THE UNIVERSITY OF CALGARY

MODELING AND MANAGING UNCERTAINTY IN OBJECT-BASED GEOSPATIAL INFORMATION SYSTEMS

by

ALI ASGHAR ALESHEIKH

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF GEOMATICS ENGINEERING

CALGARY, ALBERTA

MAY, 1998

ABSTRACT

This thesis investigates uncertainty in object-based Geospatial Information Systems (GIS). This issue is of vital importance to users since uncertainty information aids users in determining the fitness of the data to their particular application.

The theories that are used to deal with uncertainty - information theory, probability theory, evidential theory, fuzzy sets theory and geostatistics - are briefly reviewed. Each of these theories can be used to analyze a specific aspect of uncertainty. A general theory to handle all aspects of geospatial uncertainty has yet to be developed.

The techniques for modeling uncertainty in GIS are classified into four categories: analytical, simulation, experimental and error descriptors. Analysis of the categories has shown that each method has associated advantages and disadvantages and that the method selection is directly related to functional approximation errors, computer efficiency, and economic constraints.

Current models of geospatial uncertainty are analyzed. It is found that they are oversimplified. Therefore, more complete uncertainty models that describe the positional variations of spatial objects are proposed and evaluated.

A data reduction strategy for representing curvilinear boundaries is proposed. This recursive spline approximation has shown that a 70% data reduction can be achieved in boundary delineation for the cases studied in this thesis. However, data reduction depends on the assumed approximation error.

Uncertainty management problems are viewed as closely related elements. The components of the framework are uncertainty identification, uncertainty modeling, uncertainty communication, uncertainty reduction, and uncertainty absorption. A framework is proposed to deal systematically with uncertainty management issues in object-based GISs.

Finally, a software package for uncertainty modeling and visualization has been developed which facilitates the implementation of the proposed uncertainty management techniques.

Further research may be directed towards studying the uncertainty implications due to scale changes, integrating temporal and topological uncertainty in the management strategies, searching for the most effective methods for uncertainty communication, and analyzing the impact of spatial database uncertainty upon decision making and risk analysis.

ACKNOWLEDGEMENTS

I would like to thank several people who assisted me during the research and writing of this thesis. My appreciation and thanks must go to my supervisor, Professor J.A.R. Blais for his continuous guidance, discussions, suggestions and also for his constant availability throughout the course of my Ph.D. program. I also wish to thank my ex-supervisor, Dr. R. Li, who directed me at the first stages of the research. The other members of my supervisory committee, Professor M.A. Chapman and Professor K.P. Schwarz are also thanked for providing valuable insight, reference material and discussions in GIS and its uncertainties. I also wish to express my deep gratitude to Dr. Ballantyne, for his consistent assistance and encouragement throughout the last period of my study.

I gratefully acknowledge the financial support provided by the Ministry of Culture and Higher Education (MCHE) of Iran for studying at the Ph.D. level. Support from the Department of Geomatics Engineering and the Natural Science and Engineering Research Council of Canada (NSERC) for accomplishing the last stages of the research are also appreciated. Without their financial help, this research would not be realized.

I wish to thank the following graduate students; Hamid Ebadi, Bruce Chaplin, Derek Lichti, Christine Pelletier, and Ahmed Amer, for valuable discussions during my program.

Finally, grateful thanks must be extended to my wife, Gita, my son, Hessam, and my daughter, Mahsa, who, like those of most other doctoral candidates, have had to put up

with considerable inconveniences during the last five years. Without their love and support, this final product would not have been possible.

DEDICATION

To my wife, Gita, who left parents and friends to live this adventure with me. For her continuous support and encouragement during my Ph.D. program.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF SYMBOLS

The list of symbols is not exhaustive but it contains the symbols that are used more frequently in the text.

A	The first design matrix
B	The second design matrix
Bel	Belief function
C_L	The variance covariance matrix of observations
C_X	The variance covariance matrix of unknowns
H[p]	The amount of information
i, j, k	Indices
L	The vector of observation
m	The number of equations
n	The number of observations
$P_i = P(i)$	The probability of observing i
Pl	Plausibility function
P(h | e)	The conditional probability that h is true given observation e
u	The number of unknown
X	The vector of unknowns; a set of hypotheses; also the universe of discourse
W	The misclosure vector, also the weight matrix
α	The slope of the line; a coefficient
β	The intercept of a line; a coefficient
δ	The difference between the updated and initial values of unknown
ε	The width of the fuzzy band; also the vector of random errors

ρ	The correlation coefficient
σ	The standard deviation
μ	The grade of membership; also the statistical mean
Θ	The uncommitted support

CHAPTER 1

INTRODUCTION

1.1 Overview

This thesis examines the management of uncertainty in object-based Geospatial Information Systems (GIS). In recent years, this issue has become of increasing importance to GIS users, as well as decision makers. Rigorous modeling, efficient algorithms, and proper communication of uncertainty to GIS users are important issues in uncertainty management and, as such, they constitute important parts of this research. However, it is argued that they represent only part of the solution, and that such work should be contained in an overall strategy that can be effectively applied. By properly modeling and communicating uncertainty, the quality of derived database products may then be assessed, so as to determine whether the products are satisfactory for the task for which they are to be applied.

This opening chapter explores the scope of the uncertainty problem, its significance, and its consequences if left unresolved. First, the importance of modeling and managing uncertainty in GIS is identified. Next, the problems that are the focus of this research study are highlighted. Finally, the research study is introduced in terms of method used, scope, objectives, and relation to the current body of knowledge in this field. The chapter concludes with an explanation of the structure and content of the thesis, giving readers a clear guide to the dissertation.

1.2 The Significance of Uncertainty Management

Geospatial Information Systems are used in almost every discipline which handles spatial data. A GIS can be conveniently defined as a software package for the storage, analysis, and presentation of geospatial data [Worboys, 1995]. GISs have thus quickly become an indispensable tool for the efficient management of spatial data and are now routinely being employed by academics, government agencies and private industry [Maguire et al. 1991; Goodchild et al. 1993; Chrisman, 1997]. The incomplete treatment of uncertainty within GIS products, however, may limit their usefulness.

Over the last fifteen years, the issue of uncertainty and GIS has gained increasing prominence at national and international conferences, resulting in much discussion of the problems but few solutions [Hunter, 1996]. Many users are now becoming anxious at the apparent inability to quantify the accuracy of system products, while others still question whether there is any real cause for concern. It is now a well recognized fact that while accuracy might not be a 'money-making' subject [Kitchen, 1989], it could well become a 'money-losing' one for those who ignore the warnings [Epstein, 1991]. The following explains the importance of uncertainty management in GIS.

1.2.1 The revolution of the digital era

The progression from hardcopy maps to digital data started in earnest several years ago [Alesheikh, 1993]. This revolution made uncertainty assume greater importance than it did with previous cartographic products [Goodchild and Gopal, 1989]. This is due, in part, to the larger variety of spatial manipulations which may now be performed with ease

to create a host of displays and tailored images. The volume of spatial analysis that GISs are capable of accomplishing raises a false feeling of security -"computers do not make mistakes"- among GIS users.

It is not just cartographers who make maps now. The creation and management of spatial data bases has passed from the domain of 'the expert few' into the hands of a far larger portion of the community, such as land administrators, emergency service managers, environmental scientists, business analysts, bankers, retailers, market analysts and many others [Hunter, 1993]. However, it is now recognized that only the cartographers possessed fair knowledge about inherent uncertainty, and took appropriate precautions in the processes involved, such as the amounts of generalization that were allowed for particular maps [Hunter and Goodchild, 1995].

Traditional paper maps often contained useful forms of accuracy indicators, such as accuracy standards for horizontal and vertical positional errors [Bureau of the Budget, USA, 1947]. These indicators represented an attempt by cartographers to convey product limitations to users. Unfortunately, in the digital age, this information is missing from many GIS outputs, and new users of GIS are often unaware of the defects that can lie in the misuse of their data and associated technology [Hunter et al. 1994].

While GIS's capability to easily change the scale of geospatial data is categorized as one of its strengths [Goodchild, 1989], the false feeling of scale-free, and hence error-free, data should adequately be addressed [Quattrochi and Goodchild, 1997].

1.2.2 Protecting GIS growth

As geospatial information is increasingly relied upon for decision support [Alexander and Waters, 1996; Karimi and Hwang, 1996], the lack of accuracy and reliability estimates have the potential to harm both the integrity of agencies and the public's confidence in them [Davis and Simonette, 1991]. This is particularly so in cases where administrative decisions are subject to judicial review and, where the use of GIS may be open to question before the courts [Epstein, 1991].

For example, geospatial data accuracy has become one of the key factors in challenging government orders to clean up contaminated land in the USA. The Environmental Protection Agency in that country has allocated $1 billion to fund court battles with affected land owners seeking to overturn decisions against them, which were made in part by using GIS [Hunter, 1993]. Clearly, when the problem has such significance, issues such as accuracy quickly come to the forefront.

1.2.3 Fitness for use

The growth in consumer awareness in recent years has fostered a culture of skepticism. Court cases are already occurring where users of geographic data are claiming compensation for damages arising from use of expensive datasets which were allegedly of a quality that was inappropriate for their needs.

Data purchasers simply should not accept this situation, and similarly, data producers should not leave themselves so open to litigation by disaffected customers [Hunter et al., 1994]. Users would not think of buying GIS software or hardware without a guarantee or

a set of manuals so why should geospatial data (which usually costs many times more than the technology) be the exception to the rule?

1.2.4 Spatial data transfer standards

One of the key reasons for establishing GISs in the first place was to enable institutional sharing of data between agencies [Guptill, 1991]. Yet the literature contains many examples where integration of data sets became difficult when their respective accuracies were undefined [Star et al. 1991]. For example, in discussing parcel-based mapping programs, Raymond [1989] cites the attitude common amongst some users that 'rubber sheeting' can solve any problem, and that data from any source and scale can be easily integrated through such methods. In his experience those same users are often disappointed when data sets from different sources are overlaid and they discover "…. *telephone poles displaced out into a road, roads overlapping wetlands or other water bodies, or infrastructure appearing disconnected between adjacent maps*" [Raymond, 1989, p. 82].

On the other hand, if products are developed with a knowledge of the errors associated with a set of data and operations, then there can be cost savings in avoiding over-engineering. Jackson and Woodsford [1991] propose that the key to the reduction of the burden of data capture costs on a project is data sharing. The most important aspect of data sharing is validation to ensure the data are of a quality acceptable to the needs of a wide community of users.

