement, typification (Figure 6, 'B' locations), enlargement or simplification are not explicitly mentioned in the specifications for buildings, but they are included in more generic specifications.

From the interactively generalized maps, we can see that cartographers mainly simplified and enlarged remaining buildings, where they applied displacement or elimination of buildings to solve cartographic conflicts.

Another observation from the interactively generalized data is that the values for minimum sizes are treated with a notion of flexibility by cartographers, as was mentioned in the section on 'Outlining the method'. The sample dataset of interactively generalized 1 : 50k data contained 69,435 ordinary buildings. Of these, 12,657 (18.23%) are below the minimum size threshold (20 × 20 m), and 237 (0.34%) are below the lower threshold of 15 × 15 m. The difference in minimum size as mentioned in the specifications and as used by cartographers can be explained by two reasons. First, it is not possible for humans to distinguish between the threshold and the threshold plus/minus a flexibility range and therefore, cartographers use the thresholds with a notion of flexibility (Ruas, 1998; Bard, 2004). Second, in specific situations, the cartographer may have chosen to relax the size constraint in order to meet a more important constraint, for example, 'keep important buildings'.

When comparing the existing specifications to the interactive solutions for detached houses, we identify other information that should be added to specifications for automated generalization. Although the specifications indicate to never aggregate detached houses to built-up

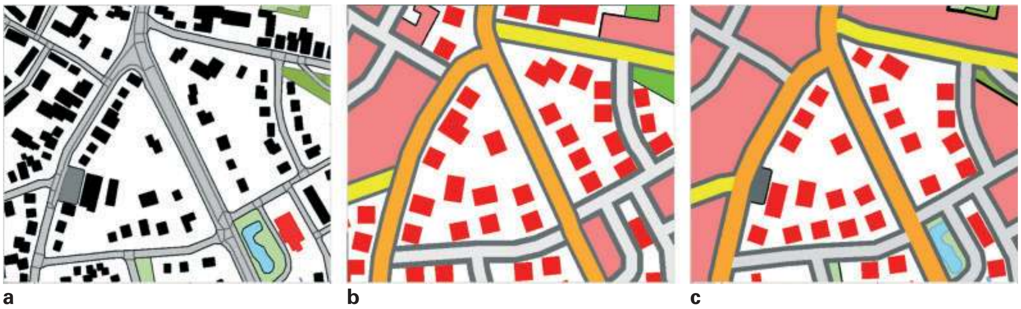


Figure 7. Optimally locating buildings in limited amount of space: (a) original buildings in TOP10NL, (b) buildings manually optimized by a cartographer, (c) buildings generated by the ESRI prototype Optimizer engine. Note that the rule to convert individual buildings to built-up area may have been applied differently in b and c

area, we see many cases where cartographers did aggregate detached houses to built-up area. This is because of the interaction between several specifications. Detached houses are often enlarged to meet the minimum size (15 × 15 m). In most building blocks with detached houses, both sides are covered with such houses. Consequently, the widths of the building blocks should be at least 40 m to accommodate enlarged detached houses at a minimum distance of 10 m. At the same time, streets in TOP50 are symbolized with line widths of 20 m, which is wider than streets widths in reality. Because of this high competition for space, cartographers did convert the original double-line pattern of detached houses to built-up area. In conclusion, detached houses are only presented in a few locations in TOP50, despite the specifications. To be more precise, an analysis of two test areas learned that cartographers converted 35% of detached houses in the rural test area and 65% of these houses in the urban test area to TOP50 built-up area. More space was available in the rural area to actually place detached houses as prescribed by the specifications. In contrast, more detached houses in urban area had to be converted to built-up area because of lack of space.

The analysis of how cartographers treat detached houses also revealed that the specifications do not address how to handle areas that are a mixture of detached and other houses, for example, by indicating a significant number of detached houses.

A final conclusion from implementing the existing building specifications into automated processes is that more than one solution is possible for locating buildings in a limited amount of space. This is true for both the interactive and automated process (Figure 7). Several algorithms are available to locate features in a sparse amount of space (for example, Sester, 2000; Monnot *et al.*, 2007b).

Generalizing roads

In TOP10NL, roads are represented by both road polygons and centrelines, while TOP50 only contains road centrelines. Consequently, the planar partition in TOP10NL consists of roads, water and land use, and in TOP50 of water and land use only, with exception of very large road features such as landing strips and car parks.

Generalizing TOP10NL road networks to TOP50 road networks consists of two main operations: collapsing road polygons to lines and selecting important roads.

