12
Data quality parameters

H VEREGIN

This chapter focuses on the definition and measurement of quality components for geospatial databases. A basic distinction is drawn between quality control and truth-in-labelling paradigms. Components of data quality – accuracy, precision, consistency, and completeness – are defined in the context of geographical data. Treatment of quality components in data standards is discussed and the implications of cartographic bias in geospatial data models are briefly addressed. The chapter ends with a discussion of the ways in which institutional values are embedded in geospatial databases and the ways that data quality documentation can help to articulate these values.

1 DEFINING DATA QUALITY

The meaning of 'quality' depends on the context in which it is applied. The term is commonly used to indicate the superiority of a manufactured good or attest to a high degree of craftsmanship or artistry. In manufacturing industries, quality is viewed as a desirable goal to be achieved through management of the production process. Statistical quality control has a relatively long history in manufacturing, where it is used to ensure conformity in products by predicting the performance of manufacturing processes.

Quality is more difficult to define for data. Unlike manufactured products, data do not have physical characteristics that allow quality to be easily assessed. Quality is thus a function of intangible properties such as 'completeness' and 'consistency'. On further reflection, however, these differences are perhaps not as great as they might initially seem. After all, data are the result of a production process, and the manner in which this process is performed clearly affects data reliability. Data consumers can therefore use the same diligence in selecting a database that they might in purchasing an automobile or a pair of shoes.

These comments also apply in the context of geospatial data. Concern for geospatial data quality has burgeoned in recent years for the following reasons:

- Increased data production by the private sector. Historically, mass production of geospatial data was the domain of governmental agencies such as the US Geological Survey (USGS) and the British Ordnance Survey (Rhind, Chapter 56). Unlike these agencies, private companies are not required to conform to known quality standards (Goodchild and Longley, Chapter 40).
- Increased use of GIS as a decision-support tool. This trend has led to realisation of the potential deleterious effects of using poor quality data, including the possibility of litigation if minimum standards of quality are not attained (Onsrud, Chapter 46).
- Increased reliance on secondary data sources. This has been fuelled by a reduction in accessibility and cost constraints resulting from network accessibility and the development of standards for data exchange (Goodchild and Longley, Chapter 40).

These trends have contributed to a reappraisal of the responsibilities of data producers and consumers for data quality. Until quite recently data quality was the responsibility of the producer, and compliance testing strategies were applied in order to sanctify databases meeting official quality thresholds. Compliance testing is a form of quality control that

seeks to eliminate error through management of the database production process. However, compliance tests are useful only in a limited range of applications environments. For some applications a particular compliance test may be too lax while for others it may be too restrictive and hence impart unnecessary costs.

Responsibility for assessing whether a database meets the needs of a particular application has therefore shifted to the consumer who is in a position to make such an assessment. This is referred to as determining 'fitness-for-use'. The producer's responsibilities have changed as well. Rather than producing authoritative databases, the producer's role has shifted to data quality documentation or 'truth-in-labelling'. The truth-in-labelling paradigm views error as inevitable and casts the data quality problem in terms of misuse arising from incomplete knowledge of data limitations.

2 DATA QUALITY COMPONENTS

Geographical observations describe phenomena with spatial, temporal, and thematic components (Berry 1964; Sinton 1978). Space, which defines geographical location, is the dominant member of this troika. This dominance is problematic on several levels. First, time is not given sufficient attention. Although poorly accommodated in conventional geospatial data models, time is critical to an understanding of geographical phenomena, not as entities that exist at some location, but as events that appear and disappear in space and time (Peuquet, Chapter 8; Raper, Chapter 5). A second problem is that geographical phenomena are not really about space, but about theme. We can view space (or more precisely space-time) as a framework on which theme is measured. It is true that without space there is nothing geographical about the data, but on the other hand without theme there is only geometry.

These comments set the stage for our discussion of data quality components. Like geographical phenomena, data quality can be differentiated in space, time, and theme. For each of these dimensions, several components of quality (including accuracy, precision, consistency, and completeness) can be identified.

2.1 Accuracy

A useful starting point for discussing accuracy is the entity–attribute–value model, which serves as the conceptual basis for most database implementations of real-world phenomena. According to this model, 'entities' represent real-world phenomena (such as streets, counties, or hazardous waste sites), 'attributes' specify the relevant properties of these objects (such as width or number of lanes), and 'values' give the specific qualitative or quantitative measurements pertaining to a particular attribute. In this model, error is defined as the discrepancy between the encoded and actual value of a particular attribute for a given entity (see also Fisher, Chapter 13). Accuracy is the inverse of error. This model can be used to define spatial, temporal, and thematic error for a particular entity as, respectively, the discrepancies in the encoded spatial, temporal, and thematic attribute values.

This definition is useful but somewhat limited. What is missing is recognition of the interdependence of space, time, and theme. Geographical phenomena are not just thematic data with space and time attached. They are instead events unfolding over space and time. A change in space or time implies a change in theme, and vice versa. Thus while accuracy can be measured separately for space, time, and theme, these measurements are not necessarily independent. Consider a database dated '1992' that depicts a two-lane road, but assume that in late 1991 the road was converted to a four-lane highway. This is both a thematic error (because in 1992 there were four lanes, not two) and a temporal error (because when the road contained only two lanes the year was at most 1991). Similar types of dependencies exist across space and theme. A classic example is the soil mapping unit delineation problem, in which a mislocated unit boundary is simultaneously a spatial error and a thematic error, since boundary location is defined by variations in thematic attribute value.

