COMMUNICATING DATA QUALITY IN A GIS ENVIRONMENT

by

Kanagaratnam Thavalingam

"This thesis was submitted to the department of Earth Resources Engineering of the University of Moratuwa in partial fulfillment of the requirements for the Degree of Master of Science"

Department of Earth Resources Engineering
University of Moratuwa
Sri Lanka

October 2001

Abstract

The GIS database is a digital representation of the real world. Any abstract of reality will contain discrepancies from its source. With traditional methods many of the problems are visible and the skilled map analyst makes the necessary adjustments and knows how far the information can be relied upon. With a Geographic Information System the equivalent operations are not transparent (the black box effect), usually the operators are no longer so skilled and the problems are largely invisible. The digital modelling has the potential to dramatically increase both the magnitude and importance of errors in the models. The results may be used for decision making and planning despite possessing levels of uncertainty that are completely unknown and usually cannot even be guessed. That is why the accuracy analysis is one of the most important problems in the development and applications of the system.

Currently there are several demands from users of data to include quality parameters in the related GIS databases. A number of researchers have done work on the derivation of data quality especially on positional or geometrical accuracies. However there has been little work done on qualitative or semantic accuracies and ways of communicating them. A major contribution toward standardizing the definition, assessment and reporting of GIS data quality has been made by the Data Set Quality Working Group of the National Committee for Digital Cartographic Data Standards.

This research, provides an overview of the data quality factors that should be considered when using geographic information, and is intended to explore the possibility of generating and communicating data quality in various ways in a GIS environment. Suitable algorithms, .mainly concerning positional and attribute accuracy assessments, were adopted from relevant literature to determine the data quality parameters at different levels of abstraction, for different data types. The levels of abstraction. considered were overall accuracy parameters at coverage level and specific accuracy parameters referring to entity level. To communicate the data

quality to the user different methods such as numerical, graphical and textual messages were adopted .The area for the case study is located in Kegalle district. The feasibility of the reported implementation was assessed by means of the referred case study. The results obtained with this case study were used to draw some conclusions and recommendations regarding the communication of, data quality in a GIS environment.

The work included in the thesis in part or whole, has not been submitted for any other academic qualification at any institution

October 2001

K.Thavalingam

Acknowledgements

Firstly, I would like to mention all effective support and guidance given throughout the research, by my supervisor Dr.U.G.A.Puswewala, Department of Civil Engineering, University of Moratuwa.

I wish to express my appreciation to my former supervisor Dr.U.G.Senerath for the initial guidance given to me to start this research project.

To Mr. Sarath Weerawaranakula, Head, Department of Earth Resource Engineering. I would like to express my sincere thanks for the help given to me throughout the period of study and contributed to this thesis with the helpful hints and suggestions.

To Prof. P.G.R. Dharmaratne, Department of Earth Resource Engineering, University of Moratuwa. I would like to thanks for the helpful suggestions he has made during the review of the research project and thesis writing.

To the staff attached to University of Moratuwa, especially to Department of Earth Resource Engineering, I would like to express my gratitude for all the support given to me during my research work.

Much of the data input were performed by the GIS Branch in Survey Department under the supervision of Mr.D.N.D.Hettiarachchi, Superintendent of Surveys (GIS); their collective efforts are warmly appreciated.

To Mr.A.Dissanayake, Superintendent of Surveys, Kegalle District, and his surveyors; I must acknowledge all help given through the field data collection.

To Mr.S.D.PJ. Dampegama, Superintendent of Surveys (Geodetic Surveys), Institute of Surveying and Mapping, Diyatalawa and his surveyors and staff, I must thank all the helps given through the GPS observations and the support given at all stages of the research projects.

For the useful and necessary information and explanation regarding Arc/Info data structure and programming, I give my thanks to Mr.S.Sivanantharajah, Asst. Superintendent of Surveys, Center for Remote Sensing.

For helpful criticism and suggestions given to me for the improvement of the research, I would like to thank Mr. K.D. Parakkum Shantha, Superintendent of Surveys (Air Surveys) and his wife Mrs. Shamily Parakkum Shantha, Superintendent of Surveys.

I would like to thank the Surveyor General and Survey Department for supporting this research and releasing the digital data and resources.

This research was funded by ADB. I gratefully acknowledge the fellowship provided me by ADB.