For automating the collapsing operation, it seems straightforward to use TOP10NL centrelines for TOP50 centrelines. However, when analysing the interactively generalized map, we observed that the concept of centrelines differs between TOP10NL and TOP50. In TOP10NL,

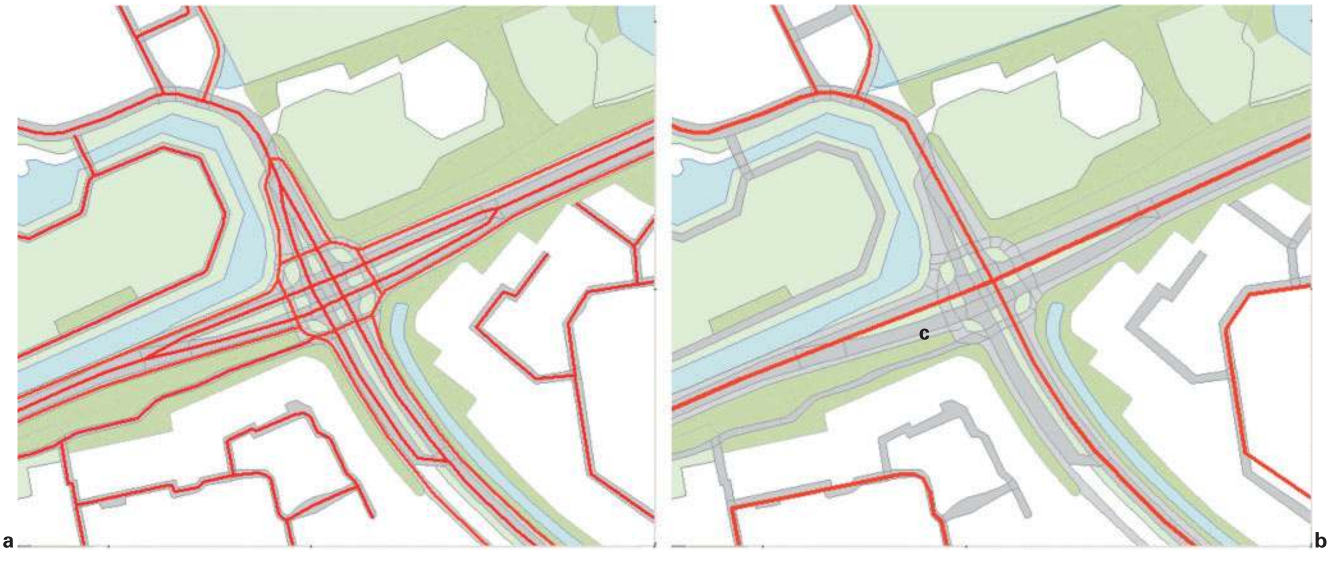


Figure 8. Centrelines in TOP10NL (left) and TOP50 (right), both projected on TOP10NL polygons

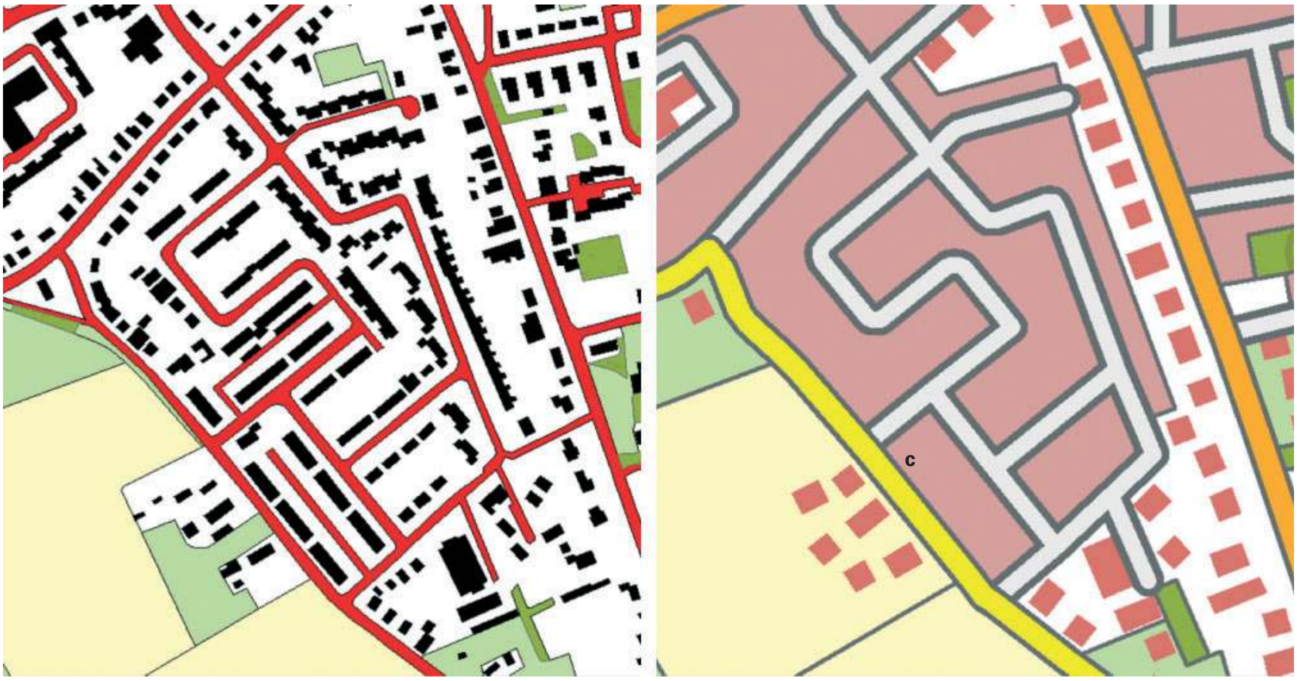


Figure 9. Selection of TOP10NL roads in TOP50: (a) TOP10NL, (b) interactively generalized TOP50

