7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.
Edited by M. Caetano and M. Painho.

Modeling spatial variation in data quality using linear referencing

MGSM Zaffar Sadiq[1], Matt Duckham[2] and Gary J Hunter[2]

1 Cooperative Research Centre for Spatial Information, Department of Geomatics,
University of Melbourne, 723, Swanston Street, Victoria - 3010, Australia.
Tel.: + 061 3 83443177; Fax: + 061 3 93495185
s.mohamedghouse@pgrad.unimelb.edu.au

2 Department of Geomatics,
University of Melbourne, Victoria - 3010, Australia.
Tel.: + 061 3 83446935; Fax: + 061 3 93472916
mduckham@unimelb.edu.au, garyh@unimelb.edu.au

Abstract

Spatial data quality (SDQ) is conventionally presented in the form of a report. The data quality statements in the report refer to the entire data set. In reality the quality of data varies spatially due to data collection methods, data capturing techniques, and analysis. Thus, the quality of spatial data for one area may not be applicable to spatial data describing other regions. The present systems for reporting and representing SDQ (data quality statements) cannot address the data user's requirements as they are not location specific. Consequently, conventional approaches to SDQ that ignore variation in quality within a data set impair the data producer's ability to correctly communicate knowledge about data quality and jeopardize the user's ability to assess fitness for use. To enable proper communication of SDQ, spatially varying data quality needs to be represented in the database. This paper discusses the representation of spatial variation of data quality in spatial databases using three models: per-feature, feature-independent, and feature-hybrid. In the per-feature model, quality information is stored against each spatial feature (object) stored in the database. In the feature-independent model, quality information is stored independently of particular features (as a field). The feature-hybrid model is derived from a combination of per-feature and feature independent models. One example of an existing data management technique that can be adapted for use as a feature-hybrid model is linear referencing. Applying linear referencing in this way is a new approach to representing spatial variation in quality. The paper concludes with a review of the relative merits of the different strategies for storing spatially varying data quality information.

Keywords: spatial data quality, uncertainty, metadata, sub-feature variation, linear referencing

1 Introduction

1.1 Spatial variation in data quality

Spatial variation is a feature of data quality that is unique to spatial data sets. The quality of spatial data for one area may not be applicable to spatial data describing other regions (Wong and Wu, 1996). Clarke and Clark (1995) state that different data capturing techniques, like field surveys, global positioning systems, aerial photography, satellite imagery, existing maps, and other documentation, are one of the key factors affecting spatial variation in spatial data quality (SDQ). According to Wong and Wu (1996), the quality of data varies spatially due to problems in data collection (for example, when socio-economic data is gathered by sampling and not all regions are sampled to the same extent), data capture (for example, if cloud cover exists in a remotely sensed image, the data for certain areas covered by clouds in the database

may be less accurate than the data describing other regions), compilation, analysis, and representation.

1.2 Motivation

Spatial data quality is conventionally presented in the form of a report. The data quality statements in the report refer to the entire data set. The data users determine the usability of the data based on these statements. The users of a data set need to use information about spatial data quality in order to be able to assess the "fitness for use" of the data set (Chrisman, 1984). In the Australian context, ANZLIC (Australia, New Zealand Information Council) identified that one barrier to implementing an Australian Spatial Data Infrastructure (ASDI) is that the spatial industry has to overcome the problem of incomplete knowledge about the availability and quality of existing spatially referenced data (ANZLIC, 2003). The producers of a data set need to effectively communicate the quality of their data for a variety of reasons, including competitiveness, data warehousing, liability, and litigation. With respect to liability and litigation, it has been noted that:

> "It becomes imperative, from a liability perspective to check the positional and attribute accuracy, logical consistency, resolution, completeness, timeliness and lineage of data." Graham (1997).

Consequently, conventional approaches to SDQ that ignore variation in quality within a data set impair the producer's ability to correctly communicate knowledge data quality and jeopardize the user's ability to assess fitness for use. Hunter et al. (2005) remarks that updated data often has better accuracy than the data it replaces as a consequence of new positioning technologies used for cadastral upgrading or DEM capture. Therefore, the clients will be interested to know where more accurate data lies within the data set. As a specific example of the importance of spatial variation in data quality, the Victorian Department of Sustainability and Environment (DSE, the government department which is responsible for maintaining and updating spatial data for the state of Victoria) have identified that the positional accuracy of their data is *higher* in some locations than the reported accuracy (Ramm, 2005). As a result, there is a need to communicate spatial variation of SDQ. To communicate spatial variation it is first necessary to represent it in the database.

