National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing (RTMC) 2011
Proceedings published in International Journal of Computer Applications® (IJCA)

ANALYSIS ON AXIOMS OF SPATIAL DATA QUALITY

Vinti Parmar, Priyanka Goyal, RahulRishi
Department of Computer Science, Banasthali University, P.O. Banasthali
Vidyapith-304022, Rajasthan, INDIA
Department of Computer Science,The Technological Institute of Textile & Sciences , P.O Bhiwani-127021
Haryana, INDIA
Department of Computer Science, The Technological Institute of Textile & Sciences , P.O Bhiwani-127021
Haryana, INDIA

ABSTRACT
Data about positions attributes and relationships of features in space are often termed as spatial data. Spatial data quality is defined as the concept of 'fitness for use'. Data sharing and data integration are evitable consequence of the widespread availability of Geographic Information System (GIS) technologies due to the high initial cost of establishing a spatial database. Because spatial data is transferred and shared by many users, the data must be trustworthy and useful. To ensure that existing digital data is appropriately used, the data producer must provide documentation about the history of spatial data. In addition, data developers and users have begun to document and implement data quality measurements which allow judgment to be made about spatial data.

Keywords
Spatial data, Spatial data quality, Geo- graphic information system

1. INTRODUCTION
Over the years, there has been a drastic increase in usage of spatial data. "Data about positions attributes and relationships of features in space are often termed as spatial data" .Spatial data quality is defined as the concept of 'fitness for use' . Some elements of the spatial data quality are the positional accuracy, temporal accuracy, attributes accuracy, logical consistency, completeness, lineage and semantic accuracy. These are produced and used by various organizations for numerous applications. Public have easy access to spatial data through various means, e.g. Google maps and Google Earth. For instance Google Map Maker is a new service that allows users to edit and contribute map information like draw, add map features. This increase in mass consumptions and production of spatial data has its related issues of which maintaining the quality is one of the significant issues.Data quality is the degree of data excellence that satisfy the given objective. In other words, completeness of attributes in order to achieve the given task can be termed as Data Quality. Production of data by private sector as well as by various mapping agencies assesses the data quality standards in order to produce better results. Data created from different channels with different techniques can have discrepancies in terms of resolution, orientation and displacements. Data quality is a pillar in any GIS implementation and application as reliable data are indispensable to allow the user obtaining meaningful results All data sources and spatial data entry methods present errors into the information that is created and used for display and analysis. The type, severity and implications of these errors inherent in a Geographic Information System (GIS) database determine the quality of spatial data. For century's cartographers, geographers, surveyors and geodesists have been involved in collection, storage, analysis and visualisation of spatial data. How to test for and how to assure data quality, has been the topic of several international conferences and workshops. This paper seeks to present a review of the description of spatial data quality, elements of data quality and sources of error at different stages in spatial data ..

2. SPATIAL DATA
Spatial data is a data that contains positional values. Occasionally the more precise phrase 'geospatial data' is used as a further refinement, which means spatial data that is georeferenced. However, information is data that has been interpreted by a human being. Humans work with and act upon information, not data. Human perception and mental processing leads to information which may result in understanding and knowledge. Geoinformation is a specific type of information that involves the interpretation of spatial data. As this information is intended to reduce uncertainty in decision making, any errors and uncertainties in spatial information products may have practical, financial and even legal implications for users. For these reasons, those involved in the acquisition and processing of spatial data should assess the quality of the base data and the derived information products.

Since the 1980s, concerns about spatial data quality have increased, as a result of two developments:(1)the emergence of GIS in the 1960s and (2) from the 1970s onwards, a strong increase of available spatial data from satellites. With the large-scale adoption of GIS, the number of users from non-spatial or analog to digital spatial data disciplines has grown. The shift from analog to digital spatial data has been rapid. The inherent nature of spatial data is cartesian, the points, lines and polygons on maps are imposed on a grid. In fact, the data is collected in the field as numbers in an X/Y, Latitude/Longitude coordinate system. Managing these numbers as digital data is easier than managing them as analog data. Digital cartographic data has been in use since the 1960s, becoming a standard in the past decade with the rapid rise in computing power, the fall in computer prices, education and training of map professionals.

