Spatial Data Quality: Problems and Prospects

Gary J. Hunter[1], Arnold K. Bregt[2], Gerard B.M. Heuvelink[2], Sytze De Bruin[2] and Kirsi Virrantaus[3]

[1] Department of Geomatics, University of Melbourne, Parkville VIC 3010, Australia
[2] Center for Geo-Information, Wageningen UR, PO Box 47, 6700 AA, Wageningen, Netherlands
[3] Department of Surveying, Helsinki University of Technology, Otakaari 1, Espoo, Finland
garyhntr@gmail.com, arnold.bregt@wur.nl, gerard.heuvelink@wur.nl, sytze.debruin@wur.nl, kirsi.virrantaus@tkk.fi

Abstract

This paper reflects upon the topic of spatial data quality and the progress made in this field over the past 20-30 years. While international standards have been established, theoretical models of error developed, new visualization techniques introduced, and metadata now routinely documented for spatial datasets, difficulties nevertheless exist with the way data quality information is being described, communicated and applied in practice by users. These problems are identified and the paper suggests how the spatial information community might move forward to overcome these obstacles.

1 Introduction

With the growth of Geographical Information Systems (GIS) and new technologies such as the Internet, the broader community is benefiting from access to datasets that were once used only by a small group of spe-

G. Navratil (eds.), *Research Trends in Geographic Information Science*,
Lecture Notes in Geoinformation and Cartography, DOI 10.1007/978-3-540-88244-2_8,

cialists. As such, there are now many more people making decisions based on information they perhaps know very little about—particularly in terms of its quality. In addition, we live at a time when there is less tolerance for poor decision-making and the consequences of 'getting it wrong' can be severe for individuals and organizations alike.

Of course, the accuracy issue was always in the minds of the map-makers who, as recently as 30 years ago, had sole responsibility for preparing our paper-based maps and charts. They met the accuracy-reporting challenge as best they could by providing estimates through reliability diagrams, special symbols, and accuracy statements based on testing to accepted industry standards. They knew their products were not perfect and users were also generally aware of this fact, so there was a degree of shared knowledge between the data collectors and users that has since disappeared with the advent of digital data.

With the introduction of digital mapping techniques in the 1960s and then GIS shortly afterwards, researchers realized that error and uncertainty in digital spatial data had the potential to cause problems that had not been experienced with paper maps (for example, see MacDougall 1975; Goodchild 1978; Blakemore 1984; Chrisman 1984; Robinson and Frank 1985; Burrough 1986; Bedard 1987; Epstein and Roitman 1987). On the other hand, the wider GIS community took far longer to realize the potential traps that existed for unwary users, and there is no doubt that the notion of 'the computer must be correct' held force for many years.

In conjunction with these warnings, an international trend started in the early-1980s to design and implement data transfer standards which would include data quality information that had disappeared from the margins of paper maps with the transformation to digital data products (Moellering 1991). The standard that clearly led the way in documenting data quality was the U.S. Spatial Data Transfer Standard (NCDCDS 1986; NIST 1992) which divided quality reporting into five parts, viz.: dataset lineage; positional accuracy; attribute accuracy; logical consistency and completeness (Chrisman 1991; Guptill and Morrison 1995). By and large, these elements have stood the test of time, although there have been recent additions and/or variants such as 'semantic accuracy' (as a broader alternative to attribute or thematic accuracy), 'temporal accuracy' (the accuracy of reporting time associated with the data), and 'metaquality' (data about the quality data, such as its reliability and confidence) (CEN 1998; ISO 2002, 2003a, 2003b).

By the late-1980s and early-90s, special attention was being given at international conferences and in the scientific literature to the importance and need for the proper treatment of spatial data quality, and the benefits that would come from its use. Moreover, leading international research centers in the US and Europe had identified key initiatives in spatial data

accuracy, the treatment of indefinite boundaries, and visualization of spatial data quality as being of fundamental scientific importance. There was even an international uncertainty visualization 'challenge' conducted in the U.S. Buoyed by this activity and the widespread publicity surrounding the topic, we believe spatial data users at the time expected they would soon be able to (1) easily interpret data producer's quality statements, (2) understand the inherent strengths and limitations of a dataset in quantitative terms, and (3) translate that information to a form suitable for assessing whether or not they should use it for their decision tasks.

While this might have seemed a utopian view, it was in fact the complete expression of the solution to the spatial data quality research problem, and it was discussed to differing degrees by research leaders such as Openshaw (1989), Burrough (1991) and Goodchild (1992). Of course there were many assumptions underlying this perfect vision of the future. For instance, it was expected that: (1) data quality statements would have appropriate content and format; (2) data consumers would possess the requisite knowledge and skill to comprehend and translate these statements; (3) commercial product developers would write the software to analyze, portray and keep track of error; (4) innovative researchers would produce new error theories, models and methods; and (5) spatial data quality would be able to be expressed in terms of its quantifiable impacts upon intended decisions in a manner that would be obvious to all concerned.

Clearly, the achievement of all these tasks was always going to be difficult, and so just like other critical reviews that have recently been conducted into topics such as space-time data representation (Peuquet 2001), computational methods for representing geographical concepts (Egenhofer et al. 1999), and the integration of GIS and spatial analysis (Getis 2000; Goodchild 2000), in this paper we reflect upon the progress made in spatial data quality over the past 20-30 years. Certainly, standards have been implemented, many datasets now carry quality statements, and researchers continue to investigate models of error and its portrayal, however we suggest that several of the original goals are still to be met. Accordingly, in this paper we revisit the topic of spatial data quality to identify the problems that remain (Section 2) and the work that needs to be undertaken to bring the original vision to completion (Section 3).

