

Data Quality And Flow Of Errors In GIS

NIRMALENDU KUMAR

Survey of India, Chandigarh.

ABSTRACT

The term quality is generally used to indicate the superiority of a manufactured good or to attest a high degree of craftsmanship or artistry, but quality is more difficult to define for data as they do not have physical characteristics. Quality is thus, a function of intangible properties such as accuracy, resolution, consistency and completeness. The quality of geo-spatial data affects the decision made with these data and bad decision can have severe consequences. Moreover, since these data are often used for regulatory purposes or to resolve dispute, the custodians of such data are clearly exposed to potential liability if the data are shown to be in error. However, much of GIS practices continue to proceed as if data were perfect and result of a GIS analysis rarely shows estimate of confidence or other indicators of effect of data quality. Awareness regarding data quality among consumer, is also lacking.

The concern for geospatial data quality has burgeoned in recent years due to increased data production by the private sectors which do not require to conform to the known quality standard as governmental agencies, be it Survey of India or US Geological Survey of USA. GIS is also now being increasingly used as a decision support system and poor quality data may cause litigation and related compensation. Increased reliance on secondary data sources due to increased network accessibility, development of standard for data exchange and interoperability also necessitates never before data quality concern.

Until recently, data quality was the responsibility of producer and compliance testing strategies were applied to sanctify databases meeting official quality threshold which may be too less for some application and too restrictive and costly for others. Responsibility for assessing whether a database meets the needs of a particular application has therefore shifted to consumers, which is now referred as fitness for use. Producers responsibility has changed as well and rather than producing authoritative databases their role has shifted to data quality documentation or truth-in-labelling, which view error as inevitable and cast the data quality problem in terms of misuse arising from incomplete knowledge of data limitation.

In geo-spatial database, combined role of storage and communication of the paper map can be more easily separated. This means that data can be collected in, as raw a form as possible and representation can be created to achieve any particular communication objective without altering the content of database. The problem of error flow would not have come to fore, if GIS were used only as an electronic map drawer. However GIS has enormously extended uses of geo-spatial 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, analysis, and decision-making models. One of the most powerful capabilities of the GIS is that it permits the derivation of new attribute from the attributes already held in the GIS database. When a map is stored in a database it can be safely said that it is not truly error free, and when this is used, as input to a GIS operation then the error in the input will propagate to the output of the operation. Therefore, output may not be sufficiently reliable for correct conclusions to be drawn from it. Moreover error propagation continues when the output from one operation is used as input to an ensuing operation. Consequently, when no record is kept of the accuracy of the intermediate results, it becomes extremely difficult to evaluate the accuracy of final result. Although user may be aware of hard facts in practice they rarely pay attention to the problem. The monitoring and documenting of these error flow is not given their due attention by the vendors / creators of GIS software and more efforts are needed in this direction.

INTRODUCTION

The meaning of **quality** depends on the context, it is used. The term is commonly used to indicate the superiority of a manufactured good or to attest a high degree of craftsmanship or artistry. But 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 accuracy,

resolution, consistency, and completeness. The same comments also apply in the context of geo-spatial data.

The quality is an important property of almost all geographical data and it certainly affects the decisions made with these data. In general the poorer the quality of data, the poorer the decision. Bad decision can have severe consequences, as ambulance can be sent to wrong location. Moreover, geographical data are often used for regulatory purposes or

to resolve dispute. The custodians of such data are clearly exposed to potential liability if the data are shown to be in error.

Despite what appears to be obvious, arguments in favour of explicit treatment of data quality in GIS, much GIS practices continue to proceed as if data were perfect. Result of GIS analysis whether in the form of tables, maps or displays rarely shows estimate of confidence or other indicators of effect of data quality. There is a general tendency to give computers more credit than they deserve to believe that because numbers or maps have emerged as if by magic from digital black boxes, they must necessarily be reliable. The awareness regarding data quality among consumers is also lacking and therefore same diligence is not given while selecting a database for GIS use as they might give in purchasing an automobile or a pair of shoes.