2 Spatial data quality representation

Spatial data quality has been an important area of research in the topic of geographic information science (GISc) for many years. This section reviews the main GISc approaches to representing spatial data quality.

2.1 Elements of spatial data quality

The overriding majority of work into spatial data quality takes as its starting point the five elements of spatial data: lineage, positional accuracy, attribute accuracy, logical consistency and completeness. Although the five elements of spatial data quality form the foundations for data quality representations, one criticism of this representation is that it is not exhaustive (cf. Worboys and Duckham, 2004, chapter 9). A further criticism is that the elements of spatial data quality may not be exclusive, in the sense that some elements overlap, and that it is not always possible to distinguish between different elements. For example, for some data it may be unclear whether an attribute has been recorded incorrectly at the correct location (attribute accuracy) or correctly recorded at the wrong location (positional accuracy) (Chrisman, 1984).

Despite their limitations, the elements of spatial data quality have gained widespread acceptance and over the years have formed the basis of most national and international spatial data quality standards (Moellering, 1997), including the US Federal Geographic Data Committee (FGDC) spatial data transfer standard (SDTS) and the ISO geographic information meta data standard (ISO/FDIS 19115, 2005). The standard organizations mainly focus on developing standards to document and represent SDQ in the database, but ignore issues connected with the spatial variation of SDQ.

2.2 Related works

Some models for storage of spatially varying spatial data quality exist. These models can be classified according to whether they adopt an object-based or field-based approach, and include: reliability diagrams (object-based), object-oriented techniques (object-based), variability diagram (field-based), the data quality matrix (field-based), and quadtrees (field-based).

Cartographers have for a long time used symbols to depict uncertain contours as reliability diagrams (Fisher, 1991). Geologists represent variation in the quality of knowledge of geological features with the help of reliability diagrams (Hunter and Goodchild, 1996). The United States Department of the Interior Bureau of Land Management has developed a database called the Geographic Coordinate Data Base (GCDB) to communicate the reliability of the features by means of reliability diagrams (United States Department of the Interior Bureau of Land Management, 2001). The approach illustrates that spatial variation can be represented by storing an additional attribute called reliability value in the attribute table. However, the approach lacks the ability to show variation within individual features. Qiu and Hunter (2002) have stored data quality information at multiple-levels within a hierarchical structure, using object-oriented techniques. Qiu's method of storing quality at feature level represents spatial variation but it is limited in its representational capabilities, since it cannot represent variation in quality within an object. Similarly, Duckham (2001) develops a formal model of object-based variation in spatial data quality, using object calculus. Although the technique does offer the potential to store sub-feature variation, the practical efficiency of storage and querying using this method is open to question.

Soil scientists have used separate variability diagrams to depict the variation in homogeneity of soils (Maclean et al., 1993). As variability diagrams were stored as a separate layer in the database, overlay operations are required each time the information is retrieved. Hetrick (1991) proposes the development of a data quality matrix for raster data in which information can be coded and stored to represent missing data. As a result of storing SDQ in each pixel, the approach increases the volume of the stored data significantly. Beard et al. (1991) have discussed the possibility of storing spatially variable data quality information in a manner similar to the quadtree method. The approach uses a quadtree to recursively subdivide cells into quadrants such that each quadrant has a homogeneous level of quality. The limitations of the method are that each time a feature is updated, the corresponding structure of the quadtree also has to be modified. Thus, data sets which need frequent updating are not suited to using the quadtree method. Moreover, if the features in the raster are non-areal, like points and lines then the quadtree is acknowledged to be inefficient for data storage.

3 Spatial variation models

A few models and some limited storage capabilities already exist for spatial variation in spatial data quality (see section 2.2). Each of these models have their limitations, and no model can claim to efficiently represent and store spatially varying spatial data quality across a range of spatial data types. One of the critical question that is not adequately addressed by these modes is the representation of *sub-feature variation* (variation in quality within a geographic feature). Sub-feature variation is not an issue for raster-based data structures, since each cell in a raster represents an atomic unit of space. Similarly, sub-feature variation is not a problem for points in vector-based spatial data. However, potentially lines and polygons in vector data sets can exhibit sub-feature variation in data quality, and it is this issue that we pay particular attention to in the remainder of this paper.

