GENERALISATION AND DATA QUALITY

N. Regnauld

1Spatial, Tennyson House, Cambridge Business Park, Cambridge CB4 0WZ, UK, nicolas.regnauld@1spatial.com

Commission II, WG II/4

KEY WORDS:
Data Quality, Quality Criteria, Automatic Generalisation, Validation, Fitness for Purpose

ABSTRACT:
The quality of spatial data has a massive impact on its usability. It is therefore critical to both the producer of the data and its users. In this paper we discuss the close links between data quality and the generalisation process. The quality of the source data has an effect on how it can be generalised, and the generalisation process has an effect on the quality of the output data. Data quality therefore needs to be kept under control. We explain how this can be done before, during and after the generalisation process, using three of 1Spatial's software products: 1Validate for assessing the conformance of a dataset against a set of rules, 1Integrate for automatically fixing the data when non-conformances have been detected and 1Generalise for controlling the quality during the generalisation process. These tools are very effective at managing data that need to conform to a set of quality rules, the main remaining challenge is to be able to define a set of quality rules that reflects the fitness of a dataset for a particular purpose.

1. WHO NEEDS SPATIAL DATA QUALITY?

The quality of spatial data is important to both the data producer and the user of the data. Someone using bad quality data will encounter difficulties in using it, introducing delays and costs if the data needs fixing before it can be used. Alternatively the user can alter the process it is using to cope with the shortfalls in the quality of the data it uses. This too has a cost. Sometimes the bad quality data is not spotted straight away, but leads to bad materials being produced from it, and potentially bad decisions made as a result. For the data producer, bad quality data means unhappy customers, who come back to them with complaints. This means additional and usually unplanned work required to fix it. Data quality is therefore a major concern to both the data producer and the data user. The ISO 19157:2013 (Geographic information - data quality) standard has been defined to facilitate the description and evaluation of quality of spatial data. The quality criteria defined by this standard are limited to checking that the data is geometrically and topologically clean, geometrically and semantically accurate, and does not contain omissions or commissions. This is however not sufficient to describe the quality of generalised data aimed for mapping, as these need to be readable at a given scale. For example, nothing in the ISO 19157:2013 checks that features are large enough to be seen on the target map or that the cartographic symbol of a road won't obliterate a building alongside it.

In this paper we discuss both the impact of data quality on generalisation and the impact of generalisation on data quality. We describe how 1Spatial products are dealing with data quality, to assess it, improve it and maintain it during processes such as generalisation. This leads us to a discussion on quality criteria, and the open challenge of finding a set of quality criteria that can reflect the fitness for purpose of a set of data.

2. DATA QUALITY AND GENERALISATION

Generalisation is a process that transforms data and is affected by data quality in two ways; Firstly it relies on the quality of the source data to operate properly: the source data must be suitable for the generalisation process. Secondly it needs to deliver generalised data fit for a purpose. The quality of the generalised data is an evaluation of how closely the data fulfils the specified requirements.

The simplest way to ensure that a dataset is suitable for a generalisation process is to design the generalisation process to cope with the state of the data, improving or enriching the data along the way when needed. For example, if the source data contains a river network which is not properly connected (no links for sections covered my manmade features, as often happens in urban environment), it will be difficult for a generalisation process to identify which river sections are part of the main channel. This could result in sections of the main channel being removed because they are small and seem isolated. One way to overcome this is to add a data enrichment process that will automatically look for breaks in the network and automatically deduce the missing links, sometimes using additional data sources. In a similar way, the easiest way to make sure the generalised data is fit for purpose, is to design a specific process. This is the approach which was followed for all the existing systems performing automatic generalisation in production today, like the production of OS VectorMap District at Ordnance Survey GB (Regnauld et al. 2013), the production of the 1:25000 maps at IGN France (Maugeais et al. 2013), the production of 1:50000 maps at the Dutch Kadaster (van Altena et al. 2013), or the production of 1:50000 and 1:100000 Maps in some German Länders (Urbanke and Wiedemann 2014). These bespoke systems are costly to develop and difficult to maintain or to reuse. The process needs to be heavily updated if the source data changes or if the required characteristics of the output changes.

Many studies to overcome this problem have been conducted, with the aim to build a system that can take formal descriptions of the requirements as input, and automatically adapt the process to achieve it. Research has been conducted in many relevant directions: how to formalise requirements and how to capture them. Assuming these requirements are available in a machine readable way, how can they be used to automatically adapt the generalisation process? Constraint based systems (Burghardt et al 2007), optimisation techniques (Sester 2005), multi agent model (Ruas and Duchêne 2011) among others have been proposed. A comprehensive review on the research

available to date on these subjects can be found in (Burghardt et al 2014). As noted by the editors of the book in their concluding section, on-demand mapping remains one of the main challenges faced by the community of researchers on the subject.

