



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

74349



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

UOM Verified Signature

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 thank 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.P.J. 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



University of Moratuwa, Sri Lanka.
Electronic Theses & Dissertations
www.lib.mrt.ac.lk

Table of Contents

Abstract	i
Acknowledgements	ii
Table of Contents	iv
List of Annexes	vii
List of Figures	viii
List of Tables	ix
CHAPTER 1 Introduction	1
1.1 Research objectives	4
1.2 Digital Spatial Datasets used	4
1.3 Research methodology	4
1.4 Thesis structure	5
CHAPTER 2 Errors in spatial database creation	7
2.1 Source of Errors	7
2.2 Details of Source of Errors	8
2.2.1 Obvious source of Errors	8
2.2.2 Errors resulting from natural variations / original measurements	10
2.2.3 Errors arising through processing	11
2.2.3.1 Numerical errors in the computer	11
2.2.3.2 Faults arising through topological analyses	12
2.2.3.2.1 Problems and errors arising from overlay boundaries	12
2.3 Problem in combining all Errors	18
CHAPTER 3 Data Quality Evaluation in a GIS	19
3.1 Components of Data Quality	19
3.2 Positional Accuracy	20
3.2.1 The methods of determining Positional Accuracy	21
3.3 The proposed methodology to determine the Positional Accuracy	26
3.3.1 Identify Control Data	27
3.3.2 Determination of the Sample for Comparison	27
3.3.3 Derivation of statistic parameters	28
3.3.4 Checking for existence of Gross Errors (Blunders)	29
3.3.5 Hypothesis testing on computed statistics	29
3.3.6 Test for the existence of significance bias error	30

3.3.7	Test for the precision	31
3.3.8	Computation of covariance to evaluates the tendency of correlation	32
3.3.9	Statistic parameter - Root Mean Square Error (RMSE)	33
3.4	Positional Accuracy assessment for linear and areal data	33
3.4.1	Epsilon band approach	33
3.4.2	Distance buffering method to determine the accuracy	35
3.4.3	"Rule of thumb" approach to determine the accuracy	37
3.5	Attribute Accuracy	38
3.5.1	The methods of determining attribute accuracy	38
3.5.2	Attribute Accuracy Assessment - Descriptive Label	39
3.5.3	Attribute Accuracy Assessment - Nominal attributes	40
3.5.4	Attribute Accuracy Assessment - Interval attributes	46
3.6	Completeness	46
3.6.1	Data completeness and Model completeness	48
3.6.2	Testing of completeness	48
3.6.3	Source of incompleteness	49
3.6.4	Example	50
3.7	Logical consistency	50
3.8	Lineage	53
3.8.1	Purpose of ' Lineage' Information	53
CHAPTER 4	Determination of the Data Quality of the Datasets	57
4.1	Production of the data set	57
4.2	Field Data Collection	57
4.3	Determination of Positional Accuracy	58
4.3.1	Positional accuracy assessment procedures- Point Entity	58
4.3.2	Positional accuracy assessment procedures - linear data	60
4.3.2.1	Rule of Thump Approach - line_entity	60
4.3.2.2	Error (Epsilon) Band Approach – line entity	61
4.3.2.3	Positional Accuracy Assessment – direct measurement	62
4.3.2.4	Positional Accuracy Assessment – distance buffering	63
4.3.3	Positional accuracy assessment procedures - areal data	64
4.4	Determination of Attribute Accuracy	65
4.4.1	Accuracy of Interval Attributes	65
4.4.2	Accuracy of Descriptive label	66
4.4.3	Attribute accuracy assessment for nominal data	67
4.4.3.1	The procedures for attribute accuracy assessment	67