Most spatial data transfer standards around the world have followed the United States

lead, and have set a mandatory requirement for data quality statements to accompany data sets upon transfer [DCDSTF, 1988]. Marketing requirements will ultimately force data producers to provide comprehensive data quality statements.

Information describing data is often called metadata or data about data. Metadata provide the necessary information on availability, fitness for use, access requirements and transfer protocols. Because of the diverse nature and multiplicity of data formats and standards, a need for some type of metadata standards harmonization is required. Throughout the world, there are efforts to develop a standard that can be used for the description of digital georeferenced data sets. Because of the importance of data uncertainty, all the metadata standards accommodate a section for the information about data uncertainty [Plunkett and McKenna, 1996].

1.2.5 Scientific advancements

Researchers in GIS should be concerned with quantifying the accuracy of geospatial data and derived products simply as part of the scientific advancement of the discipline [UCGIS, 1994]. If geographic information science is to become a recognized field of scientific endeavor, then it is imperative that its researchers be able to describe how close the information is to the truth it represents.

Goodchild [1992a] noted that the spatial information research community, in general, is now handling intellectual and scientific questions which go well beyond the limited technical capabilities of current technology in geospatial information systems. He suggested that the study of geographical data will become a science in its own right,

containing its own set of scientific questions, studied by researchers who are motivated by intellectual curiosity. Achieving these aims is a critical component of any program of scientific inquiry and researchers cannot ignore the responsibilities.

1.2.6 Information completeness

To represent fully an item of spatial information, the object, attribute, relationship, and uncertainty must all be included. The object identifier specifies the spatial object type, such as a point, line, or polygon. The spatial and other attributes are used to describe the characteristics of a spatial object. Spatial attributes position the object, and other attributes specify its value, color, size and other features [Blais, 1996]. Relationships refer to the object's relation to other objects, and the uncertainty indicates the accuracy of the item of information whether it be an attribute, or the relationship of the spatial object [Zhou, 1995a]. Therefore, the spatial information is incomplete without a specification of uncertainty values [Chrisman, 1991].

1.3 The Problem of Uncertainty Management

Uncertainty management problems reviewed by this research study are discussed here,

1.3.1 Local uncertainty measures of vector data

Although research on uncertainty has made some progress [Goodchild and Gopal, 1989], there are still many unsolved problems. Currently, most uncertainty representations and uncertainty propagation methods for spatial analysis are based on the raster data structure. Goodchild and Min-Hua [1989] discuss the difficulty of modeling errors in spatial databases, and suggest that error models based on fields are fundamentally easier to

construct than models based on objects. Dunn et al. [1990] indicate that vector data uncertainty management is much more complex, and that it is unclear whether general expressions for accuracy may be found.

Furthermore, current error models in GIS are inherited from hard-copy maps and are of a global nature. For example, map accuracy standards are designed as a global measure of uncertainty and fail to answer such questions as "what is the positional accuracy of a particular point?" Other examples of global measures of uncertainty are the digitizing errors in GIS, registration errors in remote sensing, and Digital Terrain Model (DTM) errors. The usually quoted error measures apply uniformly to an entire region, and do not take into account local variability in these measures [Edwards and Lowell, 1996].

1.3.2 Uncertainty modeling and representation

Modeling and representation of uncertainty are the most important issues in uncertainty management. Currently used uncertainty models lack analytical support (e.g. epsilon band model), and oversimplify the uncertainty behavior of GIS objects (e.g. error band model). If the true uncertainty of spatial data is not explicitly represented, data from different input sources with different levels of innate accuracy can be mixed, integrated, and manipulated, which totally disguises the likely reality of the situations. Since most GIS operations are not transparent, the users are not aware of all the limitations of the data, as well as the influence on them of spatial operations.

1.3.3 Data reduction

Boundaries of geospatial objects are frequently approximated by straight-line segments. Consequently, the uncertainty models applied are specific to straight-line segments [Dutton, 1992]. Regardless of the jagged nature of the boundary and its uncertainty, approximating a curvilinear boundary by small straight-line segments imposes a heavy load on databases. The problem is compounded if the uncertainty information of each small segment must also be saved.

1.3.4 Uncertainty handling strategies

Even when geospatial data uncertainty is adequately modeled, and properly communicated to GIS users, the problem of "how to deal with uncertainty?" is not fully answered. It is argued that the modeling and communication of uncertainty must be embodied in an overall strategy that can be effectively and operationally applied [Chapman et al. 1997]. Such strategy should track the uncertainty from its sources, and continue to the propagation of uncertainty, communication of uncertainty, and finally the decision that might be made in the presence of a given uncertainty.

1.3.5 Proper software packages

Current GIS software operates under the assumption that data are free from uncertainty [Lam, 1992]. The preparation of prototype software packages that can handle uncertainty is important [Hunter, 1996], because they pave the way for quantifying the errors, and because they may be used to educate untrained users and to make them aware of the probable uncertainty in GIS products.

1.4 The Scope and Objectives of the Thesis

There are many different aspects of the problem of handling error in GIS [Chrisman 1991]. Therefore, any research that aims at obtaining concrete solutions should necessarily start with selecting a particular segment within the whole research area of error management in GIS. This study restricts itself to the modeling, communicating and management of uncertainty of geospatial primitives: points, lines, and polygons in a 2D object-based GIS.

Since this thesis deals only with the problem of positional accuracy, it ignores the remaining four error components of attribute accuracy, lineage, logical consistency and completeness [DCDSTF, 1988]. This does not mean that these are not significant issues, but they appear to be of secondary practical importance to many applications in Geomatics.

The uncertainty information of geometric primitives depends on the scale and resolution of the original maps or images that the primitives are derived from. For instance, small polygons may not appear in small scale maps. Moreover, the issue of resolution is broader than the spatial component and can encompass the spectral and temporal domains. This research implicitly assumes that the proper scales are used in presenting the geospatial data.

The research ceases with the communication of uncertainty to users, which means that it does not continue into the area of quality assessment, that is, testing for 'fitness for use'. It is really up to the users to determine what is required for their needs. However,

once uncertainty is communicated to a particular user community, the thorough report may assist in deciding what level of uncertainty they are willing to accept. Obviously, there is a need for greater research in determining how uncertainty in GIS transmutes to uncertainty in decision making, and how such tools may be embodied in Spatial Decision Support Systems (SDSS). But these issues cannot be dealt with until the basic topic of managing spatial database uncertainty has been first addressed.

The main objectives of this research are:

i) To model rigorously the spatial distribution of positional uncertainty of linear and areal objects in a GIS, and respect the effects of all the error parameters.

ii) To reduce the volumes of data required for representing boundary lines and their uncertainties.

iii) To propose and test a strategy for managing uncertainty in a GIS.

iv) To develop a prototype software package that facilitates the management of uncertainty in an object-based GIS.

Once the uncertainty of the geometric primitives is properly modeled, the metadata file can be updated. The completed metadata file facilitates the process of decision-making.

1.5 Structure of the Dissertation

This thesis is composed of seven chapters. Following this introduction, in Chapter 2, a

taxonomy of errors in GIS is developed, which helps to synthesize current knowledge and clearly distinguishes between sources of errors, forms of errors and the resultant error which may reside in a product. It also analyzes the major theories on which most of today's uncertainty analysis methods are based. The limitations of each theory are discussed and the theories used in this research are indicated.

In Chapter 3, a review of the methods and models generally used in error analysis are presented and analyzed. The applicable fields of these models and methods are examined. Based on these analyses, the models that should be further developed for handling positional uncertainty are identified.

Chapter 4 describes the proposed general uncertainty models of line and polygon objects in GIS. In identifying such models, major error factors are considered. The derivation of the polygon uncertainty model is also presented here. Implementation of the polygon uncertainty model in solving the point-in-polygon problem is explained. A technical comparison between the proposed models and the currently used ones is carried out and, then, the generality and completeness of the proposed models are discussed.

Chapter 5 deals with curvilinear features and their uncertainty. Ways to reduce the volume of data required for presenting curvilinear boundary lines are examined. Algebraic polynomials are used to represent the boundary segments, and the Akaike Information Criterion (AIC) is introduced to determine the statistically best-fit function to a given set of observations. A recursive spline approximation is then proposed to approximate boundary lines. The proposed method reduces the data volume and

preserves the smoothness of the boundary. Uncertainty indicators of the curvilinear boundaries are proposed and examined.

Chapter 6 introduces the proposed uncertainty management strategy for GIS as a Spatial Decision Support System (SDSS). The dissertation ends with conclusions and recommendations in Chapter 7.

1.6 Chapter Summary

In this chapter, the importance of management of uncertainty in GIS was identified. The problems associated with managing uncertainty in an object-based GIS were recognized. It is recommended that dealing with uncertainty in GIS should accompany any spatial analysis. The scope, objectives and limitations of this study are presented. The chapter ends with a brief description of the organization of the thesis.

CHAPTER 2

IDENTIFICATION OF UNCERTAINTY AND RELEVANT THEORIES IN GIS

2.1 Overview

This chapter deals with the basic problem of identifying sources of uncertainty and classifying them. A three-level taxonomy of uncertainties is presented that establishes a connection between uncertainty sources, uncertainty forms and final product uncertainty. The objectives of the taxonomy are to: (i) clarify the relationship between the sources of uncertainty and the forms that uncertainty may take (cause and effect), (ii) identify the potential sources of uncertainty that may exist in a database (whether the uncertainty was there when the database acquired from producers, or whether it was caused by some subsequent actions), and (iii) help focus research on specific forms of uncertainty. Identifying sources of uncertainty is the first step in modeling and managing, as such, it composes a part of this chapter.

Many theories have been developed for dealing with uncertainty in spatial databases. However, a comprehensive theoretical framework that can handle all the existing uncertainty forms in GIS has not been developed, yet. Existing theories such as spatial statistics, fuzzy set theory, probability theory and mathematical theory of evidence can be adapted and applied to certain forms of inexactness in databases. Every theory is based on specific assumptions and as such can only be effectively applied to those uncertainties that conform to the *a priori* assumptions. In this chapter, the theories that are most frequently used in uncertainty handling in GIS are introduced, and the uncertainty

measures in each theory are specified. Based on the analysis of these theories, and the uncertainties that this research is concerned with, the theories that are applied in the course of this research are selected.

2.2 Classification of Uncertainty in GIS

An error taxonomy in uncertainty handling in GIS is essential, since it clarifies the framework for greater focus on the problems at hand. In fact, it algorithmically partitions the problem into elements that are manageable for modeling.

