Chapter 18

DATA QUALITY IN MASSIVE DATA SETS

Michael F. Goodchild
National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA
good@ncgia.ucsb.edu

Keith C. Clarke
National Center for Geographic Information and Analysis, and Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA
kclarke@geog.ucsb.edu

Abstract All data contain errors, and large spatial data sets are especially prone because they contain data from multiple sources, and use different assumptions about structure and semantics. The general issue is one of data quality assurance, defined in terms of lineage, completeness, logical consistency, attribute accuracy, and positional accuracy. We review a series of quality metrics suitable for empirical description of data quality, and consider some of the special issues of quality related to spatial data, especially the need to incorporate visualizations of data quality into graphics and maps. We conclude that data quality is an essential component of software for spatial data handling, including geographic information systems.

Keywords: Data quality, Metrics, Spatial data, Geographic information systems.

1. The Quality Problem

Data sets capture facts and enable their management and retrieval. Yet in almost all instances there exists some external reference to which the data set's version of the facts can be compared. A data set created by digitizing the contents of a published book can be checked against the original; a measurement of temperature recorded in a data set can be checked against an independent measurement; and a real-estate agent's

J. Abello et al. (eds.), Handbook of Massive Data Sets, 643–659.

record of the price of a house purchase can be compared against the local tax assessor's data set. Data *quality* is quantifiable from the results of such checks, and data set contents that fail to match their own or independent reference sources are said to contain *errors*, or to be of poor quality. Measures of data quality can be devised, based on the frequency of errors, or on their magnitudes. Where there is no reference source of equal or superior quality, then the fact being recorded is based on inadequate definitions and is inherently vague and error-prone. For example, there is no way of checking the quality of the statement "it is cold here" against measurements of temperature, or location, although the statement may be a correct representation of what was said, and captured into the data set. Since a perfect match between data and the real world is generally impossible, we conclude that error as defined here must be endemic in data sets.

Many different terms are associated with quality, or the lack of it, and there is little consensus about their precise meanings. Imprecision, inaccuracy, inexactness, vagueness, uncertainty, unreliability, and incompleteness all imply lack of quality in some sense. Nevertheless, the terms capture divergent forms of variation between data set and reference, or different sources of difference.

Data quality is an important issue for massive data sets, because poor quality implies that decisions based on data set content will also be poor, and because massive data sets may have been assembled quickly, from multiple sources, at multiple scales, from sources with inherent vagueness, or with little concern for quality. Massive data sets once gloriously isolated by their size or complexity now find themselves open to searching and use by millions over the World Wide Web, regardless of their quality. High quality can be expensive, particularly if it involves human intervention in verification and if many or all data set records have to be checked.

Poor-quality can itself result in high costs, which may exceed the costs of correction. Data sets may be used for regulation, where poor-data quality may be the cause of costly litigation, particularly if it can be shown that the developers and users of data sets failed to take adequate actions to maximize quality or to deal with the known consequences of poor quality. Cartographic examples of missing map features or mislocated buildings abound, as in the case of the 1998 ski-lift accident in Italy, or the 1999 accidental bombing of the Chinese Embassy in Belgrade. Poor quality data sets used for scientific research cast doubt on the quality of the resulting scientific conclusions. Users of poor quality data sets quickly become frustrated once products are found to be unreliable. Errors and uncertainties *propagate* from the data set to products

and decisions derived from it, including answers to queries, results of analysis, and transformations. Users of data sets need to know something of the inherent quality of a data set's contents in order to assess the fitness of the data set for specific purposes, and to determine the quality of products derived from the data set. Such information can be communicated in the form of text, but visualization also provides an important tool for informing users about quality.

This chapter is structured as follows. The next section deals with the description and representation of quality in data sets, and the techniques that have been devised for communicating knowledge of quality through visualization. This is followed by a section on the implications of quality, with discussion of the state of knowledge in propagation. The chapter uses the example of spatial data sets frequently, in part because research on them has advanced to a significant degree, and many results have been incorporated into standards and practice; and in part because quality has added dimensions and significance for spatial data.