2 Problems

In reflecting upon our present level of understanding of spatial data quality, we believe the problems still being experienced in this subject lie in five fundamental areas, viz.: data quality reporting, description and visualization;

error propagation; and the application of data quality information in practical decision-making environments.

2.1 Poor Quality Reporting

Starting with the issue of data quality information content, we believe the current lack of detail provided in many data quality statements makes them ineffective for any subsequent use. To demonstrate this point, the examples of poor data quality reporting presented in Table 1 have been selected from actual data quality statements recently collected.

Taking the positional accuracy examples first, obvious questions soon arise with these statements such as, "Exactly how variable are the observations?", "Where are the 1000m errors located?", "Where does the urban/rural quality transition occur in the dataset?" and "Where were the deliberate cartographic offsets made?" In other cases there may be little or no actual basis for making these statements—for instance when the positional accuracy of a very small sample of well-defined point features is tested, and the results are then inappropriately assigned as an accuracy indicator to all objects, regardless of their type (such as linear and areal features).

As for attribute accuracy, to state that this is not relevant for a vegetation map is clearly unacceptable, while claims of 'high' accuracy and '100%' accuracy that carry no indication of what was tested, how it was tested or the sample size used, do little to inspire trust in a prospective data consumer. There are also other deficiencies with attribute accuracy reporting that are not listed in Table 1 and which need correction. Firstly, the accuracy of all attributes should be reported separately, since it is not possible to assign a single accuracy value to describe multiple attributes in a database (and indeed, if it were possible it would be an outstanding breakthrough in the data quality research agenda). Secondly, the scale of measurement for each attribute (for example, nominal, ordinal, interval and ratio) should be included in the data quality report as an aid to its later use in conjunction with error modeling and visualization tools.

Moving to the logical consistency examples in Table 1, data with different lineage should be tagged with appropriate identifiers if there are different accuracies present—so that users might learn which features can be expected to possess poorer logical consistency. In addition, there seems to be a common misconception that logical consistency consists only of ensuring polygon boundaries are closed and that linear features meet where intended, however in practice there are many different tests for logical consistency that need to be undertaken with spatial datasets. Reporting of completeness suffers similarly and data quality statements rarely state what information is actually missing. However, stating that some (unidentified)

features are missing or that "street address details are partially complete", provides little useful information to potential users.

Table 1. Examples of poor data quality reporting.

Positional Accuracy
"Variable", "100m to 1000m"
"+/- 1.5m (urban) to +/-250m (rural)"
"No feature in error by more than 100m"
"90% of all points lie within 1mm at plot scale"
"Cartographic offsets may be present"
Attribute Accuracy
"Not relevant" (for a vegetation map)
"100% accurate"
"High attribute accuracy"
Logical Consistency
"Between 1% (new data additions) and 5% (pre-maintenance contract data)"
Completeness
"Some features have been eliminated"
"Street address details partially complete"
Currency
"From aerial photography 1965-1992"

Finally there is the reporting of currency (temporal accuracy) and the example given in Table 1 would surely have a potential consumer wondering exactly which parts of the dataset referred to are derived from 40-year old photography and which ones have been updated from more recent material. In addition, currency tends to be described for datasets as a whole and not as it should be for each data quality element where appropriate. For example, the date at which a feature's position is observed may often be different to the date that its attributes were recorded—and coupled with this is the need to record database transaction dates for maintenance and update purposes.

2.2 Incomplete Quality Descriptions

While the problems associated with poor data quality reporting are relatively minor, there are several other problems that will have greater impact in the future if left unresolved—and they relate to incompleteness in spatial data quality descriptions.

The first of these is that data quality information suffers generally from being presented at a generic global level rather than at more detailed levels of granularity. While global information will always be required in data quality statements, there is also a need to report any spatial variation in data quality. This divergence might reside naturally in the data, or else it

might come about as a result of spatial operations—such as when two datasets with different positional accuracies are overlaid or merged. There is also the need to report any spatial uncertainty or spatial correlation of local data quality. This is important, for instance, if the areas of continuous regions are to be estimated from raster data or when slope gradients, viewsheds or watersheds are computed from DEMs.

Another fundamental problem with data quality descriptions is that they tend to work far better with data representing crisply-defined objects usually found in the built environment, rather than with the more abstract and vaguely-defined data representing phenomena occurring in the natural environment (for instance, see Burrough and Frank 1996). This is hardly surprising since we are the ones who have designed the coordinate systems, built the technology to measure positions, and defined the terms and values used to describe their attributes. However, the natural world is not of our making, and trying to observe and represent its processes are difficult enough to achieve in practice without also having to describe the accuracy with which we define and model it. For example, when we perform soil sampling we must limit our testing to points to minimize soil damage, and then (to make the observations fit our relatively simple computational models) we allocate crisply-defined boundaries to polygons having homogeneous consistency to represent something that is inherently heterogeneous and known to possess widely varying transition zones. Describing this difference, between the models we use to depict the real-world and the real-world itself, is a continuing problem and continued research will clearly be required in this area.