The definition of error given above assumes that there is some objective, external reality against which encoded values can be measured (Chrisman 1991). This definition requires not only that 'truth' exists but that it can be observed. Quite apart from any philosophical problems that it raises, this definition is problematic for several reasons. First, the truth may simply be unobservable, as in the case of historical data. Second, observation of the truth may be impractical (because of data cost, for example). Finally, it is possible that multiple truths exist because the entities represented in the database are abstractions rather than real-world phenomena. Indeed many phenomena of interest belong to perceived reality (sometimes referred to as *terrain*

nominal: Salgé 1995). Examples include entities that are highly variable (e.g. shorelines) or subjective in nature (e.g. land cover classes interpreted from air photos). In these cases inexactness is a fundamental property of the phenomena under observation (Goodchild 1988b).

Fortunately, objective reality does not need to be articulated in order to perform accuracy assessment. This is because geospatial data are always acquired with the aid of a model that specifies, implicitly or explicitly, the required level of abstraction and generalisation relative to real-world phenomena (Figure 1; Martin, Chapter 6). This conceptual model defines the database 'specification' and it is against this reference that accuracy is assessed (Brassel et al 1995). Accuracy is a relative measure rather than an absolute one, since it depends on the intended form and content of the database. Different specifications can exist for the same general types of geospatial data. To judge the fitness-for-use of the data for some applications, one must not only judge the data relative to the specification, but also consider the limitations of the specification itself (Comité Européen de Normalisation (CEN) 1995).

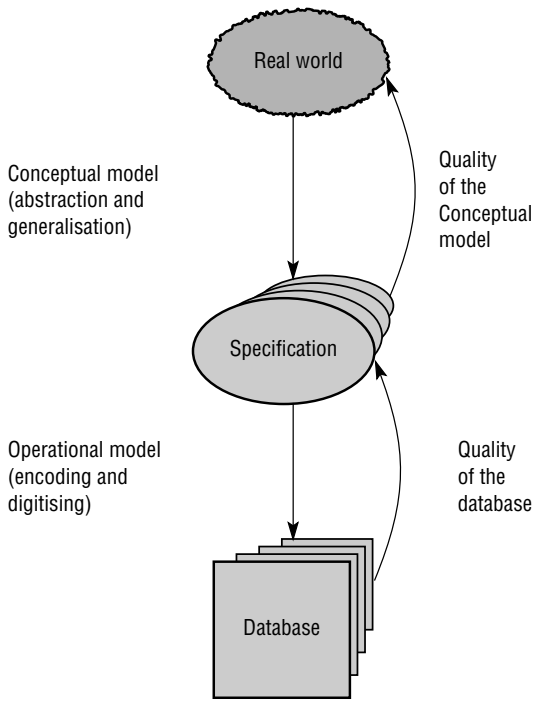

Fig 1. The mediating role of the database specification in assessing data quality.

2.1.1 Spatial accuracy

Spatial accuracy (or 'positional accuracy') refers to the accuracy of the spatial component of a database. Measurement of spatial accuracy depends on dimensionality. Metrics are well defined for point entities, but widely accepted metrics for lines and areas have yet to be developed. For points, error is usually defined as the discrepancy (normally Euclidean distance) between the encoded location and the location as defined in the specification. Error can be measured in any one of, or in combinations of, the three dimensions of space. The most common measures are horizontal error (distance measured in x and y simultaneously) and vertical error (distance measured in z) (Figure 2).

Various metrics have been developed to summarise spatial error for sets of points. One such metric is mean error, which tends to zero when 'bias' is absent. Bias refers to a systematic pattern of error (e.g. error arising from map misregistration). When bias is absent error is said to be random. Another common metric is root mean squared error (RMSE), which is computed as the square root of the mean of the squared errors (see Beard and Buttenfield, Chapter 15). RMSE is commonly used to document vertical accuracy for digital elevation models (DEMs). RMSE is a measure of the magnitude of error but it does not incorporate bias since the squaring eliminates the direction of the error.

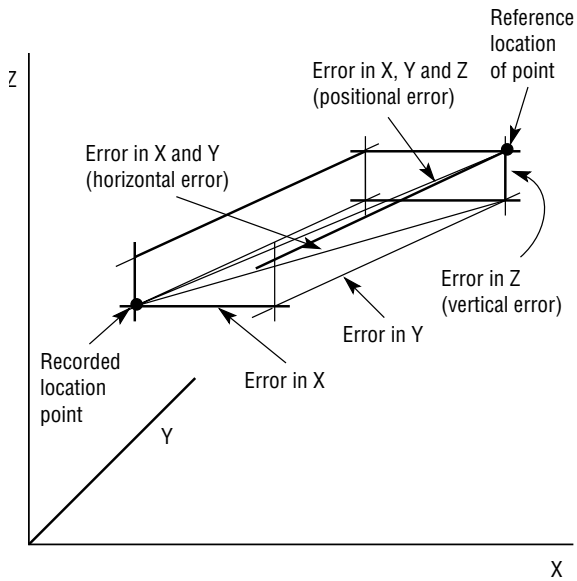

Fig 2. Measuring components of spatial error.

There is a close analogy between classical approaches to error and error in the location of a point. Horizontal error is a 2-dimensional extension of the classical error model in which error in position is defined in terms of a bell-shaped probability surface (Goodchild 1991a). Thus it is possible to perform statistical inference tests and derive confidence limits for point locations (American Society of Civil Engineers 1983; American Society for Photogrammetry 1985). For lines and areas the situation is more complex since there is no simple statistical measure of error that can be adopted from statistics. Errors in lines arise from the errors in the points that define those lines. However, as these points are not randomly selected the errors present at points cannot be regarded as somehow typical of errors present in the line (Goodchild 1991b).