2.1 Quality management during generalisation: 1Generalise approach

At 1Spatial, we also believe that we are not yet ready to produce a system that can automatically adapt to the source data and a formal description of the expected result. We definitely see it as an exciting area of expansion in the future though. We do have an Agent-based optimisation technology available in 1Generalise that can be used, but we are lacking the formal description of expected generalised results to use it effectively. 1Generalise provides predefined generalisation processes based on reusable components, working in both nationwide creation and change only update modes. The whole system is designed to facilitate the quick reuse of parts of existing systems (simple or complex) to build new ones. This also provides the building blocks to go further in the future and add to these components, the metadata that will allow more intelligent mechanisms to use them when they are relevant, progressing towards a system to build maps on demand.

In its current version, 1Generalise encapsulates a full generalisation process in a Flowline. Typically a Flowline will be built to derive a specific type of map, for example a general purpose topographic 1:25k map. The default Flowlines coming with 1Generalise are generic, they are data provider independent. They can then be extended or modified by a Flowline designer to meet their specific requirements. These Flowlines are made of a sequence of Subflows. Each Subflow is a sequence of Steps. Both Subflows and Steps are reusable components. A Subflow typically contains the steps required to generalise a theme. For example, a Subflow to generalise vegetation could be made of three steps: 1) amalgamate touching polygons, 2) remove small holes and 3) simplify the outline. Each of the steps encapsulates the logic that decides where the generalisation algorithms should be applied and validates the results. In this example, the amalgamation step would identify adjoining vegetation features and perform the amalgamation of their geometries. This step could also contain a rule that prevents the amalgamation of the two adjoining vegetation features if a fence runs along their boundary. The third step of this Subflow, in charge of simplifying the outline, could have a rule that checks that the result does not intersect other features and if so, either apply some corrective action or reject the result and simplify only parts of the features which are not close to the conflict area. The steps are editable using a rule language, so that the logic used to trigger algorithms and validate their results can easily be changed by the Flowline designer, without having to write code in a programming language like C or Java.

In this way, the steps can encapsulate pre and post conditions that allow the Flowline designer to integrate quality checks, and possibly corrective actions at any stage in the generalisation process. The actual generalisation algorithms, usually coded in C or Java, are triggered by these rules, so the Flowline designer has a high level of control over them without needing to understand how they are implemented.

2.2 Quality management pre and post generalisation: 1Validate and 1Integrate approaches

While managing the quality during generalisation is a good idea, it is not always possible or practical. Usually, during the generalisation, after a specific generalisation algorithm, we add quality checks to trap known potential side effects of the algorithm. If we were to track all possible problems after each application of each algorithm, the performance would quickly become unacceptable. Similarly, checking the conformance to standard for the input data for each algorithm would result in the same hit on performances. Not performing systematic full checks opens the door to problems though, as a slightly odd result from one algorithm ends up as input to the next which can then produce unexpected results. So it makes sense to have an initial pre-process that checks that the source data is of suitable quality, and a post-process that checks that the result meets the requirements. These are completed by targeted local checks performed within generalisation steps.

Both initial and final quality checks can be handled by 1Validate, 1Spatial's dedicated software to check the conformance of a dataset against a set of rules. These rules can be authored by the user, using a rule editor and a number of functions performing geometric and topologic checks. The screenshot in Figure 1 shows a 1Validate rule that checks that two distinct buildings do not intersect unless they simply touch (one special case of intersection where the interior of the geometries do not meet). The rules used in 1Generalise use the same syntax.