The major issues that have challenged the spatial data community over the years have had to do with data quality and error. Quality is about "fitness for use." It has to do with the extent to which a data set, or map output or a GIS satisfies the needs of the person judging it .Error is the difference between actual data and true data. Error is a major issue in quality and it is often used as an umbrella term to describe all the types of effects that cause data to depart from what they should be. These errors must be recognized and properly dealt with. It is virtually impossible to eliminate spatial data errors altogether. However, GIS users can reduce and manage errors effectively thus improving the quality of data. Identifying and assessing data errors are not the only factors which determine data quality. Data quality includes all of the processes involved with developing, utilizing and maintaining a spatial database. These include data collection, data input, positional and

attribute accuracy, data storage, data manipulation, data conversion and quality control procedures.

3. DATA QUALITY PARAMETERS

Data quality parameter provides an insight to the user on the dataset fitness-for-use. Essentially the data quality of a particular dataset is described by the data quality elements and subelements. The quality parameters are completeness, logical consistency, positional ,thematic and temporal accuracy. Depending on these elements the user takes decisions whether data is of relevant to their application.

A. Completeness

Completeness is number of committed and omitted objects, their attributes and relations in the dataset to the reference. Completeness is of feature completeness, attribute completeness and value completeness .The quality measure could of Boolean, integer, percentage or ratio depending on the method of evaluation. Data quality sub-elements are Commission and Omission. Commission is of extra committed objects in a dataset. Example: If there are 20 lakes present in reference dataset, and there 25 lakes existing in our dataset ,then the quality is "commission 5%". Omission is the omitted objects in a dataset. Example: If there are 100 buildings in reference dataset, and there 95 buildings exist in our dataset, then the quality is "omission 5%".

B. Consistency

Consistency is of four types: conceptual consistency, domain consistency, format consistency ,topological consistency. The data quality measure of consistency could be of boolean , integer, ratio ,and percentage .

1*) Logical Consistency:* It refers to the degree of adherence to logical rules of data structure, attribution and relationships.

2*) Conceptual Consistency*: It refers to the degree of adherence to rules of the conceptual schema .Example: If the dataset is not consistent with the application schema, then it is conceptual consistency error.

3) *Domain Consistency*: It is degree of adherence of values to the value domains .Example: If the attribute value ranges from 1 to 5.But if the value does not fall in the range, and then it is a domain consistency.

***4) Format Consistency*:** It is degree which data is stored in accordance with the physical structure of the data set.Example: If the data format is of ESRI shape file and if does not contain database file (.dbf file), then it is format consistency.

5*) Topological Consistency*: It is the ratio of items out of rules of topological characteristics like overlapping, undershoot, overshoot and contain. Example: number of overshoots and undershoots.

C. ACCURACY

Accuracy is one of the primary components in the data quality. Accuracy can further subdivided into accuracy of attribute values spatial and temporal references .

1) *Positional Accuracy*: It shows the deviation of geographical feature location in a dataset to its ground truth. It is how well the true measurements of a object on ground match with same object in the database. It also relates to the relative or absolute positional accuracy of Positional accuracy refers to the accuracy of the spatial component of a database the features. The data quality measure for this element is error statistics i.e. Root mean square error RMSE.

***2) Absolute Accuracy*:** It is the closeness of the reported coordinate values in a dataset to the values accepted as or being true.
Example: If the RMSE of co-ordinate value of our dataset from the value of reference dataset is 0.45,then the "Absolute accuracy is 0.45m"

3) Relative Accuracy: It is the closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true. Example: The difference between relative distances of two control points of reference dataset from the control points of our dataset is 0.27m, "then the relative accuracy is 0.27m".Example:Gridded data positional accuracy is defined as closeness of the gridded data position values to the values accepted as or being true . Example: RMS error of the TIN grid point elevation values and value of the reference dataset is 1.2m then the "gridded positional accuracy is 1.2m".

4) *Temporal Accuracy* : Temporal accuracy is of accuracy of time measurement, temporal consistency and temporal validity .Temporal accuracy is correctness of temporal reference of an item. The quality measures are generally of Boolean values or error statistics .Data sub-quality element are following: Accuracy of time measurement "Difference between time attribute recorded in our dataset to the reference dataset". Example: If there are 6 days difference between data of construction completed to the recorded data of construction completed in our dataset. 'Accuracy of time measurement is of 6 days'. Temporal consistency is correctness of ordered events .Example: If the data of demolition is earlier than the construction data, then it is temporal consistency error. Temporal validity is the validity of data with respect to time". It is treated with the same data quality measures used for domain consistency.