Although the terminology varies in the literature on this subject [Veregin, 1989a], it is generally agreed that there are several categories of errors which may contribute to the overall accuracy of products derived from GIS. What is usually unclear are the relationships among these errors and there is often confusion between their sources and the forms they may take [Veregin, 1989b]. Classification would help to clarify this confusion.

While taxonomic study is a science in itself, it need not always be complex to be effective [Obermeyer, 1989]. Calkins and Obermeyer [1991] have elaborated on the application of taxonomies to describe the use and value of spatial databases, and make it clear that though the classification process is not a simple task, it is nevertheless an achievable one.

Efforts to classify spatial database uncertainty have been made by many authors, such as Dangermond [1983], Hudson [1988], Goodchild [1992b], and Zhou [1996] to name but a few. Other authors have also tried to understand database errors through their

classification. For instance, Burrough [1986] organizes errors into three groups, (i) obvious sources of errors (such as the age of the data and the extent of area coverage), (ii) errors resulting from the original measurement or through natural variation, and (iii) errors resulting from computer processing (such as rounding). Error classification can also be based on the GIS functionality [Alai, 1993]. The most widely used error taxonomy is embedded in the U.S. Spatial Data Transfer Standard (SDTS) [NIST, 1992]. Although SDTS places emphasis on the data quality elements, it does not classify the sources of error nor does it clarify the cause and effect relationships. According to the standard, a quality report should accompany all spatial data and must cover the following aspects:

- Lineage, describing the source material, how data were derived, information of transformations and dates of updates;
- Positional accuracy, providing quality measures (for example, of control surveys) based on established geodetic standards and procedures;
- Attribute accuracy, providing numerical estimates of expected variation of non-spatial characteristics, in a manner similar to positional accuracy measures;
- Logical consistency, describing the integrity of relationships of data; and
- Completeness, describing the exhaustive extent of spatial and taxonomic properties, the consistency with which features have been assigned, the selection criteria, and standards employed.

Given the variations among respected researchers in organizing uncertainties, the question of uncertainty classification in this research is answered by the following three-level taxonomy (see Figure 2.1).

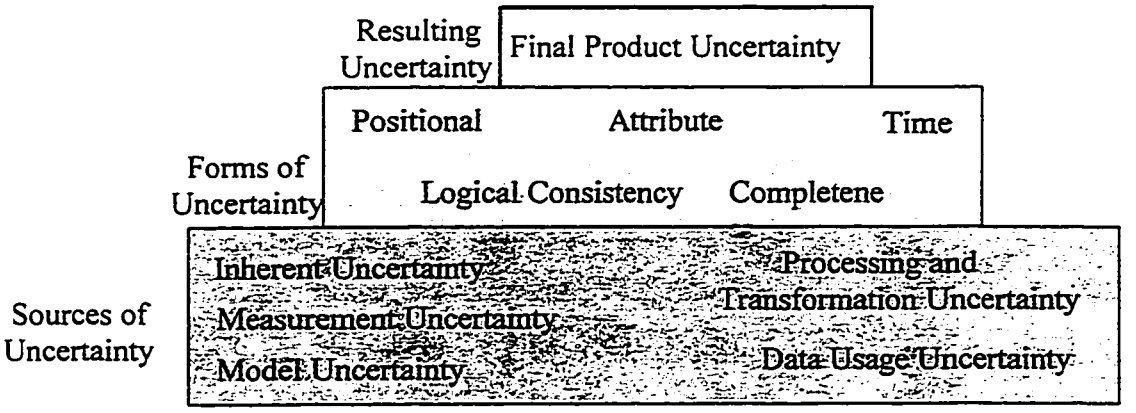


Figure 2.1 A taxonomy of uncertainty in GIS

First, it is clear that several operations cause uncertainty in GIS. However, these are not the forms that uncertainty takes - they are only the contributing factors. Uncertainty manifests itself in several ways. Aronoff [1989] divides the data quality elements into (i) micro level components of positional accuracy, attribute accuracy, resolution and logical consistency, and (ii) macro level components of completeness, time and lineage. He argues that the former apply to individual data elements, while the latter apply to the data set as a whole. In other words, the consideration of micro level components takes place prior to studying the macro level components, when determining whether a data set is suitable for the task to be applied.

While lineage constitutes a critical component of the data quality element, it is not a form of uncertainty - it is merely a historical report on what has happened to the dataset - and as such, it does not form part of the taxonomy, though it remains extremely valuable.

2.3 Major Sources of Uncertainty

Uncertainty may arise from the very first step of conceptualizing the real world (conceptual data models) through the last operations of the decision making process. In this research, uncertainty will be tracked until the decision to be made using uncertainty management strategy (Chapter 6). Starting at the bottom of the classification, the various sources of error are now described.

2.3.1 The inherent uncertainty of phenomena being mapped

Unlike geographic information data structures, the real world is not always distinct and clear, but changes gradually from one object to another. Topographical features are usually surveyed to a very high degree of accuracy, that is appropriate for the specific application purpose, and tested against well defined objects such as roads, houses and land parcel boundaries. In contrast, the position of soil or forest boundaries often reflect the judgment of the interpreter about where a dividing line should be placed. A boundary line represents a change from one coverage to another. Locating a line involves discrimination of the adjacent features. The border between some features might not be a line at all. It might be a fuzzy zone of transition in spatial phenomena. Carefully drawn soil or forest boundaries are elegant misrepresentations of changes that are often gradual and fuzzy [Burrough, 1989; Goodchild et al. 1994a].

Seddon [1971, pp. 20-21] elaborated on this issue, and states that: "Any attempt ... to demarcate vegetation boundaries must be a crude simplification which ignores the transitional changes determined by micro climate and topography within a border zone. It is excusable for practical reasons in small scale maps, but should not be allowed to

disguise the real nature of the changes represented." Thus, the current model of natural phenomena introduces uncertainty into GIS. More accurate modeling of the phenomena will require modification of the models.

2.3.2 Measurement uncertainties

Another major uncertainty in GIS stems from the nature of the measurements of spatial phenomena. All observations are of limited accuracy. A survey done for the purpose of producing a database is prone to various kinds of measurement errors, the two common ones being limitations of measuring equipment, ranging from leveling to photogrammetry to GPS, and human errors [Lichti, 1998].

The scale at which the measurements are made leads to another potential source of measurement uncertainty. Scale determines the smallest area that can be drawn and recognized on the paper maps, and influences the level of detail included on thematic maps [Goodchild and Quattrochi, 1997]. The issue of scale is not limited to the spatial domain and can be extended to the temporal as well as the spectral domain. A coarse image taken from a satellite platform may not exhibit the small features at all. Searching in color-blind images for features that exhibit the prevalent red radiance may raise measurement uncertainty too [Larouche, 1995]. The frequency of updating databases can also introduce uncertainty.

The assignment of a class of vegetation to an area depends on the degree of generalization made, which is often influenced by the scale of the map. A larger scale map reveals more detail than the smaller scale map, and shows fewer dominant species.

A digital database may at first glance appear to be scale independent (through zooming functions). However, if the data originated from a map, then the scale of the source determines the size of the minimum mapping unit and the material included or omitted. Scale change capability of GIS is one of its major benefits over hardcopy maps. However, it may well be misleading too, as noted in the description of the process leading to the establishment of National Center for Geographic Information Analysis (NCGIA), " Two GIS capabilities that excite enthusiasm among potential users are the ability to change map scales and the ability to overlay maps at random. ... both capabilities may mislead decision makers who are untutored in the ways errors compound when map scales are changed or when maps are merged" [Abler, 1987 p. 305].

2.3.3 Model uncertainty

The third major source of uncertainty arises due to the models that are used to communicate the measurements. A model refers to an abstraction of the real world. Since spatial objects are inherently complex, preserving their entire complexity is impossible, the objects are then simplified for analysis and inference purposes. Model uncertainty is implied by this simplification process. Simplification operations may take place at the very beginning of the conceptualization stage (conceptual model), before any physical measurement is accomplished, for example, in assuming a homogeneous distribution of attributes over space. Model error may also happen at the operational level, for example, in spatial linear interpolation. When a concept is formed, the uncertainty has already been introduced into the model [Zhou, 1995a].

GIS literature is frequently focused on analyzing the effects of model errors in empirical processes [Heuvelink, 1993]. It is because the model is calibrated for specific conditions which are not respected in practice. However, it should be emphasized that the simplification of the model is not restricted to the empirical models [Chapman et al. 1997].

2.3.4 Processing and transformation uncertainty

Raw spatial data frequently undergo some processing to answer GIS queries. Uncertainty resulting from the processing of data refers to the secondary uncertainty caused during computer manipulation of data, following the data measurement. These can be categorized in two classes: numerical errors in the operations, and aggregation errors due to computer processes.

Computers have a limited precision to store and manipulate spatial data. If the precision of the computer is not sufficient for the required mathematical manipulations, particularly multiplying or subtracting two large numbers, erroneous results may be derived. Burrough [1986] elaborated on this issue and discussed the consequences of rounding and truncation errors using examples.

Uncertainty of the input data seldom resembles the uncertainty in the GIS output, due to the uncertainty contribution of the processes themselves. GIS users commonly think of overlay operations as the main source of uncertainty in processing, but the GIS operations are composed of thousands of data manipulations, and each process has the potential to introduce uncertainty. The operation could range from raster-to-vector conversion, data

adjustment, editing and generalization, data input, surface modeling, display and analysis. Chapter 3 of this thesis analyzes methods of computing the uncertainty propagated during GIS processing. Table 2.1 presents typical data processing functions which potentially introduce errors.

Table 2.1 Potential sources of uncertainty in data processing (after [Hunter 1993]).

Coordinate Adjustment
- Rubber Sheeting - Transformation
- Edge Matching - Mosaicking
- Projection Changes - Datum Conversion
- Scaling - Rotating

Feature Editing
- Line Snapping
- Extension of lines to intersection
- Elimination of Spurious Polygon

Attribute Input and Editing
- Data Base Modeling
- Attribute Entry
- Attribute Modification
- Attribute Extraction
- Numeric Calculation

Boolean Operations
- Overlay
- Polygon on Polygon
- Point on Polygon
- Line on Line

Generalization
- Line Simplification
- Spline - Curve Fitting
- Addition - Deletion of Vertices

Raster - Vector Conversion
- Pixels to Polygons
- Polygons to Pixels
- Thresholding
- Assigning Attribute to Pixels
- Post-Scanning Line Thinning

Data Input
- Split/extract/append/merge Files
- Topological construction
- Dissolving Polygons with same attributes

Surface Modeling
- Contour/Lattice/TIN Generation
- Cross Section - Profile Generation
- Line of Sight - Visibility Determination
- Slope - Aspect Determination

Display and Analysis
- Cluster Analysis
- Perimeter Length - Area Calculation
- Shortest Route
- Buffer Creation
- Adjacency - Contiguity
- Spatial Autocorrelation
- Areal I nterpolation
- Raster/Vector Integration
- Spatial Statistics

2.3.5 Data usage uncertainty

Data usage uncertainty, which has only recently received attention among researchers, is concerned with the manner in which spatial data are used [Beard, 1989]. This is probably because of the fact that while the traditional producers and users of spatial data have been aware of the limitations, inexperienced users are now being encountered, who are not necessarily familiar with the accuracy, techniques or processes involved in product creation. Proper creation of metadata is thus of vital importance in reducing data usage uncertainty.