Given the importance of features and sub-feature variation, three different models of spatial variation in data quality have been identified and defined: per-feature, feature-independent, and feature-hybrid. Quality information is stored against each feature in the per-feature model. In the feature-independent model, quality information is independent of the feature. The feature-hybrid is derived from a combination of the other two models.

3.1 Per-feature

The spatial data quality of a feature can be stored along with the feature as an additional attribute as for many of the object-based models of spatial variation in data quality discussed above (e.g., Hunter and Qiu, 2003; United States Department of the Interior Bureau of Land Management, 2001). Per-feature (object based) quality can be modeled as a function $f: O \rightarrow Q$ where O is a set of objects (features) and Q is a quality codomain. Features can include points, lines and polygons. The quality codomain might comprise information about any of the elements of spatial data quality (e.g., Q might represent the set of all positional accuracy values for features in a data set). Because each feature maps to a single quality value, the per-feature model cannot represent variation *within* a feature (sub-feature variation).

3.2 Feature-independent

The second option is to store spatial variation in spatial data quality as a separate theme or layer in a spatial database, as for many of the field-based models of spatial variation in data quality discussed above (e.g., Maclean et al., 1993; Heuvelink, 1996), termed here the feature-independent model. Feature independent (field-based) quality can be modeled as a function $g: S \rightarrow Q$ where S is the spatial framework and Q is the quality codomain. Spatial objects, such as lines and polygons are not explicitly represented in this model. Instead, data quality is represented as a field, independent of features in the data set. Although sub-feature variation can be represented in the feature-independent model, the model may require additional storage space to store quality information. Further, in order to retrieve sub-feature variation, it is necessary to perform spatial joins within stored spatial features.

3.3 Feature-hybrid

Finally, the third option, termed the feature-hybrid model, is to store quality information on a per-feature basis, but augment the stored quality with some additional spatial structure. Feature-hybrid (object-and field-based) quality can be modeled as a function $h: O \rightarrow Q^S$ where Q^S denotes the set of functions (fields) that map from S to Q, i.e., $Q^S = \{f \mid f: S \rightarrow Q\}$. One example of a technique that can be used as the basis of a feature-hybrid model is linear referencing. Applying linear referencing in this way, examined in the following section is a

new approach to represent spatial variation in quality, although linear features (lines and polygon boundaries) can be represented by this model.

4 Implementation

To illustrate, the three models to represent spatial variation of spatial data quality for vector data were implemented as a prototype using ArcGIS 9.0 software.

4.1 Scenario

Consider a highway that was constructed in stages. Data about the highway in the database may have varying positional accuracy at each stage of its construction. Here we assume that the highway has three varying positional accuracies of 1m, 3m and 5m across four segments. In our example, the first and third segments have 1m accuracy, the second segment has 3m accuracy, and the fourth segment has 5m accuracy. To represent the spatial variation of positional accuracy in the highway three models discussed (section 3), per-feature, feature-independent and feature-hybrid were tested.

