IMPROVING DATA QUALITY AS THE BASIS FOR AUTOMATED GENERALISATION

Martin Gregory
martin.gregory@1spatial.com

Steven Ramage
steven.ramage@1spatial.com

1Spatial
Cavendish House
Cambridge Business Park
Cambridge CB4 0WZ
United Kingdom

Abstract

Data quality is a significant concern for all involved in information technology and the software business globally. The Data Warehousing Institute estimated that data quality problems cost U.S. businesses more than $600 billion per year. In Europe, for those working with spatial data, PIRA (Commercial Exploitation of Europe's Public Sector Information, 2002) estimated that ten years ago it would cost the European Community countries €36 billion to replace its geographical information assets. This amount was estimated to be growing at €4.5 billion per annum. Similar costs for the U.S. were estimated at $375 billion with a $10 billion growth per annum. How does data quality then relate to delivering against business objectives of automated generalisation?

Generalisation is concerned with the detection and resolution of conflicts between map objects for representation at the target scale. Generalisation has historically been part of the day-to-day operational processes within regional and national mapping agencies. Since the early adoption of digital cartographic techniques, these agencies have aspired to implement automated processes to achieve generalisation, thus achieving efficiency, improving concurrency between products and allowing more up-to-date products for their target markets. As this paper will outline the underlying data quality is of paramount importance if these aspirations are to be met.

In order to achieve the benefits of digital cartographic techniques, agencies have made major investments in the capture and maintenance of digital spatial data. Governing authorities are seeking increasingly to obtain a greater financial return from these investments and reuse the data for purposes other than those originally intended. One of the main repurposing requirements is related to geospatial data intended for map publishing now being used for electronic delivery across the Web. There are numerous reasons why spatial data may not meet user expectations, for example the use of GPS has introduced a more rigorous accuracy for reference datasets, transformations alter

information or it may just be that there is a lack of understanding or recognition that data may be inaccurate. Whatever the reason it is essential that prior to generalising data, it represents the highest quality achievable; otherwise problems relating to poor data are just expounded.

Data quality has been referred to as a forgotten common sense (Informatica 2007) and there is relatively little material available relating to data quality in generalisation research. Generalisation influences many aspects of data quality, such as attribute accuracy, location accuracy, consistency and completeness (Muller, 1991). There may also be unpredictable effects of generalisation on metrics, such as topological and semantic accuracies of map products (Haunert & Sester, 2008). Quality also relates to the specification of any generalised dataset, i.e. what was the planned outcome or purpose of the generalisation process and do the results match the desired outcome? In this instance it is important to clarify the quality objectives in terms of model or cartographic generalisation. As will be explained, a Generalisation Framework approach enables the context to be clearly established; although both areas are related to fitness for purpose, the end purposes are different. Quality therefore also refers to product suitability for its intended use.

Finally, the level of automation possible will also be dependent on the quality of the data. There are many operational instances of generalisation systems worldwide that utilise semi-automated techniques and proven instances of increasing levels of automation. Increasing the automation improves the productivity and consistency and provides the basis for producing more specialised and targeted products. However, the level of automation possible is affected directly by both the information available to define the target requirements, such as data specifications and the quality of source data used to meet those requirements or specifications.

Introduction

Historically, generalisation is the task of deriving small-scale maps of digital data from existing, more detailed mapping or source data. It involves removing irrelevant detail that would clutter and confuse, and exaggerating those aspects which are important for the particular purpose and scale. Today, generalisation is required to support the provision of digital data sets at smaller scales, as well as maps. It has progressed significantly from purely research to an operational business activity; 1Spatial has experience working in this domain with customers worldwide.