INCREASING CONCERN REGARDING DATA QUALITY

Concern for Geospatial data quality has burgeoned in recent years due to the following factors:

Increased Data Production by the Private Sector

Historically mass production of the Geospatial data was the domain of Governmental agencies be it Survey of India, US Geological Survey or British Ordinance Survey. Unlike these agencies private companies are not required to conform to the known quality standards. More over with the advent of hand held GPS and high-resolution imageries data generation has become a child play.

Increased Use of GIS as a Decision Support tool

This trend has led to the realization of the potential deleterious effects of using poor quality data, including the possibility of litigation and compensation if minimum standards of quality are not attained.

Increased Reliance on Secondary Data Sources

This has been fuelled by a reduction in the accessibility and constraints resulting from network accessibility and the development of standards for data exchange and *interoperability* of GIS.

Changing Responsibility of Data Producer/User

These trends have contributed to a reappraisal of the responsibilities of the data producer 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 threshold. However these compliance tests can be used for a limited range of application as they may be too less for some application and too restrictive and hence impart unnecessary cost for the other.

Responsibility for assessing whether a database meets the needs of a particular application has therefore shifted to the consumers, who are now 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 the data quality documentation or *truth-in - labelling*. The truth in labelling views error as inevitable and cast the data quality problem in terms of misuse arising from incomplete knowledge of data limitations.

DATA QUALITY COMPONENTS

Geographical observations describe phenomena with spatial, temporal and thematic components. Historically, space, which defines geographical location, is the dominant member of this troika, which is problematic on several levels. Firstly, time is not given sufficient attention which may be critical to an understanding of a geographical phenomena not as an entity that exist at some location but as events appear and disappear in space and time. Secondly geographical phenomena are not really about space but theme. We can view space-time as a framework on which theme, is measured. It is true that without space there is nothing Geographical about 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* and for each of these dimensions several components of quality including accuracy, precision, consistency and completeness can be identified.

Entity - Attribute - Value Model

This model, which serves as the conceptual basis for most database implementations of real world phenomena may be a good starting point for discussing accuracy. According to this model *entities* represent real - world phenomena (such as street) *attributes* specify the relevant properties of those objects (such as width or No. of lane) and *values* gives the specific qualitative or quantitative measurements pertaining to a particular attribute . In this model *error* is defined as the discrepancies between the encoded and actual value of a particular attribute for a given entity. *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.

Limitations of Entity- Attribute-Model

The above definition is limited, as it does not recognize 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.

Secondly, the definition of error given above assumes that there is some objective external reality against which encoded value can be measured. This definition required not only that **truth** exist but also it can be observed. This definition is problematic for several reasons. Firstly; truth may simply be unobservable e.g. historical data. Secondly, observation of the truth may be impractical e.g. because of data cost. Finally it is possible that multiple truth exists because the entities represented in the database are observations rather than real world phenomena. Thus accuracy is a relative measure rather than absolute one since it depends on the intended form and content of the database.

Spatial Accuracy

Spatial accuracy (or positional accuracy) refers to the spatial component of a database. Measurement of spatial accuracy depends on dimensionality. Metrics are well defined for point entities but widely accepted metrics of line or areas are yet to be defined. For points, error generally depend on the discrepancy between the encoded location and location as specified in the specification. Error can be measured in any one of or in combination of the three dimensions of space.

Various metrics have been developed to summarize spatial error for set of points. One such metrics is mean error, which tends to zero when bias is absent. Bias is a systematic pattern of errors (e.g. error arising from map mis-registration) when bias is absent error is said to be random. Another such metrics is Root mean square error, which is computed as square root of the mean of the squared error and is commonly used to document accuracy of digital elevation model.

Thus, it is easy to perform statistical inference tests and derive confidence limits for point location for lines and areas the situation is more complex, since there is no simple statistical measure of error that can be adopted from statistics. Error in the line arises from error in the points that define those lines. However as the points are not randomly selected the error present in the points cannot be regarded as somehow typical of errors present in the line.

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.

