Methodologies for the evaluation of generalised data derived with commercial available generalisation systems

Dirk Burghardt*, Stefan Schmid*, Cecile Duchêne, Jantien Stoter***, Blanca Baella[+], Nicolas Regnauld[++], Guillaume Touya****

* University of Zurich, Department of Geography, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland – tel: 0041 44 635 63 54 – firstname.lastname@geo.uzh.ch

** Institut Geographique National, Laboratoire COGIT, 2 avenue Pasteur, 94165 Saint-Mandé Cedex, France – tel: 0033 1 43 98 85 43 – firstname.lastname@ign.fr

*** International Institute for Geo-Information Science and Earth Observation, P.O. Box 6, 7500 AAEnschede, the Netherlands, lastname@itc.nl

[+] Institut Cartogràfic de Catalunya, Parc de Montjuïc, E-08038 Barcelona – firstname.lastname@icc.cat

[++] Ordnance Survey, Research and Innovation, Romsey Road, Southampton, UK – firstname.lastname@ordnancesurvey.co.uk

Abstract. The paper investigates methodical questions on the analyses and evaluation of automated generalised maps. The maps are produced with commercially available out-of-the-box generalisation systems, in a way that every system was tested by several persons on four test cases. The requirements on the generalised maps were described by cartographic constraints in a formal way. In addition, manually generalised maps were provided to give further reference information for the tester.

The analyses of the generalised maps are to be based on empirical and automated evaluation methods. The paper will present these evaluation methods in detail with objectives, related research, how the methods are realised and expected outcomes. Possible interchanges and synergies between the evaluation methods will also be described. The work published within this paper contributes to research on formal descriptions of cartographic requirements on generalised maps. It supports the development of methods for the situation and context dependent application of generalisation functionality and serves on the evaluation of existing generalisation products, to derive future research and development potential.

1. Introduction

This paper reports on an on-going work that takes place in the context of the EuroSDR project studying the "state-of-the-art of commercial out-of-the-box generalisation software". The aim of this project is to test commercial generalisation software systems on "benchmark" generalisation cases [Stoter et al., 2008; Burghardt et al., 2007; Stoter, 2007]. Four commercial available generalisation software systems, relying on different approaches for generalisation, have been tested on four test cases or *generalisation problems*. A *generalisation problem* is a large sale source dataset provided by a National Mapping Agency (NMA) together with specifications describing the expected output of the generalisation as well as the symbolisation of the output. The specifications are expressed as a set of constraints that the generalised data should respect. The datasets have been provided by ICC (Catalonia),

IGN (France), OS (Great Britain) and Kadaster (the Netherlands). [Stoter et al., 2008] describe the project in more detail.
Every software system has been tested from June 2007 till February 2008 by different testers on each generalisation problem. This has resulted in about ten outputs for each *generalisation problem*. Apart from structured documents in which experiences of the testers were captured, the outputs are the generalised data, consisting of all output layers in Shape formats as well as the symbolised version of the layers (i.e. output maps). After the testing stage the project has now entered its evaluation stage. This paper presents the methodology for evaluating the generalised data. The evaluation of generalised data aims at answering the question of "how much automated generalisation is available in commercial software" as well as "how different are generalisation solutions for the same generalisation problem". The evaluation methodology for generalised data of the project was designed in 2007 in an initial state. Based on test evaluations with the first versions of the methodology and based on a project meeting in April 2008, where these initial experiences were further discussed, the methodology was improved and better aligned with the research questions of the project.

The next section describes the general objectives of the evaluation task on generalised data within the EuroSDR project and presents the three main evaluation questions of this task. It also introduces the evaluation methodology aiming at answering the three evaluation questions. Sections 3, 4 and 5 detail the three evaluation procedures on which this methodology relies. The paper ends with concluding remarks and perspectives in section 6.

2. Evaluation of generalised data within the EuroSDR project

2.1 How many outputs have we got to evaluate?

In the testing stage of the project, for every defined *generalisation problem*, the four generalisation systems were tested by 2 till 3 testers from the project team. All testers are familiar with generalisation but not necessarily with all the tested generalisation systems. Therefore we can distinguish between *novices* and *experts* of the systems. Moreover, the software suppliers were invited to produce outputs in parallel tests where they were allowed to design additional developments to their systems, in contrary to the regular testers who only used out-of-the-box versions of the generalisation systems. Consequently for some of the *generalisation problems*, there are also outputs available produced by improved and customised versions of the tested systems. Theoretically there could have been 16 outputs per *generalisation problem* (12 from regular testers and 4 from software suppliers). Because in practice not all the expected tests could be done, there are about 10 different generalised outputs per *generalisation problem*.

2.2 The objective of evaluating generalised data

Evaluating the generalised data produced with the tested generalisation systems aims at:

- assessing the quality of generalisation output that current generalisation systems are able to provide
- knowing more about the domain of application (strengths/weaknesses) of the tested generalisation systems

In other terms, we evaluate the generalised data in order to evaluate the systems, as described in [Ruas, 2001, p.15] as part of a "loop in evaluation between system and output". [Mackaness and Ruas, 2007] distinguish between three types of evaluation: *evaluation for tuning* the generalisation system (prior to generalisation), *evaluation for controlling* the

generalisation process (during generalisation), and *evaluation for assessing* the quality of generalised data (after generalisation). The evaluation task described in this paper falls into the last category, but is specific because (1) we use the results of the evaluation in order to get insights into the systems used to generalise the data, and (2) we will perform comparative evaluation on several outputs that are supposed to meet the same specifications to learn more about generalisation processes. Regarding *evaluation for assessing*, [Ruas and Mackaness, 2007] distinguish three further subtypes:

" – *evaluation for editing* which aims to identify errors and mistakes (...).
– *descriptive evaluation* which provides summary information on what has been removed, emphasised or altered,
– *evaluation for grading*, which seeks to derive an aggregated value reflecting the quality of the solution overall" [Ruas and Mackaness, 2007, p.105].