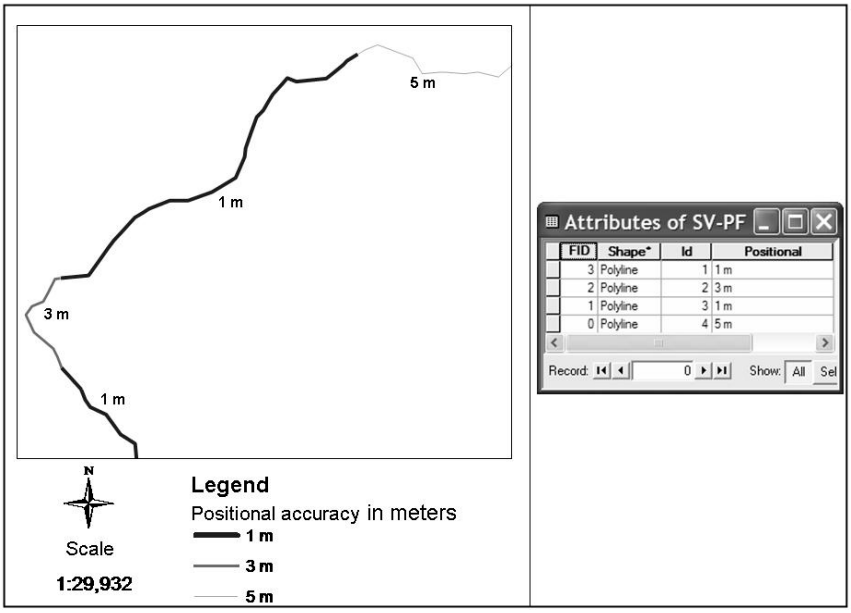

Figure 1 Per-feature model for linear features.

4.2 Per-feature model

In the per-feature model the positional accuracy of the highway can be stored as an additional attribute called “Positional” (Figure 1) in the highway (SV-PF attribute table) table. As per the scenario, the single record of highway feature has to be segmented into four records to store positional accuracy of each segment which will represent spatial variation. To retrieve all the

features with a positional accuracy of between 1 and 4m, for example, we might use the following SQL statement:

SELECT * FROM sv-pf WHERE ((‘Positional’ >= 1m) and (‘Positional’ <= 4m))

Spatial variation of quality can be represented in polygon features in the same way. For example, the land parcel in Figure 2 has two attribute accuracy values as 1% and 4%. The only way to record this sub-feature variation of attribute accuracy is to again split the feature into two records (see Figure 2, attributes of parcel second table). To retrieve the features with attribute accuracy greater than 1%, we might issue the following SQL statement:

SELECT * FROM parcel WHERE (‘attribute’ > 1%)

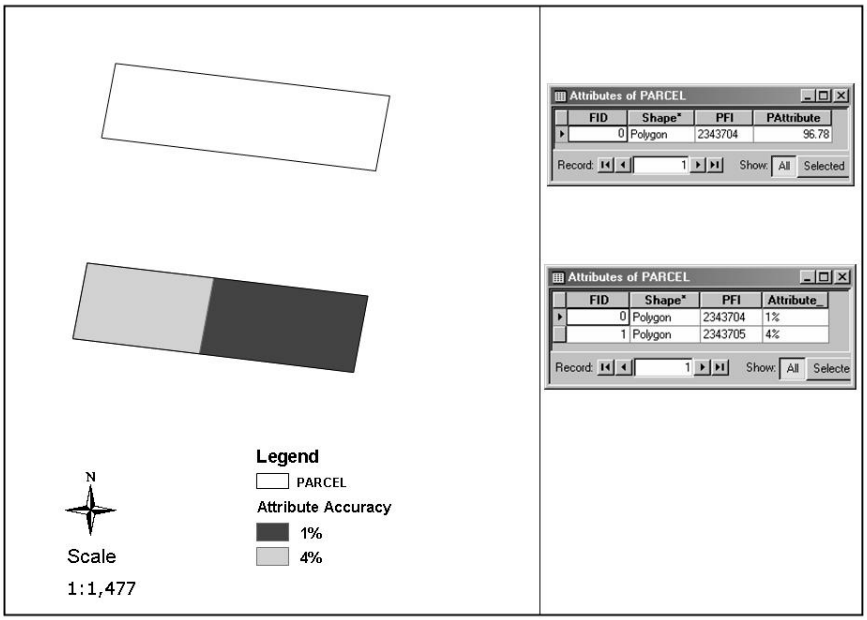

Figure 2 Per-feature model for polygon features.

In summary, in the per-feature model the data structure must be altered for the sake of storing sub-feature variation in quality information. For example, the highway is segmented based on the positional accuracy information. In turn one record is split into four records. Changing the data structure in this way causes fragmentation of the data set and leads to increases in the volume and complexity of the database.

