CHAPTER NINE

An evaluation matrix for geographical data quality

Howard Veregin and Péter Hargitai

Defining Data Quality

The purpose of this chapter is to outline a logical schema for data quality assessment in the context of geographical databases. The schema is based on two main concepts.

- Geographical observations are defined in terms of space, time and theme. Each of these dimensions can be treated separately (but not always independently) in data quality assessment.
- Data quality is an amalgam of overlapping components, including accuracy, resolution, completeness and consistency. The quality of geographical databases cannot be adequately described with a single component.

The combination of geographical data dimension and data quality component gives rise to an "evaluation matrix" for data quality assessment (Fig. 9.1). Each cell in the matrix refers to a particular data dimension and a particular data quality component. This provides a logical structure for measuring, documenting and communicating data quality information. The goal of this chapter is to describe the evaluation matrix and explain its uses in the context of data quality assessment. Tools appropriate for each cell in the matrix are presented and described. For the most part these tools are discussed in greater detail in other chapters in this book. Reference to the appropriate chapter is provided where appropriate. This chapter does not present radically new tools for quality assessment so much as it provides an alternate view of the way in which data quality information could be organized and structured.

The three dimensions of geographical data identified above are the same dimensions that form the basis of the "geographic data matrix" model formalized by Berry (1964) and others. According to this model, geographical observations can be located in a coordinate system based on

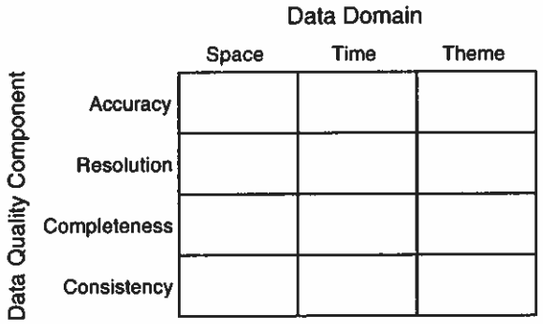

Figure 9.1 An evaluation matrix for data quality assessment.

their spatial, temporal and thematic coordinate values. In data quality assessment, different domains may stand out as being of particular importance in different contexts. The spatial domain is typically more dominant in cadastral mapping, for example, while in natural resource management the thematic domain is often more critical. The geographical literature tends to ascribe special significance to the spatial and thematic domains. The temporal component of data quality is important in a variety of contexts, but has unfortunately received little attention in the literature. Moreover, it is not always possible to treat space, time and theme as independent dimensions of data quality.

It has been argued that geography is distinct from geometry because in geography, space is indivisibly coupled with time (Parkes and Thrift, 1980; Langran, 1992). One could go a step further: geography's distinctiveness lies in its concern with thematic features and attributes and their relationships in space and time. Geographical data are not simply spatial, but neither are they simply spatio-temporal. Geographical data are derived from observations of *things* (Wood, 1992). Theme defines the set of features or entities that are of interest, and space and time serve as a framework for thematic representation. Without theme there is only geometry. Without space and time there is nothing geographical in the data.

The structure of geographical databases reflects this thematic bias. Geographical features represent real-world entities encoded as spatial objects (Moellering, 1992). These objects acquire meaning only through their association with thematic information. In this way the database achieves a representation of the multiplicity of relations among real-world entities. The thematic attributes encoded in a database provide the conceptual mapping between the real world and its computer representation.

At best the set of thematic attributes encoded in a database can encompass only a fraction of attributes and their relationships that exist in the real world. Any geographical database is therefore an abstraction of the real world, incomplete and generalized.

The degree to which the objects encoded in a database are assumed to unambiguously represent real-world entities is the usual starting point for assessing the accuracy of geographical databases. Accuracy is conventionally interpreted in terms of discrepancies between database objects and the real-world entities that these objects are assumed to represent. The conceptual gap between real-world complexity and the conventional vector database representation of this complexity in terms of abstract geometric objects (points, lines, polygons) underscores the inability of accuracy alone to provide a complete description of the reliability of a database. As shown by the chapters in this book, accuracy is only one component of data quality for geographical databases. Data quality also includes aspects of consistency, completeness, semantic integrity, currency, etc. The schema adopted in this chapter defines three components of data quality in addition to accuracy: resolution, completeness and consistency. Resolution refers to the amount of detail observable in the database. Completeness refers to the degree to which the database achieves success as a model of the real world. Consistency refers to the degree to which the database is free from internal inconsistencies.

Accuracy, unlike these other components, is a relative measure of quality. That is, the accuracy of a database is assessed with reference to another database (or "reference source") that is assumed to be more accurate. Given different reference sources, multiple accuracy indices can be computed. This might be the case, for example, when a database is used in different applications with different accuracy requirements. For in-house demo projects it might be sufficient to assess positional accuracy relative to the paper source map (i.e., assess how accurately the database depicts the source document), while for more demanding applications it might be necessary to assess positional accuracy using ground survey data. Indices of accuracy cannot reliably be interpreted without a complete description of the reference source and an assessment of its limitations. This requirement underscores the need for detailed lineage information when performing accuracy assessment (see Chapter 2). The implication for the evaluation matrix is that accuracy is more appropriately conceived as a two-dimensional array, one dimension of which is defined by space, time and theme, and the other by the reference source used for accuracy assessment (Fig. 9.2).

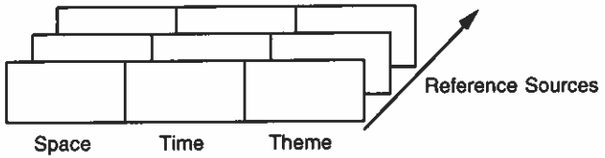

Figure 9.2 The dimensions of accuracy.