2. Elements of Quality

One of the most comprehensive analyses of data set quality is found in Federal Information Processing Standard 173, otherwise known as the Spatial Data Transfer Standard (`www.fgdc.gov`; for a more extensive discussion of the elements of spatial data quality see Guptill and Morrison (1995). Devised in the mid 1980s, it identifies five components of quality for spatial data, as follows:

- *Lineage*, defined as information about the process of creation of the data set, such as the instruments used to make measurements, the identities of individuals and agencies responsible for creation, and the standards used to define the data set's contents. By knowing such details, it is possible in many cases to make inferences about quality. For example, knowing the identity of the instrument used to acquire measurements often allows the user to make meaningful estimates of their accuracy. Lineage also serves another useful purpose by providing feedback – for example, if serious errors are found in data it might be possible to link them to specific faults in the production process. It is the data set lineage that answers science's call for documentation permitting repeatability of experimental results, and therefore the independent confirmation of findings necessary for the scientific method.

- *Completeness*, or the degree to which the data set captures all of the expected data. Completeness is often linked to the currency of the data, or the degree to which the data represent current condi-

tions, or conditions that existed at some point in the past and for which the data set is intended to form a complete representation. Currency is a significant problem for digital data sets, especially when the date for which the data are intended to be valid differs from the date of construction of the data set, or if either of these dates are not precisely defined, or if different versions of the data set are not clearly identified. Completeness can also refer to spatial extent, the number of available attributes actually included, and to known missing data. Many data sets for the United States, for example, actually exclude Alaska, Hawaii, and the United States Territories, and variable numbers of attributes for each state.

- *Logical consistency.* This refers to the internal consistency of the data, and the data set's adherence to its own defined rules. For example, logical consistency is violated if an object has two unique identifiers, or if the value of an attribute falls outside its defined domain. In spatial data sets, there can be logical inconsistency between the geometric content of a data set (a point lies inside the boundary defining California) and the topological content (the point has an attribute indicating it is in Nevada). If the rules are well-defined, then it is in principle possible to detect errors of logical consistency without human intervention, and it may also be possible to correct them. Such corrections require their own rules (does geometry over-ride topology, forcing the attribute to be changed to California, or does topology over-ride geometry, forcing the point to be moved to the geometric center of Nevada?), and it is difficult to avoid rules that create their own conflicts (moving the point may be problematic if it is connected to another object – for example, if the point is part of a lake shoreline).

- *Attribute accuracy.* This refers to the accuracy of the recorded attributes associated with each object. In a spatial data set, each object – a road, a mountain, a lake, a city, a house – will have certain defined attributes. These might include a unique identification number, a name, or in the case of a city the current population. Attributes can be differentiated in various ways by type. They may be *qualitative* (e.g., name) or *quantitative* (e.g., population count), and more elaborate schemes exist (see, for example, Chrisman (1997)). From the perspective of quality, it is important to distinguish between cases where an attribute can be only *right* or *wrong*, and cases where it is possible to define degrees of correctness. In the former instance, quality is best measured by the proportion of errors, but in the latter case many methods are

available for measuring quality, and many of these are discussed in the next section. For example, a misspelled name of a city is more right (and possibly open to automatic correction) than a name that is completely wrong (e.g., in the case of Pittsburgh, *Pittsburg* is less erroneous than *Pittston*). Finally, correctness may be defined with reference not to reality but to the measurement process. The debate over the use of sampling in the Year 2000 U.S. Census, for example, has led to legislative prohibition of the methods that could have provided the most accurate results given the available budget. Yet the census itself assumes that the population's street addresses on April 15th, 2000 are their actual "places" as far as the federal government is concerned.

- *Positional accuracy.* The position recorded for an object on the Earth's surface can never be perfectly accurate, since the instruments available for measuring position (surveying instruments, or the Global Positioning System) have limited accuracy, and the positions of the reference objects (the Poles, Equator, and Greenwich Meridian) are also not perfectly defined. Even well-known positional reference systems, such as the latitude and longitude of geographic coordinates, require, at the minimum, knowledge of the Earth model, its size and shape, and the vertical datum in use. Standard coordinate systems such as the Universal Transverse Mercator have inherent positional accuracies of about 1 part in 2000, with systematic error depending on position. In some cases it may be impossible to separate positional accuracy from attribute accuracy. For example, it may be impossible to determine in the case of a measurement of the elevation of a point above sea level whether the correct elevation has been recorded at the wrong point, or whether the wrong elevation has been recorded at the correct point. Nevertheless, spatial data with only limited positional accuracy or precision, such as digital versions of coarse-scale maps, can still have immense scientific value and may need to be used in combination with data of different levels of quality.