Furthermore, for the estimation of error propagation to be successfully achieved (see section 2.4) we need considerably better information to be provided than we now receive. Taking DEMs as an example, the elevation error in a DEM is typically conveyed by a Root Mean Square Error (RMSE), however that on its own is not always sufficient. For error propagation to be estimated (such as when we derive a slope map or a viewshed from a DEM) we also need to know the spatial autocorrelation in the error. Ideally, we should have the full joint probability distribution but this is not available in practice so we tend to get, at best, a parameterized model of the joint probability distribution. This means that someone else has chosen a particular model, with its inherent assumptions, such as stationary random variables.

Finally, some comments should be made about error modeling, because if we cannot define error then we cannot measure it or describe it. Certainly, ten years ago few theoretical error models existed and Goodchild (1993) noted at the time that the known and accepted techniques we possessed for describing and measuring error were essentially limited to: the locational accuracy of a single point (through the circular normal model of

positional error); the accuracy of a single measured attribute; the probability that a point at a randomly chosen location on a map has been misclassified (through the misclassification matrix); the effects of digitizing error on measures of length and area; the propagation of error in raster-based area class maps through spatial operations such as overlay; and the error associated with measures of area derived from dot counting. Since then, the development of error models has progressed and numerous models have now been proposed in areas as diverse as: positional error in vector data; thematic uncertainty in the integration of Remote Sensing and GIS; the accuracy of TINs and Lattices; elevation error in DEMs; errors in point-in-polygon analysis; fuzzy representation of boundaries; field attribute error; errors in buffering operations; probabilistic viewsheds, and in cartographic generalization processes. However our knowledge of error remains relatively immature, although it is not due to lack of effort.

2.3 Barriers to Communicating Quality

Moving away from how we describe the fundamental spatial data quality elements, there is a range of issues relating to how quality is being communicated to spatial data users. While data producers have generally accepted the need for data quality reporting, consumers of their products do not seem to have embraced the spatial data quality issue to the same extent. This could be due in part to reasons such as: (1) the fact that many users have never received formal education in GIS; (2) that there is no commonly taught approach to the critical analysis of results in geographic information science (unlike in other disciplines such as surveying and geodesy); (3) that clients who commission a data product may not necessarily be interested in its quality; and, perhaps, (4) that users have become disillusioned with the lack of progress in this area. Of course, even if we were able to overcome each of these difficulties, there remains the issue of how to effectively communicate data quality to different types of users. For instance, while an analyst may readily understand the meaning of linear regression statistics, standard deviations and semi-variograms, such concepts can be bewildering for both the novice at one end of the user-spectrum and the senior decision-maker at the other.

Another communication problem, this time associated with spatial database structure and design, is that we do not possess the tools to manipulate, query, analyze or display data quality information—as we already do for spatial data. Similarly, we are unable to update data quality information in real-time as changes occur in a database. For example, while we can easily take two separate point datasets and combine them to form a new dataset through a simple ‘merge’ or ‘append’ operation, if they each have their

own data quality statements we are currently incapable of automatically integrating their respective data quality information to yield a new data quality report for the merged data product. Similarly, we are unable to produce a quality report for a slope or aspect map that might be derived from a DEM—even though the DEM will in all likelihood have its own quality information readily available (albeit in a relatively simple form such as a global RMSE). So while we can easily update spatial features and their attributes, it remains a challenge to researchers to provide an effective means of updating attached data quality information 'on-the-fly' when spatial processes are applied and new datasets are created—yet this is something that will obviously have to occur in future GIS.

Effective communication of data quality also means having the visualization tools to help do the job, and while we would appear to already have the techniques necessary to perform the task they have yet to be implemented in commercial GIS packages (although there are numerous examples of their implementation in proof-of-concept form). This is partly due to the fact that data quality information is not normally coupled with the data to which it refers, and so there is no capability for subsequently linking it to error modeling and visualization software. While the software developers are naturally the people best able to implement these visualization techniques, the task still does not seem to have a high priority for their companies. Informal discussions suggest there is still not a strong enough level of demand from the user community for this product functionality to justify the expense of incorporating it into commercial systems. On the other hand, the drive by vendors and third parties over the past five years to provide easy-to-use metadata entry tools has been rapidly achieved in response to demands by data producers (particularly government agencies)—so the industry has certainly demonstrated its capacity and technical skill to act quickly when the need arises.

2.4 Keeping Track of Error

Another key issue impediment in dealing with spatial data quality is that our knowledge is still deficient in the way error propagates in many spatial operations. Although we have approximate methods of error propagation in the area of quantitative modeling with continuous data based on the principle of propagation of variances (Heuvelink et al. 1989), and simulation methods in which the effect of perturbation of the input data is observed and quantified in the outputs, these are methods that become impractical when dealing with chains of complex operations and when dealing with categorical rather than continuous data. Furthermore, the error propagation techniques we do possess are inevitably applied by expert ana-

lysts, with the result that once the 'average' GIS user studies the data quality statement for a dataset there is little else that can be done to translate that initial information into quality descriptors for any secondary products they might create. So in essence our progress beyond the current body of knowledge in modeling, reporting and communicating spatial data quality lies frozen at this point.

2.5 Application of Data Quality Information

Finally, users are experiencing problems applying data quality information in real-world, everyday situations—and we should remember that the notion of quality is concerned with 'fitness-for-use' or suitability for a task, not just the degree of error in the source data. At the present time data quality reporting could be said to generally be characterized as governed by producer-driven standards and requirements rather than user applications. Of course, from a producer's perspective this is reasonable since there is no way of controlling how users might apply their data. So their products are tailored to meet certain specifications to satisfy particular past user-demands, but these do not necessarily help other consumers assess whether an information product is actually suitable for a given function. On the other hand, we believe that users would like to be provided with the technical capability to take the initial data quality information and use it to determine what output accuracy will result from the use of a given set of inputs, models and spatial operations—preferably before the task is actually undertaken so that alternative data, algorithms and models can be investigated if required.