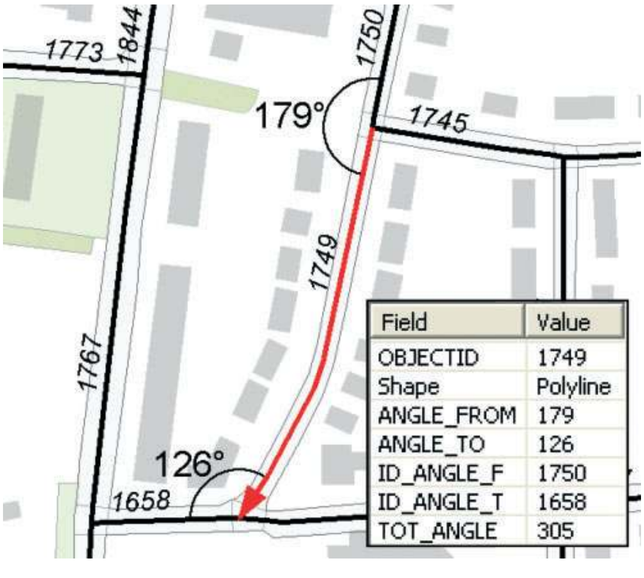
centrelines represent separate lanes, whereas in TOP50, centrelines represent complete road constructions, which may include verges, adjacent cycle tracks, etc. (Figure 8). The specifications do not explain these differences. Only when comparing the two maps, these differences become clear. Consequently, for roads with more than one lane, TOP50 centrelines need to be generalized from either TOP10NL road polygons or TOP10NL centrelines, which is possible in the case of dual lines (for example, Thom, 2007).

Comparing the results of collapsing road polygons to the interactive solutions shows again that the cartographer applies interpretation to the process which is not easy to automate. Ideally the collapsing should be applied to aggregated road features that cover the full construction of roads. However, a TOP10NL grassy area that is part of a road construction (i.e. verges) cannot automatically be identified as being part of the road. The reason is that all grassy areas in TOP10NL are classified as land use without any information on their function, for example, 'grass' as verge, 'grass' as pasture, 'grass' in parks, etc. Consequently, TOP10NL 'land use' actually contains information on 'land cover' rather than on the use. If this 'use' information were available, collapsing could be applied to the aggregation of all features that constitute a road, including the verges. This would avoid generating parallel roads in the collapsing process when two lanes are separated by verge. In addition, collapsing large crossings or roundabouts would be much simpler.

To enrich the data for automated processes, we added an extra attribute to land use areas specifying their functions, for example, 'separation of lanes'. This extra information enabled to generate TOP50 centrelines representing the full road construction instead of only the road surfaces of separate lanes as in TOP10NL.

The second operation for generalising roads is selecting important roads to be able to present symbolized roads in TOP50 without cartographic conflicts (Figure 9).

The specifications contain several guidelines for the selection of important roads, for example, 'dead ends' should be removed and 'through roads' should be kept. These two types of roads can be identified by the cartographer based on the overall map view. However, this information is not encoded in the data and is therefore not available for automated processes. We implemented an algorithm to detect 'through roads' and 'dead ends' to support the selection of the important roads. The algorithm assigns the angles of the most collinear connected 'from' and 'to' segments to each individual road segment, as shown in Figure 10. Thom (2007) and Touya (2007) also developed algorithms for road selection. Despite the promising results of these algorithms, our experiments show that importance



Field	Value
OBJECTID	1749
Shape	Polyline
ANGLE_FROM	179
ANGLE_TO	126
ID_ANGLE_F	1750
ID_ANGLE_T	1658
TOT_ANGLE	305