This five-component scheme is recognized by being written into a major U.S. standard, but many other terms have been proposed, often with conflicting definitions, to capture the elements of data quality. Many forms of data are inherently *vague*, because it is impossible to decide with certainty what the correct value should be. For example, it is impossible to determine when something should be described as cold. Such evaluations are often termed *subjective*, because there is no reason to expect any two people to agree on the correct value – they are not

replicable. Many scientists would argue that such data have no value, but others would argue that vagueness of communication is an indispensable part of human existence.

Empirical scientists often distinguish between *accuracy*, or the degree of agreement between a recorded observation and its true value, and *precision*, or the degree of detail with which the measurement is recorded. A widely recognized principle holds that precision should never exceed accuracy. For example, if a thermometer is capable of measuring to the nearest Celsius degree, then recorded measurements should never include decimal places (e.g., 21 is acceptable but 20.986 is not). But precision is also used to refer to the variation among repeated measurements of the same phenomenon with the same instrument.

3. Description of Quality

3.1. Numeric values

Consider a measuring instrument such as a thermometer, and suppose that it is being used to measure a temperature whose correct value is 21.0 Celsius. The thermometer is inherently inaccurate, and returns a value of 23. By repeatedly comparing true and measured values it is observed that the thermometer's measurements are in error by amounts ranging from -2 to $+2$ Celsius. So a straightforward way to record quality would be by specifying the *range*. In a data set, this could be recorded in the form of additional attributes – for example, as $\langle 23, +2, -2 \rangle$. The query "Is the temperature greater than 26?" would return "no", but the query "Is the temperature less than 22?" would return "maybe".

Range provides an easy means of responding to simple queries, but it is problematic because it provides no information on the relative frequency of large and small errors. In reality, it is almost always true that the thermometer will produce small errors more frequently than large ones. Moreover, if large errors are rare, it will be difficult to provide an accurate estimate of range without making a very large number of tests. Fortunately, it is known that under a wide range of circumstances the relative frequencies of large and small errors are consistent with a simple model, known as the *Gaussian* or *normal* distribution, the *error function*, or the *bell-curve*, and shown in Figure 18.1. As a *probability density function*, the probability of an observation lying between any two values of the x-axis is defined by the area under the curve between those limits. The width of the curve is best defined by the distance between the center and the points of inflection, and is known as the *standard deviation.* The instrument is said to be biased if the mean error is not zero. Finally, the *standard error* or *root mean squared error* (RMSE) is

defined as the square root of the mean squared difference from the true value:

$$\text{RMSE} = \left[\frac{1}{n}\sum_i (x_i - X)^2\right]^{1/2},$$

where X is the true value, n is the number of observations, and x_i denotes an observation, when the number of such observations is very large.

The points of inflection shown in Figure 18.1 represent one standard deviation on either side of the mean. Approximately 68% of errors will be smaller than one standard deviation, and 32% will be larger. More useful perhaps is the fact that 95% of errors will lie within 1.96 standard deviations of the mean, or approximately 2 standard deviations. This is the basis for the *confidence limits* commonly heard in association with opinion polls – for example, that the true value "will lie within 2 percentage points 19 times out of 20".