Existing Data Quality Standards

Numerous standards for geographical data have been developed in various countries in support of mandates for data acquisition, storage and dissemination. In the United States, an important standard is SDTS (Spatial Data Transfer Standard), recently adopted by the National Institute of Standards and Technology as a Federal Information Processing Standard (FIPS) to serve the U.S. federal geographical data processing community. SDTS has inspired other efforts aimed at developing transfer standards, metadata standards and associated data quality components. The structure of the SDTS data quality model is reflected in other attempts to design data quality reporting strategies, including the strategy adopted in the organization of this book. Thus the data quality component of SDTS is a useful starting point for assessing the current level of understanding of spatial data quality for geographical databases.

Models of Data Quality

The SDTS data quality model is based on the principle that users should be able to characterize fitness-for-use for a given application based on data quality documentation. According to this "truth-in-labeling" approach, the data quality report is not intended to provide guidelines defining fitness-for-use. The report is not a compliance standard like the National Map Accuracy Standard (NMAS) that defines the minimum acceptable level of spatial accuracy for U.S. Geological Survey topographic maps. According to SDTS, the producer is responsible for documenting data quality, and the user is responsible for determining whether or not the database is of sufficient quality for a particular application (Fegeas *et al.*, 1992).

Five components of data quality are identified in STDS: positional accuracy (the accuracy of the spatial component of the database); attribute accuracy (the accuracy of the thematic component of the database); logical consistency (the fidelity of the relationships encoded in the database); completeness (the external validity of the database); and lineage (the processing history of the database, including sources, data capture methods and data transformations techniques).

While specific tests and metrics are not provided, SDTS allows for various assessment methods: comparison of the database to the source documents from which it was derived; comparison of the database to a source of higher accuracy; internal evidence (e.g., identification of internal inconsistencies in the database); and deductive estimates based on knowledge of the manner in which error accumulates as a result of the techniques used for data encoding, editing and processing. Each of these assessment methods has a different goal and connotes a different meaning of the term accuracy. For example, comparison to the source document is used to identify errors introduced during the digital encoding process, while comparison to a source of higher accuracy is used to evaluate the degree to which the database conforms to a more accurate standard that is presumably closer to reality. A database may faithfully represent the source from which it was derived while simultaneously exhibiting numerous discrepancies relative to ground survey data.

Similar models of data quality have been adopted in other data transfer and metadata standards. For example, the Draft Content Standards for Spatial Metadata developed by the U.S. Federal Geographic Data Committee (FGDC, 1992) contains a data quality section designed to assist potential users in determining whether a database is suitable for a particular application. The metadata elements for data quality include positional accuracy, attribute accuracy, model integrity (i.e., logical consistency) and completeness. The positional and attribute accuracy components include a numerical measure, a description of the assessment method (i.e., deductive estimate, internal evidence, comparison to source, or comparison to source of higher accuracy), and a textual explanation of how the assessment method was applied.

Analogous approaches have been adopted in other countries in the development of geographic data standards. An example is the Hungarian standard, developed by a working group of the Cartographic Department of Eötvös Loründ University, Budapest (Divényi, 1991). This standard is designed to facilitate the evaluation and transfer of data quality information for geographical databases. The data quality component includes positional accuracy, attribute accuracy, state of maintenance, course of processing completeness and methods of data acquisition. Data quality assessment can be done in several ways, including comparison to the source document or a document of higher accuracy, or deductive methods that assess the effects of different processes on data quality. Deductive methods are seen as less reliable, but may be the only option in some situa-

tions. Where possible, quantitative expression of data quality components is preferred, due to the enhancement in comparability that is thus afforded.

Although these standards have been developed for different purposes and constituencies, they show a degree of consistency in the structure of the data quality description. For example, there seems to be general agreement about the desirability of separating the spatial (positional) and thematic (attribute) dimensions, and the need for flexibility in accuracy assessment methods. Moreover, both internal and external aspects of data quality are usually recognized (e.g., consistency and completeness). Finally, aspects of lineage (including data acquisition and processing history) are consistently seen as exerting an impact on data quality.

Many other data standards have been developed or are in the process of being developed. Interested readers are referred to Moellering (1991).

Limitations of Data Quality Models

In the absence of long-term empirical assessment, it is possible only to speculate on the performance of such data quality standards. On the surface, they seem like workable models for organizing knowledge about data quality. They appear to offer the ability to assess, document and communicate data quality information and allow for comparison of data quality components for different databases. However, it is argued here that these standards suffer from a number of limitations that seriously affect their ability to act as a general model of data quality for geographical data. These limitations are discussed below.