Cartographers have traditionally performed spatial data generalisation manually in order to produce maps and charts. They follow some guiding principles and rules, but also introduce their own context understanding to the process. This manual process leads to high costs, difficulty of replication and slowness to market. Issues with data quality and accuracy arise due to the possible introduction of factual errors and inconsistency of presentation. More than 10 years ago 1Spatial was involved in a project called AGENT:

Automated Generalisation - New Technology, this was funded under a European Framework Programme. The focus of this project was to define and validate multi-agent methods applied to geographical modelling, algorithms, constraint modelling, process control and solution evaluation in map generalisation. 1Spatial has since embedded these software algorithms into their products and has focussed on automating as much as possible of the generalisation process. A Generalisation Framework offers a methodical approach to address the issues associated with manual vs. automated generalisation while tackling data quality iteratively.

Objectives

The objectives of this paper are to:

1. Highlight the impacts of poor quality data
2. Examine practical and implemented approaches for solving quality issues
3. Consider further options for quality control and improvement

The first objective uses examples of situations encountered during the analysis or processing of data and associated quality information obtained while generalising data over a number of years. Examples are drawn from 1Spatial's collaboration with organisations such as Institut Géographique National (IGN) in France and Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik (ADV) in Germany.

The second objective is to examine practical strategies for solving problems via a Generalisation Framework. This focuses on software solutions that have been applied in pursuit of automating the generalisation process.

The third objective considers those situations where an algorithmic solution is difficult and where such an implementation is not currently considered cost effective. These are usually examples where the current recourse is to improve the data or find the missing information.

Methodology

The approach is based on a Generalisation Framework that consists of several stages to address the data management issues: data preparation, model and cartographic generalisation, text placement and then product finishing.

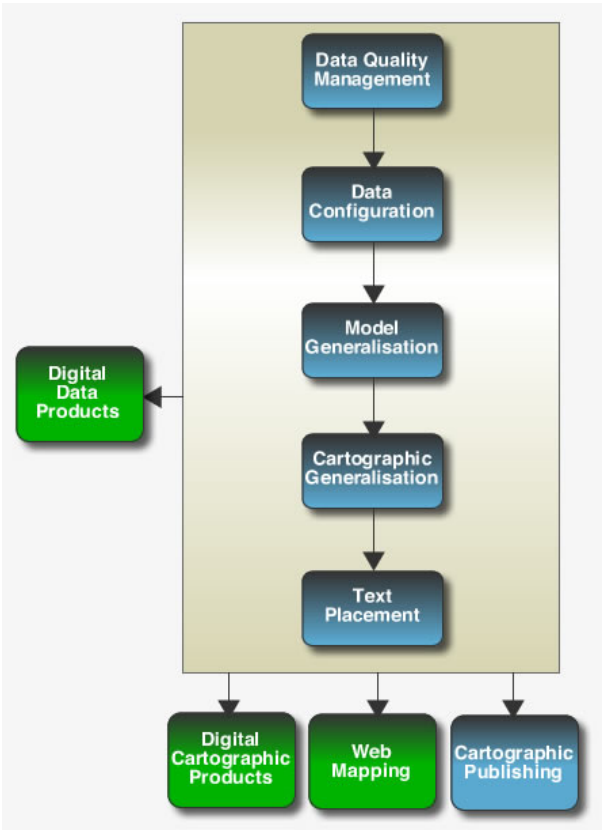


Figure 1: A Generalisation Framework

Data preparation is the initial stage to ensure the data is quality assured, for example through topology management. Generalisation then starts from a source dataset, within which data of a particular scale and accuracy have been gathered, stored and maintained. It applies a set of business processes and decisions, to select and transform this data in order to derive a preferred structure and/or legibility for the target product. The simplification and re-modelling aspects are separate processes to the enhancing of data and the final product portrayal/encoding. This involves model generalisation and it seeks to reduce and simplify the data for the required scale. Model generalisation precedes the cartographic generalisation process, which seeks to create legible presentation for the data. The text placement can be described as a special form of symbolisation (placing labels) on the map where the text must match up the location and geometry of the map features, like rivers and roads. However, at the same time the text must be placed in such a way it doesn't cause resolution-conflicts. This process will include identifying candidate locations, evaluating those candidate locations by considering the neighbouring map features and or text features and identifying the best location of each label. This is achieved using AGENT technology. Finally, product finishing refers to the post-generalisation processes before turning the data into a cartographic end product. These include the following actions: allocation of identifiers, rounding, conformance checking, manual edits and pre-press symbolisation.