As we do not seek for ranking the tested systems, we will not concentrate on *evaluation for grading*. In order to learn more about the systems, we will rather perform *descriptive evaluation* as well as *evaluation for editing* (to detect well and badly generalised cases).

2.3 What do we evaluate?

In order to meet the evaluation objectives, an evaluation methodology was designed. Figure 1 presents a schematic view on the evaluation methodology of generalised data within the EuroSDR project. The ellipses presents the data related to one *generalisation problem*: one *initial data* set, the *specifications* describing the expected output data, and several *outputs* (1 to n), all of them intended to meet the specifications as best as possible. The two-directional arrows present the three main evaluation questions and show what data is involved to answer these questions.

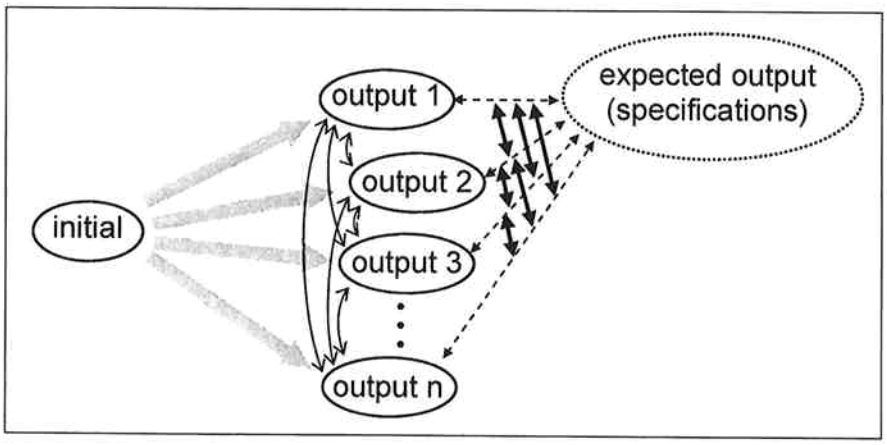


Figure 1: What do we evaluate for one generalisation problem ?

The three main evaluation questions are:

(1) How does each output respect the expected specifications (dashed arrows, right)? Here we consider one output at a time with respect to the specifications. This question will give insight into possibilities and limitations of commercial out-of-the-box software for automated generalisation with respect to NMA requirements. More precise questions are: are the provided solutions globally good? are the provided solutions good when looking in more detail to specific, local problems?

(2) How different are the outputs (curved arrows, left)? Here we consider all the outputs with respect to each other *without* regarding the specifications. We are interested if the provided solutions are very different as they are supposed to meet the same specifications. More precise questions are: are there cases that are handled very differently from one generalisation system to another and from one tester to another?
(3) How differently do the outputs respect the specifications and, more importantly, why (straight continuous arrows, right)? Here we consider all the outputs with respect to each other and with respect to the specifications. More precise questions are: which generalisation systems are or are not able to handle which kind of problems? Are there cases handled in the same way by all the software? Are there cases that were not handled at all by any software? Is the quality of generalised data significantly better when the generalisation system has been enriched with dedicated developments? Are the results with a same generalisation system very different if the tester is novice or expert? etc.

Apart from these questions on characteristics of the generalised data, the evaluation will provide insight into the test protocol itself. For example how understandable and precise the specifications are that are expressed as a set of constraints by the four NMAs.

2.4 How do we evaluate the generalised data: three interconnected evaluation procedures

Three parallel but interconnected evaluation processes are set up to cover the list of *what we evaluate* presented in the previous section.

- An *expert evaluation*, where experts of the NMAs that provided the four tests assess the cartographic outputs. The conclusions of experts for different outputs are compared to get insight into the different quality of different outputs. The assessments performed by the experts will contribute to answer question (1) and the comparative analysis contributes to the answer of question (3).
- An *automated constraint-based evaluation*, where the cartographic outputs are analysed to quantify the satisfaction of cartographic requirements by means of automatically computed constraint values and statistical indicators. These derived values are compared to give insight into differences in generalisation outputs. The computation of constraint values contributes to evaluation question (1) and their comparison across the outputs contributes to question (3).
- An *evaluation to compare generalised data*, where the different cartographic outputs obtained for a given generalisation problem are directly compared, especially by investigating several specific local situations. This procedure enables to answer evaluation question (2) and contributes to evaluation question (3).

The methodologies for the expert evaluation, the automated constraint-based evaluation and the evaluation on comparing generalised data are described in section 3, section 4 respectively section 5, by addressing the specific objectives and related research, the methodology in more detail and the outcomes of the evaluation.