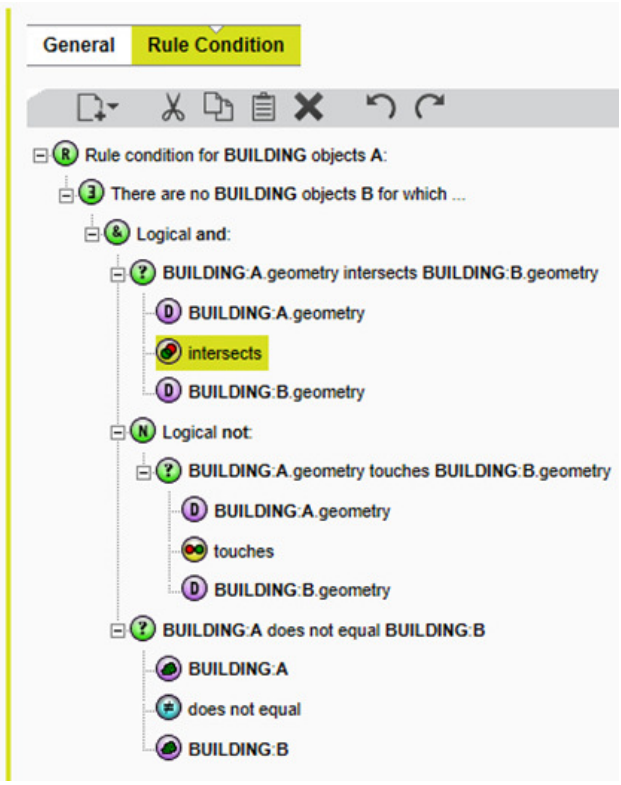

Figure 1: 1Validate rule editor

When validation failures are found in the source dataset, it must be corrected. This can either be done by manual intervention, using an editor, or sometimes automatically. For fixing issues automatically, 1Integrate can be used. 1Integrate is an extension of 1Validate that lets the user write actions, using the same but extended rule language as in 1Validate. The same rules can be

written in 1Integrate, and they can be mapped to corrective actions. So for each object in the database that does not pass the rule, the mapped action will be triggered on it. These are particularly useful to fix topological issues, for example when network edges don't exactly meet, creating little gaps or overshoots. These are often not identifiable visually, unless using a level of zoom which will make the analysis of the whole dataset unpractical. While these issues are difficult to see, they will disrupt processes that perform network analysis tasks. For example, a generalisation process focusing on pruning a path network will usually preserve paths which are part of long routes. Bad connectivity could make the algorithm remove sections apparently disconnected to the rest while they are in fact important links in the network. 1Integrate can fix such issues in two steps. First a rule identifies all edges that end very close but not on another edge. For each instance of this, it triggers an action that moves the culprit edge end on the closest position on the nearby edge.

For validation failures on the generalised data, correction is rarely possible without the source data as reference. Usually, if the evaluation of the generalisation is not satisfactory, then either the generalisation process needs to be altered to avoid creating the issue, or a manual editing process can be added to correct the issue. Which way is the most cost effective depends on the frequency with which the problem occurs, the complexity of fixing the generalisation process to prevent it, and the cost of fixing it manually for the life time of the system.

3. QUALITY CRITERIA

While we have tools to express quality rules and to evaluate them against a set of data, identifying what quality criteria are required to ensure the input or generalised data are fit for purpose remains a challenge.

3.1 For input data

The ISO 19157:2013 standard categorises the quality criteria into five main families: Completeness, Topological consistency, Positional accuracy, Temporal consistency, Thematic accuracy. It also proposes methods to measure these. The main difficulty is to aggregate these measures in a way that can help decide if a set of data is fit for a particular use. While the criteria related to clean geometries (no double points, well formed polygons, no self intersection, etc.) and topology (no overshoots or undershoots at junctions, no overlaps, no double points; well formed polygons, no self intersection, etc.) can fairly easily be checked and sometimes automatically fixed, the others are more difficult to check, and would require a reference dataset. For example, looking at a single dataset, it is often impossible to detect that a feature is missing or misclassified. Exceptions exist, when semantic inconsistencies can be detected, for example, when a section of river has been misclassified as a road. The combination of the fact that a section of road is isolated from the rest of the road network and the fact that this section also connects two dead end river nodes could lead to the conclusion that this section of road has been misclassified and should be reclassified as a river. Writing such rules is extremely time consuming though, as combinations of potential misclassifications and the context in which they occur are almost endless. Such rules are worth writing once we have identified that a particular type of misclassification occurs frequently. Some simpler types of checks can still be done though, like checking that all the features have a valid height attribute if one is expected.

3.2 For generalised data

While the above criteria are also relevant to generalised data, interpreting the results can be difficult. By its nature, generalisation will reduce the content, so completeness will be affected. Formulas like the ones proposed by (Töpfer and Pillewizer 1966) can be used as a guide to how much features should be preserved. This is by no means a universal law, the thematic focus of the map can have a strong influence on the ratio of features that should be preserved. Positional accuracy is also affected by generalisation, as map features sometimes need to be exaggerated and displaced to make them readable at a smaller scale. So accuracy must be interpreted according to the scale of the map.. However, these criteria designed for reference data are not enough for evaluating the quality of generalised data. The main concern is to make sure that the characteristics of the data meet the requirements of their expected use. In particular, for generalised data aimed at mapping, the readability of the map also needs to be evaluated (Burghardt et al 2008).

For generalised data, the same set of criteria is therefore relevant but needs to be extended by a number of additional ones, which vary based on the expected use of the product. [João 1998] provides a complete review on the effect of generalisation, which is a good source of information to define additional criteria.