There are challenges with generalisation because the issues being addressed often involve complex interactions between many different features. For example, if a road junction is enlarged to make it clearly readable, the buildings alongside it must also be displaced retaining their relative positions to the roads. If the buildings are densely packed and there is not sufficient space for them to move into, it might be necessary to merge or delete some of them. A cartographer presented with such a situation would use his skill and judgement to produce a satisfactory solution in the vicinity of the junction. An automated generalisation system must be able to apply equivalent intelligence to this judgement and be able to reach a compromise that adequately fits the solution.

For automation to be introduced into the generalisation process, it is imperative for the source data to be in a qualitative state. It must be known to be geometrically clean and to conform to supporting business conditions. The key to achieving automated generalisation processes therefore begins with assembling and maintaining high quality source data. This is based on conformance checking using business rules, for example these could be data specifications pertaining to the data. Conformance of the data is checked against the rules. This approach helps to ascertain if rules are being broken or if there are exceptions to the rules, prior to any fix up. This process determines the fitness for purpose of the data.

A Generalisation Framework enables an automated generalisation production process offering the following advantages:

- Produce consistent and reliable results, increasing confidence in the data and the reputation of your organisation
- Quickly derive new products from existing data, reducing the time to market and opening up new sales opportunities
- Reduce maintenance costs - only one source of data must be up-to-date
- Reduce time and costs for manual checking and processing
- Free cartographers to work on mission-critical tasks
- Produce a documented set of known rules for the manufacture of small-scale products from large-scale products, saving time in recording internal processes

A Generalisation Framework approach was developed as a result of designing and implementing solutions for customers or advising potential customers regarding automation strategies. Often quality issues were encountered late in the project. When the solution was implemented and tested using an example test data set, it did not reflect the quality issues that arose when the real data was used.

Once such issues have been highlighted then it is necessary to identify:

1. Programmatic solutions to quality issues encountered
2. Interactive solutions to address quality issues

Collaboration with AdV

1Spatial has been working in partnership with AdV on one of the largest generalisation projects in Europe to improve business efficiencies in Germany's state mapping agencies. Using Radius Clarity™, 1Spatial has delivered an automatic generalisation flowline to generate digital landscape models at 1:50,000 scale. This enables AdV to derive digital topographic maps at their target scale and make time and cost savings alongside being able to release more frequent updates of their products.

AdV wanted to enable the group to release more frequent updates of their products and hoped this could be achieved by automating the process. The process of updating their products previously meant maintaining several datasets for different projects, but AdV wanted this to be limited to maintaining a single dataset (the BasisDLM). The ATKIS-GEN project aims to develop a software system that automates production of data, from BasisDLM (1:25,000) to DLM50 (1:50,000), and consequently the digital topographic maps DTK50 (1:50,000). 1Spatial has delivered an automated BasisDLM to DLM50Model and Cartographic generalisation workflow incorporating GML3 (NAS) import/export. The implementation for the German flowline processes 4.1 million objects for the state of Rheinland-Pfalz in approximately 5 days. This is a fully automatic process that matches the early stages of the solution proposed. The original success criteria required 90% automation with a 5-week generalisation process. 100% automation has since been achieved with clean source data.

The combination of both automated model and cartographic generalisation flowlines allows AdV staff to concentrate on data maintenance, reducing the amount of time between data creation and product publication, thus creating more up-to-date products. This new flowline will enable AdV to deliver products almost simultaneously, allowing AdV to avoid previous criticisms that different features were present at different scales.