Error is usually defined for lines using some variant of the epsilon band. The epsilon band is defined as a zone of uncertainty around an encoded line within which there is a certain probability of observing the 'actual' line. As yet there is no agreement as to the shape of the zone and the distribution of error within it. Early models assumed that the zone was a uniform 'sausage' within which the distribution of error was uniform (Blakemore 1983; Chrisman 1982). More recent studies show that both the distribution and the band itself might be non-uniform in shape (Caspary and Scheuring 1993; Honeycutt 1986) (Figure 3).

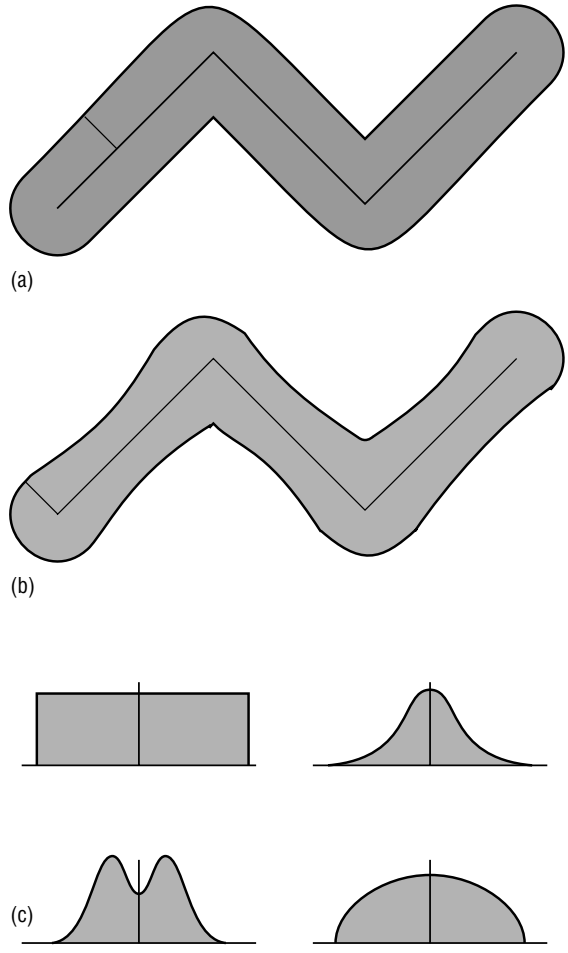

Fig 3. (a) Early models of the epsilon band show a uniform 'sausage' of width epsilon, ε, surrounding the encoded line; (b) more recent studies suggest that the band may be non-uniform in width; and (c) four of the many possible distributions of error around the encoded line.

2.1.2 Temporal accuracy

Temporal accuracy has not received much attention in the literature, just as time itself is not dealt with explicitly in conventional geospatial data models. Temporal accuracy is often equated with 'currentness' (Thapa and Bossler 1992). In fact the two concepts are quite distinct. Temporal accuracy refers to the agreement between encoded and 'actual' temporal coordinates. Currentness is an application-specific measure of temporal accuracy. A value is current if it is correct in spite of any possible time-related changes in value. Thus currentness refers to the degree to which a database is up to date (Redman 1992). To equate temporal accuracy with currentness is to state, in effect, that to be temporally accurate a database must be up to date. Clearly this is not the case since a database can achieve a high level of temporal accuracy without being current. Indeed historical studies depend on the availability of such data.

Assessment of temporal accuracy depends on the ability to measure time objectively using a standard temporal coordinate system. However, standards are not universally accepted (Parkes and Thrift 1980). Another impediment to the measurement of temporal accuracy is that time is often not dealt with explicitly in geospatial databases. Temporal information is often omitted, except in databases designed for explicitly historical purposes. This assumes that observations are somehow 'timeless' or temporally invariant. The implications of this omission are potentially quite significant, especially for features with a high frequency of change over time.

2.1.3 Thematic accuracy

Metrics of thematic accuracy (or 'attribute accuracy') vary with measurement scale. For quantitative attributes, metrics are similar to those used to measure spatial accuracy for point features (e.g. RMSE). Quantitative attributes can be conceived as statistical surfaces for which accuracy can be measured in much the same way as for elevation. For categorical data most of the research into data quality has come from the field of classification accuracy assessment in remote sensing. This work was carried out initially to devise methods to assess the accuracy of classification procedures. Accuracy assessment is based on the selection of a sample of point locations, and a comparison of the land cover classes assigned to these locations by the classification procedure with the classes observed at these locations on a reference source (usually 'ground truth'). A cross tabulation of the results (the 'classification error matrix') permits accuracy assessment (Aronoff 1985; Genderen and Lock 1977).

Various metrics summarising the information in the error matrix have been developed (proportion correctly classified, kappa, user's and producer's accuracies, etc.). These metrics are useful for assessing overall thematic accuracy. The classification error matrix contains additional information on the frequency of various types of misclassification, e.g. which pairs of classes tend most often to be confused. In addition, the matrix permits assessment of errors of omission (omission of a location from its 'actual' class) and errors of commission (assignment of a location to an incorrect class).

2.2 Precision or resolution

Precision refers to the amount of detail that can be discerned. It is also known as granularity or resolution. The latter term is commonly used in GIS and related fields, and is adopted here to avoid confusion with the statistical concept of precision as observational variance. All data are of limited resolution because no measurement system is infinitely precise. Resolution is also limited because geospatial databases are intentionally generalised. Generalisation includes elimination and merging of entities, reduction in detail, smoothing, thinning, and aggregation of classes. Generalisation is inevitable because, at best, geospatial databases can encompass only a fraction of the attributes and their relationships that exist in the real world (Weibel and Dutton, Chapter 10).

Resolution affects the degree to which a database is suitable for a specific application. The resolution of the database must match the level of detail required in the application. Low resolution does not have the same negative connotation as low accuracy. Low resolution may be desirable in certain situations, such as when one wishes to formulate general models or examine spatial patterns at a regional level.