At present this has only been performed in a limited way by skilled analysts, and this functionality does not generally exist in commercial software packages. In particular, the problem of verifying model outputs is currently causing concern amongst leading scientists as they discover that governments are increasingly reluctant to commit to highly sensitive policy decisions without any knowledge of the validity of the models being used. For example, Beven (2000, p. 2605) reported that a proposal to establish an underground repository for radioactive waste at Sellafield in the UK was refused permission to proceed after the results from simulated groundwater flow studies "…differed drastically between modelers on both sides of the argument." At the same time, there have been calls for new research efforts into how we might generally describe the quality of models, and how we might derive a set of model quality elements in a fashion similar to the data quality elements we now possess.

Even if we could propagate error in spatial data and quantitatively assess its effect upon derived outputs, ultimately what users really want to

know is what risk is associated with using information of a given quality—in other words "What can go wrong?" and "Will my decision remain unaltered?" The answer to these questions may well require users to be better trained in decision-making and risk management to foster a fundamental change in the way they perceive their information, and a more probabilistic approach will probably need to be adopted in terms of their interpretation of spatial information.

3 Future Prospects

If these are the fundamental problems associated with spatial data quality that remain after 30 years, then what are the prospects of overcoming them? Certainly, the problems do not rest solely with any particular sector of the spatial information industry and solutions must be found jointly by the data producers, consumers, system developers, educators and researchers. Indeed, there are valuable messages for each of these groups to take from the following discussion.

3.1 Enhanced Quality Reporting

Looking at the data quality reporting problem first, it is clear that poor reporting can be overcome relatively easily with experience and advice, and there are excellent examples of comprehensive data quality reports and technical user guides available on the Internet—such as the Geoscience Australia (2006) and Ordnance Survey (2009) digital data products. The latter provides a good illustration of completeness reporting in its user guide where it lists several pages of real-world features not included in its 'MasterMap' dataset (for example, buildings below a minimum size are not shown, telephone lines and poles are only shown when they are of outstanding significance, and roads on private property are only shown when longer than 100m).

This is what Brassel et al. (1995) would refer to as 'model completeness' information, as opposed to 'data completeness' information which would report whether all existing roads greater than 100m in length have actually been included in the database with their correct attributes. Similarly, for reporting logical consistency the user guide for the Geoscience Australia 'TOPO 250K' product details some 60 tests that are performed and reported (for example, label points have only one coordinate pair, road tunnels and bridges are coincident with nodes in the road network, coastline is cloned as a zero height contour, features labeled as an island or reef are completely surrounded by water), together with the test sample size

and the acceptable quality level for each test. So, as stated previously, the prospects for good quality reporting are excellent and will improve with time.

3.2 Improved Quality Descriptions

Moving to the next issue of how quality is described, we have already seen producers provide multi-level data quality information, and an example of this that has existed for over 10 years can be found in the product metadata described in Geoscience Australia (2006). In this instance, data quality information is presented at four levels for the product, viz.: the dataset; data layer, feature class and individual feature levels—with the quality information for the three highest levels being stored in hardcopy narrative form, while the quality information at the feature level is held in attribute form against each object.

As an exercise in describing this multi-level data quality information, Qiu and Hunter (2002) took a sample set of the Geoscience Australia product and its associated data quality information, converted all of the latter to digital form, and then attached it at all four levels within a commercial GIS (ArcView)—so that it became possible to select and display both the spatial data and the data quality information from within the GIS environment. While the trial was successful, in the longer-term they believe a more elegant solution would be achieved by adopting an object-oriented approach in order to make full use of the inheritance, classification and encapsulation capabilities available to more effectively model the data and its quality information. Subsequent research in this area has been conducted by Sadiq and Duckham (2009) who successfully implemented a data quality module in Oracle Spatial software to cater for individual feature- and even sub-feature-level quality variation reporting and querying. In this area, the proposed US ANSI Metadata Standard expected to be introduced in 2009 includes the provision for metadata descriptions at different data levels.

Moving next to the description of data quality particularly when we attempt to represent natural phenomena, some researchers are now suggesting that if we were to ‘step back’ and focus more on describing the quality of the original observations (instead of the models we consequently create from them), then we may well find it easier to provide data quality statements that meet our current standards. For instance, we could more ably define the positional accuracy of soil test sites and better state the variation that occurred in their observed attributes. In essence, we would simply provide source data that had been quality tested, and then let users take responsibility for what happens to it from that point onwards. However, this

assumes they possess the tools necessary to describe the quality of the outputs of subsequent spatial operations on the original data—and at this time they do not. Furthermore, it could be argued that providing only the original test site data may not be sufficient as we could only confirm how well the data used for calibration of the model were represented by the model. In the case of kriging, the fit would be perfect, which could lead a less-skilled user to assume the whole model was perfect. On the other hand it might be more useful to supply users with the model data plus a set of independent data test sites which would allow them to validate the model they propose using.