Figure 10. Calculated angles of most collinear connected segments

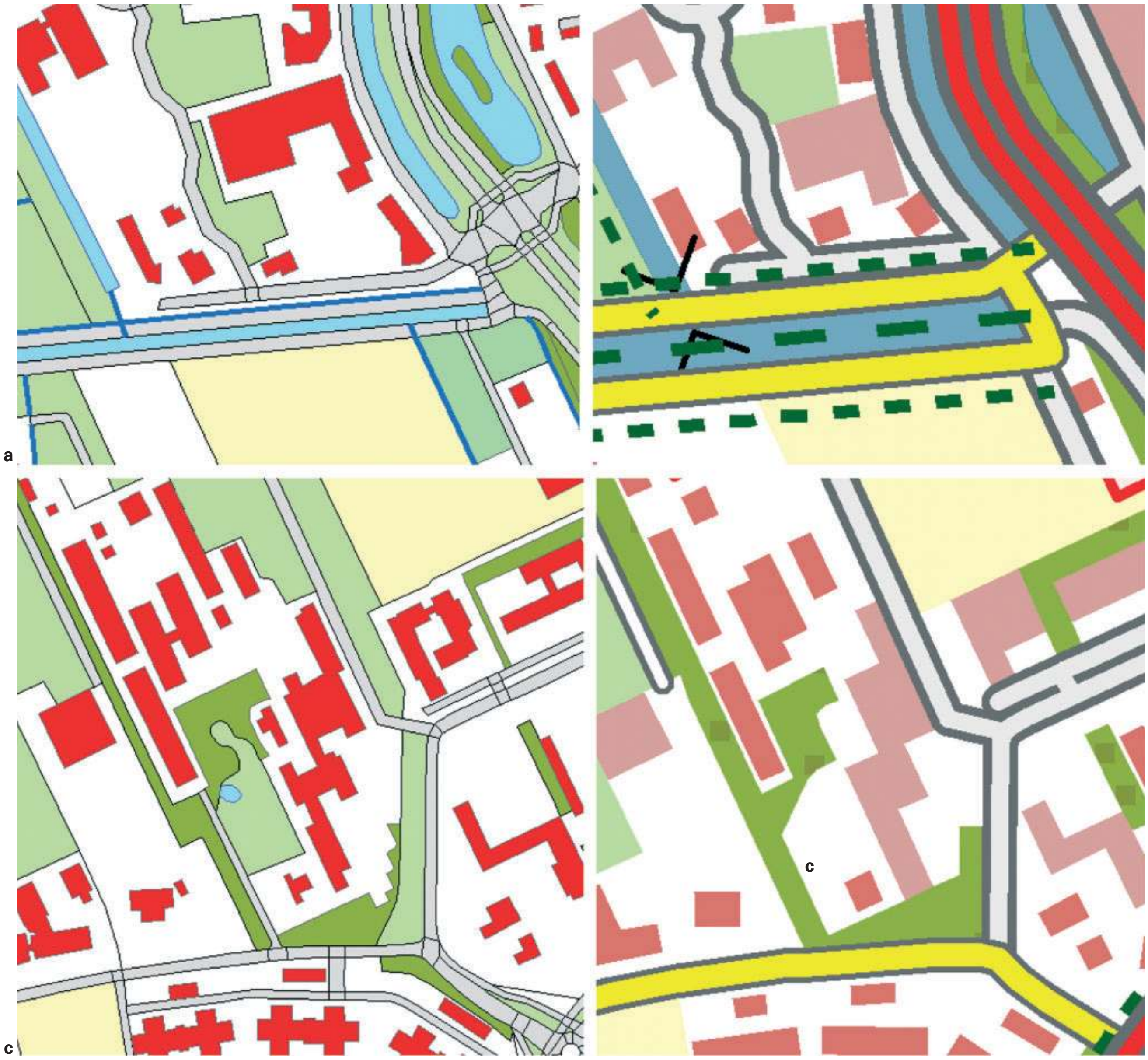


Figure 11. Examples of generalization of TOP50 water areas and woodland: (a) TOP10NL – roads alongside thin canal (centre), and roads between lakes (top right), (b) interactively generalized TOP50 – canal (centre) is enlarged, so roads is displaced. Lakes (top right) did move for enlarged roads, (c) TOP10NL – long thin woodlands between road and building (top left), and two similar woods (bottom and mid right), (d) interactively generalized TOP50 – long thin wood is enlarged, and buildings are moved for room. Mid right wood is kept and enlarged, but bottom right wood is deleted

information encoded on road features would yield more consistent solutions. This is also true for interactive generalization as can be seen in Figure 9, where cartographers selected roads in different ways.

Generalizing land use

For generalization of land use, we specifically studied two aspects: generalization of water areas and woodlands (both land use areas) and assigning former road areas to neighbouring land use.

The existing specifications do not address woodlands and water areas specifically. Therefore, we analysed specifications addressing general land use areas. These are:

- minimum size of patches is 15 × 15 m, otherwise they are eliminated and their areas are assigned to the neighbouring areas with the largest common boundary;
- exceptions are woodland patches between the lanes of a highway; these are always exaggerated to a minimum width of 15 m;
- patches should be at least 15 m wide; parts that are narrower than 15 m are widened to 15 m;
- woodlands and water areas should not overlap with road symbols. In the case of conflicts, they should be moved away from the roads;
- patches should snap to the edge of the road symbol if closer than 15 m.