Resolution is also important because it plays a role in interpreting accuracy. For example, two databases may have approximately equal spatial accuracy levels, but if their spatial resolutions are significantly different then the accuracy levels do not denote the same level of quality. One would generally expect accuracy and resolution to be inversely related, such that a higher level of accuracy will be achieved when the specification is less demanding.

2.2.1 Spatial resolution

The concept of spatial resolution is well developed in the field of remote sensing, where it is defined in terms of the ground dimensions of the picture elements, or pixels, making up a digital image (Figure 4). This defines the minimum size of objects on the ground that can be discerned. The concept is applicable without modification to raster databases. For vector data, the

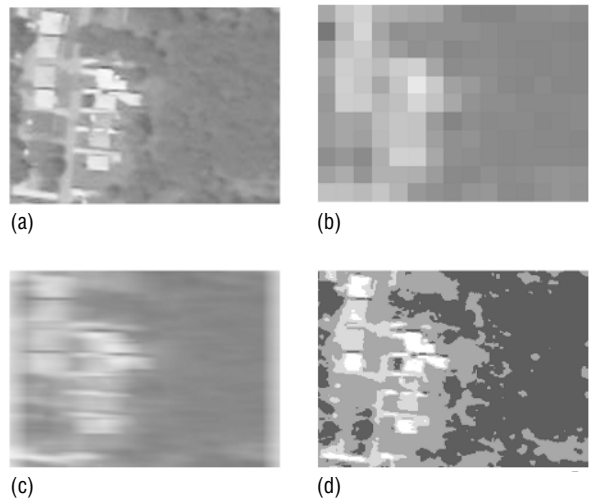

Fig 4. (a) A portion of a video image (Akron, Ohio) with spatial resolution of 1 metre, temporal resolution of 1/30 of a second, and thematic resolution of 8 bits (255 unique values); (b) the same image as in (a) but with spatial resolution degraded to 10 metres; (c) the same image as in (a) but with temporal resolution degraded, thus showing the effects of along-track movement of the sensing platform; and (d) the same image as in (a) but with thematic resolution degraded to four unique values.

smallest feature that can be discerned is usually defined in terms of rules for minimum mapping unit size which depend on map scale.

Spatial resolution is related to, but distinct from, the concept of the spatial sampling rate. Resolution refers to the fineness of detail that can be observed while the sampling rate defines the ability to resolve patterns over space. For remotely sensed images, resolution refers to the pixel size (ground area resolved) and sampling rate to the spaces between pixels. Thus in theory one could mix high spatial resolution with low sampling rate (small pixels with large gaps between them) or low spatial resolution with high sampling rate (large pixels that overlap). Normally, resolution and sampling rate are approximately equal.

2.2.2 Temporal resolution

Temporal resolution refers to the minimum duration of an event that is discernible. It is affected by the interaction between the duration of the recording interval and the rate of change in the event. Events with a lifetime less than the sampling interval are generally not resolvable. At best they leave a 'smudge' like pedestrians on nineteenth-century daguerreotypes. This has been referred to as the 'synopticity' problem (Stearns 1968). A shorter recording interval implies higher temporal resolution, just as faster film has given us the ability to photograph quickly moving objects (Figure 4).

For geospatial data, the situation is more complicated because interactions between spatial and thematic resolution must also be considered. In general one cannot resolve any event which, during the time interval required for data collection, changes location in space by an amount greater than the spatial resolution level. Likewise, one cannot resolve any event for which theme changes to a degree that would be discernible given the thematic resolution level (Veregin and Hargitai 1995).

There is a clear distinction between resolution and sampling rate in the temporal domain. Sampling rate refers to the frequency of repeat coverage while resolution refers to the time collection interval for each measurement. For example, motion pictures have a resolution of perhaps a thousandth of a second (one frame) but a sampling rate of 24 frames per second. Geosynchronous satellites are capable of much higher sampling rates than sun-synchronous satellites (repeat coverage several times per minute vs several times per month). Resolution, however, is a function of the time required to obtain spectral reflectance data for one pixel.

2.2.3 Thematic resolution

In the thematic domain, the meaning of resolution depends on measurement scale. For quantitative data, resolution is determined by the precision of the measurement device (Figure 4). For categorical data, resolution is defined in terms of the fineness of category definitions. Land cover classification systems used in remote sensing are useful models to illustrate resolution. These systems define the level of detail in taxonomic definitions in terms of the spatial resolving power of the remote sensing system. This illustrates the interdependence between space and theme when extracting spatial information (land cover class boundaries) from thematic information (spectral reflectance data).

2.3 Consistency

Consistency refers to the absence of apparent contradictions in a database. For geospatial data the term is used primarily to specify conformance with certain topological rules (Kainz 1995). These rules vary with dimensionality; for example, only one point may exist at a given location, lines must intersect at nodes, polygons are bounded by lines, etc. Elimination of topological inconsistencies is usually a prerequisite for GIS processing (Dowman, Chapter 31), such that most databases are topologically 'cleaned' before being released.

Topological consistency is one aspect of consistency in the spatial domain. Spatial inconsistencies can also be identified through redundancies in spatial attributes. For example, an entity might have the value 'Delaware' for the attribute 'state' but the value 'Lincoln' for the attribute 'county'. This is inconsistent since there is no Lincoln county in Delaware. In this case redundancy is partial: the state 'Delaware' eliminates the possibility of the county 'Lincoln', but the county 'Lincoln' does not necessarily imply the state 'Maine' since Maine is only one of 24 states containing a Lincoln County. On the other hand, redundancy may be complete (e.g. state is implied completely by the Federal Information Processing Standard (FIPS) state code), since there is a unique state code for each state. Non-redundancy implies that there is independence between two attributes such that meaningful consistency constraints do not exist (Redman 1992).