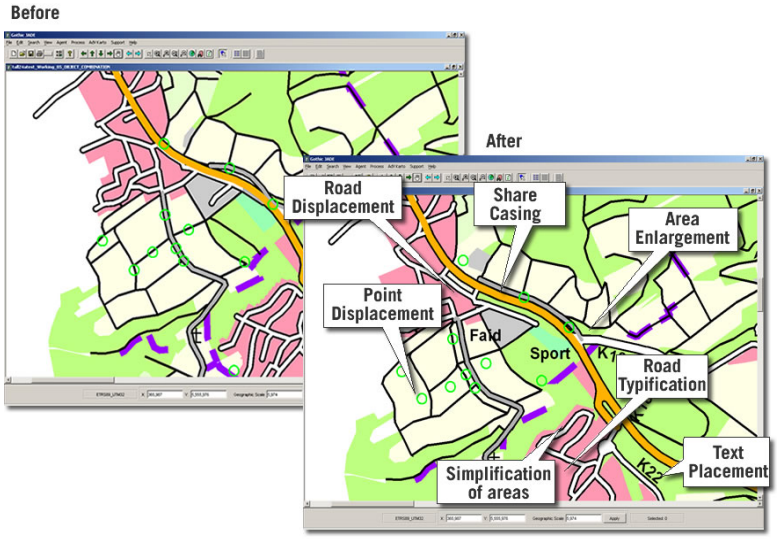


Figure 2: AdV generalisation results

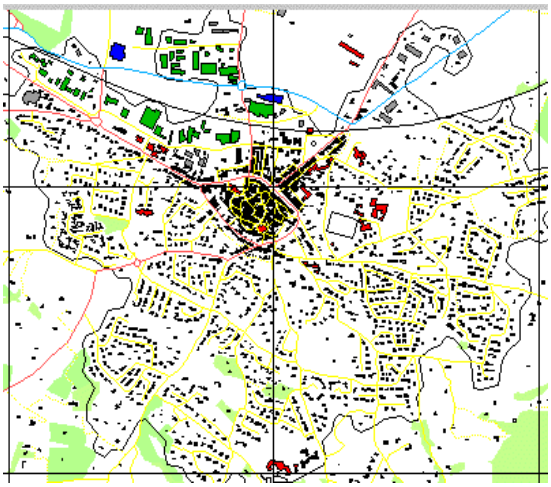
Collaboration with IGN France

In 1999, the Carto2001 project was launched by Institut Géographique National (IGN), with IGN selecting 1Spatial's automated generalisation technology to achieve its target of creating 1:100,000 mapping (Top100) from the BD Carto® database at 1:50,000. Since the BD Carto® database does not hold buildings as elementary objects, the generalisation flowline needed to focus on road networks. The project was initiated on the back of IGN's leading involvement in the European AGENT project (Automated GEneralisation New Technology).

1Spatial's solution has allowed IGN to automate generalisation in their production lines. Generalisation problems were mainly concerned with road and railway networks, and retaining consistency between the various themes of the map. The automated generalisation allows IGN to derive products from those that already exist and reduces the cost of maintaining multiple datasets, which can be costly, time consuming and produce inconsistencies in the data. COGIT Laboratories at IGN France achieved fantastic results from their Carto2001 project on automated generalisation with 1Spatial. The automated generalisation solution enabled the original estimates of 1000 hours of operator work per map to be cut dramatically to just 150 hours.

This automated generalisation technology was previously used to create 1:100,000 mapping from its database at 1:50,000 and for research and development work. Today 1Spatial will deliver a new version of Radius Clarity (v2.7), featuring algorithms for building and road generalisation that have been tailored specifically to be fit for use within an automated production flowline. It will also deliver enhancements to the current user interface and will consolidate the developments made during previous projects, into a single version of the product. A framework contract between IGN France and 1Spatial allows for a potential 100 users within the next four years.