Alternatively, we could try to search for a more rigorous and exhaustive means of storing uncertainty information about the continuous and categorical forms of spatial data that tend to characterize natural resources. For this to succeed we will need to know not only the (spatially varying) variances but also the distribution type, the spatial autocorrelation, and any cross-correlation with errors in other attributes—and this information will also have be derived for any new attributes that are created in the database. Taking all of this into consideration is clearly a major task and will impact significantly on the database design, which suggests we might need to settle for a less rigorous approach—but the question then arises "What choices do we make?" Unfortunately, when we come to deal with categorical variables the situation becomes worse—since not only is there the question of "How do we manage and control the parameters needed to fully convey uncertainty?" but more importantly "How can we estimate them?"

So we continue to experience considerable difficulty in describing and measuring error in these types of data, which might explain why many of the examples of (what is taken to be) good data quality reporting tend to relate to digital versions of traditional products such as topographic maps. However, these products are often not the ones used for decision-making, and instead are inclined only to be used along the way to derive secondary information which is what users will ultimately want to know more about in terms of quality. Of course, some of the deficiencies mentioned here are not just confined to data representing natural phenomena, and indeed information on distribution types, spatial autocorrelation and cross-correlation is required for all field data (as opposed to object data) regardless of what they represent.

3.3 More Effective Quality Communication

Dealing next with the problems relating to communicating data quality, clearly the next generation of spatial data consumers will be better educated in issues such as quality. Whereas 10 years ago many GIS courses

tended to focus on GIS technology and its applications, it is now quite common for students to be introduced at an early stage in their studies to the legal and institutional issues of GIS which inevitably connect with matters of quality. So in that respect, the problems associated with the lack of education and awareness in the subject can be expected to be overcome with the passage of time.

In addition, new forms of metadata description are now under development. For example, Boin and Hunter (2007a) have found that to begin with consumers have little or no understanding of the terms 'metadata' or 'data quality'—instead preferring the term 'product description'. Furthermore, their review of many hundreds of data complaint emails coupled with a detailed survey of data consumers, revealed a much greater need for information on what users could expect a dataset to contain. Thus, in terms of understanding how potential datasets are chosen for user applications, their surveys revealed that published metadata played little or no part in the selection process since its content was considered too technical in its nature—even by professionals such as engineers, architects and planners who are using spatial data on a daily basis. Instead they relied upon colleague opinions of which datasets to use for a given purpose or else learnt through experience which datasets could be trusted to meet their needs. The implication of this is that data producers are creating metadata and populating data directories around the world with information that has little or no benefits to data consumers. So in an effort to make the metadata that is collected more meaningful, Boin and Hunter (2007b) reports on the design and testing of a new graphical style of describing the contents of a spatial dataset which consumers found more interesting and informative. Devillers et al. (2007) have also been working in this area of communicating metadata in new ways and their dashboard-style of presentation is similarly finding interest amongst consumers.

A more serious impediment to better communication of data quality, however, is the fact that we are not making life any easier for data users by introducing notions of error and uncertainty given their negative connotations. For instance, urban and regional planners, civil engineers, real estate appraisers and others have all used soil maps for decades to make effective decisions without being aware of the uncertainties about inclusions and map unit variability—a view supported by the work of van Oort and Bregt (2005). Surprisingly, this approach has worked well for many years (for example, see Hudson 1990). So for groups such as these who have learned to live quite comfortably with the usual binary outcomes in GIS processes ('one' or 'zero', 'in' or 'out', 'yes' or 'no', 'black' or 'white'), we are not necessarily seen as doing them any favors by introducing grayness or a 'plus/minus' to their decision-making—even though we know it is something they should be taking into consideration.

What would undoubtedly help most in promoting the importance of data quality to users would be a series of well-documented case studies describing the perils associated with ignoring data quality. While there are already several well-known examples of proven legal liability associated with the provision of erroneous nautical charts and topographic maps, in such cases the relationships between data error and its adverse consequences tends strongly to be both obvious and severe in its impact. For example, if a reef in the middle of a shipping channel is missing from a nautical chart or assigned the wrong depth sounding, then it will only be a matter of time before a ship runs aground on it resulting in expensive claims for compensation being made against the data provider, and (all too often today) major environmental harm.

On the other hand, to the average GIS user the range of possible adverse consequences due to using poor quality data never seem to be quite as dramatic in terms of their impact. Of course for the mistakes that do happen, the reality is that their news tends to be suppressed, arising out of a sense of shame and often also as a condition of any out-of-court compensation payments made—which are preferable to the publicity of a court hearing and the establishment of a legal precedent. Thus, the effect is that any prospective authors who decide to report such cases in the literature do so at considerable risk of initiating legal action against themselves. So it seems unlikely that any "Greatest Failures in GIS" texts will ever appear, although we could certainly develop a 'best practice' handbook in spatial data quality which cites positive outcomes—similar in essence to what Marble (2000) called for to encourage greater interest and interaction between the GIS and spatial analysis communities. In addition, it is interesting to note that there is still no textbook dedicated solely to the way we treat spatial data quality, although a valuable and practical handbook on positional accuracy is available on the web (Minnesota Planning 1999).

3.4 Better Error Tracking

One way of increasing data quality awareness would be to communicate the changes that actually occur to quality in real-time as users combine and process datasets using GIS. However the methods that have been developed by researchers to date are invariably time consuming and complex to use. For example, the effort required to run a Monte Carlo uncertainty propagation analysis is at least an order of magnitude greater than running the basic analysis itself (not only in terms of computing time, but also in terms of parameter estimation and management of the process). In addition, these operations tend to have something of a prototype air about them, meaning 'it only works when I run it'—which implies they are prob-

ably still far too context-specific to be of much value to the broader GIS community. Nevertheless, some of these communication problems have already been recognized by software developers—with the IDRISI product designers clearly taking the lead in the mid-1990s when they introduced a suite of uncertainty management and decision-making tools (Eastman 1997). Other products are also showing increased functionality in this area such as ESRI's ArcGIS software which caters for basic metadata management (which includes data quality) in its ArcCatalog module (ESRI 2000).