Little work has been done on consistency in the temporal domain, although a framework for

temporal topology has been developed (Langran 1992). For example, since at a given location only one event can occur at one time, an inconsistency exists if a different entity appears at the same location on two maps of the same date. Since events have a duration, this idea can be extended to identify events that exhibit temporal overlap.

In the thematic domain, the ability to identify inconsistencies requires a level of redundancy in thematic attributes – for example, the three sociodemographic variables 'population', 'mean household size', and 'total number of households'. Of course, the identification of an inconsistency does not necessarily imply that it can be corrected or that it is possible to identify which attribute is in error. Note also that the absence of inconsistencies does not imply that the data are accurate. Thus consistency is appropriately viewed as a measure of internal validity. Despite the potential to exploit redundancies in attributes, tests for thematic consistency are almost never carried out.

2.4 Completeness

Completeness refers to the relationship between the objects in the database and the 'abstract universe' of all such objects. Selection criteria, definitions, and other mapping rules used to create the database are important determinants of completeness. This definition requires a precise description of the abstract universe since the relationship between the database and the abstract universe cannot be ascertained if the objects in the universe cannot be described. The abstract universe can be defined in terms of a desired degree of abstraction and generalisation (i.e. a concrete description or specification for the database). This leads to the realisation that there are in fact two different types of completeness. 'Data completeness' is a measurable error of omission observed between the database and the specification. Data completeness is used to assess data quality, which is application-independent. Even highly generalised databases can be complete if they contain all of the objects described in the specification. 'Model completeness' refers to the agreement between the database specification and the abstract universe that is required for a particular database application (Brassel et al 1995). Model completeness is application-dependent and therefore an aspect of fitness-for-use. It is also a component of 'semantic accuracy' (Salgé 1995).

Additional distinctions are required. The definitions of completeness given above are examples of 'feature or entity completeness'. In addition we can identify 'attribute completeness' as the degree to which all relevant attributes of a feature have been encoded. A final type of completeness is 'value completeness' which refers to the degree to which values are present for all attributes (Brassel et al 1995).

Feature completeness can be defined over space, time, or theme. Consider a database depicting the locations of buildings in the state of Minnesota that were placed on the National Register of Historic Places as of 1995. This database would be incomplete if it included only buildings in Hennepin County (incompleteness in space, since Hennepin County covers only a portion of Minnesota), or only buildings placed on the Register by June 30 (incompleteness in time, since buildings may have been added after June 30), or only residential buildings (incompleteness in theme, due to the omission of non-residential buildings).

As this example shows, completeness is typically defined in terms of errors of omission. However, completeness may also include errors of commission (CEN 1995). Following on the previous example, errors of commission would occur if the database contained buildings in Wisconsin, buildings added to the Register in 1996, or historic districts as well as buildings.

3 DATA QUALITY STANDARDS

A concern for data quality issues is clearly expressed in the development of data transfer and metadata standards. Such standards have been developed at both national and international levels in support of mandates for data acquisition and dissemination. Data quality documentation plays a key role in many standards due to the realisation that an understanding of quality is essential to the effective use of geospatial data (see also Salgé, Chapter 50).

US readers will be most familiar with SDTS (the Spatial Data Transfer Standard) and the Content Standards for Digital Geospatial Metadata developed by the FGDC (Federal Geographic Data Committee). SDTS is a data transfer standard designed to facilitate dissemination and sharing of data. It provides standard definitions of data elements, a standardised format for data transfer,

and descriptive metadata about database contents. In 1992 SDTS was adopted by the National Institute of Standards and Technology as a Federal Information Processing Standard (FIPS-173) (Fegeas et al 1992).

The FGDC was established to promote coordinated development and dissemination of geospatial data. Its membership includes numerous US federal government departments and independent agencies. The FGDC has been involved in several activities related to geospatial data quality, including the development of the metadata content standards. Metadata describe the contents of a database. The FGDC standards provide a common set of terminology and a common structure for geospatial metadata (FGDC 1994). The FGDC standards were approved in 1994, and use of these standards is one of the minimum requirements for serving as a node in the National Geospatial Data Clearinghouse of the National Spatial Data Infrastructure (NSDI) (Morain and Budge 1996).

The FGDC standards follow SDTS in terms of recommendations for data quality information to be reported and tests to be performed. The five components of data quality in SDTS are listed in Table 1. Text-based documentation is the norm, although other formats are also permitted including numerical measures and even interactive graphics through online resources.

Many organisations have also created internal standards that contain data quality information. For example, the USGS DEM standard includes descriptors of horizontal and vertical accuracy. Standards have been adopted or are in development at national and international levels as well. Examples include the National Transfer Format (NTF) developed by the Association for Geographic Information (AGI) and adopted as the official British standard (BS7666) in 1992; the Digital Geographic Information Exchange Standard (DIGEST) developed by military service agencies from a number of NATO countries; the International Hydrographic Organisation (IHO) standard for nautical charts; and the draft standard of the CEN. Interested readers should consult Salgé (Chapter 50), Cassettari (1993), and Moellering (1991) for more details.

A major limitation of data quality standards is that they do not necessarily lend themselves to specific software implementations (see Guptill, Chapter 49). Standards provide models for data documentation but not a mechanism whereby users of disparate GIS packages can implement these models for database documentation. A related problem is that standards treat data quality as essentially static. While some accommodation is made for changes in quality as a result of data transformations, there is no mechanism to automatically update quality components as data are

Table 1 Data quality components in SDTS.