5*) Thematic Accuracy*: Thematic accuracy is the accuracy of either spatial or thematic attribute of feature .Data quality sub-elements are following: "Classification correctness indicates the correctness of classification items ."Example: If a rail is classified as road, then it is a thematic error ."Non-quantitative attribute accuracy indicates the correctness of non-quantitative attributes ."Example: If 5% road names of our dataset are incorrect to the reference dataset road names then the "Non-quantitative attribute accuracy is 5%".
"Quantitative attribute accuracy is the accuracy of quantitative attributes."RMS error, by comparing the attribute "length" in the dataset to the length in the reference dataset, if it is 10m then the "Quantitative attribute accuracy is 10m".

4. ASSESSMENT OF DATA QUALITY

Data quality is assessed using different evaluation techniques by different users.

A. the First Level of Assessment It is performed by the data producer. This level of assessment is based on data quality check based on given data specifications.

B. Second level of Assessment
It is performed at consumer side where feedback is taken from the consumer and processed. Then the data is analyzed rectified on the basis of processed feedback.

5. SOURCES OF SPATIAL DATA DISCREPANCY

A .Data Information Exchange

Data information exchange is basically the information about the data provided by the client to organization. The degree of information provided by the client defines the accuracy and completeness of data.

B. Typeandsource

Data type and source must be evaluated in order to get appropriate

data values. There are many spatial data formats and each one of them is having some beneficiary elements as well as some drawbacks. For example: In order to use CAD data on GIS platform, data must be evaluated and problems must be rectified otherwise resultant values will show the high extents of discrepancies. Conventional data formats are quiet specific to data storage technique and functionl compatibilities . Example: Topology can not be created on shapefiles. This can be created only on the latest geospatial storage format.So, data type and source must be identified and evaluated before proceeding towards any analysis.

1*) Data Capture*: There are many tools that incorporate manual skills to capture the data using various softwares like ArcGIS. These software allows user to capture information from the base data. During this data capture, the user may misinterpret features from the base data and captures the features with errors. For example: A user misinterprets two buildings as single building and capture as a single feature. But in real world, there are two features. So, the correct interpretation of features in base data must be performed. However, there are many tools that enables user to find and fix those errors, but still these tools are not used frequently due to lack of awareness. Data capture must be performed on a perfect scale where one must be able to view the features distinctly.

2) *Cartographic Effect:* After capturing the data, some cartographic effects like symbology, pattern, colors, orientation and size are assigned to the features. This is required for a better representation of reality. These effects must be assigned according to the domain of the features. Like for Forestry application, forestry domain specific cartographic elements must be used. Elements of any other domain used for a particular domain degrades the output of results.

3)*Data Transfer*: Some discrepancies may occur while transferring the data from one place to another. For example: Data transferred from a web source to the standalone, web disconnected machine. Sometimes, In order to make the accurate data more accurate, user tries to apply different advanced rectification technique but as a result the less accurate data changes into highly degraded data. "There is no bad or good data. There are only data which are suitable for a specific purpose." So, Data must be evaluated according to the domain for which it is supposed to be used.

4)*Metadata*: Sometimes metadata is not updated according to the original features. For example: Few features are edited on some software platform but the edited information is not updated like name of the editor, reason for editing and some more relevant information. So, metadata must be updated with the original data.

6. CONCLUSIONS

The following conclusions can be deduced from the discussions It is pertinent that data quality be maintained at all stages of a GIS database especially during data development and data maintenance. It is important to note that any type of manipulation done on data affects the quality of that data. So caution should be taken when adding, editing and updating spatial data. Errors induced at any stage will change the outcome of spatial data analysis and this can undermine the whole purpose and functionality of a GIS. Data developers should have documented rules and guidelines to follow when creating and updating data layers. This documentation will help to eliminate any future questions concerning data creation, quality and/or data analysis and avoid any duplicate efforts made of creating specific data layers. This information is not only useful for in-house data development, but data customers and users are able to determine the validity of data by checking the sources and procedures used to create the data.

7. REFERENCES

[1] Danko, D., 2000. ISO 19115 Geographic Information - Metadata, ISO/TC211 Geographic information/Geomatics.

[2] Devillers, R. and Beard, K., 2006. Communciation and Use of Spatial data Quality Information in GIS.Fundamentals of Spatial Data Quality. ISTE, London, 237-250 pp

[3] Hunter, G.J. et al., 2005. Next-Generation Research Issues in Spatial Data Quality,Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis, Melbourne.

.[4] Minton S., Nanjo C., Knoblock C., Michalowski M.,and Michelson M. (2005) ―A Heterogeneous Field Matching Method for Record Linkage" The Fifth IEEE International Conference on Data Mining

[5] Yang, T., 2007. Visualisation of Spatial Data Quality for Distributed GIS, The University of New South Wales, Sydney, 199 pp.