- Space, time and theme are not dealt with consistently across different components of data quality. Accuracy is typically divided explicitly into a spatial component (positional accuracy) and a thematic component (attribute accuracy). However, other dimensions of data quality (i.e., consistency and completeness) are seen as general characteristics undifferentiated over space, time and theme.
- Time is not defined explicitly for any data quality component. This absence bespeaks a view of time as either a self-evident truth requiring no elaboration or an aspect of the world that is of no consequence for geographical observation. In fact, time is of fundamental importance in geographical data and the quality of the temporal dimension can have significant implications for spatial and thematic data quality. The meaning of location in space is always bound up with location in time, since all entities are in fact events that move, change or disappear over time (Parkes and Thrift, 1980).
- Resolution is not defined explicitly as a data quality component. Resolution refers to the degree of detail observable in space, time or theme. It exerts considerable impact on accuracy, particularly when accuracy is assessed by comparing a database to a higher-resolution reference source. Many standards consider resolution as an aspect of completeness (e.g., mapping rules, minimum size of features mapped).
- Lineage is considered to be a dimension of data quality. Lineage refers to the data acquisition and processing history of a database. Lineage information includes source materials from which the database was derived, dates of source materials, dates of any ancillary data used for updates, methods of derivation, processing steps performed, locations of control points used for coordinate transformations, methods of coordinate transformation and methods used to avoid roundoff error. Although these factors certainly *affect* data quality, note that they do not define a particular *component* of data quality. Lineage is not a dimension of data quality so much as it is a prerequisite for assessing data quality based on deductive estimation methods. The amount and quality of lineage information is not an index of data quality, since a database may be of high quality even if no lineage data exists. Rather, the amount and quality of lineage information is an index of *metadata* quality (i.e., the quality and completeness of the data that describe the database).
- Data quality is defined as a static attribute of a database. However, it is well known that data quality characteristics can change as data are transformed in GIS. The ability to model changes in data quality as data transformation functions are applied is referred to as error propagation modeling. Error propagation modeling is important because data quality documentation can easily become obsolete as data are transformed. Data transformation functions change error characteristics in different ways. Some functions primarily affect spatial error, while others affect thematic error. Some functions accentuate error, while others smooth or eliminate error (Veregin, 1989). The tools used to measure and document data quality affect the nature and effectiveness of error propagation models. In part, this is because little is currently known about error propagation mechanisms for many classes of data transformation functions. Although a number of error propagation systems have been developed (e.g., Heuvelink *et al.*, 1989; Lanter and Veregin, 1992; Carver, 1991), these systems are limited to selected subsets of data transformation functions and are dependent on assumptions and conditions that are often unattainable outside of the laboratory environment.

In the following section, a general geographical model of data quality is proposed that addresses many of the problems discussed above.

A Geographical Model of Data Quality

Evaluation of data quality for geographical databases should conform to the manner in which real-world data are encoded and represented in these databases. At the most abstract level, geographical data are organized in terms of three dimensions.

- The *spatial* dimension defines the horizontal and vertical coordinates (x, y and z) of a location, P. Note that while we regard z as a spatial coordinate, it is often viewed as an attribute that can change with time.
- The *temporal* dimension defines the coordinates of P in time (t).
- The *thematic* dimension defines a value for P for some theme or attribute.

At an operational level, this model may be translated into a model specific to a particular application tool (i.e., a GIS system) with its own demands in terms of data encoding and storage. These issues do not have much impact on the development of the abstract data quality model.

Geographers and others have long debated the nature of the relationships between space, time and theme. In our view, the spatio-temporal structure (x, y, z and t) provides a framework for the collection, encoding and digital representation of themes (attributes) in geographical databases. Space and time alone proffer no information about the real-world relationships encoded in geographical databases; rather they supply the framework upon which such information is imposed (see Wood, 1992). The notion of location in time is critical in geographical observation and cannot be divorced from location in space. Geographical entities are in fact events that move, change or disappear over time (Parkes and Thrift, 1980). Due to this dynamism, accurate mapping of events in space depends on accurate mapping of events in time, and vice versa. Geographical events portrayed on maps and in digital databases are measured over finite spatial and temporal domains. The spatial domain is almost always well-articulated, either in the form of coordinate values (e.g., latitude and longitude) or in the map title or database description. The temporal domain is not always so clearly explicated; it is often necessary to assume that the map or database represents conditions at the time that it was published, unless there is a clear indication to the contrary (as in the case of historical data).

Data quality can be assessed in terms of three dimensions similar to the definition of geographical data. This serves as a useful starting point for classifying data quality components. In the schema described above, the "true" quality of P is a point in the three-dimensional space defined by space, time and theme. This point is approximated by a volume of

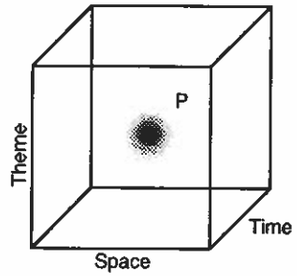

Figure 9.3 Uncertainty in space, time and theme.

uncertainty that contains the true quality location (Fig. 9.3). The size and shape of this volume can vary as a function of the amount of uncertainty associated with the spatial, temporal and thematic coordinates of P. Such uncertainty has many sources, including measurement error, limitations imposed by resolution and precision, and inherent inexactness in the spatial and temporal locations of real-world entities and their thematic composition. The volume in effect defines a three-dimensional probability distribution of P. There is a finite probability of defining the quality of P as being somewhere within the volume. The probability at any quality location within the volume depends on the characteristics of the probability distribution (Giordano *et al.*, 1994).

A major drawback is that this model suggests that space, time and theme are independent dimensions when in fact they are interdependent. Space and time are interdependent due to the indivisibility of space and time in defining the locations of geographical events. The thematic dimension is also dependent on the spatio-temporal structure. A change in the spatial or temporal coordinates generally leads to a change in theme; indeed this is the basis behind the ability to identify patterns over space and time.

Components of Data Quality

The geographic model of data quality discussed above suggests that all components of data quality can be differentiated by space, time and theme. The result is the evaluation matrix model shown in Fig. 9.1. This section of the chapter discusses the individual components, or cells, in the evaluation matrix, and presents some possible tests and metrics that might be applied in each case.