Data usage uncertainty occurs when a user misapplies processes given the types and limitations of the data. Level of measurement, scale and spatial variation of data can all dictate which processes or operations are valid [Chrisman, 1997]. Data may be considered accurate and the algorithm mathematically correct, but in misapplying a process a user can generate incorrect results.

Examples of this can happen during Boolean operations of raster data, where it usually occurs that ordinal valued layers are added, generating invalid results, or else interval or ratio values that are not independent are added, which may overestimate results [Hopkins, 1977]. In these cases, data misuse is revealed as attribute uncertainty. However, in other examples, vector overlay is misapplied in a case where data of very different scales are overlaid, and misapplication of interpolation routines can occur when users attempt to derive dense contours from sparse data sets. In these instances, positional errors might result. Common uncertainty arising from data misuse is listed in Table 2.2.

Table 2.2 Potential sources of uncertainty due to inappropriate data usage

- Lack of knowledge about scale of data (spatial, spectral, temporal) and its uncertainty implications
- Lack of user training and education
- Differences in terminology
- Lack of adherence to map accuracy standard
- Convenience of digital data regardless of scale
- Inappropriate usage compared to initial purpose
- Map space limitation
- Easy accessibility of data
- Extent of coverage
- Cost considerations

2.4 Forms of uncertainty

Uncertainty in GIS databases may take one or a combination of several forms, as depicted in Figure 2.1. Positional data occur when one tries to answer questions of the "where?" type, for instance "where is Calgary?", while attribute data are concerned about "what?" question types, e.g. "what feature may be found at the specified location?". Information about time is supplied when questions of the "when?" type, such as "when the sea height reaches a particular elevation?", are raised. Positional, attribute, temporal uncertainty are those uncertainties that arise during answering the above questions precisely. Veregin [1989a] refers to the first two types of errors as "cartographic" and "thematic" uncertainties, while Bedard [1987] calls them "locational" and "descriptive" errors.

Logical consistency, and completeness tend to be more global and affect the database as a whole, while positional, attribute, and temporal errors are assigned to each data element. Logical consistency deals with topological integrity of the database. Typical concerns that might be raised about a database are, (i) Do lines intersect where intended?

(ii) Are there any dangling lines? (iii) Are there any overshoots or undershoots in the network? (iv) Do all of the polygons have attributes? Fortunately, topological tests are frequently done automatically, but users must be aware of the effects of inappropriate tolerances being set.

Completeness is related to data collection procedures and policies. It is concerned with aspects such as the selection of minimum mapping areas (in soil and land cover mapping), minimum widths or lengths (for roads and rivers in topographic maps).

2.5 Resulting Uncertainty

The separation of the final product uncertainties from the forms of uncertainties is done because of the manner in which they may occur in the product. For instance, a GIS product may be composed of several different forms of uncertainty (e.g. positional and attribute) within it but the relationship between them may not be apparent. It is constructive to decompose the final product errors into its components, and report them individually [Alesheikh, 1994]. The rationale for this approach is that different applications may have different sensitivities to different forms of errors. For example, the development of a computer-base land tax register may have 100% sensitivity to attribute errors and be robust to positional errors.

2.6 General Theories of Reasoning about Uncertainty

Several theories have been adapted to handle and reason about uncertainty in GIS. Each theory has basic assumptions, and can only deal with certain aspects of uncertainty in GIS [Zadeh, 1995]. The types of uncertainty that may occur in the database were reviewed

before. Analysis of the corresponding theories will aid in the selection of an appropriate theory for the uncertainty problems dealt with in this research.

2.6.1 Information theory

Information theory provides a mathematical definition of uncertainty. Information theory has its origin in the theory of communication, that studies the transmission of electronic signals from a source to its recipient. The measure of information proposed by Shannon and Weaver [1959] is based on the probability of selection of each of the decision alternatives facing an information source. The amount of information, H[p], is given by:

$$H[p] = - k \sum_{i=1}^{n} p_i \log p_i . \qquad (2.1)$$

where, p_i represents the probability of the signal occurring and k is an appropriate constant. If the probability of a signal occurring is zero, then $p_i \log p_i = 0$, and no information results. When the probability of the signal is one, then again, $p_i \log p_i = 0$, and no information results. These two extremes represent states of certainty. On the other hand, if probabilities of signals lie between zero and one, then some amount of uncertainty is present. This demonstrates the notion that information is the dispelling of uncertainty, and that if there is no uncertainty, there can be no information [Blais, 1988].

Although Shannon provides a measure of information (or uncertainty), he did not specify it. He was concerned about the technical aspect of information transmission, and did not include semantic aspects of information. Information theory and other information measures have been used in digital image processing and pattern recognition

[Blais and Boulianne, 1988] and appear to be valuable in studying complex systems. The information theory has been used to some degree in cartographic communication and also been applied to GIS. However, little research has been done with respect to the application of information theory in information systems such as GIS [Zhou, 1995b]. Information theory, as it is applied to digital images, may be able to provide measures of uncertainty that can be adopted in conjunction with GIS. For instance, the current attempt in the GIS industry to integrate vector and raster data structures may benefit from information theory. Blais [1991b] pointed out that information theory has a real potential in spatial information processing, and suggested that it can help in quantifying information in terms of ambiguity, fuzziness, and other types of uncertainties.

2.6.2 Probability theory

Probability theory can be applied to model uncertainty caused by random errors rather than imprecision or incompleteness. Consequently, it is appropriate for positional and attribute uncertainties where the quantitative or qualitative characteristics have a statistical basis. In this theory, uncertainty is modeled as the conditional probability that a hypothesis is true given some observations. Bayes' rule uses conditional probability to estimate or update the probability of a hypothesis by combining the prior probability of the event, and the likelihood of the evidence given that hypothesis is true. The probability of a hypothesis indicates a quantitative measure between 0 and 1 of the belief in that hypothesis. The following equations presents the Bayes' theorem.

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}, \tag{2.2}$$

where,

P(h) is the probability that hypothesis h is true (the *a priori* probability);

P(e | h) is the conditional probability of observing e given that hypothesis h is true (likelihood of the event);

P(e) is the probability of observing e

p(h | e) is the conditional probability that hypothesis h is true given observation e (*a posteriori* probability)

The above equation can be extended to handle multiple competing hypotheses and several pieces of evidence, as shown in the following equations. This causes serious issues, since it is required to know all possible combinations of all possible hypotheses, hence it is unworkable.

$$P(h_i \mid e_1, \ldots, e_m) = \frac{P(e_1,\ldots,e_m|h_i)P(h_i)}{\sum_{j=1}^{n} p(h_j)P(e_1,\ldots,e_m|h_j)} \tag{2.3}$$

As an example, in the classification of remote sensing data, h_i might indicate the hypothesis that a given pixel belongs to class "i". "e" can be the vector of responses for the pixel in different spectral bands. The uncertainty is then reflected by the conditional probability $P(h_i \mid e_1, \ldots, e_m)$, which indicates the degree to which one might assume the class to be correct given the vector of responses.

This model assumes that a predefined and uniform distribution of values is known, and that probabilities $P(h_i)$ can be assigned correctly, which is not true in many applications. The commitment of partial belief to hypothesis leaves the remaining belief to its negation or alternative hypothesis, which is counter intuitive, and thus it cannot distinguish between uncertainty and ignorance. The use of conditional probability, which assumes mutually exclusive and exhaustive hypotheses and conditional independence of evidence, is not always true [Cohen, 1985].

2.6.3 Mathematical theory of evidence

The evidential theory [Shafer, 1976] is a generalization of probability theory, and differs from Bayes' theorem in some aspects. In evidential theory, the probability measure that the hypothesis h_i is true is substituted by a measure of probability that the existing evidence supports the truth of that hypothesis. Evidential theory accepts degrees of belief in hypothesis or their subsets. In other words, a set of all propositions about the exclusive and exhaustive possibilities in the domain is defined as a frame of discernment, Θ. This frame of discernment is given a basic probability assignment with values between 0 and 1. Therefore, Θ has a maximum potential of a probability of 1. Contrary to Bayes' theorem, probability assignment for the subset of hypotheses need not sum to one. The choice of assigning values between one and zero to Θ, in this theory, represents what is not known about the situation, and this allows a degree of doubt or ignorance to exist. It also allows the modification of Θ when new evidence becomes available.

Evidential theory is based on an interval defined by a belief function and a plausibility function. A belief function measures the probability that the evidence

supports the hypothesis, and is a lower bound on the probability of the truth of that hypothesis. The lower bound of the interval is sometimes termed the support or necessity for that hypothesis. A plausibility function on the other hand measures the degree to which the evidence fails to refute the hypothesis and, therefore, it is the upper bound of the interval. The inferential interval [Bel(h_i), Pl(h_i)] presents the uncertainty due to the incompleteness of evidence for hypothesis h_i because of uncommitted support (Θ).

For example, let us assume that X is a set of possible hypotheses and the probability that each hypothesis is true is given. In mathematical terminology;

$$X = \{P[h_1], P[h_2], P[h_3], \ldots, P[h_n], \Theta\}, \tag{2.4}$$

where: $P[h_i]$ is the probability that hypothesis h_i is true

Θ is the uncommitted or distributed support.

By definition:

$$\Theta + \sum_{i=1}^{n} P[h_i] = 1 \tag{2.5}$$

The belief function is simply the probability of h_i $Bel[h_i] = P[h_i]$, and the plausibility function is given as ;

$$Pl[h_i] = 1 - \sum_{j=1}^{n} P[h_j] = Bel[h_i] + \Theta \qquad i \neq j. \tag{2.6}$$

When Θ is zero, then $Bel[h_i] = Pl[h_i] = P[h_i]$, that is equivalent to probability theory. Therefore, the probability theory is a special case of evidential theory.