3. Expert evaluation

3.1 Objectives and related research

Quality assessment has always been an important aspect of map generalisation since derived data must satisfy various requirements in order to be a satisfactory generalisation solution: it should be fit for the desired map purpose, it should represent reality and it should be readable by the user after generalisation. Traditionally, generalisation results have been assessed visually meaning that domain experts have been evaluating whether the (manually) derived data sets meet the underlying requirements. Most researches on evaluating the quality of generalised data focused on evaluating the effect of generalisation on one object or on one feature class. Examples are [Ehrliholzer, 1995] and [Bard, 2004]. Quality measures of interrelations of one feature class with other feature classes and of different requirements need further study. The objective of evaluating the generalised data by asking experts to assess it is related to these interrelations. More precisely the objective of the expert evaluation comprises 1) the assessment of the complete output maps and 2) the assessment of solutions for specific requirements. The key question in this evaluation is to what extent the outputs respect the requirements, where the requirements are considered to be laid down in the constraints. The global part of the expert evaluation will answer the question of what are we able to achieve with commercial software in automated generalisation addressing quality aspects at macro level. The second, detailed part of the expert evaluation has four objectives. Firstly, it will answer in what way current commercial software can handle specific constraints. Secondly the experts' assessments are compared to the output of the automated evaluation (see section 4). This comparison will yield insight into the (in)consistencies of both methodologies. Thirdly the detailed evaluation will focus on the interaction of several constraints. Evaluating single constraints is not an absolute measurement for the quality of generalised data since it does not take into account that violating constraints might have been necessary in order to solve other more important constraint and also not that good outputs for constraints might be due to violating others, e.g. respecting minimal dimension between all buildings is possible because (too) many buildings were eliminated. To address the interaction of constraints, the expert evaluation will focus on specific locations of the map taking all involved constraints into account. A last objective of the detailed part of the expert evaluation is to compare the different assessment outputs in order to see how differently the generalised data respect the specifications. This will show which software is capable to handle which kind of problems (evaluation question 3).

3.2 The expert evaluation in more detail

To get insight into how experts assess the generalised data, a survey has been designed which extends an earlier experts' survey of the AGENT prototype (AGENT, 2000). A first version of this survey was tested on outputs that were available in December 2007. Based on the first experiences, the survey was improved. Experts selected for the survey are experts that are familiar with the specific data in order to assure that they understand thoroughly what is expected in the outputs. For the global part of the survey the experts are asked to assess the whole map on the following aspects:

- Legibility
- Level of manual editions required to meet the specifications
- Deviation from initial data (ungeneralised data set)
- Preservation of the geographic characteristics of the test area (urban, mountainous, rural or costal area)
- Seriousness of main detected errors

- Frequency of main detected errors
- Number of occurrences of positive aspects
- Information reduction (undergeneralisation / overgeneralisation)

From the first version of the survey it could be concluded that it is impossible for experts to assess the solution for each constraint separately because these are too detailed. In addition there are constraints which cannot be assessed visually at all, such as orientation and position constraints. Consequently, for the improved version of the survey experts will be asked to assess the constraints on higher concepts. These are:

- For one object:
  - minimal dimensions (legibility)
  - granularity (amount of detail)
  - shape preservation
- For group of objects:
  - quantity of information (e.g. number of buildings preserved)
  - spatial distribution
- For two objects:
  - spatial separation between features (distance)
  - relative position
  - inconsistencies between related themes

The experts will be asked to assess these aspects for several feature types: buildings, roads, water, coastal features, relief (contour lines and spot heights) and land use. For the first version of the survey, the experts were asked to grade all aspects described above from 1-5 to be able to average grades for summarising the results.

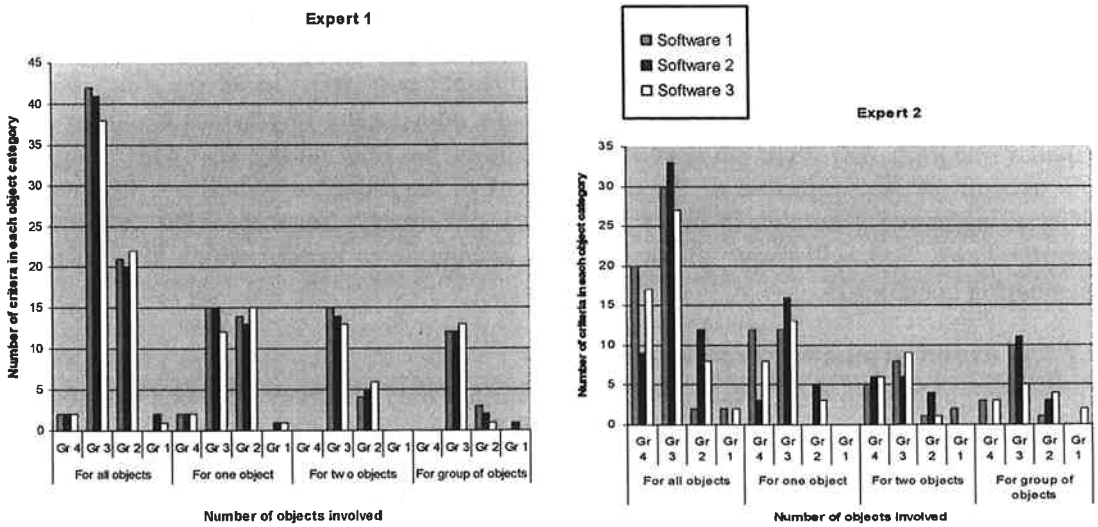
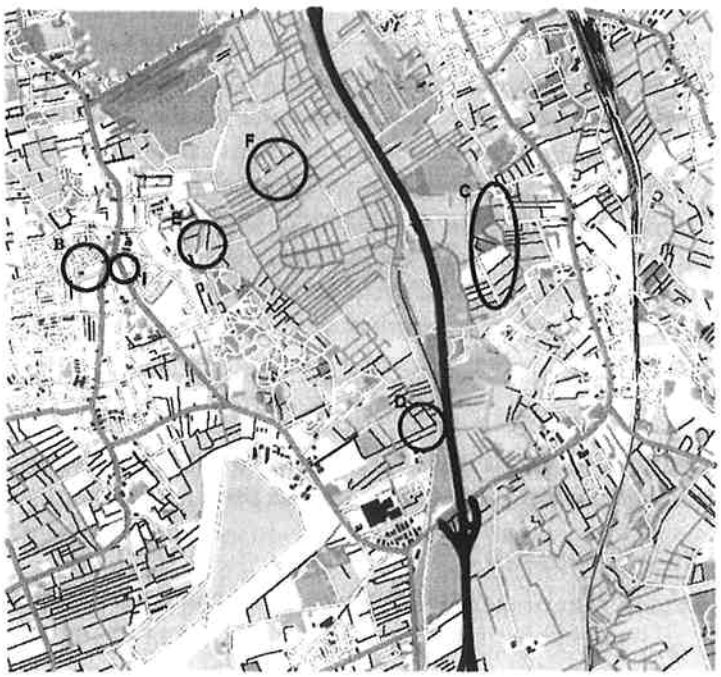