Accuracy

Accuracy is probably the best understood component of data quality. Whether it refers to space, time or theme, accuracy is conventionally defined in terms of discrepancies between a database and a reference source, where the reference source is usually a database of known higher accuracy. (See the definitions given for positional and attribute accuracy in Chapters 3 and 4.) This definition has much in common with the statistical treatment error, in which error is seen to result from imperfect data acquisition methods. According to this view, error can be reduced through the use of more refined data acquisition techniques and can be characterized statistically based on methods of repeated sampling. Error characteristics include bias (mean magnitude and direction of discrepancies) and precision (dispersion or variation in discrepancies). Both bias and precision are measures of accuracy, since each is based on observed discrepancies between the database and the reference source.

To the extent that geographical databases are often derived by encoding existing documents (e.g., paper maps), errors in databases may be seen to arise from two main sources. The first of these is associated with the data encoding process, resulting in discrepancies between the digital database and the source documents from which it was derived (Chrisman, 1982). The second source of error is associated with the inadequacies of the source document itself. Thus no matter how faithfully the database mirrors the source document, the database will still contain some degree of error. Each source of error is the result of a transformation of some formal representation of the world (Bedard, 1986). One such transformation occurs when the cartographer's mental representation is used to create a map. Another transformation occurs when this map is encoded to create a digital database. Each of these transformations is significant because it results in discrepancies between digital data and the real-world phenomena that these data are presumed to represent.

Accuracy can be defined in terms of the discrepancies between the database and a variety of reference sources. The ideal source in most situations is objective reality, since discrepancies in this case reflect an imperfect knowledge of the real world. Measurement of discrepancies relative to a source of higher accuracy results in a less exacting standard. Measurement of discrepancies relative to the source documents used for database encoding reveals whether or not the database truthfully depicts the source documents from which it was derived. Different reference standards thus give rise to different conceptions and meanings for the term accuracy. Each reference standard may be appropriate in a different context. A database might be deemed accurate for certain purposes (e.g., demonstration projects) as long as it faithfully depicts the features and relationships among features exhibited in source documents. In other situations (e.g., research and development projects), it might be necessary to assess accuracy in terms of discrepancies between the database and a source of higher accuracy.

Accuracy in the Spatial Domain. As noted in Chapter 3, spatial accuracy assessment tests depend on the positional component and feature class under consideration. Tests for spatial accuracy may refer to horizontal accuracy (i.e., planimetric accuracy, or the accuracy of the x and y coordinates) or vertical accuracy (i.e., the accuracy of elevations, or z-values). Some tests treat the x and y components of horizontal accuracy separately, while other tests combine them. Some tests account for the fact that horizontal and vertical accuracy are interrelated, while others treat these components independently. Tests for positional accuracy can be categorized depending on the feature class to which they refer. Following convention, positional accuracy may be defined for points, lines and areas.

For point data, accuracy assessment is quite straightforward. As noted in Chapter 3, it is possible to define error in the location of a point in the x, y and z dimensions. The significance of this error can then be inferred using standard statistical tests. Such tests are detailed in standards such as EMAS (Engineering Map Accuracy Standard), which originated as an accuracy specification for large-scale engineering and topographic maps (Rosenfield, 1971; Kellie and Bryan, 1981; Merchant, 1982; American Society of Civil Engineers, 1983; American Society of Photogrammetry, 1985; Merchant, 1987). EMAS uses a statistical accuracy test based on errors in the x, y and z coordinates for a sample of points. The test is based on a comparison of the encoded coordinates for these points relative to the coordinates for the same points as derived from a reference source. This source typically is of known higher accuracy, but may also be the source document. The test allows for compliance testing based on hypothesis tests using acceptable error thresholds known as limiting errors. This makes it possible to define subcategories of accuracy for any database category. The test provides information about the actual amount of error present in the database and at individual point locations. Test results can therefore be used to estimate how much effort might be required to obtain a more accurate map and how much this effort might cost.

For linear and areal data, accuracy assessment is more difficult. The main problem is that there is not an unambiguous correspondence between vertex locations encoded in the database and the locations of these vertices on the reference source. Models of error for point locations

are of little use in this context. Of the various strategies that have been suggested to overcome this problem, only the epsilon band seems to have attained widespread acceptance (Chrisman, 1982; Blakemore, 1983; Chrisman, 1983; Lee, 1985; Honeycutt, 1986). The epsilon band model defines error in terms of a zone of uncertainty surrounding the encoded line location. In the most basic form of the model, error is assumed to vary uniformly around the line. Variants of the epsilon band exist that account explicitly for non-uniformity (see Veregin, 1994).

The epsilon band procedure for data quality assessment is based on the superimposition of the encoded and reference databases with the aim of identifying discrepancies in the locations of linear features. This approach was originally put forward by MacDougall (1975) with reference to the propagation of horizontal accuracy through map overlay. MacDougall argued that the total horizontal error in a map could be computed from two parameters: a measure of mean horizontal error and a measure of the total length of the lines on the map. In the worst-case, the total error would be a product of these two parameters. The result is an estimate of the total area of a map that is inaccurate due to horizontal error.

Application of the epsilon band is not appropriate for all types of linear and areal data. It is best applied to data for which linear features in the database have an unambiguous real-world meaning. Examples are roads, hydrologic features, political boundaries and property boundaries. Many natural or interpreted features, including boundaries between soil types, vegetation communities and land cover types, are abstract features without precise real-world locations. As noted in Chapter 4, for these features, it may be more appropriate to assess thematic accuracy. Little work has been done on the interactions between spatial and thematic error.

