Towards Quality-Aware GIS: Operation-based retrieval of spatial data quality information

Rodolphe Devillers[1,2], Amin Zargar[1]

1 Geography, Memorial University of Newfoundland, rdeville@mun.ca
2 Centre for Research in Geomatics, Laval University

Abstract

Despite the fact that all geospatial data are, at different levels, imperfect, GIS technologies do not offer much support to GIS users to make them aware of the imperfections and potential limitations in the use of these data. Recent research progresses have been made towards the design of GIS that would take into account data quality. These approaches extend typical GIS functionalities to manage, use and communicate information describing spatial data quality. This paper presents a project that aims at linking quality information with GIS operations to provide users with the most relevant quality information.

Background and Relevance

All geospatial data are at different levels imperfect (Goodchild, 1995, Devillers and Jeansoulin, 2006). This results from a number of factors, such as human errors in the recording of the data or interpretation of the world features, errors arising from the manipulation of the data, representation of complex phenomena with an inappropriate data model, imprecision of the equipment used to measure the phenomena, inappropriate sampling, etc. To record this information and release it to GIS users, a number of geospatial data producers provide metadata (i.e. data about data) that describe some aspects of the data quality of the datasets. Such metadata are through to be helpful to the users for assessing the fitness of the data for their specific usage. Standardization bodies, such as the Federal Geographic Data Committee (FGDC) in the USA or the International Standard Organisation (ISO) internationally, provide standards for documenting such metadata. These metadata are however criticized by a number of GIS experts for different reasons. They are often too technical to be understood by the intended audience (Timpf *et al.*, 2006, Boin and Hunter, 2006), they do not describe all the types of data quality issues (Comber, 2007), they are usually too general (e.g. describe the dataset as a whole and not individual objects) (Devillers, 2005, Sadiq, 2008), they are rarely linked properly to the data they describe, etc. In practice, metadata end-up to be ignored or not understood by the majority of GIS users. In addition, to encourage their adoption, metadata standards request only few metadata to be documented (ISO TC/211, 2003). These minimal metadata are often referred as “discovery metadata” and typically include, for instance, the spatial and the temporal extents of the dataset and the dataset title and language, but exclude all aspects of data quality information. As a consequence, most metadata files provide no information about data quality at all and GIS users are left without any information about the quality of the data they are using. Some commercial GIS (e.g. ArcGIS) now provide the ability to manage, update and display metadata. However, these metadata are hard to find for users with little expertise in GIS and also lack of any real connection with GIS operations. Metadata are stored but do not affect the functioning of the GIS. For instance, the result of a GIS operation (e.g. a distance measurement between two objects on a map) will not change if the user has data of very good quality or data of very poor quality. In a context where geospatial information gets increasingly used by the

general public, the fact that GIS work as if they were handling perfect data can result in poor decisions being made.

For these reasons, a number of research projects have been directed towards improving GIS to have them provide additional functionalities related to spatial data quality. Some work focused on improving the management of data quality information and allowing, for instance, its storage using different database models and at various levels of details (e.g. Devillers *et al.* 2005, Sadiq, 2008). Other works looked at communicating quality information to users through different visualization techniques, such as thematic maps of certain aspects of data quality (e.g. see Beard and Buttenfield, 1999 for an overview of such approaches), display of quality indicators (Devillers *et al.* 2005, Huth *et al.* 2007), display of warning messages (Reinke and Hunter, 2002, Levesque *et al.*, 2007).
This paper focuses on an ongoing project that looks at relating the data quality information stored in the metadata to GIS operations.

Method and Data

A survey of data quality elements suggested by metadata standards has been performed in order to select the metadata that would be communicated to the GIS users. Data quality elements and some of the sub-elements provided by the ISO 19115 have been selected, although other elements could easily be added to the list if one wanted to cover quality issues not described by these standards. Linking the right quality information to a specific GIS operation implies identifying which quality elements are of relevance for the operation. This task has been done by an expert in spatial data quality, with the support of existing literature on uncertainty propagation related to the operations. For instance, some authors described the impact of data uncertainty on specific GIS operations such as overlay and buffer analysis (Veregin 1994, 1995). Relationships between data quality elements and operations were stored in a separate database that is accessed by the extended quality-aware operation. An extension to the current GIS has been designed to be able to add a layer between the metadata and the GIS operators. This new layer allowed these two components to interact more dynamically.

Experimentation

The method described in the previous section has been tested using a software prototype developed as an add-on to the software ArcGIS. The “distance measurement” and the “select by attribute” functions available from most GIS have been modified into quality-aware operations that can query the metadata as they operate. The new operations have been programmed in Visual Basic and can replace the existing GIS operations. When using these new operations, the program checks for the relevant quality information related to the given operation (e.g. spatial accuracy will be important for a “distance measurement”, while attribute accuracy may be more important for a “select by attribute” operation). It then retrieves the relevant metadata from the xml files that store metadata in ArcGIS and finally displays metadata into an interface that appears together with the result of the operation. Users can then see the output of their operation at the same time they visualize relevant quality information related to this output. Topographic data and their associated metadata provided by Natural Resources Canada (NRCan) have been used for the test.
The usefulness of the prototype has been validated through tests done by about 30 GIS users having various levels of expertise. Users had to complete a test using the existing GIS operation and the same test using the modified quality-aware operation. For each test, users had to answer questions about the quality of the data they were using. The prototype has proved to improve significantly user’s awareness of quality issues.