Evidential theory may address types of uncertainty that concern incompleteness of the information and, as such, can be applied to both attribute and positional uncertainty. This theory is becoming popular in expert system design. In the field of GIS, where integration is the topic of the day, this inferencing technique may have merit. However, this theory relies on assumptions of independence of evidence that may not be justified in real applications. The theory cannot distinguish between different hypotheses having similar levels of support.

2.6.4 Fuzzy set theory

The theory of fuzzy sets is concerned with a subset A of the universe of discourse X, where the transition between full member and non member is gradual, rather than abrupt. The fuzzy set has no distinctive boundaries where the universe of discourse, X, covers a definite range of objects. Fuzzy classes are often encountered in the real world, and are applied frequently in those fields that are subjective to some degree.

Fuzzy set theory, introduced by Zadeh [1956], is a generalization of abstract set theory [Kaufmann, 1975]. In other words, the former always includes the latter as a specific case; definition, theorems, and proofs of fuzzy-set theory always hold for non-fuzzy sets. Intuitively, a fuzzy set is a class that allows the possibility of partial membership in it. A fuzzy set A in its universe of discourse X is a set of ordered pairs,

$$A = \{x, \mu_A(x) \mid x \in X\} \tag{2.7}$$

where, $\mu_A(x)$ is termed the grade of membership or the characteristic function for x in A.

This theory is well suited to handle uncertainty of an imprecise or vague type, such as the linguistic imprecision of natural language (e.g. "steep", "moderate" or "flat" slope). Fuzzy set theory is effectively applied in handling positional and thematic uncertainties, where the concept or measurement of either quantitative or qualitative characteristics involves an interpretation of natural language.

A number of fuzzy set operations such as union, intersection, and complement, are available and are suitable for many types of geographic operations in GIS. Lam [1992] has successfully applied fuzzy set theory to classify soil types and Leung [1988] used the theory in a study of the vagueness of climatic boundaries. He categorized geographic areas to its core (non-fuzzy zone), boundary (zone of transition) and edge (the outermost extremity) and therefore demonstrated that the fuzzy concepts can model the transition of one class to another.

2.6.5 Spatial statistics

Over the last decade or so, statistics has evolved from the collection of numbers that summarize a complex phenomenon to the science of uncertainty [Cressie, 1991]. Spatial data can be classified into three different categories:

- geostatistical data, that describe spatial processes indexed over continuous space;
- lattice data that deal with spatial processes indexed over lattices in space (the spatial analogue of time series), and
- point pattern data that describe spatial point processes.

The term geostatistics is now widely applied to a special branch of applied statistics, originally developed by Georges Matheron of the Centre de Morphologie Mathematique in Fontainebleau, France. Geostatistics was devised to solve problems encountered when conventional statistical theory is unable to estimate changes in ore grade within a mine. Because geostatistics was a general theory of statistical behavior, it is applicable to many circumstances in different areas of natural sciences.

In an attempt to generalize a model for geostatistical data, lattice data, and point pattern data, Cressie [1991] proposed his general spatial model. Heuvelink [1993] proposed a similar model to account for continuity or discontinuity of spatial variations. Shi [1994] used spatial statistics to model the uncertainty in classifying remotely sensed images. He argued that the classical non-spatial model is a special case of a spatial model. Spatial statistics is a more general theory than non-spatial statistics; it may be assumed as a spatial extension to non-spatial variables. Statistics for spatio-temporal data have to be even more general. The theory has been extended from the non-spatial field, and could be further extended to spatial-temporal models.

2.7 Chapter Summary

In this chapter, a taxonomy of uncertainty was presented. The objectives of the taxonomy are to understand the specific uncertainty that this study deals with, and to focus on the related theories that can be applied to handle the uncertainties. Sources of uncertainty have been classified into five categories: (i) the inherent uncertainty of the phenomena being mapped, (ii) measurement uncertainty (iii) model uncertainty (iv) processing and transformation uncertainty, and (v) data usage uncertainty.

In the GIS literature, several theories have been proposed to handle uncertainties. The theories evaluated in this chapter are five examples of the most frequently used theories in modeling uncertainties in the GIS field. Each has its own merit. Based on the specific problems that this study deals with, and the analysis of the existing theories, two perspectives have been selected in this research. They are probability theory, and (spatial) statistics. There is, however, a considerable ongoing debate about the theoretical strength of each of these, and the respective applications.

The modeling problem considered in this study is to devise uncertainty models for line and polygon objects in GIS. The nature of this problem is to deal with pseudo-random errors in the spatial domain. Consequently geostatistical theory will be used, and to quantify the quality of the results, probability theory will be applied.

CHAPTER 3

UNCERTAINTY MODELING AND VISUALIZATION IN GIS

3.1 Overview

The purpose of this chapter is to survey and analyze the methods by which the uncertainty of geospatial objects (points, lines, and polygons) can be modeled, and communicated to users. Based on the analysis, the study identifies the needs for further development of the existing uncertainty models, and uncertainty communication.

Since this study involves modeling, visualizing, and managing uncertainty in an object-based GIS, formal definitions of space partitioning data models are presented first. Next, methods for uncertainty modeling are classified and evaluated. The applications of the methods in computing the uncertainty of geospatial objects are examined and current models are evaluated. Approaches that can be employed to communicate uncertainty to GIS users are discussed.

It is argued that the methods of uncertainty modeling differ in cost, time, computational efficiency, and rigor. As such these factors should be considered in choosing one method. It is also elaborated that while the uncertainties of geospatial objects may entirely be presented in their respected covariance matrices, visualization tools readily help users to appreciate the uncertainty information in the matrices. Since most GIS users are familiar with maps; static visualization, graphic representation of uncertainty is adopted to represent errors in this study.

This chapter provides a basis for presenting general uncertainty models in the subsequent chapters. As mentioned in Chapter 1, and which will be elaborated upon in Chapter 6, modeling and reporting uncertainty constitute a major part in uncertainty management. Therefore, this chapter will pave the way for the discussion of uncertainty management to be presented there.

3.2 Partitions of Space

Any measurement of spatial data in GIS may involve three basic components: location, theme and time [Chrisman, 1997]. A specific item of spatial information is always related to a certain theme of specific time at a particular location. The critical development in any measurement is the recognition of the three distinct roles that location, theme and time may play. While the goal here is to obtain uncertainty measurement, measurement of one component can only be made with explicit constraints on the other components. The three roles that basic components can take are: fixed, controlled and measured. In order to carry out a measurement, one component is fixed, the second one is controlled leaving the third one to be measured. For instance, in the case of maps, it is time which is normally fixed leaving space and attribute for control and measurement. The map of the 1988 landuse/landcover of Banff National Park contained the spatial information at a fixed time: 1988. The control is the location: Banff National Park, and the measured attribute is the landuse/landcover. In this categorization, the fixed component does not allow any variation. The variation of a controlled component is restricted and predicted [Sinton, 1978].

Based on the controlled component, the space on which spatial phenomena exist can be partitioned or classified into location-based known as raster data, and feature-based which mostly exhibits itself in vector data models.

The controlled component in location-based partition is the location. Attributes are measured at the specific locations. Tessellation of space is an example of a location-based partition. The most frequently used tessellation for space partition is uniform cells: square grids [Goodchild, 1989]. Most remote sensing data sources fall into this general category using nearly regular cells as control. These remote sensing sources include satellite platforms and photogrammetric equipment used to create digital elevation arrays.

The other way to partition the space is to control the theme. That is, in feature-based models, attribute is used to measure location. The geometry of spatial objects is delineated by uniform attribute distribution. For instance the geometry of the spruce coverage in the 1988 map is determined by the areal range of the spruce trees. Since the geometric data in this model is derived from its attribute, the model is called an object-based partition [Zhou, 1995a].

Space partitioning data models impose structure and pattern on geospatial data. They are closely related to the two broad and opposing classes of models of geographic information, namely field-based and object-based models [Worboys, 1995].

The class of field based models treats such information as collections of spatial distributions, where each distribution is formalized as a mathematical function mapping

from a spatial framework (for example, a regular grid placed upon Earth's surface) to an attribute domain. Patterns of topographic altitudes, rainfall and temperature manifest this view [Laurini and Thompson, 1992].

The class of object-based models treats the information space as populated by discrete, identifiable entities, that exhaust the geographic space. These two model types result in two GIS implementation methodologies, namely vector and raster [Masry and Lee, 1988].

Geospatial primitives in a 2-dimensional object-based data model can be categorized in points (0-simplex), lines (1-simplex) and triangles (2-simplex) [Foley et al. 1990]. Points are considered the base for constructing higher order simplexes, since lines, triangles, and polygons can be assumed as an extension of the points into higher dimensions.

It is important to note that currently applied object-based partitions in GIS software does not comply with the gradual transition between different objects [Lam, 1992]. In this model the object has sharp boundaries and totally ignores the fuzzy boundaries of spatial objects [Burrough, 1989; Alesheikh and Li, 1994].

3.3 Methods of Uncertainty Modeling

An error model refers to a stochastic process capable of simulating the range of possibilities known to exist for spatial data. These possibilities may exist because measuring instruments are known to be of limited accuracy, or because vital information,

such as datum or map projection, is missing [Blais, 1998].

The methods by which geospatial data uncertainty can be modeled may be categorized into four classes: analytical, simulated, experimental, and uncertainty descriptors.

The basic geometric components in 2-D object-based GIS are points, lines and polygons. They are considered as different objects since their spatial properties and their uncertainty behavior are different [Zhou, 1995a]. Consequently, the uncertainty models that describe the behavior of the objects should be different. The following sections classify the methods of uncertainty modeling and the application of the methods to various geometrical primitives.

3.3.1 Analytical method of uncertainty modeling

Analytical methods refer to obtaining the stochastic characteristics of (functionally) dependent variables, given the characteristics of the independent variables and the functional relationships relating to the two sets of variables.

Let $L = [l_1, l_2, ..., l_n]$ be a set of random variables with a known n-dimensional probability density function $f(l_1, l_2, ..., l_n)$. Let $X = [x_1, x_2, ..., x_u]$ be another set of random variables that is related to L by $F = [f_1, f_2, .., f_m]$ such as:

$$F(X, L) = 0 \tag{3.1}$$

The task is to determine the stochastic properties of X from those of L, i.e., to determine the probability density function $g(x_1, x_2, ..., x_u)$ [Krakiwsky, 1992].