Figure 2: Two outputs of expert evaluation for one data set. The graphs show the percentage of constraints for which solutions were graded as 1, 2, 3 or 4; divided between solutions for constraints on one, two and group of objects.

However, comparing the outputs of two different experts on the same output reveals the (expected) subjectivity of experts (see figure 2): the overall assessments are similar, but there are many differences when looking at the assessments in more detail. For example in the category "for all objects" expert 1 evaluated only two constraints with grade 4, while expert 2

rated between 9 and 20 constraints with grade 4 depending on the software system. In interviews afterwards, the experts of the first survey also indicated that the grades cannot be used as hard values; they depend on some obvious side effects, such as time of the day or whether the question was posed at the start of the survey, where the expert was still fit, or at the end. For the improved version it was therefore decided to only use nominal values: very bad, bad, good and very good. A medium category was deliberately excluded in the improved version of the survey in order to 'force' the experts to choose.

The final part of the survey asks the experts to annotate the map with both good and bad examples, by specifically taking into account the interaction of several constraints (see figure 3 for an example output of the first version of the EuroSDR expert survey). The automated evaluation (see section 4) and the evaluation on comparing generalised data (section 5) will perform a pre-study so that experts will be pointed at situations with well or badly solved constraints are at situation that are of interest for a comparative analysis.



(A)

ID of identified example	Quality level			Comments on how far the constraints are satisfied. (e.g. reason for weakness, suggestion for improvement, description of the strength of good solutions)
	Good	Medium	Bad	
B			X	No buildings or built-up area on the map. There should be one of them.
C			X	Hardly any building left. It looks like this is an empty village.
D		X		Very small areas of forest.
E			X	Too crowded (coalescence).
F		X		Road is not connected.
I		X		Still (too) much detail in this building

(B)

Figure 3: Annotated map with good and bad examples (A) and explanation of examples (B)

The experience with the first version of the survey showed that care should be taken in assessing the solutions for constraints on preservation (shape preservation, relative position and inconsistencies between themes). At scale transition, these types of constraints are not violated assuming that the input data is correct. Good assessment by experts indicates that the initial situation is not deteriorated by the generalisation process. This might be either because the situation was not touched or because the system carefully took the preservation constraints into account.
In order to address evaluation question (3) a comparative analysis will be performed considering the outcomes for one test case. The assessment values for the four systems will be compared to see what software is appropriate to address which kind of problems and to identify cases handled in the same way by all the software and cases that were not handled by any software.

3.3 Expected outcomes

The assessments of the complete maps will result in a descriptive analysis addressing specific criteria. For the detailed part, the outputs of the expert evaluation will be summarized in tables per test case showing per constraint type, per software system, the assessment values of experts. Constraint types are classified based on the classification of [Burghardt et al., 2007]. In addition further distinction is made between constraints on natural phenomena (expected to be more irregular) and manmade objects, as well as between polygons representing small objects (such as buildings), polygons representing land cover objects, points, and lines. Consequently the expert evaluation will be able to answer the question on what generalisation functionality is available and missing in current software detailing on all these aspects. It will also give insights into differences of the outputs produced by different systems. The annotated maps will provide a way to assess the interaction of several constraints.