Geostatistics provide one means of quantifying certain types of uncertainty in a fairly complete fashion through the use of standard deviations, variograms and cross-variograms. It also offers possibilities for generating realisations of uncertain spatial attributes needed for Monte Carlo uncertainty propagation analysis (and here we could think of sequential Gaussian simulation as the simplest example). However, geostatistics deals primarily with quantitative field data, although there are some extensions to categorical field data using indicator approaches. When it comes to handling uncertainty in object data we lack an equivalent set of tools, although the spatial statistics software, S-Plus, has some functionality in this direction through point pattern analysis techniques. While there has been some introductory work in perturbing the locations of spatial boundaries (for instance, Hunter et al. 1999), it remains to be seen whether these techniques can be easily incorporated or connected to GIS packages.

One solution would be for commercial GIS to have a range of error models available to run on datasets, coupled with error propagation functions that would automatically operate whenever a spatial operation or model was initiated, plus a suite of error communication options, which would all work towards providing input for a set of decision management tools. Of course, such a solution could also exist within a third-party software product, and the move from closed proprietary GIS to open system architectures is an important advancement that will obviously facilitate this. This is starting to occur with links between statistical software and GIS modules (for example, through OLE/COM), and it is expected that a broader group of users will take advantage of tools that were previously restricted to specialists. Unfortunately, third-party products suffer from requiring a separate, deliberate purchase on the part of data users, and therefore may not become as widely adopted as the mainstream GIS package with which they operate in conjunction. Nonetheless, third-party products can eventually become indispensable components of popular software, and the spellcheckers we employ in our word-processing packages are an obvious example.

Clearly, the sensible approach would be for software developers to start small and introduce some simple error models. For instance, it is a comparatively easy task to calculate the standard deviations of polygon areas

from the horizontal positional error estimate in a dataset. Similarly, for grid-cell data an error propagation tool could be developed for the numerical modeling case, while for error communication in vector data a drop-down menu of visualization options could be made available. Such tools would still require users to have a sound knowledge of their application, in much the same way that the well-known Microsoft Excel Charting module offers a wide range of graph types but the responsibility for the outcome ultimately rests with the user. While some of these ideas were promoted over a decade ago by Burrough (1991), to this day they remain such an obvious part of the solution to the spatial data quality problem that they need to be repeated—although clearly the respective forces of demand (from users) and supply (from software developers) have been far too small to bring about their introduction.

3.5 Complete Utilization of Quality Information

Finally, there is the matter of how the data quality application problems described above might be overcome. While easy to express, they will most likely prove difficult to resolve in the short-term given that their solutions depend on how we deal with the more fundamental problems occurring with data quality description and communication. Nonetheless, some researchers suggest that risk management theory might be usefully applied here to assist decision-makers (Agumya and Hunter 1999), and case studies using spatial data are already starting to appear in the literature (De Bruin et al. 2001). Importantly, the risk management approach links into cost-benefit analysis so users can determine whether it is worth improving their data or else taking other provisions to cover their risks (for instance, they might limit the reliance placed upon the outputs of a GIS). Other researchers are combining uncertainty assessment with sensitivity analysis to estimate the comparative quality of different GIS-based models. For example, the excellent work of Crosetto and Tarantola (2001) describes a study in which 15 different types of error residing in seven separate datasets were assessed to judge the reliability of different hydrologic models for flood forecasting.

However, it is possible that these types of studies may be far more detailed than some consumers actually need, and in addition the majority of GIS users are neither experts nor highly-skilled analysts, so we must learn to cater for their needs. Indeed, there may be little point in ‘disturbing’ a large group of users with questions they can neither answer nor understand regarding the specification of data quality parameters. Instead, we should consider developing tools that cater for different user backgrounds and supply default parameter values for non-experts. These users can identify

themselves at the time of log-in and let the system interact in the most appropriate manner from then on.

Similarly, for users browsing spatial data directories on the Internet, we could have a system that requests information about the intended application and then consults a library of spatial data usage histories to suggest possible datasets, algorithms, models and methods to meet the user's need. Alternatively, in some situations it might save time and money if we were able to audit or pre-certify the quality of a spatial dataset as being suitable for a particular task. Users could then be provided with simpler product descriptions coming from organizations that they could trust without having to undertake their own detailed analysis. While data producers have traditionally been reticent to document the applications their digital data might possibly be used for, an agency such as the UK Ordnance Survey handles the matter quite openly and not only states for each of its products the professional groups who are likely to use it and the general application areas for which it is used, but it also provides Internet-based case studies of the actual application of the data (Ordnance Survey 2009).

3.6 Final Remarks

This paper now closes by posing several questions that might help us to understand why we have still not resolved our spatial data quality problems. Firstly, are our difficulties with spatial data quality the result of having to work with legacy system structures designed almost 40 years ago? In essence, our commercial software packages are still based on concepts derived in the 1960s and 1970s when the situation was one of applications searching for computer-based solutions. The tide then turned in the mid-1980s when the software and hardware we needed finally arrived, and continued to improve to the extent that by the 1990s we witnessed technology in search of applications. Perhaps the ebb and flow of scientific and technological development needs to turn again, and we need to see a second generation of geographic information science concepts and systems designed and developed that will handle spatial data in new ways—such as object-oriented, error-aware GIS (Duckham and McCreadie 1999).