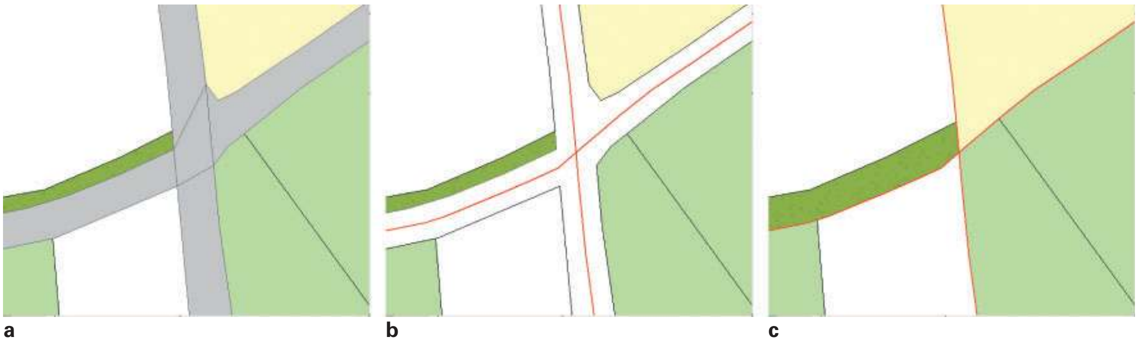


Figure 12. Extending TOP10NL land use areas to touch road centrelines: (a) TOP10NL, (b) collapsing TOP10NL roads causes gaps in partition, (c) TOP50

A more general specification for woodlands, grass and arable area is applicable here. This specification prescribes that 'character of the terrain should be maintained', where woodlands have higher priority than grass or arable area. In addition, as observed in the section on 'Visual analysis of existing maps', 'land use' has the lowest priority compared to all other feature classes in the case of displacements.

To learn more on how to automate these specifications, we studied the interactively generalized map with these requirements in mind. Also here we can conclude that the specifications are not sufficient to result in the current map and that extra information is needed. First, the applied generalization seems to differ between isolated woodlands/ water areas and woodlands/water areas next to a road. Woodlands and water areas next to roads are often kept and enlarged to support navigation (to inform drivers that they are passing woodlands), while others of similar size in the interior are discarded. Second, often the solutions lack consistency. Sometimes, they even contradict the displacement hierarchy as prescribed by the specifications: roads and buildings are displaced to make room for enlarged woodland and water areas (Figure 11). A final insight obtained from studying the interactively generalized map, is that enlarged woodland and water areas in TOP50 may lead to overestimating their importance when applying generalization of 1 : 100k map from TOP50, as is currently done.

Based on these insights, we formulated specifications for automatically generalizing woodland and water areas:

- patches are aggregated with a tolerance of 15 m (the minimum distance between features at 1 : 50k);
- patches that are smaller than 15 × 15 m after the aggregation process are deleted, and these areas are assigned to the neighbouring areas with the largest common boundary.

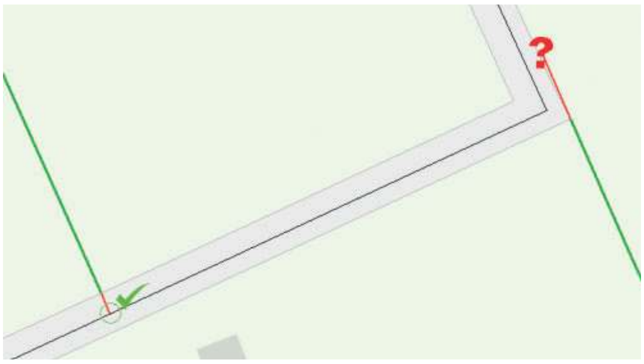


Figure 13. Example of an extended land use polygon boundary never meeting a road centreline

The second aspect of land use that we addressed in our generalization experiments is the process of assigning former road areas to neighbouring land use. This is needed because collapsing road polygons into TOP50 road centrelines causes gaps at the location of the former TOP10NL road areas (Figure 12). This 'repairing' of the data is not written down in the current specifications and must be added to the specifications for automated generalization.

We developed an algorithm for extending the original land use polygon boundaries until they touch the new road centrelines. The experiments show that this approach only works if the extended land use boundaries ever meet a centreline. This is not the case when the land use boundary is in line with a bend of road. In those cases, the land use boundary will extend to a line parallel to the road centreline and will never meet the road centre line (Figure 13).

Furthermore, the TOP10NL centrelines that we used for TOP50 centrelines yielded problems for the extension operation. Missing centrelines or centrelines that were not connected to the next centreline resulted in disappearing roads and in mistakenly combining two adjacent land use polygons.

For TOP50 data, it is important to realize that because of the extension of land use areas, these areas are larger than in TOP10NL (also true for the interactively generalized data).

SPECIFYING MAP REQUIREMENTS FOR AUTOMATED GENERALIZATION

From the experiments, of which the final results are shown in Figure 14, we can conclude that map specifications meant to support interactive generalization are suitable to trigger automated generalization, especially when they deal with aspects that are easy to measure (e.g. minimal area) and for isolated features. Although the automated process results in more consistent solutions than the interactive process, in many cases it is not straightforward to use existing specifications in automated processes. Often