4. Automated constraint-based evaluation: automatically comparing generalised data to what they are expected to be

4.1 Objectives and related research

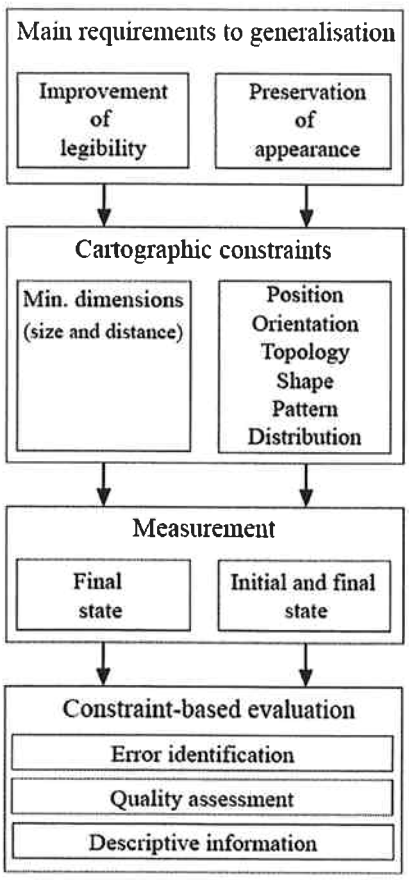
To date, with the development of automated generalisation and the intention for modelling the overall process, it is remarkable that the automation of quality assessment still lacks in techniques. Beside a few studies, which only focus on sub-areas of quality assessment [Brazile, 2000; Skopeliti and Tsoulos, 2001; Cheung and Shi, 2004], [Bard, 2004] was the first who contributed to development of a holistic assessment model proposing a general method to assess the quality of generalised map objects, as was already mentioned in section 3.
An automated quality assessment system as part of an overall generalisation process has many advantages, not only for research, but also for map producers as National Mapping Agencies (NMA). Firstly, in contrast to visual assessment, an automated system allows to reduce both time and cost for the necessary quality control within a production process. Nevertheless setting up such a system and preparing the data for the evaluation takes time, thus within the EuroSDR project even more time had to be invested on the automated evaluation than on the expert evaluation. Secondly, automated evaluation may support the improvement of generalisation processing with the identification of commonly occurring cartographic errors, the provision of further meta-data (e.g. descriptive information about modifications resulting from generalisation), the integrated comparison between alternate generalisation solutions and the improvement in specifying requirements to generalisation.

Automated evaluation of generalisation results can be performed on the basis of cartographic constraints. Crucial for that is the degree of constraint formalisation, which should give as much as possible implementation details. A cartographic requirement such as "initial and generalised shape should be similar" can be evaluated through very different shape measures and is less formal than the requirement on "preserving the width-length ratio" for example on a building shape.
On the basis of specified cartographic constraints provided by the NMAs, a typology of constraints was derived which reflect the main requirements to generalised data (see [Burghardt et al., 2007]). As a result, sets of constraints build the basis for the quality assessment of generalised test data sets.

The *automated constraint-based evaluation* pursues following purposes within the EuroSDR project:

a. Identification of constraint violations reflecting cartographic conflict situations.
b. Derivation of aggregated and average values representing the quality of a part or the overall generalisation result.
c. Provision of summary information on modifications resulted from generalisation (e.g. statistical analysis)

Both, purpose a) and b) are applied against specific cartographic constraints. Figure 4 illustrates the interaction of the components within the automated evaluation framework.



A set of cartographic constraints for a specific data set is principally derived from two main requirements to generalisation, namely the *improvement of legibility* and the *preservation of appearance*. In case of legibility, constraints are specified independently from initial data set, that is, they are defined by thresholds. On the contrary, constraints, which aim at preserving the appearance, are defined subject to the initial data set (reference data set) which is assumed to be correct. Legibility constraints as well as preservation of appearance constraints relate to specific and measurable map object properties as for instance size, position or orientation. On the basis of the measured *property values* in the initial (ungeneralised) and final (generalised) state, the actual evaluation procedure can be applied for preservation constraints. The evaluation contains the comparison of the measured *final value* with a calculated or predefined *ideal final value* as explained at the end of section 4. The difference results in a constraint violation between 0 and 1 whereas a maximal violation is equated with a constraint violation = 1. The resulting degree of constraint violations can then be qualitatively interpreted for grading the whole generalisation result or local solutions.

Figure 4: Framework of automated constraint-based evaluation

4.2 Automated evaluation in more detail

Implementation. The prototype for automated constraint-based evaluation of generalisation solutions has been implemented in OpenJump which is an open source and vector GIS software. The workflow is coded in Java language and it is presented in a user-interface form. The interface is composed of three main parts:

a. *Presentation*: It includes two windows, one for the presentation of the ungeneralised data set and another for generalised data set. Further, evaluation results (degree of constraint violation) are visualized in the last mentioned window on a separate layer.

b. *Application*: It involves the main menu for the execution of the evaluation procedure including data loading, attribute selection, constraint selection, and parameter setting.

c. *Output*: The corresponding window is integrated in the main menu and contains the results from the evaluation process in a tabular structure. Following values are displayed: measured property value in the initial and finale state, ideal finale value, difference between final value and ideal finale value, degree of constraint violation.