Accuracy in the Temporal Domain. In the geographical literature, temporal accuracy has received much less attention than spatial accuracy. Temporal accuracy may be defined in terms of the discrepancies between an encoded temporal coordinate and the temporal coordinate as obtained from a source of higher accuracy. Essentially this is a measure of positional accuracy in which position is measured in time rather than space. Such a definition of temporal accuracy runs counter to the notion of currentness, which is often suggested as an appropriate measure of temporal accuracy. The problem with currentness as a measure of temporal accuracy is that it is application-specific. A database can achieve high temporal accuracy without being current. Indeed historical and longitudinal studies (e.g., change detection) depend on the availability of such data.

Measurement of temporal accuracy in terms of temporal location depends on the ability to measure time objectively (i.e., with reference to some agreed-upon standard origin and unit of measurement). In the same way, measurement of spatial accuracy depends on objective measurement of location in space. Like objective space, objective time is easier to define in theory than to measure in practice. Objective time can be based on a clock or calendar synchronized to an accepted temporal coordinate system. Historically, this coordinate system has been based on the earth's rotation and revolution relative to the sun and other celestial objects. More recently a standard has been adopted based on subatomic radiation transitions defining the duration of a second of time (Parkes and Thrift, 1980).

An objective, application-free temporal accuracy standard requires, first, that geographical databases be stamped with temporal coordinate data. Such information is often omitted except in explicitly historical or longitudinal studies. Omission of temporal information can lead users to conclude that the data contained in the database represent current conditions, even though it may have taken a long period of time to acquire and publish the database (see Chapter 8). The implications for temporal accuracy are potentially significant, especially for events that move, change or disappear with fairly high temporal frequency. Temporal accuracy is also affected by temporal resolution, as discussed in a subsequent section.

Accuracy in the Thematic Domain. As noted in Chapter 4, thematic accuracy assessment tests depend on the measurement scale of the attribute under consideration. Nominal scale refers to attributes for which specific attribute values have no inherent mathematical meaning (e.g., land cover types). Ordinal scale refers to rankings, where numerical attribute values indicate relative, but not absolute, quantities (e.g., suitability scores). Interval and ratio scales refer to quantities for which a given interval or ratio has the same meaning for all values of the attribute (e.g., temperature and precipitation).

For nominal data, the most widely-used accuracy indices are derived from the classification error matrix. The classification error matrix is a cross-tabulation of encoded and reference values for a thematic attribute for a sample of locations. The information contained in a classification error matrix can be summarized using an variety of accuracy indices, including *PCC* (proportion correctly classified) and κ (kappa). These indices are described in detail in Chapter 4. The classification error matrix itself contains information that is lost when such indices are computed; this information might be retained to assist in accuracy assessment. It is possible, for example, to identify systematic errors (classes that are most often confused), which may assist in developing strategies for editing data.

Additional information on the classification error matrix is available in van Genderen and Lock (1977), Congalton *et al.* (1983), Aronoff (1985), Rosenfield and Fitzpatrick-Lins (1986), and Hudson and Ramm (1987).

For ordinal data, a modification of the classification error matrix is needed to account for the relative significance of misassignment of observations to ordinal classes. In this context, a weighted kappa statistic might be employed (Greenland *et al.*, 1985). Little work has been done on this topic. For interval and ratio data, accuracy assessment can be achieved using techniques similar to those for assessing vertical positional accuracy, as shown in Chapter 4.

Resolution

Resolution refers to the amount of detail that can be discerned in space, time or theme. A finite level of resolution implies some generalization will be present in the database. Generalization refers to the elimination of small features, smoothing and thinning of features, merging or aggregation of features in close proximity to each other, elimination and collapsing of categories to create more general categories, etc. Generalization is inevitable in geographical databases; at best such databases can only encompass a fraction of the attributes and their relationships that exist in the real world. Database consumers need to be made aware of the degree of generalization in order to determine whether a higher level of resolution is required.

The concept of resolution is well-developed in the field of remote sensing, and many of the principles applied there are also applicable in the context of data quality. In remote sensing, spatial resolution is defined in terms of the size of the objects that can be discerned on a digital image. This is affected by the ground dimensions of the picture elements, or pixels, making up the image. The concept is applicable without modification to raster databases. For vector data, the smallest feature that can be discerned is usually defined in terms of mapping rules for minimum mapping unit size, which is often affected by map scale. (While scale is not fixed for digital data as it is in the analog realm, scale can still affect resolution, particularly when digital data are derived from analog sources such as paper maps.) Similar rules often apply in terms of the minimum length of features and the minimum separation required to display features as separate and distinct.

Temporal resolution refers to the minimum duration of an event that is discernible in the database. It is affected by the length of the sampling interval required for data acquisition and the rate of change of the features being measured. The effects of a long data acquisition interval on the ability to resolve events is sometimes referred to as the synopticity problem (Stearns, 1968). The problem can be illustrated with reference to 19th-century daguerreotypes, on which moving objects (e.g., pedestrians, carriages, etc.) do not appear due to the lengthy time exposure required. For geographical data, it is necessary to consider change in both space and theme. In general one cannot represent 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 represent any event for which theme changes to a degree that would be discernible given the thematic resolution level (i.e., the detail that can be discerned in thematic attributes). This issue is discussed in this book under temporal accuracy in Chapter 8 and semantic accuracy in Chapter 7.