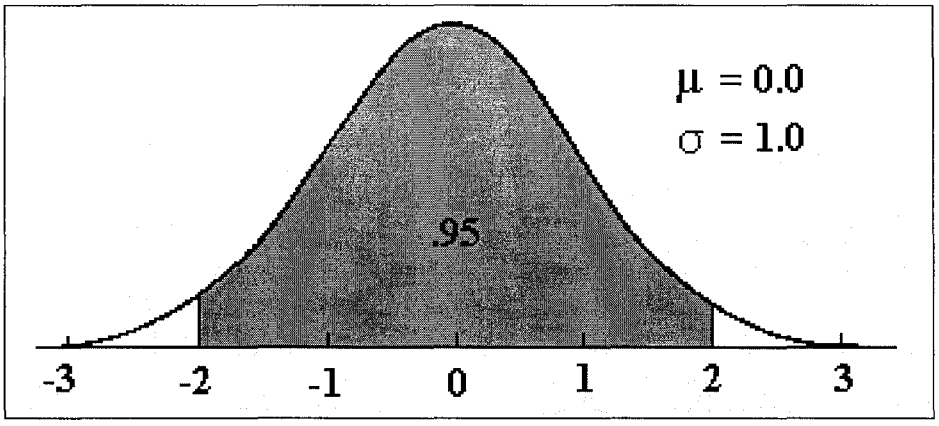

Figure 18.1. The Gaussian distribution with a mean of 0 and a standard deviation of 1, showing the probability of a value lying between 2 standard deviations on either side of the mean. Note the points of inflection (change of curvature from upward-facing to downward-facing) at 1 standard deviation on either side of the mean.

While the Gaussian distribution is often a very accurate model of errors, it is not as easy to apply to the resolution of queries. But with a little effort, it is possible to replace the "maybe" response of the earlier example with a precise estimate of the probability that the temperature is less than 22, given knowledge of the recorded observation, and of the mean and standard deviation of the error distribution.

3.2. Qualitative values

The previous section discussed attributes that have numeric qualities. Suppose now that a data set contains a qualitative attribute, such as the names of streets. In most such cases a simple approach is to estimate the proportion of such attributes that are correct, and the proportion that are in error, and to attach these proportions to the results of queries. A common instance of errors in qualitative data occurs in the accuracy assessment of certain types of maps. For example, a map of land use might be prepared by classifying a remotely sensed scene from a satellite. A scene from the Enhanced Thematic Mapper Plus instrument on the Landsat 7 satellite consists of an array of picture elements (*pixels*) that are approximately square and $15m$ on a side on the Earth's surface. To test the accuracy of the automated classification, a random sample of locations is selected on the ground, and visited to determine the actual land use. Table 18.1 shows a hypothetical result of checking 100 such points.

Table 18.1. Results of accuracy assessment of a map of land use (rows indicate recorded values, columns indicate ground truth.

	Residential	Open space	Agriculture	Woodland	Water	**Totals**
Residential	33	3	0	1	0	**37**
Open space	2	24	4	0	0	**30**
Agriculture	0	1	17	2	0	**20**
Woodland	0	0	2	5	0	**7**
Water	0	0	0	0	6	**6**
Totals	**35**	**28**	**23**	**8**	**6**	**100**

From Table 18.1, it is apparent that the recorded value (row) agreed with ground truth (column) in a total of 85 cases out of 100, since 85 cases lie on the main diagonal of the table. Thus a convenient way of summarizing accuracy is to say that the probability of a randomly chosen point having the correct recorded value is 0.85. But the table clearly contains much more useful information. Water is never confused with

any other class, since it is easy to identify correctly in satellite images. Of the 35 points that are truly residential, 33 are correctly classified, but 2 are confused with open space. Agriculture is also sometimes confused with open space, and sometimes with woodland.

Suppose that a user queries the data set to determine the class of land use at a point. In general, we can say that the probability of a correct response is 0.85. But if the land use recorded in the data set is residential, we know from the table that a better estimate of correctness is 33/37, or 0.89, with probabilities of 0.08 that the truth is open space, and 0.03 that the truth is woodland.

3.3. Fuzziness

The previous section was based on an implicit assumption – that the class at a point must be one of the five recognized options. In reality, the area covered by a single pixel may be a mixture, for example at the edge of a lake, so that the truth is not 100% water or 100% woodland but some mixture of the two. Recently, there has been much research on so-called mixture methods, which attempt to identify the percentages of various pure classes present in a mixed pixel (see, for example, Gillespie (1992)).