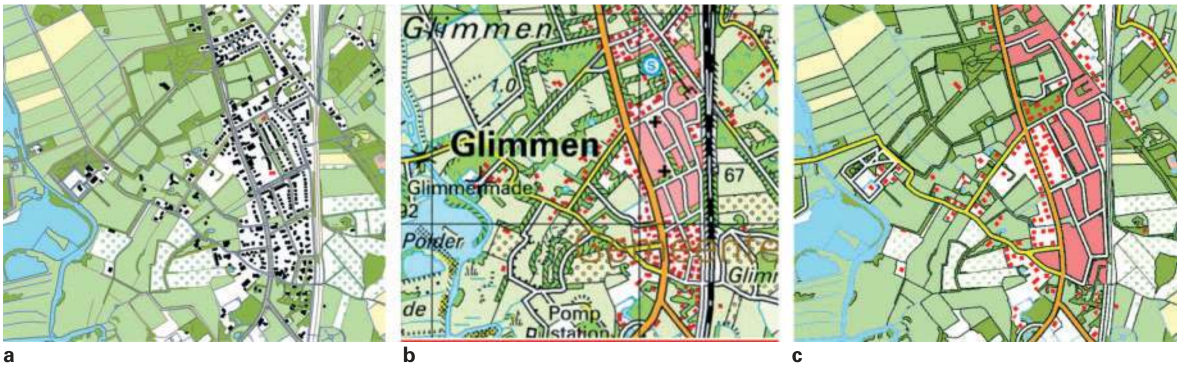


Figure 14. TOP10NL (left), interactively (middle) and automatically (right) generalized TOP50. Note that the automated solution only followed a selection of specifications as discussed in this paper and therefore, some basic operations (e.g. enlarging railways) were omitted

cartographers add very important knowledge to the interactive process. This knowledge, which was revealed by our experiments, should be made available for automated processes. From our experiments, several conclusions can be drawn to expose this knowledge and to specify map requirements for automated generalization.

Completing specifications and data

Although not complete, the current specifications are sufficient for interactive generalization since cartographers can add missing information to the process. For automated processes, missing information should be added to assure expected outputs.

First, missing or incomplete specifications that were revealed by the experiments should be improved as in the case for woodland and water areas. Because of the unclear specifications, cartographers chose non-consistent solutions which cannot be reproduced by automation. Second, specifications should be refined with cartographers' knowledge. Examples are which area to consider when converting areas exceeding a certain building coverage to built-up areas; when to treat an area as an area with detached houses; and the extension of land use areas to fill the gaps of former road areas. Another example of implicit cartographers' knowledge is the possible different meaning of concepts at different scales, as for road centrelines in TOP10NL (representing individual lanes) and in TOP50 (representing the middle of a complete road construction).

Apart from missing information in the specifications, the source data should be enriched with information necessary for generalization, such as 'detached houses', 'industrial areas' or 'verges'. Another solution could be to consult other data sources in the case of missing information. Finally, richer classifications of land use, road and water are required to better support the generalization process, e.g. pruning of artificial networks.

Formalizing specifications

The specifications may be understandable for humans, but should be formalized for automated processes. Formalizing specifications means reformulating them into measurable specifications that can be understood by automated processes. An example of an immeasurable specification is 'the character of the terrain should be maintained'.

Formalizing specifications also implies formalizing concepts that are not encoded in the data but are inferred by cartographers during the process. Examples are 'urban extent', 'character of the terrain', 'shape of buildings' and 'building pattern'.

Apart from formalizing the concepts themselves, their allowed changes at scale transitions should be mathematically described. Previous research has succeeded rather well in formalising requirements on isolated features. However, formalizing contextual concepts is much more complex (Regnauld, 1998; Weibel and Dutton, 1998; AGENT, 1999; Veltkamp and Hagedoorn, 1999; Sadahiro, 2000; Christophe and Ruas, 2002; Mackaness and Edwardes, 2002; Steiniger, 2007; Ai *et al.*, 2008). Lüscher *et al.* (2007, 2008) obtained promising results for interpreting high-level cartographic concepts from a reasoning process on formalized low-level knowledge.

Formalizing specifications require a formal language to express map requirements in a way closer to computer language. A possibly suitable language is the object constraint language (OCL) as shown in Stoter *et al.* (2008b) and Edwardes and Mackaness (1999). OCL is a language that can be used in combination with Unified Modelling Language (UML) to enrich a data model with additional knowledge.

Dealing with flexibility of specifications to assure the best output

Although specifications are meant as guidelines to obtain the best generalization results, addressing one individual specification should be treated with a notion of flexibility. This notion should somehow be included in the specifications for automated generalization.

For example, the specifications should address the interaction of multiple requirements and the intentional (partial) ignorance of specifications to meet more important ones. Weighting and prioritizing different specifications was previously addressed in the domain of constraint-based optimization (Ruas, 1998; Bard, 2004; Mackaness and Ruas 2007).

Other cartographers' knowledge related to flexibility that should be made explicit in the specifications is the notion of flexibility around the threshold values, for instance, minimum size for area features. Defining a 'sensitivity' range around these values may support implementing these values in a similar way as they are used in interactive processes.

Finally, specifications resulting mainly in exceptions (for example, 'detached houses should never be converted to built-up area') should be avoided. Although cartographers can treat such specification with flexibility by just ignoring it, automation is difficult when violating the specification is in most situations a good generalization solution.