Secondly, do we need an entirely new stimulus to drive the data quality issue to a satisfactory conclusion? One possible incentive could come from the many large spatial data infrastructure initiatives being developed worldwide at local, regional, national and global levels. As different agencies (often from different countries) contribute to the development of Geospatial Data Service Centers (GDSCs), Doucette and Paresi (2000) contend that data providers who have taken due care with their quality assurance methods are becoming anxious not to attract unnecessary liability

through their cooperative arrangements with other producers who may not have been so prudent. While there are potential rewards for the participants, the pitfalls are waiting there as well unless they can more effectively deal with data quality reporting and communication.

Finally, have we simply been expecting to achieve too much, too soon, with too many unexpected problems having occurred along the way (as Peuquet 2001 suggests may be the case with developments in space-time data representation)? Or (dare we ask) do we as a scientific community lack the necessary intellectual capacity to overcome our conceptual and technical problems? Certainly, there may be elements of truth in both these propositions—especially when compared to other scientific endeavors. In astronomy, for example, some of our planet's best minds have been continually engaged in refining our knowledge of the universe for several thousand years now, with many wrong assumptions and theories being proposed along the way. Yet it is only in the last 300-400 years that we have started to get things right—even though there are still many unexplained mysteries of the universe to be answered. If people like Galileo, Newton and Hawking—coupled with technology such as the Hubble Telescope—are needed to resolve some of astronomy's fundamental questions, perhaps we need our own equivalents to solve our more modest problems in GIScience concerning spatial data quality.

4 Conclusion

This paper has critically reviewed the problems and prospects associated with the treatment of spatial data quality during the past three decades. While the early years were characterized by warnings from leading researchers and the subsequent development of international standards that included data quality provisions, the original notion of having the tools and techniques needed to assess data quality has generally not yet been achieved. This paper has examined the current problems associated with the description of data quality, its communication and its application in real-life, and it is argued that that we still have a long way to go to fulfill our original visions in each of these areas. The solutions in some cases will slowly occur with time as user education and awareness grows with each generation of spatial data consumers. In other cases, however, greater cooperation and common focus will be needed between the different sectors of the spatial information community if we are to one day see the necessary tools and techniques either embedded in or attached to the commercial systems we now use.

References

Agumya A, Hunter GJ (1999) A Risk-Based Approach to Assessing Fitness for Use of Geographical Information. Journal of the Urban and Regional Information Systems Association 11(1): 33–44

Bedard Y (1987) Uncertainties in Land Information Systems Databases. In: Proceedings of the ACSM-ASPRS Auto Carto 8 Conference, Baltimore, Maryland, pp 175–84

Beven K (2000) On Model Uncertainty, Risk and Decision Making. Hydrological Processes 14: 2605–2606

Blakemore M (1984) Generalization and Error in Spatial Databases. Cartographica 21(2): 131–39

Boin AT, Hunter GJ (2007a) Facts or Fiction: Consumer Beliefs About Spatial Data Quality. In: Proceedings of the Spatial Science Institute Biennial International Conference (SSC 2007), Hobart, Tasmania, 14-18 May 2007, pp 721–727

Boin AT, Hunter GJ (2007b) What communicates quality to the spatial data consumer? In: Stein A (ed) Proceedings of the 2007 International Symposium on Spatial Data Quality (ISSDQ 2007), Enschede, The Netherlands, 8 pp

Brassel K, Bucher F, Stephan E-M, Vckovski A (1995) Completeness. In: Guptill SC, Morrison JL (eds) Elements of Spatial Data Quality, International Cartographic Association (ICA) Commission on Spatial Data Quality, Elsevier Science, Oxford, pp 81–108

Burrough PA (1986) Principles of Geographic Information Systems for Land Resources Assessment, Clarendon Press, Oxford

Burrough PA (1991) The Development of Intelligent Geographical Information Systems. In: Proceedings of the 2nd European Conference on GIS (EGIS '91), Brussels, Belgium, vol. 1, pp 165–174

Burrough PA, Frank AU (eds) (1996) Geographic Objects with Indeterminate Boundaries, London, Taylor & Francis, 345 pp

CEN (Comité Européen de Normalisation) (1998) European Prestandard ENV 12656: Geographic Information - Data Description - Quality, dated October 1998, CEN Secretariat, Brussels, 46 pp

Chrisman NR (1984) The Role of Quality Information in the Long-term Functioning of a Geographic Information System. Cartographica 21(2 & 3): 79–87

Chrisman NR (1991) The Error Component in Spatial Data. In: Maguire DJ, Goodchild MF, Rhind DW (eds) Geographical Information Systems: Principles & Applications, Longman, London, vol. 1, pp. 165–174

Crosetto M, Tarantola S (2001) Uncertainty and Sensitivity Analysis: Tools for GIS-based Model Implementation. International Journal of Geographical Information Science (IJGIS) 15(5): 415–447

De Bruin S, Bregt AK, Van de Ven M (2001) Assessing Fitness for Use: the Expected Value of Spatial Data sets. International Journal of Geographical Information Science (IJGIS) 15(5): 457–471