For a map product, criteria must be defined to reflect the readability of the map (minimum size, width of features, minimum distances between features, maximum density of features). This may not be sufficient though, a readable map might have been overly generalised and not contain the information required for its intended use. This requires additional criteria to be defined, they could relate to the positional accuracy of the features, criteria for their selection (based on individual characteristic or density measure over a given theme), the level of detail required (geometric and semantic). The difficulty is to find a set of criteria that reflects a specific targeted use and can be interpreted. Generalisation is an abstraction of the reality that produces a dataset which is a trade-off between the preservation of the reality and the readability of the result. What trade-off is adequate for what usage?

4. CONCLUSION

1Spatial has many experts in spatial data management and data quality. Through the 1Spatial Management Suite, 1Spatial proposes a set of software to capture spatial data, maintain a spatial database, integrate data from several sources, generalise data and publish map. In all these steps, validation plays a key role to ensure that the quality of the data is always under control.

Evaluating the quality of generalised data is an area where a lot more could be done. We are able to let our users define the criteria that they want and check them, but we would like to propose a set of predefined standard criteria that collectively provide a good insight on the quality of the generalised dataset. This could then be tweaked and extended to satisfy the specific requirements of each customer. Such standard criteria could then be integrated with the generalisation process itself, to progress towards the goal of building a system capable of performing on demand mapping. This is an active field of research. A workshop was organised in March 2015 by the International Cartographic Association commission on Generalisation and Multiple Representations to study the use of

ontologies to formalise the knowledge required to support on demand mapping [Mackaness et al. 2015]. We did participate in this workshop and are most interested in pursuing this collaboration. One of its outcomes could be the definition of standard quality criteria and user requirements for generalisation.

REFERENCES

Burghardt, D., Duchêne, C., Mackaness, W. (Editors), 2014. *Abstracting Geographic Information in a Data : Methodologies and Applications of Map Generalisation*. Springer, Lecture Notes in Geoinformation and Cartography, Publications of the International Cartographic Association (ICA).

Burghardt, D., Schmid, S., Stoter, J. 2007. Investigations on cartographic constraint formalization. *In Proceedings of the 11th ICA workshop on generalisation and multiple representation*. Moscow, Russia, 2007.

Burghardt, D., Schmid, S., Duchêne, C., Stoter, J., Baella, B, Regnauld, N., Touya G, 2008. Methodologies for the evaluation of generalised data derived with commercial available generalisation systems. *In Proceedings of the 12th ICA workshop on generalisation and multiple representation.* Montpellier, France, 2008.

João, E.M., 1998. Causes and consequences of map generalization. London: Taylor & Francis

Mackaness, W., Gould, N., Bechhofer, S., Burghardt, D., Duchene, C., Stevens, R., Touya, G., 2015. Thematic workshop on building an ontology of generalisation for on-demand mapping. http://generalisation.icaci.org/images/files/workshop/ThemWorkshop/ThematicOntologyOnDemand_Paris2015.pdf

Maugeais, E., Lecordix, F., Halbecq, X., Braun A., 2011. Dérivation cartographique multi échelles de la BDTopo de l'IGN France: mise en œuvre du processus de production de la Nouvelle carte de base. *In Proceedings of the 25th International Cartographic Conference*, Paris, France, July 2011.

Regnauld, N., Lessware, S., Wesson, C., Martin, P., 2013. Deriving products from a multi-resolution database using automated generalisation at Ordnance Survey. *In Proceedings of the 26th Internationall Cartographic Conference*. Dresden, Germany, August 2013.

Ruas, A., Duchêne, C., 2011. A Prototype Generalisation System Based on the Multi-Agent System Paradigm. *In Generalisation of Geographic Information : Cartographic Modelling and Applications*, Mackaness, Ruas and Sarjakoski eds, pp. 269-284.

Urbanke, S., Wiedemann, A. , 2014. AdV-Project « ATKIS : Generalisation » - Map Production of DTK50 and DTK100 at LGL in Baden-Wurttemberg. *In Abstracting Geographic Information in a Data Rich World*. Burghardt, Duchêne and Mackaness Eds, Springer, pp. 369-373.

Sester, M., 2005. Optimization approaches for generalization and data abstraction. *International Journal of Geographic Information Science*, 19 (8-9), pp. 871-897.

Töpfer, F., Pillewizer, W., 1966. The principles of selection: a means of cartographic generalization. *The Cartographic Journal*, 3(1), pp. 10-16.

Van Altena, V., Nijhuis, R., Post, M., Bruns, B, Stoter, J., 2013. Automated generalisation in production at Kadaster NL. In *Proceedings of the 26th International Cartographic Conference*, Dresden, Germany, August 2013.