Temporal Accuracy

Temporal accuracy has not received much attention, as time itself is not dealt with explicitly in conventional geospatial data models. Temporal accuracy is often equated with correctness. In fact the two concepts are quite distinct. Temporal accuracy refers to the agreement between encoded and actual temporal co-ordinates. Correctness is an application specific measure of temporal accuracy.

Thematic Accuracy

Metrics of thematic 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.

Precision or Resolution

Precision refers to the amount of detail that can be discerned. It is also known as resolution or granularity. The term resolution is commonly used in GIS and related field and is adopted here to avoid the confusion with 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 generalized. Generalization includes elimination and merging of entities, reduction in details; smoothing, thinning and aggregation of classes. Generalization is ineluctable because at best geospatial database can encompass only a fraction of the attributes and their relationship that exist in real world. 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 connotations as low accuracy. Low resolution may be desirable in certain situation, such as when one wishes to formulate general model or examine spatial pattern at regional level.

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 digital image. This defines the minimum size of the object on the ground that can be discerned.

The concept is applicable without modification to the raster databases. For vector data, the smallest feature that can be discerned usually defined in the terms of rule for minimum mapping unit size, which depends on map scale.

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 rate of change in the event. Events with a life time less than the sampling interval is generally not resolvable. A shorter recording time implies higher temporal resolution. For geospatial data, the situation is more complicated because interaction 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 at the thematic resolution level. Resolution however is a function of the time required to obtain spectral reflectance data for one pixel.

Thematic Resolution

In the thematic domain, the meaning of resolution depends on measurement scale. For quantitative data the resolution is determined by the precision of the measurement device. For categorical data the resolution is defined in terms of the fineness of category definitions.

CONSISTENCY

Consistency refers to the absence of apparent contradictions in a database. For geospatial data the terms is used primarily to specify conformance with certain topological rules. 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. Such that most database are topologically cleaned before being released.

Topological consistency is one object of consistency in the spatial domain. Spatial inconsistencies can also be identified through redundancies in spatial attributes. Non-redundancy implies that there is independence between two attributes such that meaningful consistency constraints do not exist.

Little work has been done on consistency in the temporal domain. Although a framework for temporal topology has been developed by. For example since at a given location only one event can occur one time, an inconsistency exists if a different entity appear at the same location in two maps of the same

date. Since events have duration the idea can be extended to identify events that exhibit temporal overlap.

In the thematic domain, the ability to identify inconsistencies require a level of redundancy in thematic attributes- for example the three socio-demographic variables. **Population, mean house hold size, total number of house holds**. In this case consistency may be appropriately viewed as a measure of internal validity.

COMPLETENESS

Completeness refers to the relationship between the object in the database and the abstract universe of all such objects. The abstract universe can be defined in the forms of desired degree of abstraction and generalization (i.e. a concrete description or specification for the database). Selection criteria, definition and other mapping rules used to create the database are the important determinants of completeness. The above discussion of completeness implies that there may be two different types of completeness, data completeness and model completeness.

Data Completeness

It is the 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 generalized database can be complete, if they contain all the objects described in the specification.

Model Completeness

It refers to the agreement between the database specification and the abstract universe that is required for a particular database application. Model completeness is application dependent and therefore an aspect of fitness-for-use. It is also a component of semantic accuracy.

The definition of completeness given above are example 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 value are present for all attributes.

Now feature completeness can be defined over space, time and theme. Consider a database depicting the location of buildings north of river Ganges in India that were placed on the register of tourism ministry as tourists places as of year 2000. This database would be incomplete if it does not include buildings of J & K (incompleteness in space, since J & K is part of India and north of Ganges) or only buildings placed on the register by June 30, 2000 (incompleteness in time since buildings may have been added after June 30) or database

shows only residential buildings (incompleteness in theme, due to omission of non residential buildings).

REQUIREMENT OF DATA QUALITY STANDARD

A concern for data quality standard is being clearly expressed in the development of data transfer and metadata standard world over. Such standards have been developed at both national and international level in support of mandates for data acquisition and dissemination. Data quality documentation plays a key role in such standard. Due to the realization that an understanding of the quality is essential to the effective use of geospatial data, U.S. has created spatial data transfer standard (SDTS) which has five data quality components as tabulated below, for data quality. It provides standard definition

of data elements, a standardized format for data transfer and descriptive meta data about database components.