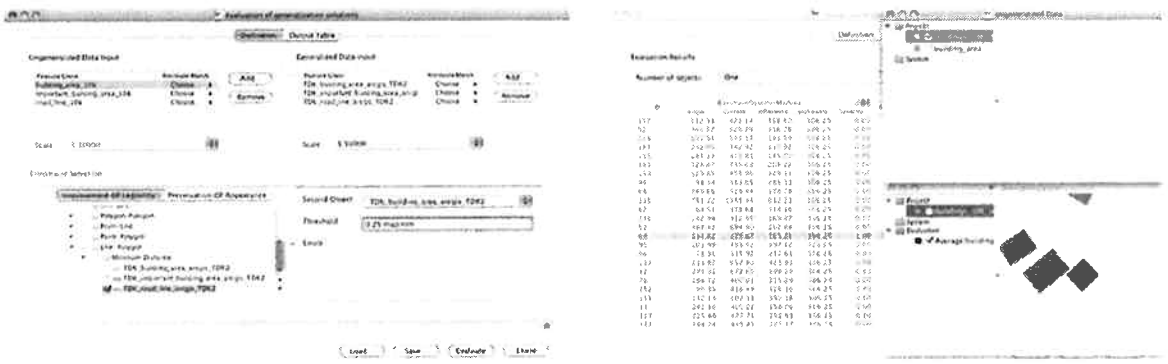


Figure 5: Prototype for automated constraint-based evaluation

Workflow. After loading of the ungeneralised and generalised data, the object properties such as symbol width and id's has to be selected necessary for the evaluation process. This is an important step since the entity relationship model (vertical relations such as 1:1-, 1:n- and n:m-relations) as well as the extent of symbolisation (e.g. width of line) are incorporated in the evaluation process in this way. The latter is necessary because the symbolisation information can not be exchanged directly by the typical vector exchange formats.
Further the cartographic constraint will be choosen for the automated evaluation with the corresponding parameter setting. The system lets the user set these parameters but in order to minimize the time exposure in case of repetition, the schema can be stored. Finally the evaluation process is carried out with the calculation of constraint violation values, which are listed in the output window with further statistical information. Due to the huge number of evaluation results and in order to enable a purposeful control of specific map objects, a visualisation tool is implemented.

4.3 Expected outcomes on the example of legibility constraints

In contrary to *preservation of appearance* constraints, *legibility* constraints are simpler to evaluate since they are only evaluated against a predefined and explicit threshold. There is no dependency on reference data (initial data set). However, the interpretation of constraint violation raises a problem. The constraint violation principally ranges from 0 to 1, whereas 0 indicates no constraint violation and 1 is maximum violation. In case of legibility, a threshold defines the minimum allowed size of part of an object or of the object itself, as well as the minimum allowed distance between objects, that is, the corresponding final value must respect these legibility thresholds. Consequently, there are two possibilities: firstly, the final value exceeds the threshold, or it is under the threshold (Boolean case). The usage of two concrete cases is advantageous for the identification of cartographic errors, but the question arises whether a Boolean approach is meaningful for grading generalisation solutions since information about the distance between the final and ideal final value is not incorporated.

With an example of evaluating three legibility constraints, we want to illustrate the expected outcomes:

a. Constraint 1: *target width of protrusion* $_{\{buildings\}} \geq 0.25$ *map mm*

Number of map objects without constraint violation CV = 0	*Number of map objects with constraint violation CV = 1*	*Mean constraint violation*	*Mean deviation from threshold in case of CV = 1 (in map mm)*
521	27	0.0493	0.08

b. Constraint 2: *target area* $_{\{buildings\}} \geq 0.1225$ *map mm*^{2}

Number of map objects without constraint violation CV = 0	*Number of map objects with constraint violation CV = 1*	*Mean constraint violation*	*Mean deviation from threshold in case of CV = 1 (in map mm)*
539	9	0.0164	0.021

c. Constraint 3: *target distance* $_{\{buildings,\ roads\}} \geq 0.25$ *map mm*

Number of map objects without constraint violation CV = 0	*Number of map objects with constraint violation CV = 1*	*Mean constraint violation*	*Mean deviation from threshold in case of CV = 1 (in map mm)*
527	21	0.0383	0.11

The examples provided here focus on building and road objects in rural areas. The data (topographic map) used in this example is the generalisation problem of TD Kadaster (Netherlands) where the final data set at scale 1:50'000 is derived from data at 1:10'000.
The evaluation results for the three constraint examples show the degree of constraint violation on all map objects, that is, the corresponding *mean constraint violations* state whether the underlying constraints are globally satisfied or not. A further value indicates the *mean deviation* from the threshold in case of constraint violations. In support of the mean deviation value, the prototype provides additional information as for instance statistical dispersion.
As mentioned above the constraint-based evaluation of *preservation of appearance* constraints is much more complicated since the ideal final value can rarely be defined explicitly. This problem holds especially true for *shape* constraints. In order to be able to

determine an ideal final value, that is, the quantitative description of target shape, various aspects must be previously considered.
Firstly, it is hardly impossible to describe mathematically the *general shape* of a map object. There are a huge number of existing measures for the quantitative description of shape specific properties such as concavity, elongation or shape index. At this point, the question arises: which existing shape measures are suited for an optimal shape description and evaluation? If this question is answered, shape of map objects can be measured meaningful to quality assessment.
A second problem, as mentioned at the beginning, is the calculation of the ideal final value of a map object's shape: which are the modifications to shape in generalisation at a certain scale transformation? Hence, scale depending modifications to shape must be formalized in order to be able to determine ideal final values for any starting situations. The extraction of this information can be carried out among others by means of reverse engineering, that is, gathering knowledge by studying existing maps on different scales [Harry, 2001]. It seeks to establish which transformations are needed to describe an optimal (or correct) solution on a specific scale. The technique we will apply for the information extraction concerning shape of individual map objects is *statistical analysis* of manual generalised maps of high quality. The results of this analysis are regression functions which describe the relation between independent variable (initial value) and dependent variables (final value).