Component	*Description*
Lineage	Refers to source materials, methods of derivation and transformations applied to a database. ● Includes temporal information (date that the information refers to on the ground). ● Intended to be precise enough to identify the sources of individual objects (i.e. if a database was derived from different source, lineage information is to be assigned as an additional attribute of objects or as a spatial overlay).
Positional accuracy	Refers to the accuracy of the spatial component. ● Subdivided into horizontal and vertical accuracy elements. ● Assessment methods are based on comparison to source, comparison to a standard of higher accuracy, deductive estimates or internal evidence. ● Variations in accuracy can be reported as quality overlays or additional attributes.
Attribute accuracy	Refers to the accuracy of the thematic component. ● Specific tests vary as a function of measurement scale. ● Assessment methods are based on deductive estimates, sampling or map overlay.
Logical consistency	Refers to the fidelity of the relationships encoded in the database. ● Includes tests of valid values for attributes, and identification of topological inconsistencies based on graphical or specific topological tests.
Completeness	Refers to the relationship between database objects and the abstract universe of all such objects. ● Includes selection criteria, definitions and other mapping rules used to create the database.

passed through GIS processing steps. While source data may be adequately documented, derived data frequently are not. Finally, because standards such as SDTS provide such a rich collection of information about data quality, users may find it difficult to ascertain fitness-for-use. Likewise the unstructured nature of text-based descriptions means that data quality documentation is difficult to update automatically in a GIS environment.

Data quality standards also fall short of providing the kinds of assurances demanded by agencies that need to limit liability risks (Goodchild 1995). For example, SDTS follows the 'truth-in-labelling' paradigm in which the data quality report makes no *a priori* assumptions about quality requirements. While SDTS documentation might contain statements that the data meet some minimum accuracy standard, SDTS itself does not provide for the definition of data quality objectives necessary in the development of quality assurance/quality control (QA/QC) programs.

Efforts are underway to establish QA/QC programs within agencies that produce geospatial data. Such programs are based on the development of standard operating procedures that allow specific data quality objectives to be realised (Stone et al 1990). To some extent such QA/QC programs mirror the way in which traditional map accuracy standards such as National Map Accuracy Standards (NMAS) are implemented. The NMAS guarantee of a minimal level of positional accuracy is achieved through standard operating procedures that are known to yield the required accuracy levels, coupled with a limited amount of actual compliance testing. Such approaches focus on managing the production process rather than on statistical measurement of quality.

4 METADATA SYSTEMS

Like data quality standards, metadata systems are concerned with documentation of data quality components. The essential difference is that metadata systems emphasise the operational component rather than conceptual issues. Most commercial GIS packages perform a certain amount of metadata documentation. Some metadata is essential in order that data are processed correctly. For example, raster systems need to record the number of rows and columns of cells in each layer, while vector systems need to record the spatial coordinate system. Often these metadata are propagated forward as new layers are derived (see also Church, Chapter 20).

Only a few commercial GIS packages offer the capability to document data quality. An example is Idrisi version 4.1 which allows users to store information on the five components of data quality defined in SDTS. These data are stored along with other metadata in the documentation file that accompanies each raster layer and are propagated forward to derived layers. The software also performs rudimentary error propagation modelling by transforming metadata for certain data quality components.

For the majority of systems, however, tracking of data quality is the responsibility of the user. This has led to the independent development of software packages that document layers with metadata, update the lineage of layers automatically and perform propagation of data quality components (Veregin 1991). Some systems are quite advanced. Geolineus is an intelligent system that intercepts GIS commands and dynamically builds a graphical representation of the data processing flow and derived layers (Lanter 1991). This allows the user to visualise the flow of data processing steps and the linkages between source and derived data. At the same time Geolineus automatically propagates metadata, including data quality elements. This replaces the traditional approach in which updating of metadata is the sole responsibility of the user, such that it is often not performed at all (Goodchild 1995). Geolineus also stores information about data dependencies to facilitate metadata analysis. Examples of metadata analysis include assessment of processing complexity, analysis of the adequacy of data sources, propagation of error, and the identification of optimal strategies for enhancing derived data quality (Lanter and Surbey 1994; Lanter and Veregin 1992; Veregin and Lanter 1995).

5 CARTOGRAPHIC BIAS

The ability to produce a geospatial database presupposes a model that defines rules for simplifying real-world complexity. Despite their apparent sophistication, geospatial databases reflect many of the same biases as analogue cartographic data. This is true not only because geospatial databases are often produced by digitising paper

maps, but because the models embedded in GIS are essentially digital translations of analogue models (Burrough and Frank 1995). Thus in the vector data model geographical phenomena are differentiated according to their dimensionality. Only points, lines, and areas are permitted and these three classes are assumed to be mutually exclusive even though the dimensionality of many entities is known to be scale-dependent (Hutchinson and Gallant, Chapter 9). Dimensionality, originally applied in cartography as a heuristic for representation and symbolisation, has been reified in GIS as a fundamental property of geographical phenomena.

The finite limits to cartographic fidelity imply that maps must represent the real world selectively – that is, they can represent only a subset of all possible entities in the real world and must portray these entities in a generalised way as a function of map scale and purpose. The model is a highly abstract one that assumes that entities exist unambiguously in the real world. In some cases this is true, as with roads, counties, and other anthropogenic phenomena. However, in many cases the phenomena of interest have imprecise geographical expression. These phenomena belong to the perceived world rather than the real world and are inherently inexact and imprecise (Burrough 1986). Such phenomena are accommodated only clumsily in the cartographic model, through the introduction of concepts such as mapping unit 'purity' and 'minimum mapping unit size' which acknowledge that the real world is more complex than cartographic data would allow.