At a more fundamental level, however, it may be impossible to define such categories as residential precisely, because the term itself implies a mixture of different surfaces: roof, concrete, asphalt, grass, water (pool). Rather, the set of pixels labeled residential is fuzzy, with poorly defined properties. Fuzzy set theory has become popular in recent years as a way of dealing with situations in which assignment to classes is overly restrictive (see, for example, Zhu et al. (1996)).

In fuzzy set theory, membership in a class is measured on a continuous scale that is often constrained to the range [0,1]. A pixel that is most like the pure concept of residential is assigned the highest membership value, while one that has nothing in common with residential is assigned 0. The memberships for a pixel can be conceived as a vector $\{m_1, m_2, \ldots, m_n\}$, where n is the number of classes, and m_i denotes the membership of the pixel in the i-th class.

Fuzzy set theory is attractive in dealing with uncertainty in categorical data because it admits degrees of belonging, and thus approximates the way humans think about classes of land use, or any categories defined by complex or subjective measurements such as soil type, flood risk, or land suitability. An observer might well be able to distinguish between areas that are more residential and areas that are less so, or to agree that the degree of "residentialness" declines as one moves away from

a city's center. Reasoning is also possible based on fuzzy sets, using certain axiomatic propositions to manipulate degrees of fuzziness. On the other hand, it seems dubious to claim that a degree of membership assigned by one observer has any meaning to another observer, when neither the class itself nor the scale of measurement of membership are well-defined.

3.4. Metadata

Metadata are defined as data about data; they include descriptions of the general properties of a data set, including its structure, format, language, and definitions; and also information about quality, ownership, and other properties that are useful to potential users. Metadata are analogous to the information in a library's card catalog, or to the information printed at the front of a book, or on the outside of a package.

If a data set is passed without explanation or documentation from one person to another, it can amount to little more than a confusing mass of binary digits. Metadata are "what make data useful" in the words of Francis Bretherton. They allow a user to assess the fitness of data for a particular application, particularly with respect to quality. Lack of metadata can also contribute to lack of quality, if a user makes the wrong assumptions about the data's meaning. For example, a user might see a set of numbers labeled "temperature," and not knowing the scale of measurement might wrongly assume that the scale was Celsius rather than the intended Fahrenheit. In effect, this would introduce an error in every value other than -40.

Quality description is an important component of metadata, especially for spatial data. The Content Standards for Digital Geospatial Metadata, created by the U.S. Federal Geographic Data Committee (`www.fgdc.gov`), include extensive and precise descriptions of quality, using the five components discussed earlier. The approach has been described as *truth in labeling*, since it attempts to elicit from the creator of the data as much useful information as possible about quality, but sets no absolute standards or thresholds of quality that must be met. Thus a data set with a quality statement that reads "This data set has no quality statement" is fully compliant with the standard, but also has information of value to the user in making decisions about the quality of the data.

Unfortunately, the metadata approach falls far short of a complete solution to the problem of describing quality, for several reasons. First, it favors descriptions that apply uniformly to the entire data set, such as a single measure of positional accuracy. In reality, however, it is

common for elements of a data set to have different levels of quality, and quality may need to be defined at the level of the class of object, the individual object, or even the individual attribute. For geospatial data, it is common for quality to vary geographically, and many topographic maps include a much smaller map inset indicating how the quality of the main map varies.

Second, the metadata approach implies that quality can be described adequately without substantial restructuring of the data, by adding appropriate *slots* to the existing data model. Consider the case of a geographic region, such as the Atlantic Ocean, represented in a data set as a *polygon*, a series of points delimiting the ocean's boundary in clockwise or counter-clockwise order. In reality the Atlantic Ocean is not well-defined, and we might wish to describe its quality by adding suitable descriptors to the data set. One way to do this is to create a fuzzy region, by conceiving of a continuous variable p such that the value of p at some point is the degree of membership of that point in the concept *Atlantic Ocean*. To represent the spatial variation of p, however, we would have to abandon the polygon representation, and adopt a raster or some other way of describing what is now a continuous surface. In other words, description of quality has forced a change of data model (Burrough and Frank 1996).