To complicate matters, observations are often generalized to produce a coarser temporal resolution than is necessitated by the duration of time required for data collection. For example, topographic map information may be derived from aerial photographs with very fine temporal resolution levels (on the order of several hundredths of a second for one photo and several minutes for the complete set of photos needed to completely cover the topographic map area). However, the temporal resolution of topographic maps is actually considerably coarser, as they are intended to represent conditions that do not change significantly over a time interval of years. The locations of rapidly-moving objects clearly resolvable on individual aerial photographs, such as automobiles, are not included on such maps. The intent is to produce a map that is, in a sense, free of time, in that the features shown on the map do not change appreciably over time. Clearly, rapidly-moving objects do not meet this criterion. Ultimately, of course, it is impossible to produce a map that is completely time-less, and topographic maps must be updated at regular intervals to account for changes that occur as a result of both natural and anthropogenic factors.

Resolution can also be defined in the thematic domain. In this domain, the meaning of resolution, like accuracy, depends on measurement scale. For quantitative data, resolution is determined by the precision of the measurement device used (e.g., a thermal sensing system that is able to resolve temperature differences on the order of 0.1°C, an 8-bit radiometer that is able to distinguish between 256 different levels of reflected energy, etc.). For data of kind or quality, resolution is defined in terms of the number and fineness of category definitions. For example soils can be classified at a relatively fine level of taxonomic resolution (e.g., soil series and phases), or at a relatively coarse level using more generalized classes (e.g., soil complexes or associations).

Completeness

As discussed in Chapter 5, completeness has been defined in terms of the relationship between the features encoded in the database and the abstract universe of all such features. This implies that, in order to be complete, the database must faithfully depict the real world. Since any database is an abstraction of the real world, this definition is difficult, if not impossible, to attain. Completeness may alternatively be defined as the degree to which all intended entries into a database have actually been encoded into the database. This implies that, in order to be complete, the database must truthfully depict what it purports to depict. From this perspective, completeness is related to the truth-in-labeling concept and cannot be divorced from the intended contents of the database and the resulting degree of generalization present. Even small-scale, generalized databases may score very highly on such an index of completeness.

In order to measure completeness in this way, a database must be precisely labeled with its intended thematic, spatial and temporal domains. However, such metadata are not always available, may be insufficiently detailed or may exhibit cultural biases limiting their use (see Chapter 5). Given sufficient metadata, completeness can be assessed by determining the degree to which a database contains all of the features it purports to contain. For example, a database purporting to depict the locations of leaking underground gasoline storage tanks in Portage County, Ohio, would be incomplete if it actually depicted only those tanks over a threshold storage capacity, only those tanks installed after 1980, or only those tanks adjacent to Interstate Highways. These omissions, while perhaps unavoidable due to data availability, need to be expressed clearly in the product label in order to achieve completeness.

Completeness can be assessed primarily through the identification of errors of omission. Such errors are identified by (i) defining precisely the purported contents of the database and (ii) determining the degree to which the purported contents match what is actually contained in the database. This assessment can be done for each domain (space, time and theme) in turn, as follows.

- Does the database description or title contain the necessary terms to correctly define the domain?
- Does the actual domain match the purported domain?
- Is there a systematic pattern evident in omissions? If so, what class or classes of features have been omitted? Can the database description be modified in accordance with these omissions?

The same approach might be adopted for assessing errors of commission. In this case, the identification of a systematic pattern of errors might be used to broaden the database product description.

Completeness is affected by sampling. The representativeness of a sample depends on the particular sampling scheme adopted. There has been little empirical work examining the implications of these various sampling schemes on the representativeness of the resulting sample. As a result, there is considerable disagreement about what constitutes the optimal sampling scheme (Congalton, 1991). The representativeness of a sample is also affected by the interaction between the sampling scheme and a number of characteristics, including the size and density of the sample, the spatial distribution of the sample, and the level of spatial variability and autocorrelation in the data (MacEachren and Davidson, 1987; Congalton, 1991). Guidelines have long been in use for determining sample size, both for positional and thematic accuracy assessment. Sample size depends on the importance of the class in the context of a particular study, and the fact that certain classes (e.g., water) show little spatial variability and are usually classified fairly accurately.

Errors of omission may result from the elimination of features that are smaller than the minimum mapping unit size, occur at subresolution time intervals, or have attribute values that are unimportant for the particular application for which the source document is being digitized. This suggests that completeness is related to resolution. However, a database can be complete, regardless of its resolution, if it contains all of the features it purports to contain. In this sense resolution is part of the database product description.

Consistency

Consistency is a measure of the internal validity of a database and refers to the fidelity or integrity of the database (DCDSTF, 1987). A consistent database is one for which there are no apparent contradictions in the relationships among the encoded features.

In the spatial domain, consistency is usually assumed to refer to the lack of topological errors (e.g., unclosed polygons, dangling nodes, etc.). These issues are addressed in some detail in Chapter 6. Topological errors tend to be interrelated, in the sense that they result from the same source. For example, an unclosed polygon may result from an unconnected arc, which may in turn result from an undershoot or a missing arc. This error will also appear as a dangling node and as a polygon with more than one label point. In the spatial domain it is usually not possible to differentiate among degrees of consistency, since most processing operations require

data to be topologically consistent before the operation can be carried out. Indeed the elimination of topological errors is an integral step in data editing and pre-processing.

Little work has been done on consistency in the temporal dimension. However, it might be possible to assess consistency in this dimension given temporal topology constructs (see Langran, 1992).