5 Comparative evaluation: comparing several generalised datasets between each others

5.1 Objectives

As presented in section 2 of the paper, for each *generalisation problem* there are up to 16 different cartographic outputs, all supposed to meet the same specifications. The *comparative evaluation* procedure compares these cartographic data in order to identify and hopefully explain the kinds of differences between them as suggested by [Ruas, 2001, p.19]. The objectives of this comparative analysis performed directly on the generalised data are:

1. To assess if the obtained generalised datasets are very different from each other. without considering the specifications, i.e. we do not try to assess if some outputs are better than others. We just try to have a rough idea, from a quantitative point of view, of how different the generalisation outputs can be given a set of specifications.
2. To describe more in depth, from a qualitative point of view, the differences noticed between the different generalised datasets. Here the specifications are taken into account: we analyse how differently the generalised datasets respect the specifications. And we try to outline possible correlations between the way a given problem is handled and the used generalisation system, the tester, etc.

The methodologies proposed for achieving these two objectives including the expected outcomes, are respectively described in sections 5.2 and 5.3.

5.2 Measuring how much the generalised datasets are different

Figure 6 shows that outputs obtained from the same dataset can be very different. Even though the symbolisation differs from one extract to another one, it is obvious that far less buildings have been kept in extract 3.

Figure 6: Extracts of outputs obtained from the IGN-France dataset by three different testers using different systems.

In order to measure how different the outputs are, we will perform some quantitative analysis as proposed by [Mackaness and Ruas, 2007], and analyse the distribution of values across the set of cartographic outputs related to the same *generalisation problem*. This analysis concerns total numbers for small objects like buildings, cumulative lengths for networks, and cumulated areas for land cover. They will be computed at the macro level i.e. regarding "all buildings", "all roads", "all rivers", etc. If the outputs visually appear to be very different in specific areas, the analysis will also be performed on parts of the datasets like meshes of the road network (meso level as defined in [Ruas, 2000]).
On top of these computed indicators, interactive visual comparison will be used to notice general trends. As mentioned above, this interactive visual comparison will also enable to know where to compute indicators in order to confirm, by means of numbers, trends that have been visually detected.

5.3 Qualitative analysis of the differences between the generalised datasets

Besides quantifying how much the outputs obtained from one dataset are different, we also wish to qualitatively analyse how they differ and, when possible, to explain the encountered differences. If two outputs are locally different (spatially or thematically), several cases can be distinguished:

- The two outputs respect the expected specifications (the constraints): it means the specifications are flexible enough to enable very different interpretations. It is known that generally, several solutions exist to one generalisation problem [Spiess, 1995]. If these solutions are very different, probably the specifications are not precise enough.
- At least one of the outputs does not respect the specifications: three main situations can occur regarding the case that does not respect the specifications: (1) the concerned tester has succeeded in translating the specifications into the system, but there is a bug in the system linked (2) the concerned tester did not succeed in translating the specifications into the system, because the system (2a) or its documentation (2b) is incomplete or (2c) the specifications were not understood by the tester, (3) the concerned tester thinks he has succeeded in translating the specifications, but actually he misinterpreted one part of them, possibly because they are not clear enough.

It should be possible to identify the situation (2) by analysing the "constraints expression template", a template filled in by every tester for every dataset and system, where the tester mentions for each constraint if he has succeeded in expressing the constraint in the system and, if not, why not. Comparing the constraints expression templates (and the outputs) of novice and an expert testers for the same system will help to discriminate between situations (2a) and (2b). Looking at the results obtained by the software suppliers in their parallel testings could also help.

In the same way, the situation (3) can be identified if the results obtained by the same system are very different whereas their constraint expression templates look the same. Situation (1) would rather correspond to the same problem appearing in all the outputs produced with a given system.

The comparative analysis of the outputs will mainly be based on careful visual investigations based on the following elements of methodology:

- Zones of particular interest will be defined for each dataset, on which the comparative analysis will focus.
- Figure 6 shows the importance of having similar symbology. Therefore the symbolisation of all outputs has been redone by one person.
- A grid to help further analyses will be filled in during the visual investigation stage:

Dataset	Zone no.	Type of conflict	Software	Tester	Novice/ Expert	Constraint expressed according to constraint expression template?	Short description of how the conflict has been handled by the system

The expected outcomes of this visual comparison are: for every "zone of interest" of every *generalisation problem*, a series of small extracts from the different cartographic outputs, annotated with a qualitative description of how they differ and an attempt of explanation of why they differ. If the counting's performed as presented in section 5.2 show significant differences from one dataset to the others on a particular aspect, we will also try to explain these differences at this stage.

6 How *automated constraint-based evaluation*, *expert evaluation* and *comparative evaluation* support each others

In the expert evaluation cartographers will assess both globally and in detail to what extent the outputs meet the requirements and if differences can be identified between systems (what system is best capable of handling which problem). The automated constraint-based evaluation provides values for the satisfaction of specific constraints (mainly legibility constraints) for every test data set for every of the four software systems and will study if quality differences can be identified between the four systems. The comparative evaluation will investigate common patterns and heterogeneity between the different output data for one test case achieved by different testers with the four generalisation systems. This last evaluation will yields insight into 1) different solutions for one generalisation problem and 2) which system is better suited for the generalisation of a specific test data set or group of constraints.
Synergy between the three evaluation procedures is accomplished because findings will be exchanged and examined to answer the three main research questions raised in section 2.3.

Example of interaction and interchange are:

- the quality results for specific constraints produced by the automated evaluation will be compared to the quality results for the same constraints evaluated by experts;

- the systematic automated and expert evaluation for specific constraints and situations will identify aspects on which a closer interactive analysis is needed in the comparative analysis (for example situations where the quality of the solutions differ considerably);
- a first run of the comparative evaluation and the automated constraint-based evaluation will identify situations to which experts will be pointed in the expert evaluation in their task to annotate the output maps with remarkable solutions (very good, very bad or very differently solved solutions);
- the automated and expert evaluations on particular constraints can confirm or invalidate trends noticed through the visual comparative analysis;
- the results achieved in automated constraint-based evaluation by standard deviation analysis will be compared to the visual comparative analysis.