Similarly Federal geographic data committee (FGDC) of USA has developed metadata contents standard in 1994, which provide a common set of terminology and a common structure for geo-spatial metadata. The use of these standards is one of the minimum requirements for serving as a node in the National Geospatial Data clearing house of the National Spatial data infrastructure (NSDI).

In India too, we have established our National spatial data infrastructure (NSDI) and are moving towards the creation of nodes for national geospatial data clearinghouse.

SL. No.	Component	Description
1.	Lineage	<ul style="list-style-type: none"> □ Refers to source materials, method of derivation and transformation applied to a database. □ Includes temporal information (date that the information refers to on the ground) □ Intended to be precise enough to identify the source of individual objects (i.e. if a database is derived from different source, lineage information is to be assigned as an additional attribute of objects or as a spatial overlay).
2.	Positional Accuracy	<ul style="list-style-type: none"> □ Refer to the accuracy of the spatial component. □ Subdivided into horizontal and vertical accuracy elements. □ Assignment methods are based on comparison to source, comparison to a standard of higher accuracy, deductive estimates or internal evidence.
3.	Attribute Accuracy	<ul style="list-style-type: none"> □ Refers to the accuracy of thematic component. □ Specific test vary as a function of measurement scale. □ Assessment methods are based on deductive estimates, sampling or map overlay.
4.	Logical Consistency	<ul style="list-style-type: none"> □ Refers to the fidelity of the relationship encoded in the database. □ Includes test of valid values for attributes and identification of topological inconsistencies based on graphical or specific topological tests.
5.	Completeness	<ul style="list-style-type: none"> □ 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.

Table-1 SDTS component of data quality.

Unfortunately, US counter part of SDTS data standard and FGDC meta data contents is not evolving at the required pace, without which functioning of NSDI will not be a sooth sailing.

ISSUES IN DATA QUALITY STANDARD

Some of the main issues / constraints are enumerated below:-

- i) A major limitation of data quality standard is that they do not necessary lend themselves to specific software implementations. These standards provide only models for data documentation but not a mechanism where by users of desperate GIS Packages can implement those models for database documentation.
- ii) These standard 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 passed through GIS processing steps. Hence while source data may be adequately documented, derived data are frequently not.
- iii) These standards provide such a rich collection of information about data quality, users may find it difficult to ascertain fitness for use.

FLOW OF ERRORS IN GIS DATABASE

Maps serve the dual purpose of storage and communication. For paper maps contents depend on the communication goals. The desire to communicate a particular message leads to selective enhancement and exaggeration of certain feature and elimination or displacement of other. In geospatial databases storage and communication roles can be more easily separated. This means that data can be collected in as raw form as possible, and representation can be created to achieve any particular communication objective without altering the contents of the database.

The problem of error flow would not have come to the fore if GIS were used only as an electronic map drawer. However GIS has enormously extended 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.

One of the most powerful capabilities of GIS is that it permits the derivation of new attribute, from the attributes already held in the GIS database. For example a DEM can be used to derive maps of gradient and aspect or digital map of soil type and gradient can be combined with the information about soil fertility and moisture supply to yield maps of suitability of growing maize. Many such basic type of functions used for derivation of such kind are often provided as standard functions or operation in GIS.

It can be safely said that no map stored in a GIS is truly **error** free. The word error used here is in its widest sense to include not only mistakes or blunders but also to include the statistical concept of error, meaning **Variation**. When maps that are stored in a GIS database are used as input to a GIS operation then the error in the input will propagate to the output of the operation. Therefore the output may not be sufficiently reliable for correct conclusions to be drawn from it. Moreover, error propagation continues when the output from one operation is used as input to an ensuing operation. Consequently, when no record is kept of the accuracy of intermediate results, it becomes extremely difficult to evaluate the accuracy of final result.

Although, user may be aware of these hard facts, in practice they rarely pay attention to the problem. The monitoring and documenting of these error flow is not given their due attention by the vendors/ creators of GIS software also and more efforts are needed in this direction.