3.3.1.1. Propagation of distribution

In most local or global GIS operations, the general Equation 3.1 can be simplified to a set of continuous, differentiable functions F [Cressie, 1991] such that

$$X = F(L) \tag{3.2}$$

where X is the output, L is the input of GIS operation, and F is the function that relates the input and output data. Let us assume that there exist unique, continuous, inverse functions, such that

$$L = H(X) \tag{3.3}.$$

If $f(l_1, l_2, ..., l_n)$ is the joint density function of the random vector L, then by substitution

$$g(x_1, x_2, ..., x_u) = f[h_1(x_1, x_2, ..., x_u), h_2(x_1, x_2, ..., x_u), .. , h_n(x_1, x_2, ..., x_u)] \, |J_{XL}| \tag{3.4}$$

is the joint density function of the derived random variable $(x_1, x_2, ..., x_u)$, and the problem is solved [Mikhail, 1976; Shi, 1994]. Here $|J_{XL}|$ is the determinant of the Jacobian matrix of the inverse transformations.

$$l_i = H_i(X) \quad \text{and} \quad J_{XL} = \partial L / \partial X \qquad i = 1, 2, ..., n \tag{3.5}$$

The difficulty with propagation of distribution is that the inverse function should be identified first, and it must be unique, continuous and differentiable. However, in the case of linear mapping of normally distributed variables the result g will also be normally distributed variables. Therefore, propagation may be limited to linear or linearized

functions, which will be discussed next.

3.3.1.2. Covariance propagation

The distribution of the unknown parameters $g(x_1, x_2, ..., x_u)$ can be uniquely identified by its first, second, and joint moments if g's distribution follows a Gaussian function. This assumption is usually applied in surveying, photogrammetric and geodetic measurements [Mikhail, 1976]. The reason for this lies mainly in the central limit theorem, which roughly states that the averaging of a sufficiently large number of random variables yields a normal distribution, no matter what the distribution of an individual random variable is [Dougherty, 1990]. Since GIS data are functions of several errors introduced in different stages (see Chapter 2) it is reasonable to assume they are following normal functions. Another reason for invoking the normality assumption is that it makes the computation simple [Vanicek, 1992].

The covariance propagation method relies on approximating the relation F(X, L) by a truncated Taylor series. In the case of the first-order Taylor method, F is linearized by taking the tangent of F in initial values (X^0, L^0) as follows :

$$F(X,L) = F(X^0,L^0) + \frac{\partial F}{\partial X}|_{X^0,L^0}(X - X^0) + \frac{\partial F}{\partial L}|_{X^0,L^0}(L - L^0) + remainder = 0 \quad (3.6)$$

Assuming the magnitude of the reminder to be negligible, it can be written as:

$$W + A\delta + Br = 0 \quad (3.7)$$

where $W = F(X^0, L^0)$ is the misclosure vector

$A = \frac{\partial F}{\partial X}|_{X^0, L^0}$ is the first design matrix, and

$\delta = (X - X^0)$ are the corrections to the parameters

$B = \frac{\partial F}{\partial L}|_{X^0, L^0}$ is the second design matrix, and

$r = (L - L^0)$ are the corrections to the input data.

The remainder of Equation 3.6 contains the higher-order Taylor series terms of F, whose contribution to the result of F is comparatively small in the neighborhood of (X^0, L^0). By neglecting the higher-order terms, the output (X) and their covariance matrix C_x can be computed by:

$$X = X^0 - [A^T(BC_LB^T)^{-1}A + C_{X^0}{}^{-1}]^{-1}A^T(BC_LB^T)^{-1}W \quad (3.8)$$

$$C_X = [A^T(BC_LB^T)^{-1}A + C_{X^0}{}^{-1}]^{-1} \quad (3.9)$$

where C_L is the covariance matrix of the input data and C_{x0} is the *a priori* covariance matrix of the output map.

3.3.1.2.1 Point uncertainty model

The application of covariance propagation to estimate the uncertainty of points in Geomatics is common. It can range from the adjustment of terrestrial observations in surveying engineering to satellite observation adjustment in GPS.

Covariance propagation in a GIS application is frequently accomplished with some simplifications. Consider Equation 3.2. If the random variables $L(l_1, l_2, ..., l_n)$ are mutually

independent, the variances of the dependent variables X are:

$$(\sigma_{Xi})^2 = \sum_{j=1}^{n} (\partial X_i / \partial l_j)^2 (\sigma_{lj})^2 \qquad (3.10)$$

Further, if the variables $l_1, l_2, ..., l_n$ are also contribute equally to x, i.e. $\partial X / \partial l_j = 1$ then;

$$(\sigma_X)^2 = \sum_{j=1}^{n} (\sigma_{lj})^2 \qquad (3.11).$$

The above equation is usually used to approximate the point uncertainty in GIS. For example assuming that the point data are gathered by digitizing a map that was compiled photogrammetrically. If it is also assumed that the point error in GIS were independently introduced by control, photography, aerotriangulation, compilation, drafting, printing, and digitization, then its uncertainty in some form of standardized units may be approximated by:

$$\sigma^2_{GIS} = \sigma^2_{control} + \sigma^2_{photograph} + \sigma^2_{triangulation} + \sigma^2_{compilation} +$$

$$+ \sigma^2_{drafting} + \sigma^2_{printing} + \sigma^2_{digitization} \qquad (3.12)$$

Similar assumptions have been made by several researchers [Alai, 1993; Chrisman, 1982; Drummond, 1995] and organizations [Canada Map Office, 1977] in different applications. It is worth mentioning that Equation 3.11 is an approximate method for estimating the variance of a variable derived from more than one different and independent measurements. Compatible units of uncorrelated components are assumed in deriving the Equation 3.12.

3.3.1.2.2 Line uncertainty model

Unlike the point uncertainty model that is studied frequently in Geomatics, the line uncertainty model has been left untouched. The most frequently mentioned line uncertainty model in GIS that is derived from the application of covariance propagation is the 'error band' model [Caspary and Scheuring, 1992; Shi, 1994; Zhou, 1996].

Under the assumption that a straight-line segment is built by connecting the endpoints, and the coordinate variances of the endpoints are independent and equal $\sigma^2 = \sigma^2_X = \sigma^2_Y$, the covariance propagation law yields the positional RMS (Root Mean Square) errors of points P_n along the line P_1P_2:

$$X_n = X_1 + D_n(X_2 - X_1)/D \qquad \text{and} \qquad Y_n = Y_1 + D_n(Y_2 - Y_1)/D \tag{3.13}$$

where D_n is the distance between point n and P_1, and D is the Euclidean distance between P_1 and P_2, and $0 \le Dn \le D$.

$$\sigma_{Xn}^2 = \sigma_{Yn}^2 = (1 - 2Dn/D + 2Dn^2/D^2)\sigma^2 \tag{3.14}$$

The envelope of the point errors is then submitted as the line error (See Figure 3.1).

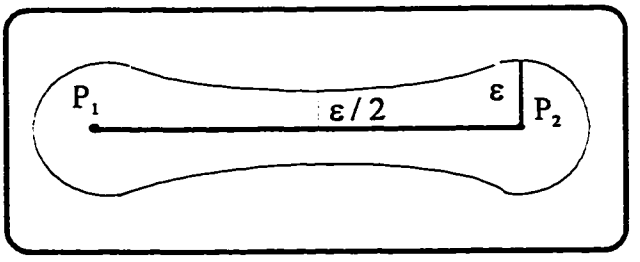


Figure 3.1 The concept of Error Band model

Under the above assumption, the standard deviation in the middle of the line, D_i = D/2, is $1/\sqrt{2}$ smaller than at the endpoints. The error band model is based on the circular error assumption at the endpoints of a line segment. Therefore, the error band model:

(i) neglects the correlation between endpoint coordinates,

(ii) assumes the magnitudes of uncertainty in perpendicular directions (X, Y) for each endpoint are similar, and

(iii) assumes model errors do not affect the errors of the intermediate points.

These restrictions may not be realized once the endpoint positions are determined using a similar method or when the linearity of the model is questionable [Chapman et al. 1997]. An error-free line segment connecting endpoints may also be doubtful, particularly when the line represents natural feature boundaries.

3.3.1.2.3. Polygon uncertainty model

Polygons are enclosed by straight line segments, and each line segment is recognized by its endpoints. If so, the uncertainty of the polygon is closely related to the statistical properties of points. An exhaustive description of the statistical behavior of a polygon is contained in the covariance matrix of the random vector formed by the coordinates of the vertices of the polygon. However, since GIS users are interested in a special aspect of accuracy, it is usual to select useful functions of the random vector and to use the corresponding variance to characterize the accuracy of the polygon. One widely used function is the area of the polygon.

The variance of the area is computed by applying the covariance propagation law. The area (A) is computed by [Davis et al., 1981]:

$$2A = \sum_{i=1}^{n}(Y_{i+1} - Y_{i-1})X_i \qquad \text{where} \qquad Y_0 = Y_n\,, \quad Y_{n+1} = Y_1 \qquad (3.15)$$

The variance of the area, $\sigma^2{}_A$, under the assumption of the independence of coordinates is then given by:

$$\sigma^2{}_A = 0.5\sum_{i=1}^{n}(X_i^2 + Y_i^2) - \sum_{i=1}^{n}(Y_{i-1}Y_{i+1} + X_{i-1}X_{i+1}) \quad \text{with} \quad \sigma = \sigma_X = \sigma_Y \quad (3.16)$$

Another function that might be used to represent the locational uncertainty of a polygon is the accuracy of the perimeter of the polygon [Caspary and Scheuring, 1992]. An alternative method is to compute the area that the boundary line uncertainty covers and divide it by the total area of the polygon. Figure 3.2 depicts the uncertainty of an areal object.

3.3.2 Monte Carlo simulation method

In principle the problems of uncertainty transmission in GIS operations can be handled by using covariance propagation [Burrough, 1986]. The lack of single continuous differentiable functions renders the use of explicit equations for error propagation impossible. Instead, it is simpler and more general to use a universal solution based on Monte Carlo simulation approach.