CONCLUSIONS

This research aimed at specifying map requirements for automated generalization starting from current data and map specifications by a combined approach of reverse engineering and machine-learning techniques.

Written specifications were experimentally implemented and compared to interactively generalized maps. If necessary, these specifications were enriched and re-implemented, also using extra information from other sources. The experiments showed that strictly implementing generalization specifications meant for interactive generalization can help to reveal cartographers' knowledge. With this obtained knowledge, we formulated recommendations to specify generalization requirements for automated processes where cartographers have only minor influence.

Although the results of the experimentally implemented specifications were more consistent than the interactively obtained results, the experiments showed that results from interactive generalization will always differ from automated results. On the one hand, this confirms that it is impossible to formalize every cartographer's interpretation. This implies allowing the map produced by automated generalization to be (slightly) different from an interactively generalized map. On the other hand, this emphasizes that we need more formalized specifications than the map specifications which were meant to be interpreted by cartographers. The research presented in this paper is an important step towards this formalization process.

BIOGRAPHICAL NOTES

Jantien Stoter (1971) graduated in Physical Geography at Utrecht University in 1995 before beginning her career as a GIS specialist with the District Water Board of Amsterdam (1995–1997). From 1997 till 1999 she worked as a GIS consultant at the Engineering Office to support the planning of large infrastructure projects. From 1999 till 2004 she was an assistant professor in GIS applications, Delft University of Technology. In February 2004, she received the prof. J.M. Tienstra research-award for her PhD on 3D Cadastre. In April 2004, she started as assistant professor at the Department of Geo-Information Processing at ITC and became associate professor at the same department in 2005. In November 2009 she will (partly) return as associate professor to Delft University of Technology having multi-scale data integration as her main research interest. At the same time she will start as consultant at the Netherlands' Kadaster.

ACKNOWLEDGEMENT

Part of the research was carried out at ITC, Enschede.

REFERENCES

AGENT. (1999). Selection of Basic Measures, Deliverable C1, http://agent.ign.fr/deliverable/DC1.html

Ai, T., Shuai, Y. and Li, J. (2008). 'The shape cognition and query supported by Fourier transform', in **Headway in Spatial Data Handling: Proceedings of the 13th International Symposium on Spatial Data Handling**, ed. by Ruas, A. and Gold, C., pp. 39–54, Springer-Verlag, Berlin.

Bard, S. (2004). 'Quality assessment of cartographic generalisation', **Transaction in GIS**, 8, pp. 63–81.

Brassel, K. E. and Weibel, R. (1988). 'A review and conceptual framework of automated map generalization', **International Journal of Geographic Information Systems**, 2, pp. 229–244.

Buttenfield, B. P. (1991). 'A rule for describing line feature geometry', in **Map Generalization: Making Rules for Knowledge Representation**, ed. by Buttenfield, B. and McMaster, R .B., pp. 150–171, Longman, London.

Christophe, S. and Ruas, A. (2002). 'Detecting building alignments for generalization purposes', in **Advances in Spatial Data Handling: Proceedings of the 10th International Symposium on Spatial Data Handling**, ed. by Richardson, D., pp. 419–432, Springer-Verlag, Berlin.

Edwardes, A. and Mackaness, W. (1999). 'Modelling knowledge for automated generalisation of categorical maps – a constraint based approach', in **GIS and Geocomputation (Innovations in GIS 7)**, ed. by Atkinson, P. and Martin, D., pp. 161–173, Taylor & Francis, London.

Foerster, T., Stoter, J. E. and Lemmens, R. (2008). 'An interoperable web service architecture to provide base maps empowered by automated generalization', in **Headway in Spatial Data Handling: Proceedings of the 13th International Symposium on Spatial Data Handling**, ed. by Ruas, A. and Gold, C., pp. 255–275, Springer-Verlag, Berlin.

Gruenreich, D. (1992). 'ATKIS – a topographic information system as a basis for GIS and digital cartography in Germany', in **From Digital Map Series to Geo-information Systems, Geologisches Jarhrbuch Series A**, ed. by Vinken, R., pp. 207–216, Federal Institute of Geosciences and Resources, Hannover.

Kadaster. (2005). **Generalisatievoorschriften TOP50vector (Generalisation Regulations)**, Topografische Dienst, Emmen.

Leitner, M. and Buttenfield, B. P. (1995). 'Acquisition of procedural cartographic knowledge by reverse engineering', **Cartography and Geographic Information Systems**, 22, pp. 232–241.

Lüscher, P., Burghardt, D. and Weibel, R. (2007). 'Ontology-driven Enrichment of Spatial Databases', in **10th ICA Workshop on Generalisation and Multiple Representation**, Moscow, Aug 2–3, http://ica.ign.fr/BDpubli/moscow2007/Luescher-ICAWorkshop.pdf (accessed 3 March 2009).

Lüscher, P., Weibel, R. and Mackaness, W. (2008). 'Where is the terraced house? on the use of ontologies for recognition of urban concepts in cartographic databases', in **Headway in Spatial Data Handling: Proceedings of the 13th International Symposium on Spatial Data Handling**, ed. by Ruas, A. and Gold, C., pp. 449–466, Springer-Verlag, Berlin.