Besides these integrations, a comparison between the outputs of the three evaluation processes will be done to detect inconsistencies between the three types of evaluations that need to be further investigated (e.g. bugs or inappropriate measuring tools in the automated evaluation process; misinterpretation of what was asked to the experts; subjectivity of the evaluations).

7 Conclusion and perspectives

This paper has presented the three main questions that are the focus of the evaluation of generalised data within the EuroSDR project. How does each output respect the constraint specifications? How different are the outputs? How differently do the outputs respect the specifications and what are the reasons for that? The evaluation is supported by several evaluation processes for which the methodologies were presented here, including some initial experiences. After an extended evaluation phase (summer 2008), it will be investigated how much the results of the three evaluation processes can be integrated further to answer the three main evaluation questions of the project. Based on these first results, the last part of the evaluation will be carried out in autumn with fine-tuned methodologies.

Acknowledgments

The authors would like to thank all participants of the EuroSDR project, in particular we wish to thank Maria Pla (ICC, Catalonia), Peter Rosenstand (KMS, Denmark), Karl-Heinrich Anders (University of Hannover, Germany), Annemarie Dortland, Maarten Storm and Harry Uitermark (TD Kadaster, The Netherlands), Magali Valdeperez, Francisco Martínez and Francisco Dávila (IGN Spain) for the discussions and contributions during the project meetings.

References

AGENT (2000), Map generalisation by multi-agent technolgy, http://agent.ign.fr/

Alt, H., Behrends, B. and Blömer, J. (1991), Approximate Matching of Polygonal Shapes. *Proceedings of the 7th ACM Symposium on Computational Geometry*, pp.186-193.

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

Bel Hadj Ali, A. and Vauglin, F. (1999), Geometric Matching of Polygons in GISs and assessment of Geometrical Quality of Polygons. *Proceedings of International Symposium on Spatial Data Quality* (ISSDQ'99), Hong-Kong, 1999, pp.33-43.

Brazile, F. (2000), Semantic Infrastructure and Methods to Support Quality Evaluation in Cartographic Generalization. *PhD thesis*, Department of Geography, Zurich, Switzerland.

Burghardt, D., Schmid, S. and Stoter, J. (2007), Investigations on cartographic constraint formalisation. In: *Workshop of the ICA Commission on Generalisation and Multiple Representation*, August, Moscow.

Cheung, C. K. and Shi, W. (2004), Estimation of the Positional Uncertainty in Line Simplification in GIS. *The Cartographic Journal*, 41(1), 37-45.

Ehrliholzer, R. (1995), Quality assessment in generalisation: Integrating quantitative and qualitative methods. Paper presented at the 17th International Cartographic Conference, Barcelona : ICC.

Harrie, L. (2001), An Optimisation Approach to Cartographic Generalisation. *PhD thesis*, Lund Institute of Technology.

Harrie, L. and Weibel, R. (2007), Modelling the overall process of generalisation. In: Ruas, A., Mackaness, W.A. and Sarjakoski, L.T. (eds.). *Generalisation of Geographic Information: Cartographic Modelling and Applications*, Series of International Cartographic Association, Elsevier, pp. 67–87.

Mackaness, W.A. and Ruas, A. (2007), Evaluation in the map generalisation process. In: Ruas, A., Mackaness, W.A. and Sarjakoski, L.T. (eds.). *Generalisation of Geographic Information: Cartographic Modelling and Applications*, Series of International Cartographic Association, Elsevier, pp. 89–111.

Podolskaya, E. S., Anders, K.-H., Haunert, J.-H., & Sester, M. (2007), Quality assessment for polygon generalization. Paper presented at the Spatial data quality 2007, Enschede, The Netherlands.

Ruas, A. (2000), The Roles Of Meso Objects for Generalisation. *Proceedings of the 9th International Symposium on Spatial Data Handling*, Beijing, 2000, pp.3b50-3b63.

Ruas, A.(2001), Automatic Generalisation Project: Learning Process from Interactive Generalisation, OEEPE Official Publication n°39.

Skopeliti, A. and Lysandros, T. (2001), A methodology for the assessment of generalization quality. *Fourth Workshop on Progress in Automated Map Generalization*, Beijing, China.

Spiess, E. (1995), The need for generalization in a GIS environment. In: Müller, J. C., Lagrange, J. P. and Weibel, R. (eds). *GIS and Generalization, Methodology and Practice*, Taylor & Francis, pp. 31-46.

Stoter, J. (2007), EuroSDR project, Research on state-of-the-art of generalisation. Oral presentation during the *Workshop of the ICA Commission on Generalisation and Multiple Representation*, August, Moscow, http://aci.ign.fr/Moscow/oral/jantien.pdf.

Stoter, J., Anders, K.-H., Baella, B., Burghardt, D., Davila, F., Duchêne, C., Pla, M., Regnauld, N., Rosenstand, P., Schmid, S., Touya, G., Uitermark, H. A study on the state-of-the-art of generalisation within commercial out-ofèthe-box softwares. *Workshop of the ICA Commission on Generalisation and Multiple Representation*, Montpellier, France.