Conclusions

This paper presented research advances towards GIS that could, at least partly, take into consideration the quality of the data being used. We described a project that looked at improving the connection between the data quality information stored in metadata and the GIS operations. A theoretical model has been presented that links metadata and GIS operations. Then, a software prototype has been described that tested the method developed. This prototype has been tested with a number of GIS users and has proved to be very useful to raise the awareness of quality issues among GIS users. Users, instead of having to look for the metadata, which was something they were rarely doing, can get access to some relevant quality information more proactively at the time they use an operation.

If this project presents encouraging results, it is fair to say that providing a fully quality-aware GIS system is an ideal goal that may never be completely reached. Such a prototype does not find an answer to all the limitations related to existing metadata but covers some of them. Also, if some operations could be easily modified to link with quality information, other ones may be too complex to be modified. Such approach also relies on metadata which are sometime absent and often incomplete. However, providing even part of these capabilities should increase the awareness of users and hence potentially decrease the risks of having them use data in an inappropriate manner.

References

Beard, M. K. and B. P. Buttenfield, 1996 “Detecting and evaluating errors by graphical methods”, in Geographic Information Systems: Principles and Technical Issues, vol. 1, P. Longley, M. Goodchild, D. Maguire, and D. Rhind, Eds., John Wiley, New York, pp. 219–233.

Boin, A. T. and G. J. Hunter, 2006 “Do spatial data consumers really understand data quality information?”, 7th international symposium on spatial accuracy assessment in natural resources and environmental sciences, M. Caetano and M. Painho, Eds., Instituto Geografico Portugues, Lisbon, pp. 215–224.

Comber, L., Fisher, P. and R.A. Wadsworth, 2007 “Approaches for providing user relevant metadata and data quality assessments”, Proceedings of the 2007 GISRUK Conference, Ireland.

Devillers, R., Bédard, Y. and R. Jeansoulin, 2005 “Multidimensional management of geospatial data quality information for its dynamic use within Geographical Information Systems”, Photogrammetric Engineering and Remote Sensing, vol. 71, no. 2, pp. 205–215.

Devillers, R. and R. Jeansoulin (Eds), 2006 Fundamentals of Spatial Data Quality. ISTE, London, UK, 312p.

Devillers, R., Bédard, Y. Jeansoulin, R. and B. Moulin, 2007 “Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use of Spatial Data”, International Journal of Geographical Information Science, vol. 21, no. 3, pp. 261–282.

Goodchild, M.F., 1995. Sharing Imperfect Data. In Sharing Geographic Information (H. J. Onsrud, and G. Rushton, eds), Rutgers University Press, New Brunswick, NJ, pp. 413-425.

Huth, K., Mitchell, N. and G. Schaab, 2007 "Judging and Visualising the Quality of Spatio–temporal Data on the Kakamega–Nandi Forest Area in West Kenya", 5th International Symposium on Spatial Data Quality (ISSDQ 2007), June 13–15, ITC, Enschede, The Netherlands.

ISO TC/211, 2003. Geographic Information – Metadata 19115. Standard published by the International Standard Organisation.

Levesque, M.-A., Bédard, Y., Gervais, M. and R. Devillers, 2007 Towards managing the risks of data misuse for spatial datacubes. International Symposium on Spatial Data Quality (ISSDQ 2007). Enschede, The Netherlands, June 14-15th.

Qiu, J. and G. J. Hunter, 2002 "A GIS with the Capacity for Managing Data Quality Information", in Spatial Data Quality, W. Shi, M. F. Goodchild, and P. F. Fisher, Eds., Taylor & Francis, London, pp. 230–250.

Reinke, K. J. and G. J. Hunter, 2002 "A Theory for Communicating Uncertainty in Spatial Databases", in Spatial Data Quality, W. Shi, P. F. Fisher, and M. F. Goodchild, Eds., Taylor & Francis, London, pp. 77–101.

Sadiq, Z. 2008. Modeling Spatial Variation of Data Quality in Databases. PhD thesis, Department of Geomatics, The University of Melbourne, Australia, 146p.

Timpf, S., Raubal, M. and W. Kuhn, 1996 "Experiences with metadata", in Advances in GIS Research II: Proc. 7th International Symposium on Spatial Data Handling, Delft: Faculty of Geodetic Engineering, Delft University of Technology, II:12B, pp. 31–43.

Veregin, H., 1994 "Integration of simulation modeling and error propagation for the buffer operation in GIS", Photogrammetric Engineering and Remote Sensing, vol. 60, no. 4, pp. 427–435.

Veregin, H., 1995 "Developing and testing of an error propagation model for GIS overlay operations", International Journal of Geographic Information Sciences, vol. 9, no. 6, pp. 595-619.