4.3 Feature-independent model

In the feature-independent model, the segmentation of highway is avoided by storing the data quality as a separate layer (see Figure 3). The positional accuracy layer is independent of the highway layer. However, to retrieve the spatial variation of positional accuracy in the highway feature an additional operation of a *spatial join* (overlay) is required. Spatial joins involve

retrieving tuples from two or more relations based on a spatial join condition. Spatial joins are amongst the most computationally expensive operations in a spatial database. In this case, to retrieve the features with a positional accuracy of 3m or less, the corresponding SQL statement will be:

SELECT highway.id, positional FROM highway, sv-fi WHERE ((highway.id = sv-fi.id) and ('positional' <= 3m))

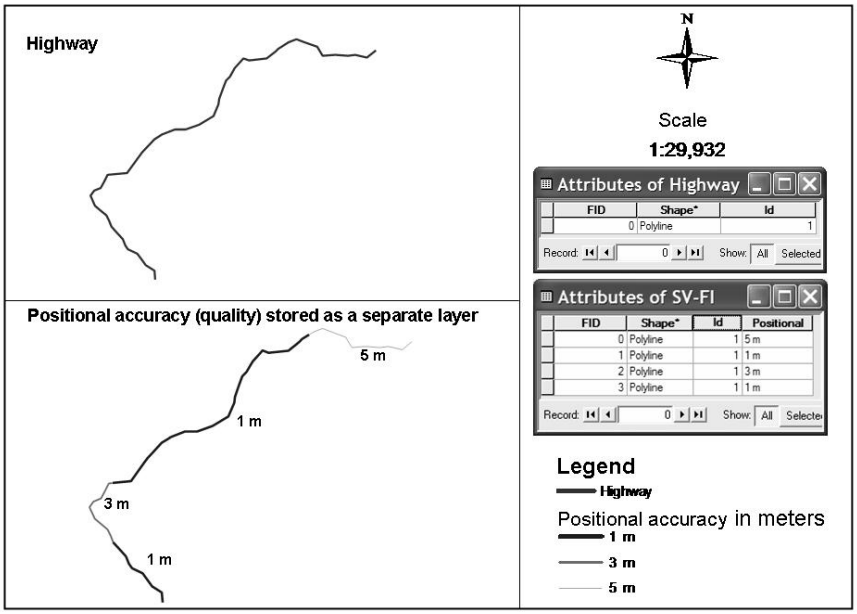

Figure 3 Feature-independent model for linear features.

The feature-independent model can equally model the quality of polygonal features. For example, consider a land parcel having spatially varying vertical accuracy. The vertical accuracy is stored as a separate layer, which is independent of parcel (see Figure 4). In order to retrieve the quality information a spatial join operation is carried out on the parcel and vertical accuracy layers. Querying an individual feature in the feature-independent model requires repeated spatial joins to determine the quality of an individual feature. Additional storage space (space complexity) and increase in processing time (time complexity) are therefore limitations of this model.

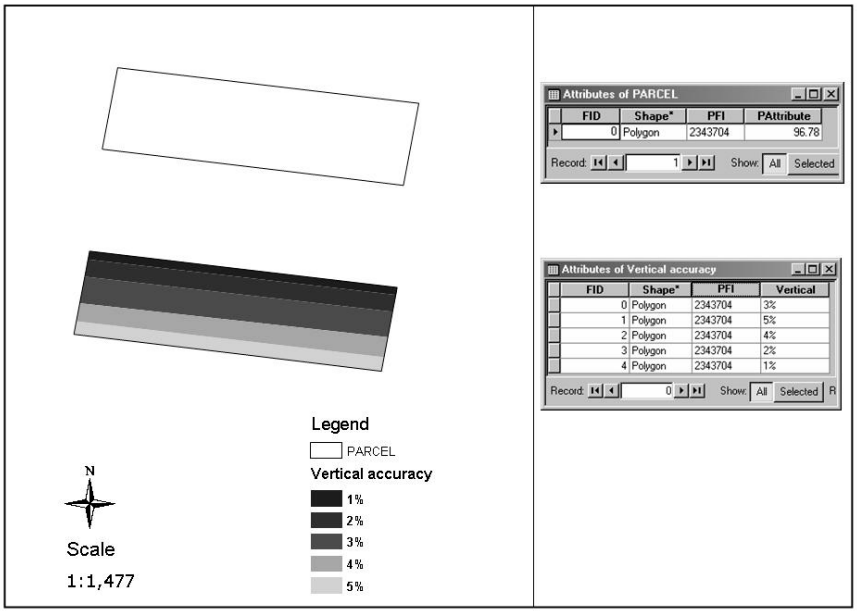

Figure 4 Feature-independent model for polygon features.