Before generalisation



After generalisation

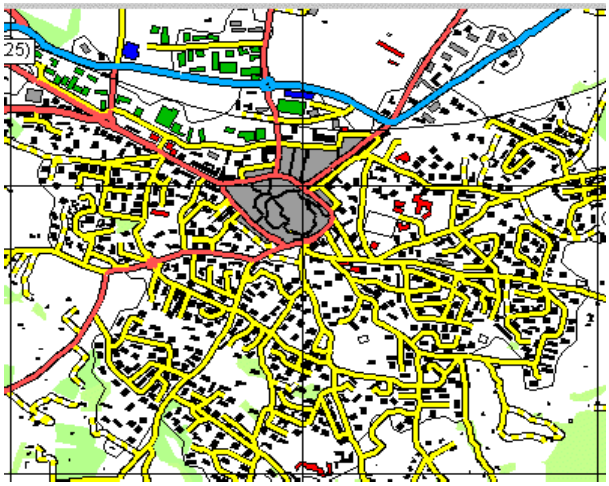


Figure 3: IGN France generalisation results

Results

The first observation from this work on data quality is that it is important to design a workflow to fix or ignore the problems once they are identified. This involves trying to identify patterns of error rather than those unexpected errors that occur due to human intervention.

The second result is that those working on generalisation should examine techniques specifically developed to examine and highlight errors that are best suited for some manual interaction or require further analysis.

Accordingly these results will highlight solutions for managing the impacts of data, which is not of suitable quality, by using practical and implemented mitigation strategies and present options for further solutions for quality control and improvement.

Conclusions

The achievement of automation is very much dependent on the quality of the information available. Due to its very nature, visual representation in cartography is vitally important, however, the factors that affect this are of equal significance. The geometric and semantic quality of the source spatial data and the quality of the specification and clear definition of the required result represent key considerations when looking at quality within the process. Whether using Commercial Off The Shelf (COTS) or bespoke solutions, achieving automated generalisation is a classic engineering task. Consequently it requires detailed requirements definition, good design, good quality source materials and continuous quality control. Arguably, the most important consideration is to define what final result is expected and to develop a programme that considers how this is to be achieved. This must highlight the importance of a data quality strategy in automated generalisation.

A Generalisation Framework facilitates a considered and structured approach to the solution. However, further thinking and research is required to address the elements that still remain difficult to automate and that, for the foreseeable future, will remain subject to manual intervention. Tackling data quality provides a good basis for such work.

References

Filippovska, Y., Walter, V., Fritsch, D., 2008. Quality evaluation of generalization algorithms, ISPRS Beijing.

Frank, R., Ester, M., 2006. A quantitative similarity measure for maps. In: Proceedings of 12th International Symposium. Progress in Spatial Data Handling, Springer.

Bard, S., 2004: Quality Assessment of Cartographic Generalization. – Transactions in GIS, Volume 8(1), Blackwell Publishing: p 63–81.

Podolskaya, E.S.,, K.-H. Anders, Haunert, J-H., Sester, M, 2008.: Quality Assessment for Polygon Generalization , Quality Aspects in Spatial Data Mining , CRC Press, Taylor & Francis Group , p. 211-220

J.-H. Haunert, J-H., Sester, M., 2008: Assuring logical consistency and semantic accuracy in map generalization , Photogrammetrie - Fernerkundung - Geoinformation (PFG) , vol. 2008 , no. 3 , p. 165-173.

Skopeliti, A., Tsoulos, L., 2001. The accuracy aspect of cartographic generalization. In: Proceedings of the GIS Research UK 9th Annual Conference GISRUK. Wales, UK

Joao E., M., 1998: Causes and Consequences of Map Generalization, Taylor & Francis

Müller, J. C., Weibel, R., Lagrange J. P., Salge F., 1995. Generalization: state of the art and issues. GIS and generalization. Methodology and practice. GIS Data I. Taylor & Francis. pp. 3 –17.