In theory geospatial data are not constrained to the same degree as paper maps. Many authors argue that geospatial technology is liberating as it facilitates new modes of representation and offers freedom from the constraints of scale imposed by pen-and-ink technology. An example is the raster model, which evolved in such fields as television and remote sensing, and represents a significant break from the traditional object-based cartographic model (Goodchild 1988a). It is perhaps not surprising then that many alternate models of geospatial data, such as the field-based model, probabilistic surfaces, and models based on fuzzy set theory, are raster based. These models are able to accommodate imprecision and uncertainty more easily than conventional cartographic models, and are thus more appropriate for many geographical phenomena.

Technology has also loosened the restriction that maps serve the dual purposes of storage and communication. For paper maps, content depends on the communication goal. The desire to communicate a particular message leads to selective enhancement and exaggeration of certain features and elimination or displacement of others. In geospatial databases the storage and communication roles can be more easily separated. This means that data can be collected in as raw a form as possible, and representations can be created to achieve any particular communication objective without altering the contents of the database. An additional advantage is that it is easier to quantify accuracy for raw data than for abstract cartographic representations (Goodchild 1988c).

These problems would not come to the fore if GIS were used only as an electronic map drawer. However, GIS has enormously extended the uses of geospatial data. Once data make their way into GIS they typically begin a process of metamorphosis in which they are transformed and merged with other data in support of queries, analyses, and decision-making models. Unfortunately there is no guarantee that the data are suitable for such applications. This problem is sometimes referred to as 'use error' (Beard 1989). Despite the advances we have made in understanding components of data quality, we have made almost no progress in the development of rules and heuristics to assess fitness-for-use and prevent use error (see Beard and Buttenfield, Chapter 15).

6 GIS, SOCIETY, AND DATA QUALITY

What is the essence of a geospatial database? Is it a faithful image of reality or a rhetorical device designed to convey a particular message? Is it an impartial representation of objective truth or a manifesto for a set of beliefs about the world? This is a central issue in the burgeoning 'GIS and society' debate in which research on data quality has many important implications (Pickles, Chapter 14).

According to some critics, technologies such as GIS have led to the ascendance of a new geospatial science focused on the goal of producing ultimately truthful and objective representations of reality. This goal is seen as a byproduct of the new technological means with its appeals to neo-positivism, reductionism, instrumentalist thinking, and naive empiricism in which 'reality' is uncontested and objectively measurable (e.g. Harley 1991; Wood 1992). According to this view, producers of geospatial

databases make no allowance for the possibility that these databases embed specific social and institutional values. As such, GIS promulgates the myth of an objective science which always produces the best delineations of reality (Harley 1989).

While there is some foundation to this critique, it would be unfair to suggest that producers of geospatial data are unaware of the limitations of these data. Like their manually-produced map counterparts, geospatial data are not intended to be miniature replicas of 'reality'. Rather they emphasise some aspects of the environment and suppress others in an effort to convey a particular message (Martin, Chapter 6; Raper, Chapter 5). What is contained in a database is a function not only of the nature of the external environment but also the values of the society and institution within which the database was constructed (Turnbull 1989). Values are embedded at the modelling stage, where they impact on database content, and at the representation stage where they affect database form.

Values are not always embedded deliberately. Broad social values are often taken for granted and may not be consciously recognised. Hence databases often unintentionally reflect and legitimate the social order (Harley 1989). Broad social values form the backdrop for more specific values that reflect institutional characteristics. Perhaps the most significant of these is institutional mandate, which defines institutional mission for data collection and dissemination. For specific databases, mandate is formalised as a set of design guidelines that outline the rules for data collection, encoding, and representation.

Unlike broad social values, values deriving from institutional mandate can be articulated, documented, and communicated to database consumers through the medium of metadata. This communication process is important since it affects the consumer's understanding of the limitations of a database and facilitates its appropriate use. Especially useful in this context is the concept of the 'specification' describing the intended contents of the database. The specification is the reference standard against which the database is compared in order to assess completeness and other data quality components. The specification concept explicitly recognises that each database has a particular set of objectives and that embedded in these objectives is the formal expression of the values associated with institutional factors.

What are the implications for the debate over values? First, geospatial databases are not intended to be accurate mirrors of reality. Rather, they are designed to conform to a database specification which could just as easily be a description of perceived reality. Second, geospatial data producers are generally aware of the significance of values. The database specification is in fact a formal statement of the values that are embedded in a given database. Third, values can be communicated to database consumers who can then use this information to assess the appropriateness of the database for a particular task. Knowledgeable map users have of course always been aware of data limitations.

These are important conclusions since the alternatives are not particularly attractive. For example, some critics have claimed that given the dependence on social values it is not possible to distinguish between competing representations of the same geographical space. Thus it has been argued that the distinction between propaganda and truth is artificial and must be dismantled, as must the arbitrary dualism between art and science (Harley 1989). According to this view, all representations are equally valid since they are all expressions of one's personal values, or the values of one's culture, or the values of one's institution, any one of which has no more claim to legitimacy than any other. This anarchistic epistemology implies that we have no agreed standard of reference and no basis for communicating biases and assumptions. On the other hand, if databases are to be more than just personal artistic diversions and are to convey information rather than simply express the values and viewpoints of their creator, then they must be able to convey their meaning to a broad spectrum of users.