4.4 Feature-hybrid model

The feature hybrid model can avoid segmentation of data and overlay operations while storing quality, which take additional processing time for data retrieval. One technique for implementing the feature-hybrid model is to use linear referencing (Cadkin, 2002) to enable spatially varying quality information to be stored on the feature without altering the feature. Each linear feature is assumed to have a unique identifier. In addition to the feature attribute table, an additional table, called an event table (see Figure 5), is used to store the quality values. Using linear referencing, modification of quality information becomes computationally straightforward, since updates only need to alter the event table. To retrieve the spatial variation of positional accuracy across the highway, the event table is queried with a similar SQL structure to that used in the per-feature model.

However, using linear referencing is only one example of a feature-hybrid model. Linear referencing offers limited representational capabilities with respect to polygon features. For example, the technique can be used to represent spatial variation in the quality of the *boundaries* of polygons (cyclic linear features) (see Figure 6), but cannot be used to represent spatial variation across the interior of a polygon. Current work is developing similar efficient storage structures for variation in quality within a polygonal area.

In the feature-hybrid model, if the features are modified the event table has to be restructured accordingly. There are inevitably computational overheads associated with managing the additional event tables. Current work is also examining the comparative processing requirements for data quality management through overlay operations, such as union or intersection of features.

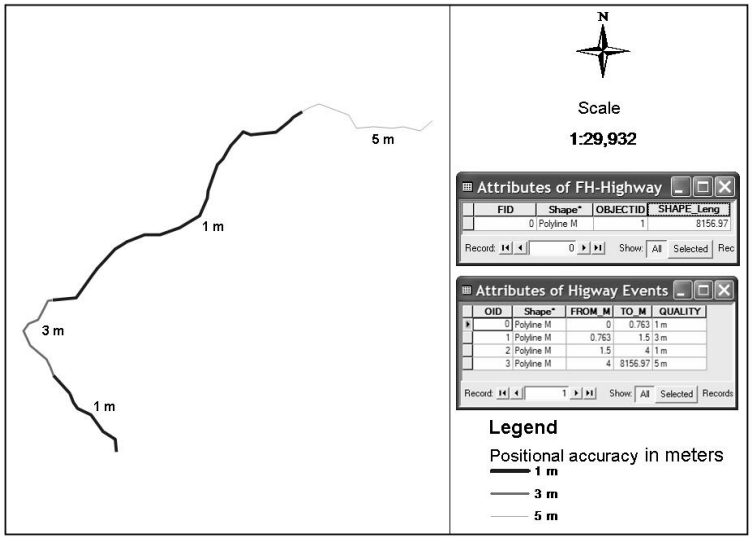

Figure 5 Feature-Hybrid model for linear feature.

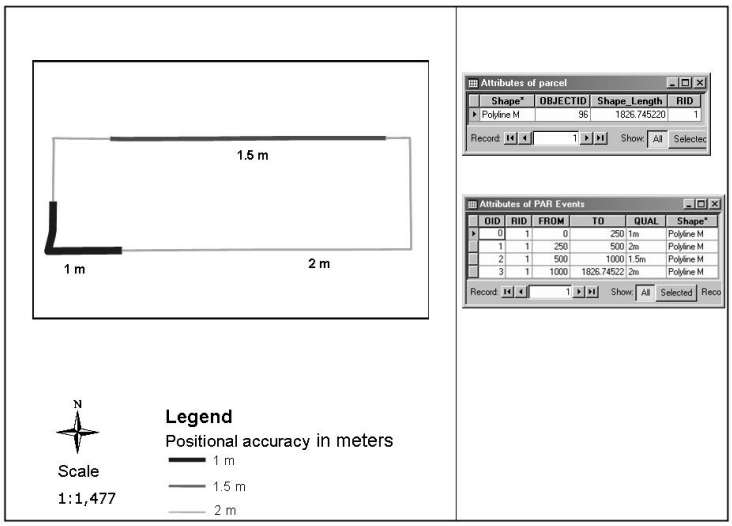

Figure 6 Feature-Hybrid model for cyclic linear feature.