The Monte Carlo method [Openshaw, 1989] uses an entirely different approach to determine the uncertainty of geospatial objects. In this method, the results of Equation 3.2

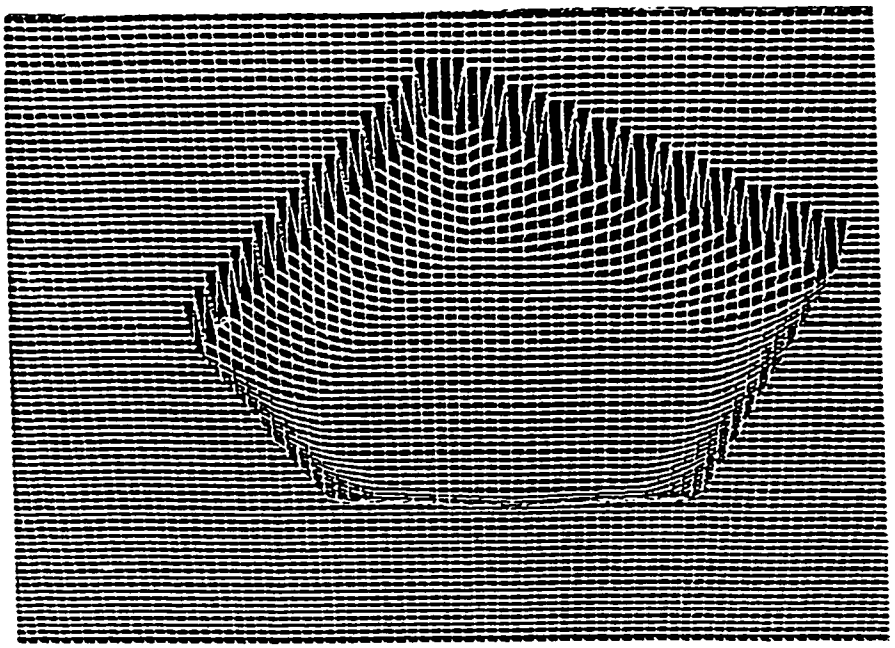


Figure 3.2 Positional uncertainty of an area object (adopted from Shi [1994])

are computed repeatedly, with input values $L = [l_1, l_2, \ldots, l_n]$ that are randomly sampled from their joint distribution. The outputs of the equation construct random samples of which parameters of its distribution, such as mean and variance, can be estimated. The basic algorithm as might be applied in GIS is as follows:

(i) Decide what levels and types of error characterize each data set as input to a GIS.

(ii) Replace the observed data by a set of (M) random variables drawn from appropriate probability distributions assumed to represent the uncertainty in the data inputs.

(iii) Apply a sequence of GIS operations to the step (ii) data.

(iv) For this set of realizations l_i, store the results x_i

(v) Compute summary statistics.

The Monte Carlo method is general and can be applied to spatial or attribute data. Heuvelink [1993] showed that the standard deviations of output mean (M_x) and output variance (S^2_x) are approximately inversely related to the square root of the number of Monte Carlo runs M. Therefore, the accuracy of the method can also be controlled. However, the method is computationally intensive.

3.3.2.1 Point data

The application of the Monte Carlo simulation method in computing the uncertainty of point data in GIS is gaining momentum as researchers realize its advantages [Keefer et al. 1988; Heuvelink 1993]. Because of its generality, most of the geomatics point densification methodologies, such as aerotriangulation, could also benefit from the approach, though little research has been carried out in randomization procedures in Geomatics [Blais, 1998].

3.3.2.2 Line data

The application of the simulation methods in representing line uncertainty can be realized by:

(i) randomly generating several endpoints from their corresponding density functions,

(ii) connecting the generated random points, and

(iii) computing the confidence region around the line segment.

By simplifying the coordinate error - the endpoints coordinates are assumed independent and their variances are equal - Dutton [1992] constructed an error model for line object based on simulation method called the "error band" model.

3.3.2.3 Polygon data

One way to determine the uncertainty of a polygon object using Monte Carlo method is to:

(i) construct one realization for each vertex based on the stochastic characteristics of the points,

(ii) connect the vertices – creating one realization of a polygon

(iii) repeat this procedures M times, and

(iv) approximate the spatial distribution of the polygon.

Dutton [1992] conducted such experiments to simulate the distribution of polygons.

3.3.3 Experimental method

The standard method of measuring accuracy of any object-based primitive is to compare the positional attribute of samples of primitives to their 'true' positions. For instance to carry out the method for point objects, first select the number and distribution of test points according to the sampling method employed. Then the coordinates of these points are compared to those that assume higher accuracy. From these two sets, the root mean square error can be obtained and presented as an uncertainty indicator for the point

objects;

$$RMS = \left(\sum_{i=1}^{n} (y_i - x_i)^2 / n \right)^{1/2} \quad (3.17)$$

where the y_i's are the coordinates of the points having higher accuracy and the x_i's are the coordinates of the points to be checked.

The method can be extended to line objects (comparing the length of the lines) and polygon objects (comparing the area of the polygons). Though rigorous, the method suffers from the cost and time involved in determining the values of the geospatial object in more accurate data sets.

3.3.4 Error descriptors

Error descriptors are not error models in the sense that they do not provide the statistical variation of geospatial objects. However, in the early quest for providing uncertainty information to users, descriptors were widely used to inform the users of possible uncertainty inherent in the data. The following section describes the two most widely error descriptors, namely the epsilon band model for describing the uncertainty of line objects, and point-in-polygon proposed as an indication of positional uncertainty for area objects.

3.3.4.1 The epsilon band

The basic concept of the epsilon band model is founded upon the principle that a cartographic line is surrounded on each side by an area of constant width, epsilon (ε),

similar in appearance to a buffer zone. The concept may be visualized as the effect of a ball with a radius ε that is rolling along the line as shown in Figure 3.3. Different interpretations exist about the shape of the band at the endpoints. While GIS software mostly represents the endpoints errors as circle (see Figure 3.3), Blakemore's [1984] model assumes a box surrounded each line segment (see Figure 3.4 and 3.5),.

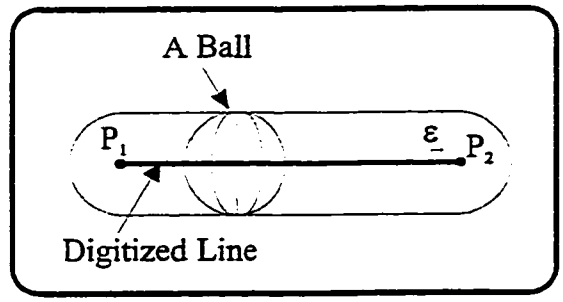


Figure 3.3 The concept of the Epsilon Band model

The model was designed to provide users with a measure of the error associated with digitizing cartographic lines. Developed and enhanced by researchers such as Perkal [1966], Blakemore [1984], and Chrisman [1989], many other authors have since applied the idea in a variety of ways.

Though many interpretations of the epsilon band exist, they can be categorized in two groups: deterministic and probabilistic. In the deterministic case, the true position of the line is considered to lie somewhere in the buffer zone. Deterministic interpretation of the epsilon band is questionable because:

(i) it provides no model of error distribution inside the band, and,

(ii) it proposes that the true line lies within the epsilon region.

In the probabilistic interpretation of the epsilon band, the width of the zone is assumed to be a function of different parameters such that their uncertainties accumulate to the final stage. For instance, Alai [1993] assumed that scale, digitization, slope and attributes of the polygons adjacent to the lines were the related variables, while Blakemore [1984] related the bandwidth to the digitizing error, round off error and generalization error.

Probabilistic interpretations of the epsilon band are apparently inconsistent with what analytical [Caspary and Scheuring, 1992; Shi, 1994] and simulation [Dutton, 1992] procedures determine.

However, in spite of the weaknesses, the epsilon band has the following advantages [Carver, 1991]:

(i) it involves little extra processing time,

(ii) it uses existing spatial operations for implementation (buffer zone operation),

(iii) different feature categories can have different epsilon values assigned to them as attributes,

(iv) the concept is easily understood and can be applied in many spatial operations, and

(v) the uncertainty can be readily integrated with the positional data and require a modest amount of storage.

3.3.4.2 The polygon uncertainty indicator

An extension of the epsilon band can be used to represent the uncertainty of a polygon object. The uncertainty of the area object is thus represented by the area covered by the epsilon band. Blakemore [1984] related the positional uncertainty of a polygon with respect to the positions of nearby points. He categorized five relationships (see Figure 3.4);

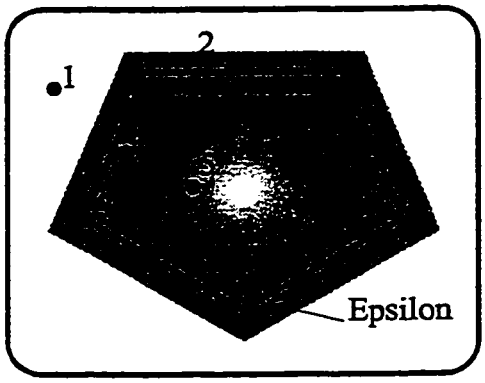


Figure 3.4 Uncertainty of an area object

(i) definitely in (point 5)

(ii) possibly in (point 4)

(iii) ambiguous (point 3)

(iv) possibly out (point 2), and

(v) definitely out (point 1)

Points in the interior (core) area are 'definitely in'; if the points lie within the inner half area of the epsilon band, they are 'possibly in'; if the points fall on the boundary line segment they are 'ambiguous'. Points are categorized as 'possibly out' if they locate in the outer half area of the epsilon band. If the points fall outside the polygon and epsilon band, they are 'definitely out'. The smaller the epsilon band area, fewer points lie in the 'possibly in' and 'possibly out' regions, and the uncertainty of the area object is lower.

The limitation of applying this error descriptor is that it cannot describe a continuous change in quantifying the location of the points with regards to a polygon. Therefore, a quantitative model which can describe the continuous change in the probability that a point belongs to an area object should be developed.

3.4 Evaluation of the Methods

The analytical method which is based on covariance propagation is modest in computational load. It is also attractive because it yields an analytical expression for the variances of the output errors, although it should be noted that the solution is approximate only. The means, variances and correlations of the input data explicitly appear in Equation 3.9, and these allow one to examine quickly how the output uncertainty changes under variations in the input error parameters.

The disadvantage of the method is that it is an approximate method only. When the function F is strongly non-linear, then the approximation error may become unacceptably large. Iteration may be required to reduce the error, however it will be at the cost of computational load. On the other hand, the approximation error is zero when F is linear.

Another disadvantage of the first-order Taylor covariance propagation method is that it requires that the function F be continuously differentiable. Covariance propagation may not respect the ease of computation if the function F is a complicated computational model that consists of many inputs. The method may not determine the real statistical properties of errors in GIS, since systematic errors may exist in the process.

The most important advantage of the Monte Carlo method is that it can provide the entire distribution of output data at an arbitrary level of accuracy. Other advantages are that the method is easily implemented and generally applicable. The method merely treats the function F as a black box, whose response to the perturbed inputs is studied from the resulting outputs.

The main disadvantage of the Monte Carlo method is that it is computationally intensive. Another disadvantage is that any sensitivity analysis requires the repeated execution of the entire process.

Experimental methods provide the best indication of accuracy. As such, they can be used to detect any possible biases in other methods. However, the results depend on the sampling procedures and the number of check points used. The most important disadvantages of this method is that it is costly, and time consuming. It is frequently used in the remote sensing discipline, where check points are needed for classification processes.