4.5	Lineage	68
4.6	Logical consistency	68
4.7	Completeness	68
CHAPTER 5	Methods of representing the Data Quality	71
5.1	Methods of representing positional accuracy	71
5.2	Methods of representing attribute accuracy	74
5.3	Completeness	75
5.4	Logical consistency	75
5.5	Lineage	75
5.6	Cartographic methods for data and information quality representation	75
CHAPTER 6	Implementing Data Quality	80
6.1	The proposed approach	80
6.2	Choosing the GIS	80
6.3	Acquiring datasets	81
6.4	Determine and storage Data Quality	81
6.5	Data Quality Presentation	81
CHAPTER 7	Conclusions and Recommendations	82
7.1	Conclusions	82
7.2	Recommendations for future	84
Annexes		I
References		XLIV

List of Annexes

Annex 1	General location of case study area	I
Annex 2	Flow diagram	II
Annex 3.1	Assessment of Completeness - Sheet No. 53/23	VIII
Annex 3.2	Assessment of Completeness - Sheet No. 53/24	X
Annex 4.1	Assessment of Positional Accuracy -point entity –Sheet 53/23	XII
Annex 4.2	Assessment of Positional Accuracy -point entity –Sheet 53/24	XIV
Annex 4.3	Assessment of Positional Accuracy -point entity –Sheet 53/23&24	XVI
Annex 4.4	Assessment of Positional Accuracy -point entity –Sheet 53	XVIII
Annex 4.5	Assessment of Positional Accuracy -line entity	XX
Annex 4.6	Assessment of Positional Accuracy -polygon entity	XXV
Annex 5.1	Assessment of Attribute Accuracy –interval data –Sheet 53/24	XXVI
Annex 5.2	Assessment of Attribute Accuracy –interval data –Sheet 53/23	XXVIII
Annex 5.3	Assessment of Attribute Accuracy –interval data –Sheet 53/23&24	XXX
Annex 5.4	Assessment of Attribute Accuracy –descriptive label	XXXII
Annex 5.5	Assessment of Attribute Accuracy –nominal data	XXXIII
Annex 6.1	Positional Accuracy Report	XXXIV
Annex 6.2	Attribute Accuracy Report	XXXVIII
Annex 6.3	Completeness Report	XL
Annex 6.4	Logical consistency Report	XLI
Annex 6.5	Lineage Report	XLII

List of Figures

2.1:	Central Grid cell coding	13
2.2a:	Digitization	15
2.2b:	Error zones with epsilon tolerance	16
2.3:	Spurious Polygons	17
2.4:	A Classification of error in Spatial Database	18
3.1:	Determination of position of the telephone pole	23
3.2:	Comparison of an edit plot to source map	26
3.3:	Error band area according to caspary	34
3.4:	Distance buffer around entities	36
3.5:	Line defines by four Coordinate pairs	37
5.1	Error Ellipses used to show positional accuracy	73
5.2	Use of curves to show positional accuracy of a line	73
5.3	Use of bar charts to show attribute accuracy	74
5.4	Simple reliability diagram	75
5.5	Geometric reliability diagram	76
5.6	Lineage presentation using quality overlay	76
5.7	Quality Overlay in the form of isolines	77
5.8	Using the Epsilon band concepts to show error in elevation	78
5.9	Presenting the quality of individual feature	79
5.10	Crop suitability Information.	79
6.1	The Proposed Approach	80

List of Tables

3.1	Results from the Buffer/ Clip Prone	36
3.2	Classification Error Matrix	40
3.3	Classification Error Matrix	43
3.4	Normalized Matrix	44
3.5	Calculation of the Kappa Coefficient	45
3.6	Comparison of the results	46
3.7	Feature Completions	48
4.1	Positional Accuracy Summary data - point entity	58
4.2	Accuracy of the line Segments	60
4.2a	Accuracy of roads	61
4.3	Accuracy of line Segments (Epsilon Band)	61
4.3a	Accuracy of roads (Epsilon Band)	62
4.4	Direct Comparison with field measurements	62
4.5	Distance buffering method	63
4.6	Error (Epsilon) Band Approach - Areal data	64
4.7	Accuracy of Interval Attributes	66
4.8	Class accuracy test results	67
4.9	Completeness of topographic data	69
4.10	Completeness of land use	69
5.1	Position accuracy -meta information for all individual objects	72