Third, the metadata approach implies that it is possible to achieve a complete description of quality that is intelligible and useful to the user. In practice, description of quality through appropriate models can be exceedingly complex. The Gaussian error model described earlier is among the simplest of statistical models of error, yet even it is a sophisticated statistical concept. In geospatial data, it is common for the error affecting one object to be similar to the error affecting other objects, especially if the two objects are close to each other and if they have been measured by the same process. For example, suppose a map is created from an aerial photograph. One of the sources of positional error is misregistration of the photograph; and this form of error will affect all objects mapped from the same aerial photograph to varying degrees. Positional errors of objects are frequently *correlated*, and the degree of correlation is found to vary inversely with distance.

Because of positive correlations, the *relative* accuracy of the positions of nearby objects can be much higher than the *absolute* accuracy, and much higher than is implied by general descriptive measures such as the RMSE that are contained in metadata. Relative properties such as ground slope can be estimated accurately from digital elevation data even though absolute elevations in the data set are of poor quality,

provided errors show strong positive correlations over short distances (Hunter and Goodchild 1997).

Many models of correlated errors exist, but they are not widely known outside the research community, and their use in metadata to describe quality is therefore highly problematic, since most users are not equipped to understand or deal with them. To address this issue, Goodchild et al. (1999) have argued that the concept of metadata should be broadened to include *methods*. Instead of the parameters of a complex error model, a producer should provide code that simulates the error model, producing a sample of versions of the data set that represent the range of variation due to uncertainty. Suppose, for example, that one wished to describe the uncertainty associated with a forecasted high temperature of 25 Celsius. Someone familiar with the Gaussian error model would understand the statement that uncertainty was characterized by a standard error of 2. But the same information is contained in the simulated set $\{26, 24, 23, 28, 21, \ldots\}$ if these are generated using an appropriate code. Goodchild et al. (1999) apply the same concept to the much more complicated case of geospatial data sets, arguing that the concept is no more difficult in the latter case. Although the models are far more complex, they need only be understood by the creators of the data and the simulation code, not by the users.

3.5. Visualization

Visualization provides an attractive medium for communication of information about quality. Visualization has already proven its effectiveness as a way of searching massive data sets for pattern and structure. The existence of uncertainty can be conveyed by removing, blurring, or greying, or by changing the visual depth of objects, bringing certain objects to the front and pushing uncertain objects to the back. Visualization of large spatial data sets as a method of communicating information about geospatial data quality has been the subject of intensive research (Beard et al. 1991, Davis and Keller 1997) and was reviewed more recently by Clarke and Teague (1998).

MacEachren (1992) investigated the use of existing map methods for uncertainty depiction, and introduced the variable of visual focus, shown by crispness, fill clarity, fog, and resolution variation used to adjust the boundaries between map features. More certain objects were depicted as "a sharp, narrow line" and less certain features as "a broad, fuzzy line that fades" toward the periphery. McGranaghan (1993) examined realism and time as potential variables for symbolization. Objects of lower data quality appear more "cartoonish" if data quality is low and

more realistic if data quality is high. Time-based methods necessarily involve animation and several methods utilizing time as a cartographic variable were considered, including blinking, fading, and moving. The amount of time a blinking object is present or absent reflects quality information. Fading can be employed by having an object on the map oscillate to reflect quality; McGranaghan showed a stream segment oscillating between green and blue (high confidence) or green and red (low confidence).

Animations showing multiple realizations of a data set, and associated with the range of uncertainty described above, have been employed by Fisher (1993) and Ehlschlaeger et al. (1997). Fisher used animation techniques based on his earlier research to depict positional uncertainty in soil maps. Soil inclusion information, provided by the data producer, is conveyed to the user through an animated soil map that uses randomization to show these inclusions within the predominant soil types. Cells are continuously and randomly selected based on given inclusion rate, producing a stochastic realization of soil type distribution at any point in time. Ehlschlaeger et al. (1997) utilized animation to display multiple stochastic realizations of output from least – cost path analyses based on coarse resolution terrain data. Using multiple possible elevation surfaces, a series of cost surfaces for a least-cost path algorithm were produced showing the resulting shortest path. Each realization was used as a frame to create a smooth animation.