Devillers R, Bedard Y, Jeansoulin R, Moulin B (2007) Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use

of Spatial Data. International Journal of Geographical Information Science (IJGIS) 21(3): 261–283
Doucette M, Paresi C (2000) Quality Management in GDI. In: Groot R, McLaughlin J (eds) Geospatial Data Infrastructure: Concepts, Cases and Good Practice, Oxford, London, pp 85–96
Duckham M, McCreadie J (1999) An Intelligent, Distributed Error-Aware OOGIS. In: Shi W, Goodchild MF, Fisher PF (eds) Proceedings of the 1st International Symposium on Spatial Data Quality, pp 496–506
Eastman JR (1997) IDRISI for Windows: User's Guide Version 2.0. Clark University, Worcester, Massachusetts
Egenhofer MJ, Glasgow J, Gunther O, Herring J, Puequet DJ (1999) Progress in Computational Methods for Representing Geographical Concepts. International Journal of Geographical Information Science (IJGIS) 13(8): 775–796
Epstein EF, Roitman H (1987) Liability for Information. In: Proceedings of the URISA 1987 Annual Conference, Fort Lauderdale, Florida, vol. 4, pp 115–25
ESRI (2000) Using ArcCatalog. Environmental Systems Research Institute (ESRI), Redlands, California
Geoscience Australia (2006) Geodata TOPO 250K Series 3 User Guide. Geoscience Australia website, http://www.ga.gov.au/nmd/products/digidat/250k.htm, accessed 1 January 2009
Getis A (2000) Spatial Analysis and GIS: An Introduction. Journal of Geographical Systems 2: 1–3
Goodchild MF (1978) Statistical Aspects of the Polygon Overlay Problem. In: Dutton G (ed) Proceedings of the First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems, Harvard University, Massachussetts, Vol. 6, 22 pp
Goodchild MF (1992) Research Initiative 1: Accuracy of Spatial Databases - Closing Report. National Center for Geographic Information and Analysis (NCGIA), University of California, Santa Barbara, 19 pp
Goodchild MF (1993) Data Models and Data Quality: Problems and Prospects. In: Goodchild MF, Parks BO, Steyaert LT (eds) Environmental Modeling with GIS, Oxford, New York, pp 94–103
Goodchild MF (2000) The Current Status of GIS and Spatial Analysis. Journal of Geographical Systems: 2, pp 5–10
Guptill SC, Morrison JL (eds) (1995) Elements of Spatial Data Quality, International Cartographic Association (ICA) Commission on Spatial Data Quality, Elsevier Science, Oxford
Heuvelink GBM, Burrough PA, Stein A (1989) Propagation of Errors in Spatial Modeling in GIS. International Journal of Geographical Information Systems (IJGIS) 3(4): 302–322
Hudson BD (1990) Concepts of Soil Mapping and Interpretation. Soil Survey Horizons 31(3): 63–72
ISO 19113 (2002) Geographic Information-Quality Principles. International Organization for Standardization, Geneva, Switzerland
ISO 19114 (2003) Geographic information-Quality evaluation procedures. International Organization for Standardization, Geneva, Switzerland

ISO 19115 (2003) Geographic Information-Metadata. International Organization for Standardization, Geneva, Switzerland

Hunter GJ, Qiu J, Goodchild MF (1999) Application of a New Model of Vector Data Uncertainty. In: Lowell K (ed) Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources, Ann Arbor Press, Michigan, pp 201–206

MacDougall EB (1975) The Accuracy of Map Overlays. Landscape Planning 2: 25–30

Marble DF (2000) Some Thoughts on the Integration of Spatial Analysis and Geographic Information Systems. Journal of Geographical Systems 2: 31–35

Minnesota Planning (1999) Positional Accuracy Handbook, Minnesota Planning: Land Management Information Centre, http://www.mnplan.state.mn.us/press/accurate.html, accessed 12 November 2001

Moellering H (ed) (1991) Spatial Database Transfer Standards: Current International Status, Elsevier, New York

NCDCDS (1986) Working Group II on Data Set Quality: Testing the Interim Proposed Standard for Digital Cartographic Data Quality. In: Moellering H (ed) Issues in Digital Cartographic Data Standards, Report #7, U.S. National Committee for Digital Cartographic Data Standards (NCDCDS), The Ohio State University, Columbus, Ohio

NIST (National Institute of Standards and Technology) (1992) Spatial Data Transfer Standard. Federal Information Processing Standard 173, US Department of Commerce, Washington, DC

Openshaw S (1989) Learning to live with Errors in Spatial Databases. In: Goodchild MF, Gopal S (eds) Accuracy of Spatial Databases, Taylor & Francis, London, pp 263–276

Ordnance Survey (2009) OS MasterMap Topography Layer. Ordnance Survey website, www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/layers/topography/index.html, accessed 1 January 2009

Qiu J, Hunter GJ (2002) A GIS with the Capacity for Managing Data Quality Information. In: Goodchild MF, Fisher PF, Shi W (eds) Spatial Data Quality, Taylor & Francis, London, pp 230–250

Peuquet DJ (2001) Making Space for Time: Issues in Space-Time Data Representation. GeoInformatica 5(1): 11–32

Robinson VB, Frank AU (1985) About Different Kinds of Uncertainty in Collections of Spatial Data. In: Proceedings of the Seventh International Symposium on Computer-Assisted Cartography (Auto Carto 7), Washington, D.C., pp 440–449

Sadiq Z, Duckham MD (2009) Integrated Storage and Querying of Spatially Varying Data Quality Information in a Relational Spatial Database. Transactions in GIS 13(1): 30–42

Van Oort P, Bregt AK (2005) Do Users Ignore Spatial Data Quality? A decision-Theoretic Perspective. Risk Analysis 25(6): 1599–1610