References

American Society for Photogrammetry (Committee for Specifications and Standards, Professional Practice Division) 1985 Accuracy specification for large-scale line maps. *Photogrammetric Engineering and Remote Sensing* 51: 195–9

American Society of Civil Engineers (Committee on Cartographic Surveying, Surveying, and Mapping Division) 1983 *Map uses, scales, and accuracies for engineering and associated purposes*. New York, American Society of Civil Engineers

Aronoff S 1985 The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing* 51: 99–111

Beard M K 1989b Use error: the neglected error component. *Proceedings, AutoCarto 9*: 808–17

Berry B 1964 Approaches to regional analysis: a synthesis. *Annals of the Association of American Geographers* 54: 2–11

Blakemore M 1983 Generalisation and error in spatial databases. *Cartographica* 21: 131–9

Brassel K, Bucher F, Stephan E-M, Vckovski A 1995 Completeness. In Guptill S C, Morrison J L (eds) *Elements of spatial data quality*. Oxford, Elsevier Science: 81–108

Burrough P A 1986 *Principles of geographical information systems for land resources assessment*. Oxford, Clarendon Press

Burrough P A, Frank A U 1995 Concepts and paradigms in spatial information: Are current geographical information systems truly generic? *International Journal of Geographical Information Systems* 9: 101–16

Caspary W, Scheuring R 1993 Positional accuracy in spatial databases. *Computers, Environment and Urban Systems* 17: 103–10

Cassettari S 1993 *Introduction to integrated geoinformation management*. London, Chapman and Hall

Chrisman N R 1982 A theory of cartographic error and its measurement in digital databases. *Proceedings, AutoCarto 5*: 159–68

Chrisman N R 1991b The error component in spatial data. In Maguire D J, Goodchild M F, Rhind D W (eds) *Geographical information systems: principles and applications*. Harlow, Longman/New York, John Wiley & Sons Inc. Vol. 1: 165–74

CEN (Comité Européen de Normalisation) 1995 *Geographic information – data description – quality* (draft). Brussels, CEN Central Secretariat

FGDC Federal Geographic Data Committee 1994 *Content standards for digital geospatial metadata (June 8)*. Washington DC, Federal Geographic Data Committee

Fegeas R G, Cascio J L, Lazar R A 1992 An overview of FIPS 173, the Spatial Data Transfer Standard. *Cartography and Geographic Information Systems* 19: 278–93

Genderen J L van, Lock B F 1977 Testing land-use map accuracy. *Photogrammetric Engineering and Remote Sensing* 43: 1135–7

Goodchild M F 1988a Stepping over the line: technological constraints and the new cartography. *The American Cartographer* 15: 311–19

Goodchild M F 1988b The issue of accuracy in global databases. In Mounsey H (ed.) *Building databases for global science*. London, Taylor and Francis: 31–48

Goodchild M F 1991a Issues of quality and uncertainty. In Müller J-C (ed.) *Advances in cartography*. Oxford, Elsevier Science: 111–39

Goodchild M F 1991c Keynote address. *Proceedings, Symposium on Spatial Database Accuracy*: 1–16

Goodchild M F 1995b Sharing imperfect data. In Onsrud H J, Rushton G (eds) *Sharing geographic information*. New Brunswick, Center for Urban Policy Research: 413–25

Guptill S C 1993 Describing spatial data quality. *Proceedings, Sixteenth International Cartographic Conference*: 552–60

Harley J B 1989 Deconstructing the map. *Cartographica* 26: 1–20

Harley J B 1991 Can there be a cartographic ethics? *Cartographic Perspectives* 10: 9–16

Honeycutt D M 1986 'Epsilon, generalisation, and probability in spatial databases'. Unpublished manuscript

Kainz W 1995 Logical consistency. In Guptill S C, Morrison J L (eds) *Elements of spatial data quality*. Oxford, Elsevier Science: 109–37

Langran G 1992 *Time in geographic information systems.* London, Taylor and Francis

Lanter D 1991 Design of a lineage-based meta-database for GIS. *Cartography and Geographic Information Systems* 18: 255–61

Lanter D, Surbey C 1994 Metadata analysis of GIS data processing: a case study. In Waugh T C, Healey R G (eds) *Advances in GIS research*. London, Taylor and Francis: 314–24

Lanter D, Veregin H 1992 A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering and Remote Sensing* 58: 526–33

Moellering H (ed.) 1991 *Spatial database transfer standards: current international status*. Oxford, Elsevier Science

Morain S, Budge A 1996 The National Spatial Data Infrastructure – why should you care? *GIS World* 9(8): 32–4

Parkes D N, Thrift N J 1980 *Times, spaces, and places: a chrono-geographic perspective*. New York, John Wiley & Sons Inc.

Redman T C 1992 *Data quality*. New York, Bantam

Salgé F 1995 Semantic accuracy. In Guptill S C, Morrison J L (eds) *Elements of spatial data quality*. Oxford, Elsevier Science: 139–51

Sinton D F 1978 The inherent structure of information as a constraint in analysis. In Dutton G (ed.) *Harvard papers on geographic information systems*. Reading (USA), Addison-Wesley

Stearns F 1968 A method for estimating the quantitative reliability of isoline maps. *Annals of the Association of American Geographers* 58: 590–600

Stone H F, Boyle S L, Hewitt M J III 1990 Development of an EPA quality assurance program for geographic information systems and spatial analysis. *GIS/LIS 90*: 814–19

Thapa K, Bossler J 1992 Accuracy of spatial data used in geographic information systems. *Photogrammetric Engineering and Remote Sensing* 58: 835–41

Turnbull D 1989 *Maps are territories*. Chicago, University of Chicago

Veregin H 1991 *GIS data quality evaluation for coverage documentation systems*. Las Vegas, Environmental Monitoring Systems Laboratory, US Environmental Protection Agency

Veregin H, Hargitai P 1995 An evaluation matrix for geographical data quality. In Guptill S C, Morrison J L (eds) *Elements of spatial data quality*. Oxford, Elsevier Science: 167–88

Veregin H, Lanter D 1995 Data-quality enhancement techniques in layer-based geographic information systems. *Computers, Environment and Urban Systems* 19: 23–36

Wood D 1992 *The power of maps*. New York, Guilford Press/ London, Routledge