The integration of uncertainty information and data into a single display without graphic overloading was explored by Wittenbrink et al. (1996) through an approach called verity visualization. This method includes uncertainty visualization using uncertainty glyphs, fat surfaces, perturbations, and oscillations. Uncertainty glyphs, using various graphic variables such as size and shape of an icon to depict data attributes, are placed on the visualization or map itself to indicate uncertainty at different locations. Fat surfaces indicate uncertainty in information by presenting a range of data values at each location on the surface. Finally, Clarke et al. (1999) have advocated using visual depth in virtual-reality-based representations of data, so that the "nearness" of the data to the viewer portrays uncertainty using some of the variables already discussed, such as color and animation. So, for example, as the data user zooms in on a feature, it wobbles more or less depending on its uncertainty.

In spite of this promising research, in the case of geospatial data, it is clear that modern cartographic practice has traditionally left little room for uncertainty, and the practices of the past – leaving areas blank, inserting mythical beasts – have now largely disappeared. Research shows

that users need to be cued to expect uncertainty, but that once appropriate instructions have been given, have no difficulty associating grayness, blurring, or even shaking with uncertainty (MacEachren 1992). As the research in this area yields results of use in everyday practice, two types of user interfaces between the data and the uncertainty seem possible. In the first, the treatment of uncertainty is as another layer of the map, subject to viewing, and use in analytical operations. In this method, the use of multiple realizations, all of equal possibility given the error bounds, is one way to show uncertainty and estimate its propagation into results (Journel 1996). Alternatively, uncertainty can be integrated into the visual display of the information, and activated by the data user when it becomes of concern during the analysis of information. Either way, the revised role for uncertainty in the use of data from massive data sets is significantly enhanced. Visualization offers a promising suite of methods for informing the data user about uncertainty.

4. Working with Poor Data

References have been made to queries based on uncertain data. More generally, the term propagation refers to the impact of uncertain or erroneous data on the results of query, analysis, and modeling. For example, consider a square parcel of land $100m$ on a side, with a true area of 1 hectare. Suppose that the corner points are inaccurately surveyed, with a mean error of 0 and a standard error of $2m$ in both coordinates. If the errors are uncorrelated, it is possible to compute the standard error in the estimate of area (Chrisman and Yandell 1989); in this case, the result is $200m^2$. If the errors have a perfect positive correlation (in other words, are identical), then the error in area is 0, since the square moves under error as a rigid body without rotation or warping. Thus the manner in which error propagates into the result – the estimate of area – depends on the nature of the error.

The classical theory of measurement provides a basis for analysis of propagation in numeric data. Suppose that some scalar measurement, such as a measurement of temperature using a thermometer, is distorted by an error generated by the measuring instrument. The apparent value of temperature x' can be represented as the sum of a true value x and a distortion δx. If some manipulation of x is required, the theory of measurement error provides a simple basis for estimating how error in x will propagate through the manipulation, and thus for estimating error in the products of manipulation (Taylor 1982) (see Heuvelink (1998), and Heuvelink et al. (1989), for discussions of this in the context of geospatial data). Suppose that the manipulation is a simple squaring,

$y = x^2$, and write δy as the distortion that results. Then:

$$\begin{aligned} y + \delta y &= (x + \delta x)^2 \\ &= x^2 + 2x\delta x + \text{terms of order } \delta x^2. \end{aligned}$$

Ignoring higher-order terms, we have:

$$\delta y = 2x\delta x.$$

More generally, given a measure of uncertainty in x such as its standard error σ_x, the uncertainty in some $y = f(x)$, denoted by σ_y, is given by:

$$\sigma_y = \frac{df}{dx}\sigma_x.$$

The analysis can be readily extended to the multivariate case and the associated partial derivatives.

In most cases, however, the analysis that results in the product y will be much too complex to represent as a single function f, and in cases where a function exists it may be non-differentiable. Simulation provides an alternative that is more general, more straightforward conceptually, and also more suited to non-numeric data. A series of inputs is generated, representing the variation in the data due to uncertainty, error, or poor quality. Each input is then analyzed, to create a series of outputs. Uncertainty in the output can be represented through some measure, such as RMSE, or by visualization.