Mackaness, W. A. and Edwardes, G. (2002). 'The importance of modelling pattern and structure in automated map generalisation', in **Joint ISPRS/ICA Workshop on Multi-scale Representations of Spatial Data**, Ottawa, Jul 7–8, http://www.ikg.uni-hannover.de/isprs/workshop/macedwards.pdf

Mackaness, W. A. and Ruas, A. (2007). ‘Evaluation in map generalisation process’, in **Generalisation of Geographic Information: Cartographic Modelling and Applications**, Chapter 5, ed. by Mackaness, W. A., Ruas, A. and Sarjakoski, L. T., pp. 89–112, Elsevier, Amsterdam.

Monnot, J. L., Hardy, P. and Lee, D. (2007a). ‘An optimization approach to constraint-based generalization in a commodity GIS framework’, in **23rd International Cartographic Conference**, Moscow, Aug 4–10, http://www.pghardy.net/paul/papers/2007_icc_moscow_monnot_hardy_lee.pdf

Monnot, J. L., Hardy, P and Lee, D. (2007b). ‘Topological constraints, actions, and reflexes for generalization by optimization’, in **10th ICA Workshop on Generalisation and Multiple Representation**, Moscow, Aug 2–3, http://www.pghardy.net/paul/papers/2007_ica_moscow_monnot_hardy_lee_workshop.pdf

Muller, J. C. and Mouwes, P. J. (1990). ‘Knowledge acquisition and representation for rule based map generalization: an example from The Netherlands’, in **GIS/LIS ’90**, Vol. 1, pp. 58–67, Anaheim, CA, Nov 7–10.

Mustière, S. (2005). ‘Cartographic generalization of roads in a local and adaptive approach: A knowledge acquisition problem’, **International Journal of Geographical Information Science**, 19, pp. 937–955.

Plazanet, C., Bigolin, N. M. and Ruas, A. (1998). ‘Experiments with learning techniques for spatial model enrichment and line generalization’, **Geoinformatica**, 2, pp. 315–333.

Regnauld, N. (1998). ‘Généralisation du bâti: structure spatiale de type graphe et representation cartographique’, PhD thesis, Université de Provence, Aix-Marseille.

Ruas, A. (1998). ‘OO-constraint modelling to automate urban generalisation process’, in **8th International Symposium on Spatial Data Handling**, pp. 225–236, Vancouver, Jul 11–15.

Sadahiro, Y. (2000). ‘Perception of spatial dispersion in point distributions’, **Cartography and Geographic Information Science**, 27, pp. 51–64.

Sester, M. (2000) ‘Generalization based on least-squares adjustment’, **International Archives of Photogrammetry and Remote Sensing**, 33, pp. 931–938.

Steiniger, S. (2007). **Enabling Pattern-aware Automated Map Generalization**, PhD thesis, University of Zurich, Zurich.

Stoter, J. E. (2005). ‘Generalisation: the gap between research and practice’, in **8th ICA Workshop on Generalisation and Multiple Representation**, A Coruña, Jul 7–8, http://ica.ign.fr/Acoruna/Papers/Stoter.pdf

Stoter, J. E., Duchêne, C., Touya, G., Baella, B., Pla, M., Rosenstand, P., Regnauld, N., Uitermark, H., Burghardt, D., Schmid, S., Anders, K.-H. and Dávila, F. (2008a). ‘A study on the state-of-the-art in automated map generalisation’, in **11th ICA Workshop on Generalization and Multiple Representation**, pp. 1–16, Montpellier, Jun 20–21.

Stoter, J. E., Morales, J. M., Lemmens, R. L. G., Meijers, B. M., van Oosterom, P. J. M., Quak, C. W., Uitermark, H. T. and van den Brink, L. (2008b). ‘A data model for multi-scale topographic data’, in **Headway in Spatial Data Handling: Proceedings of the 13th International Symposium on Spatial Data Handling**, ed. by Ruas, A. and Gold, C., pp. 233–254, Springer-Verlag, Berlin.

Thom, S. (2007). ‘Automatic resolution of road network conflicts using displacement algorithms orchestrated by software agents’, in **10th ICA Workshop on Generalization and Multiple Representation**, Moscow, Aug 2–3, http://aci.ign.fr/BDpubli/moscow2007/Thom_ICA_Workshop.pdf

Touya, G. (2007). ‘A road network selection process based on data. enrichment and structure detection’, in **10th ICA Workshop on Generalization and Multiple Representation**, Moscow, Aug 2–3, http://aci.ign.fr/BDpubli/moscow2007/Touya-ICAWorkshop.pdf

Veltkamp, R. C. and Hagedoorn, M. (1999). **State-of-the-art in Shape Matching**, Technical report, Utrecht University, Utrecht.

Weibel, R. and Dutton, G. (1998). ‘Constraint-based automated map generalization’, in **8th International Symposium on Spatial Data Handling**, pp. 214–224, Vancouver, Jul 11–15.