In the thematic dimension, inconsistencies include values for one attribute that are inconsistent with the values for another attribute related to the first. For example, if attribute A is the total population of a census tract, and attributes B and C are the mean household size and total number of households, respectively, then the attributes are inconsistent if the product of attributes B and C is not equal to the value of attribute A. Thematic consistency is simple to define in theory, but difficult to apply in practice. An obvious problem is that there are numerous relationships that must be examined among thematic attributes. General tests for such inconsistencies do not exist. In some cases, attributes may be wholly independent, such that tests of logical consistency need not be carried out.

Evaluation and Assessment

The evaluation matrix described above provides a coherent, logical structure for organizing, documenting and communicating information about data quality for geographical data. In this section we present a brief empirical assessment of the utility of the matrix. For this assessment we refer to a series of database production projects carried out by Geometria GIS Systems House (Budapest, Hungary). These projects represent different applications of geographical information systems and were commissioned by different user communities (e.g., a utility company, a mapping agency, a forestry office). During the production process, a number of data quality issues are addressed, including error correcting following preprocessing, digitizing and attribute data retrieval and encoding. Data quality for the completed database depends on the database specifications and the way in which error is treated during the data production processes.

Table 9.1 lists the database production projects considered in this example. The columns labeled spatial, temporal and thematic indicate the relative importance of each of these three domains in data quality assessment. The following classification scheme is used.

1 = The importance of this domain for data quality is known, but for one of many possible reasons it has not been considered in database production.
2 = This domain is of importance in assessing data quality.

Table 9.1 Relative importance of the spatial, temporal and thematic domains in data quality assessment

Database	Country	Scale	Spatial domain	Temporal domain	Thematic domain
Road network	Germany	1:25,000	2	1	4
Electrical utility system network	Hungary	1:2000	2	2	4
Water utility system network	Hungary	1:4000	2	2	4
Elevation database (Contours spot heights)	Netherlands	1:10,000	3	2	4
Forestry	Austria	1:10,000	2	1	4
Topographic	Netherlands	1:25,000 and 1:10,000	4	2	3
Sewage utility system network	Germany	1:1000	1	1	4
National GIS	Hungary	1:1,000,000 to 1:100,000	3	1	3
Cadastral	Hungary	1:1000	3	2	3
Landuse	Germany	1:1000	2	1	4

3 = This domain is of major importance in assessing data quality.
4 = This domain is the most important in assessing data quality for the database.

The classification scheme is relative and in application necessarily somewhat subjective. Perusal of Table 9.1 suggests that it is the thematic domain, and to a lesser extent the spatial domain, that is of most importance in data quality assessment for the selected projects. Temporal components of quality are, unfortunately, of relatively little perceived significance. Unfortunately there are not enough examples at hand to arrive at any clear generalizations. Table 9.1 is only the glimpse of a deep well, suggesting a possible future direction in data quality research.

Conclusion

One aspect of data quality assessment that is likely to increase in importance over the near-term is standardization of data quality information. This is seen in recent efforts by national agencies to develop workable data quality standards for geographical databases. Increased standardization would serve to enhance the ability to communicate the data quality characteristics of transferred data. There is, however, unlikely to be complete standardization, due to the different needs of users as reflected in different data quality objectives. Thus the continued interest in data quality issues is likely to be accompanied by the development of hybrid data quality standards. While information about data quality is required, it is not always possible, or even preferable, to predict in advance what those requirements might be. Flexible standards are required that allow for different levels of quality to be acceptable given the intended use of the database. Thus there is likely to be a move away from rigid compliance testing strategies like the National Map Accuracy Standard of the U.S. Geological Survey.

Above all, it is important that users of geographical databases move beyond simple awareness of the issue of data quality. Only the most naive of users are still unaware that data quality can have significant effects on the reliability of data processing operations. While awareness is a necessary prerequisite to the development of strategies for data quality assessment and communication, a simple statement that geographical data contain errors can no longer suffice.

References

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.

American Society of Photogrammetry (Committee for Specifications and Standards, Professional Practice Division) (1985). "Accuracy specification for large-scale line maps", *Photogrammetric Engineering and Remote Sensing*, **51,** 195–199.

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

Bedard, Y. (1986). "A study of the nature of data using a communication-based conceptual framework of land information systems", *The Canadian Surveyor*, **40,** 449–460.

Berry, B. (1964). "Approaches to regional analysis: A synthesis", *Annals, Association of American Geographers*, **54,** 2–11.

Blakemore, M. (1983). "Generalisation and error in spatial data bases", *Cartographica*, **21,** 131–139.

Carver, S. (1991). "Adding error handling functionality to the GIS toolkit", *Proceedings EGIS '91*, pp. 187–196.

Chrisman, N. R. (1982). "A theory of cartographic error and its measurement in digital data bases", *Proceedings, Auto Carto 5*, pp. 159–168.

Chrisman, N. R. (1983). "Epsilon filtering: A technique for automated scale changing", *Technical Papers of the 43rd Annual Meeting of the American Congress on Surveying and Mapping*, pp. 322–331.

Congalton, R. G. (1988). "A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data", *Photogrammetric Engineering and Remote Sensing*, **54,** 593–600.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment*, **37,** 35–46.

Congalton, R. G., Oderwald, R. G. and Mead, R. A. (1983). "Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques", *Photogrammetric Engineering and Remote Sensing*, **49,** 1671–1678.

Digital Cartographic Data Standards Task Force (DCDSTF) (1987). "Draft proposed standard for digital cartographic data", *The American Cartographer*, **15.**

Divényi, P. (1991). "Standardization efforts in Hungary", in *Spatial Database Transfer Standards: Current International Status*, edited by H. Moellering. London: Elsevier, pp. 111–122.