5. Final Comments

Quality remains a major issue for users of massive data sets, especially when the data were created by people or processes remote from the user. Humans are often faced with having to take information at face value, and have developed complex arrangements and conventions as the basis for trust. For example, we trust information we read in certain newspapers because we trust the newspaper's staff and news-gathering processes.

Many of these conventions fail in the case of digital data. Electronic networks make it easy for data sets to be copied and transferred without identification of the creator, and make it easy for data from different sources to be merged, creating products with heterogeneous quality. Metadata are expensive to create, and owners of data often lack the motivation to create them in advance of use. Finally, few software products offer the ability to handle information on quality, or to propagate it appropriately to new data or results of analysis. Nevertheless, much

progress has been made in recent years, and new products now becoming available are much more likely to provide metadata services, and to support handling, visualizing, and propagating information about quality.

Bibliography

M.K. Beard, B.P. Buttenfield, and S.B. Clapham. NCGIA Research Initiative 7: Visualization of spatial data quality. Technical Report 91-26, National Center for Geographic Information and Analysis, 1991.

P.A. Burrough and A.U. Frank. *Geographic objects with indeterminate boundaries.* Taylor and Francis, 1996.

N.R. Chrisman. *Exploring geographic information systems.* Wiley, 1997.

N.R. Chrisman and B. Yandell. Effects of point error on area calculations. *Surveying and Mapping*, 48:241–246, 1989.

K. Clarke, P.D. Teague, and H.G. Smith. Virtual depth-based representation of cartographic uncertainty. In W. Shi, M.F. Goodchild, and P.F. Fisher, editors, *Proceedings of the International Symposium on Spatial Data Quality '99*, pages 253–259, 1999.

K.C. Clarke and P.D. Teague. Cartographic symbolization of uncertainty. In *Proceedings, ACSM Annual Conference*, 1998. CD-ROM.

T.J. Davis and C.P. Keller. Modelling and visualizing multiple spatial uncertainties. *Computers and Geosciences*, 23:397–408, 1997.

C.R. Ehlschlaeger, A.M. Shortridge, and M.F. Goodchild. Visualizing spatial data uncertainty using animation. *Computers and Geosciences*, 23:387–395, 1997.

P.F. Fisher. Visualizing uncertainty in soil maps by animation. *Cartographica*, 30:20–27, 1993.

A.R. Gillespie. Spectral mixture analysis of multispectral thermal infrared images. *Remote Sensing of Environment*, 42:137–145, 1992.

M.F. Goodchild, A.M. Shortridge, and P. Fohl. Encapsulating simulation models with geospatial data sets. In K. Lowell and A. Jaton, editors, *Spatial accurary assessment: Land information uncertainty in natural resources*, pages 131–138. Ann Arbor Press, 1999.

S.C. Guptill and J.L. Morrison. *Elements of spatial data quality.* Elsevier, 1995.

G.B.M. Heuvelink. *Error propagation in environmental modelling with GIS.* Taylor and Francis, 1998.

G.B.M. Heuvelink, P.A. Burrough, and A. Stein. Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, 3:303–322, 1989.

G.J. Hunter and M.F. Goodchild. Modeling the uncertainty of slope and aspect estimates obtained from spatial databases. *Geographical Analysis*, 29:35–47, 1997.

A.G. Journel. Modelling uncertainty and spatial dependence: Stochastic imaging. *International Journal of Geographical Information Systems*, 10:517–522, 1996.

A.M. MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, 13:10–19, 1992.

M. McGranaghan. A cartographic view of spatial data quality. *Cartographica*, 30:8–19, 1993.

J.R. Taylor. *An introduction to error analysis: The study of uncertainties in physical measurements.* University Science Books, 1982.

C.M. Wittenbrink, A.T. Pang, and S. Lodha. Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 2:266–279, 1996.

A.X. Zhu, L.E. Band, B. Dutton, and T.J. Nimlos. Automated soil inference under fuzzy logic. *Ecological Modelling*, 90:123–145, 1996.