Thank you to all concerned.

K.Thavalingam

Table of Contents

List of Annexes

List of Figures

List of Tables

CHAPTER 1 Introduction

The Geographical Information System (GIS) is a powerful tool for handling spatial data, which are maintained in digital format (Aronoff, 1989). This data format is more physically compact than that of paper maps, tabulation or other conventional types. Large quantities of data can also be maintained and retrieved at greater speeds and at lower costs per unit. (Aronoff, 1989).

The Geographical Information System (GIS) gives complete freedom to combine, overlay and analyze data from many diverse sources regardless of scale, accuracy, resolution and quality of original map documents (Carver, 1991). This ability to mix geographical information from various map scales and sources is a key aspect of GIS functionality but raises the question as to what effect the combination of different levels of data accuracy has on GIS outputs. It is however recognized that combining mixed data sources in such a way is often necessary, but a problem arises because GIS packages fail to offer any means of keeping track of the effects of uncertainty and error propagation throughout a sequence of operations (Carver, 1991).

The spatial data quality information is very important to aid the user to make good decisions. Openshaw (1990) quotes "It is critically important that data specific error and uncertainty details should be carried forward and stored with the data. Without this information it will be impossible in subsequent years to utilize emergent technology for handling error propagation, an aspect which only really becomes important when data from many different sources, with varying accuracy and uncertainty characteristics, are integrated."

In the analog era, widespread concern for accuracy was almost a concern for the accuracy of the planimetric position of the feature being mapped relative to its position on the earth. In 1934, a standard set of specifications for mapping city areas at 1:2,400 scale was adopted by the American Society of Civil Engineers. These specifications recommended that all elevations obtained from the maps be accurate to within one-half the contour interval, and that the error in horizontal position of well-defined map points be within a tolerance of 0.01 inch. The standards were revised in 1943 to provide the present tolerance of 1/50 inch for scales of 1:20,000 and smaller, and 1/30 inch for scales larger than 1:20,000. The revision of June 17, 1947, which is still in effect, is called "United States National Map Accuracy Standards." It specifies that maps, which comply with the specifications, should carry the statement "This map complies with the National Map Accuracy Standards". The standards provide for only one

class of standard-accuracy maps. The standards also make it clear that each mapping agency is responsible for determining which of its maps should be designed to meet the standards and for labeling those that do.

Today's technology allows the easy creation of many more data sets and visualization through GIS. Errors are introduced at every step in the process of generating and using geographic information, from data collection of the source data to the interpretation of the results of the completed analysis (Aronoff, 1989). Common sources of errors encountered in using a GIS will be explained further in the chapter-2.

The basic problem is classifying or combining the errors introduced in each step. Goodchild (1989) classifies the errors into two categories; "cartographic error", or error in positional features such as points, lines, and "thematic error", or error in the values of the thematic attribute. Chirsman (1987) refers to these two errors as "positional" and "attribute" error, respectively, while Bedard (1987) calls them "locational" and "descriptive" error. Some authors have also distinguished between "categorical " or "qualitative" error for nominal and ordinal data, and "numeric" or "quantitative" error for interval and ratio data (Bedard,1987).

Next problem is error detection and method of assessing accuracy levels in spatial data. A number of researchers, have done work on the derivation of data quality; Blakemore 1984, Burrough (1986,1991), Drummond (1989,1991), Chrisman 1990, Lodwick 1989 to name a few. Despite the amount of research that has been carried out Burrough (1986; pp103) states that "It is remarkable that there have been so few studies on the whole problem of residual variation and how errors arise or are created and propagated in geographical information processing, and what the effects of these errors might be on the results of the studies made". Though, this statement is more than fifteen years old, it still seems a valid one.

The need for data quality information has been recognized by a number of people. Besides individual researchers, sub-committees in many data standard committees have been formed to research and made recommendation on this important issue. Examples are: the National Joint Utilities Group (NJUG) joined with the Ordnance Survey of Great Britain (OSGB) in devising a Quality Control (QC) procedures for digitizing by third party contractors (Newby,1990); and in 1987 in the United States a National Committee on Digital Cartographic Data Standards, (NCDCDS), submitted a report entitled "A draft Proposed Standard for Digital Cartographic Data" (Moellering, 1987).