5 Conclusions

The three models, per-feature model (object-based quality), feature-independent (field-based quality) and feature-hybrid (object- and field-based quality) represent spatial variation in data quality. The per-feature model is easy to implement but has limitations in representing sub-

feature variation and alters the data structure for the sake of storing quality information. Additional storage space (space complexity) and increase in processing time (time complexity) are the limitations of the feature-independent model, in spite of the model's ability to represent sub-feature variation. The feature-hybrid model integrates both per-feature and feature-independent models, and potentially overcomes the limitation of the previous models. However, using linear referencing as a technique for implementing the feature-hybrid model has limitations with respect to representing areal features. Each model varies in its data representational, querying, and modification capabilities. The data used in this paper is hypothetical. Hence, current work is using actual data from the Vicmap Property dataset provided by Australian DSE (Department of Sustainability and Environment) to evaluate the models experimentally. Additionally, Oracle spatial database is being used to implement the different models in a relational database. In addition to experimental evaluation, current work is also investigating analytical evaluation of how the time or space complexity of different procedures varies across the different models. Ultimately, one product of these investigations will be the development of an evaluation matrix that summarizes the varying representational, querying, and modification mechanisms of associated with each model, providing the basis for recommendations for the use of each in specific circumstances.

Acknowledgements

This work has been supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centers Program.

References

ANZLIC, 2003. Implementing the Australian Spatial Data Infrastructure, Action Plan 2003–2004. Tech. rep., Australia New Zealand Land Information Council (ANZLIC).

Cadkin, J., 2002. Understanding dynamic segmentation. ArcUser October-December, pp. 40–43.

Chrisman, N., 1984. The role of quality information in the long-term functioning of a geographic information system. Cartographica 21, pp. 79–87.

Clarke, D. G., Clark, D. M., 1995. Lineage. In: Guptill, S. C., Morrison, J. L. (Eds.), Elements of Spatial Data Quality. Elsevier Science Ltd, pp. 13–30.

Duckham, M, 2001.Object Calculus and the Object-Oriented Analysis and Design of an Error-Sensitive GIS, Geoinformatica, Vol 5.3 pp. 261-289.

Fisher, P. F., 1991. Spatial data sources and data problems. Geographical Information Systems : Principles and Applications 1, 179–189.

Graham, S., 1997. Products liability in GIS: Present complexions and future directions. GIS Law 4 (1), 12–16.

Hetrick, V. R., 1991. An approach to spatial data quality using standard visualization tools. Report of the Initiative 7 Specialist Meeting: Visualization of Spatial Data Quality, pp. C85 to C91.

Heuvelink, G. B. M., 1996. Identification of field attribute error under different models of spatial variation. International Journal Geographical Information Systems 10. No.8, pp. 921–935.

Hunter, G. J., 1996. Management issues in GIS: Accuracy and data quality. In: Conference on managing Geographic Information Systems for success, Melbourne, Australia, AURISA.

Hunter, G. J., Goodchild, M. F., 1996. Communicating uncertainty in spatial data bases. Transactions in GIS, Vol 1.1 pp. 13-24.

Hunter, G. J., Hope, S., Sadiq, Z., Boin, A., Marinelli, M., Kealy, A., Duckham, M., Corner, R. J., 2005. Next-Generation Research Issues in Spatial Data Quality. In: Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Sciences Institute, September, 2005. Melbourne: Spatial Sciences Institute. pp. 865–872.

Hunter, G. J., Qiu, J., 2003. Automatic updating of spatial data quality information. In: Proceedings of the 2nd International Symposium on Spatial Data Quality '03. Advanced Research Centre for Spatial Information Technology, The Hong Kong Polytechnic University, Hong Kong, pp. 210 – 214.

ISO/FDIS 19115, 2005. Text of 19115 geographic information-metadata. Tech. rep., ISO.

Maclean, A. L., D'Aveilo, Thomas, P., Shetron, G. S., 1993. Use of variability diagrams to improve the interpretation of digital soil maps in a GIS. Photogrammetric Engineering and Remote Sensing 59 (2), 223–228.

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

Ramm, P., 2005. A question of accuracy. Position, 80–81.

United States Department of the Interior Bureau of Land Management, 2001. GCDB Reliability Diagram. Tech. rep., Land & Resources Project Office.

Wong, D. W. S., Wu, C. V., 1996. Spatial Metadata and GIS for Decision Support. In: Proceedings of the 29th Annual Hawaii International Conference on System Sciences. IEEE.

Worboys, M., Duckham, M., 2004. GIS: A Computing Perspective, 2nd Edition. Boca Raton, FL: CRC press.