Federal Geographic Data Committee (FGDC) (1992). *Draft Content Standards for Spatial Metadata*.

Fegeas, R. G., Cascio, J. L. and Lazar, R. A. (1992). "An overview of FIPS 173, The Spatial Data Transfer Standard", *Cartography and Geographic Information Systems*, **19,** 278–293.

van Genderen, J. L. and Lock, B. F. (1977). "Testing land-use map accuracy", *Photogrammetric Engineering and Remote Sensing*, **43,** 1135–1137.

Giordano, A., Veregin, H., Borak, E. and Lanter, D. (1994). "A conceptual model of gis-based spatial analysis". Unpublished manuscript. Department of Geography, Kent State University, Kent, Ohio.

Greenland, A., Socher, R. M. and Thompson, M. R. (1985). "Statistical evaluation of accuracy for digital cartographic data bases", *Proceedings, Auto Carto 7*, pp. 212–221.

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

Honeycutt, D. M. (1986). "Epsilon, generalization and probability in spatial data bases". Unpublished manuscript.

Hudson, W. D. and Ramm, C. W. (1987). "Correct formulation of the kappa coefficient of agreement", *Photogrammetric Engineering and Remote Sensing*, **53,** 421–422.

Kellie, A. C. and Bryan, D. G. (1981). "A comparison of field methods for testing the vertical accuracy of topographic maps", *Technical Papers of the American Congress on Surveying and Mapping*, pp. 275–284.

Langran, G. (1992). *Time in Geographic Information Systems*. London: Taylor & Francis.

Lanter, D. and Veregin, H. (1992). "A research paradigm for propagating error in layer-based GIS", *Photogrammetric Engineering and Remote Sensing*, **58,** 526–533.

Lee, Y. C. (1985). "Comparison of planimetric and height accuracy of digital maps", *Surveying and Mapping*, **45**, 333–340.
MacDougall, E. B. (1975). "The accuracy of map overlays", *Landscape Planning*, **2**, 23–30.
MacEachren, A. M. and Davidson, J. V. (1987). "Sampling and isometric mapping of continuous geographic surfaces", *The American Cartographer*, **14**, 299–320.
Merchant, D. C. (1982). "Spatial accuracy standards for large scale line maps", *Technical Papers of the American Congress on Surveying and Mapping*, pp. 222–231.
Merchant, D. C. (1987). "Spatial accuracy specification for large scale topographic maps", *Photogrammetric Engineering and Remote Sensing*, **53**, 958–961.
Moellering, H. (ed.) (1991). *Spatial Database Transfer Standards: Current International Status*. London: Elsevier.
Moellering, H. (1992). STDS, *ACSM Bulletin*, No. 137, 30–34.
Parkes, D. N. and Thrift, N. J. (1980). *Times, Spaces, and Places: A Chronogeographic Perspective*. New York: John Wiley.
Rosenfield, G. H. (1971). "On map accuracy specifications: Part II. Horizontal accuracy of topographic maps", *Surveying and Mapping*, **31**, 60–64.
Rosenfield, G. H. and Fitzpatrick-Lins, K. (1986). "A coefficient of agreement as a measure of thematic classification accuracy", *Photogrammetric Engineering and Remote Sensing*, **52**, 223–227.
Stearns, F. (1968). "A method for estimating the quantitative reliability of isoline maps", *Annals, Association of American Geographers*, **58**, 590–600.
Veregin, H. (1989). "Error modeling for the map overlay operation", in *The Accuracy of Spatial Databases*, edited by M. Goodchild and S. Gopal. London: Taylor & Francis, pp. 3–18.
Veregin, H. (1994). *Accuracy Tests for Polygonal Features*, Technical Report, Environmental Monitoring Systems Laboratory. U.S. Environmental Protection Agency, Las Vegas, Nevada.
Wood, D. (1992). *The Power of Maps*. New York: Guilford.

CHAPTER TEN

Looking ahead

Stephen C. Guptill and Joel L. Morrison

The democratization of the use of spatial data is one of the results of society's wholesale adoption of electronic technology at the end of this century. The reader, who has worked through the preceding nine chapters dealing with the overall importance and detailed dissection of the elements of the quality of spatial data, may, by this time, have lost their perspective about the role of spatial data on post twentieth century Earth, and may be either completely depressed at the current state of affairs dealing with digital spatial data, or have adopted the attitude of "so what". In an attempt to deal with either relief or depression, or to answer the "so what" attitude, the editors of this volume append this chapter to speculate about the future and the role which digital spatial data may play in the civilization of twenty-first century Earth.

Cartographers, surveyors, and other spatial scientists, but primarily cartographers and maybe surveyors, must recognize that our current technology allows any person to go to a point on the surface of the earth and to record the position of that point to a degree of precision (chapter 3) that will serve well over 99% of the possible uses of that data. The need for surveying, in the traditional use of that term, has been superseded by technology that can be easily employed by any person, not to mention robots, in the next century. The recording of attributes for that point can also be increasingly precisely measured, but the true degree of precision depends upon the attribute (chapter 4) being measured and upon its definition (chapter 7). Completeness (chapter 5) and Logical Consistency (chapter 6) as well as Lineage (chapter 2) follow directly from the definitions and the measuring activities. The remaining element is the temporal element of spatial data quality and it is our belief that for most of the twenty-first century the temporal aspects of spatial data quality will require and receive the most attention.

Use of Spatial Data

It is postulated, that by sometime early in the twenty-first century, positions and attributes will be routinely recorded in digital files of sufficiently