One of the four sections of the report submitted by the National Committee on Digital Cartographic Data Standards, (NCDCDS) was devoted to digital cartographic data quality. This represents the first comprehensive statements on spatial data quality in the electronic age. Quoting from the report's statement of spatial data quality;

> "The purpose of the Quality report is to provide detailed information for a user to evaluate the fitness of the data for a particular use. This style of standard can be characterized as 'truth in labeling', rather than fixing arbitrary numerical thresholds of quality. These specifications therefore provided no fixed levels of quality because such fixed levels are product dependent. In the places where testing is required, several options for different levels of testing are provided. In this environment the producer provides the quality information about the data and the user makes the decisions of whether to use the data for a specific applications." (Moellering, 1987, p.8).

The National Committee on Digital Cartographic Data Standards identifies five components of data quality; **Positional accuracy, Attribute accuracy, Lineage, Completeness, & Logical consistency.** The meaning of each one of these parameters will be explained further in the chapter-3. The International Cartographic Association (ICA) Commission on Spatial Data Quality, along with other groups, has accepted these five elements as important aspects of the spatial data quality. Since 1987, a modified version of the proposed standard for the exchange of spatial data created by the Moellering committee has been accepted by the National Institute of Standards and Technology as the Federal Information Processing Standard - 173 (NIST, 1994). Further these five elements have been adopted as the aspects of the spatial data quality by, cartographers in South Africa (Clark et al., 1992); in the United Kingdom (Walker, 1991); in Australia; and the Technical Committee 287 of the Center European des Normalization, 1992.

A review of the available literature shows that, there has been some work done on ways and means of tracking data quality and communicating it to the user. Some examples are: procedures to assess positional error (Carpary,1992; Carver,1992; and Drummand,J, 1992); procedures to assess attribute error (Greenland,1985; F. Goodchild, 1995); procedures for the analysis of error propagation in GIS overlay and modeling (Bevington and Robinson 1992, Heuvelink and Burrough,1993).

1.1 Research Objectives:

The objectives of this research are as follows:

- to identify means by which data quality can be determined;
- to identify and/or devise methods by which data quality can be represented;
- to identify cartographic and/or non-cartographic means by which data quality can be communicated in a GIS environment.

1.2 Digital Spatial Datasets Used:

The digital spatial datasets used for this research were:

- Topographic data (1:10000) of an area in Kegalle District.
- Topographic data (1:50000) of the same area.
  (see annex-1 for the general location of this area)

1.3 Research Methodology:

The research was carried out in the form of four tasks. Each of these tasks as well as the methodology used to accomplish it is given below.

Task 1: To review (identify) the various methods available to determine the data quality for a digital dataset.

Methodology:

- Discuss the errors in spatial data base,
- Discuss the techniques used in determination of the data quality;
- Establish whether or not these techniques can be used to determine data quality; and
- Choose the most appropriate data quality parameters; taking into consideration the data sets that are being used.

Task 2: For a digital dataset, which was chosen, develop quality assurance procedures.

Methodology:

- Identify any existing positional quality statements and verify these statements. If these statements do not exist, devise and use a method for determining positional quality;

- Identify any existing attribute quality statements and verify these statements. If these statements do not exist, devise and use a method for determining attribute quality;
- Identify any existing logical consistency statements and verify these statements. If these statements do not exist, devise and use a method for determining logical consistency;
- Identify any existing completeness statements and verify these statements. If these statements do not exist, devise and use a method for determining completeness;
- Identify any existing lineage statements and verify these statements. If these statements do not exist, devise and use a method for determining lineage;

Task 3: To identify the various non-cartographic / cartographic methods available to communicate data quality.

Methodology:

- Identify the techniques used in communication of data quality;
- Establish whether or not these techniques can be used to communicate data quality; and
- Choose the most appropriate techniques for use; taking into consideration the data sets that are being used.

Task 4: To implement in a GIS environment; means of data quality determination and representation; and, the communication of data quality to the user.

Methodology:

- Become familiar with the Arc/Info software package, learning the techniques used to manipulate tables, to perform digitizing, and devise ways to implement means of data quality determination and representation; and, the communication of data quality to the user.

1.4 Thesis structure:

This thesis begins (in this chapter) by giving the reason for the choice of the research topic. Different types of errors that may occur in spatial databases are then discussed. A theoretical framework for the data quality is established. The particular research is