It is worthwhile mentioning that only those methods that have found their way into GIS are described. Approaches such as second-order Taylor [Heuvelink et al., 1989], and

Rosenblueth methods [Heuvelink, 1993] are excluded.

3.5 Communicating Uncertainty

As stated before, the uncertainty of geospatial objects can be entirely represented by their respective covariance matrices, which are derived from statistical processes. Statistical terms, particularly in text format, attempting to convey uncertainty to GIS users are often misunderstood by those users.

With advances in computer technologies and more specifically in GIS software and hardware, new techniques are becoming available for communicating uncertainty. These include visualization, animation, sound, and holography [Fisher, 1994]. Deciding on the most appropriate method of uncertainty communication is the subject of perception and involves neuro-psychological tests that are beyond the scope of this thesis. However, since most GIS users are familiar with maps (static visualization), the 2-D graphical representation is adopted in relaying the uncertainty messages to users. DeFranti et al. [1989] added to this point and claimed that the visual presentation of uncertainty measures has the ability to:

(i) organize abstract concepts into a meaningful display,

(ii) transform numerical values into understandable images, and

(iii) permit manipulation of geometry, color and motion.

Moreover, it speeds pattern recognition, eases the motion and facilitates change detection. Figure 3.5 demonstrates the difference between the visual approach to

presenting uncertainty and its statistical modeling counterpart. It is important to notice

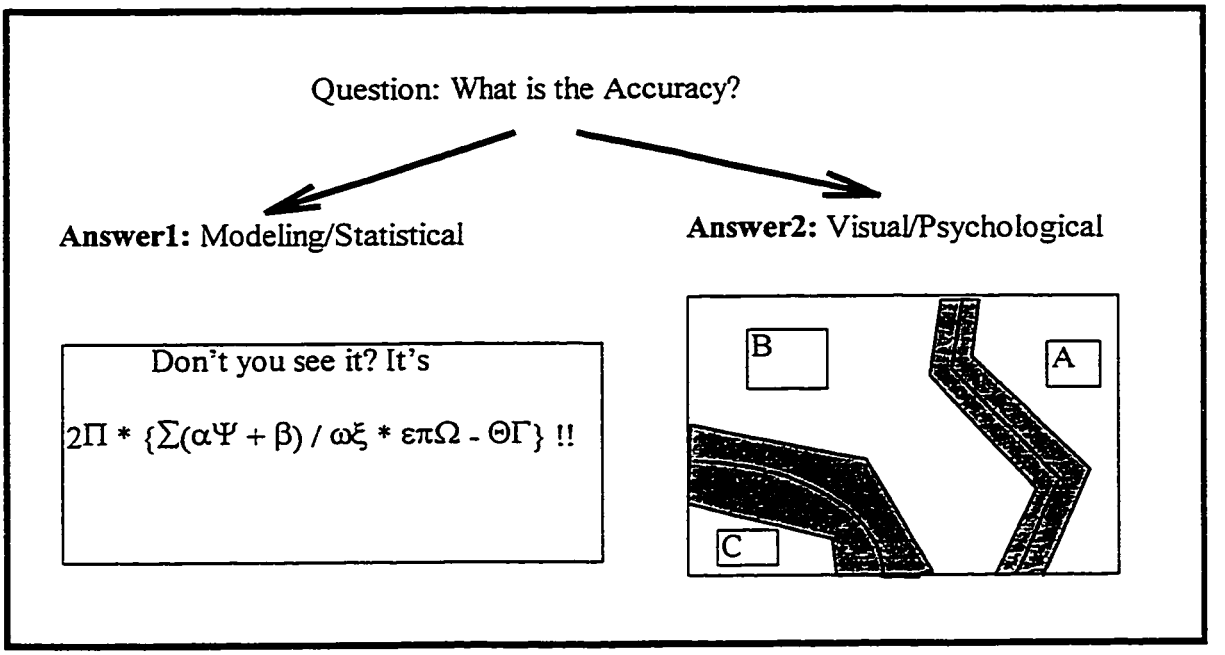


Figure 3.5 Alternative ways of answering similar question depending upon the user needs (adapted from Csillag [1991]).

that visualization changes the text format of the statistical terms to the graphical display. Both approaches are, thus, valid; it only depends on who the user is and how the information is required to be presented.

In the static mode, cartographic representations of uncertainty should be guided by application of Bertin's [1985] visual variable, including, size, shape, texture, color and orientation. With respect to data uncertainty, the visual variables seeming to offer best opportunities for display include color and texture [Goodchild et al., 1994b]. The visual metaphor of fog has been proposed [Beard et al., 1991] to alter viewers to uncertainty in data position. Fog may be created graphically by manipulating color or defocusing a display for visual impact.

Dynamic visualization, may provide several advantages, for instance when data and metadata displays are toggled at some speed, or glued together. Another advantage is that for positional accuracy, symbol movement overcomes the problem of viewers attaching too much weight to particular cartographic representations [Fisher, 1994].

3.5.1 Point uncertainty visualizations

If the coordinates of a point in (X, Y) plane are random variables having a bivariate normal distribution, their joint density function can be expressed as:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_x}{\sigma_x})^2-2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})+(\frac{y-\mu_y}{\sigma_y})^2]\}, \tag{3.18}$$

in which μ_x and σ_x are the mean and standard deviation, of X; μ_y and σ_y are the mean and standard deviation of Y and ρ is the normalized correlation coefficient of X and Y as defined by:

$$\rho_{x,y}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}\ . \tag{3.19}$$

This density function has a bell-shaped surface over the x, y coordinate plane, centered at $x=\mu_x$, $y=\mu_y$, as depicted in Figure 3.6.

A plane that is parallel to the x, y coordinate plane will cut the bivariate density surface in an ellipse. A statistical interpretation of this ellipse is that it is the locus of points having equal probability. To obtain the mathematical formula of the ellipse, it is sufficient to assign a constant value to f(x, y) = K in Equation 3.18. After simplification it reduces to:

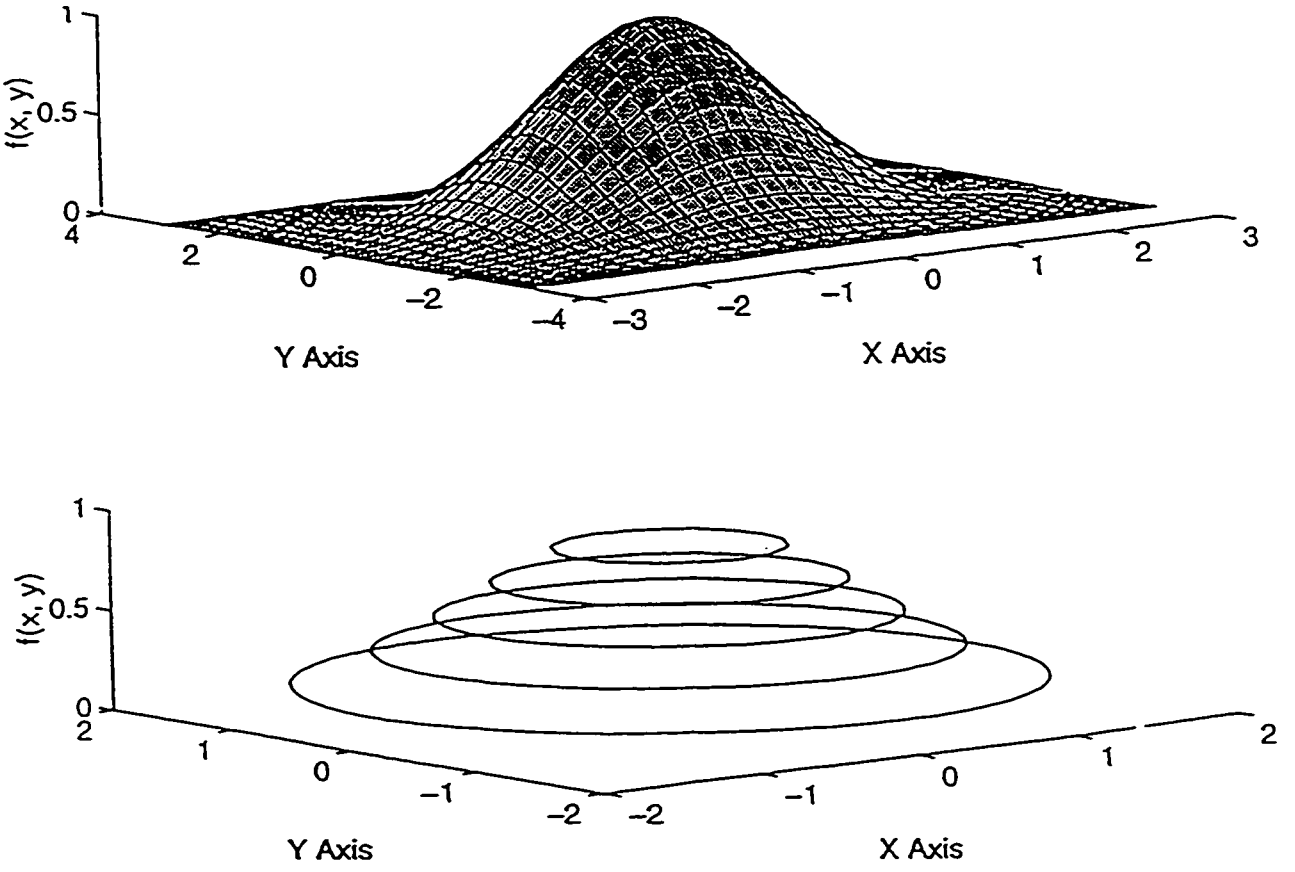


Figure 3.6 2D normal function together with its contour plots (error ellipses).

$$(\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2 = (1-\rho^2)c^2 \tag{3.20}$$

where c is a constant. The probability associated with an error ellipse $P[E \leq c^2]$, is represented by the volume under the bivariate normal density surface. $P[E \leq c^2]$ for various values of c can be determined from χ^2 distribution tables.

If there is a need to focus on the random error component only, as is the case in surveying, μ_x and μ_y can be set to zero to get a probability distribution that centers on the origin of the x, y coordinate plane. When $\mu_x = \mu_y = 0$, Equations 3.18 and 3.20 reduce to:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x}{\sigma_x})^2 - 2\rho(\frac{x}{\sigma_x})(\frac{y}{\sigma_y}) + (\frac{y}{\sigma_y})^2]\}, \tag{3.21}$$

and

$$(\frac{x}{\sigma_x})^2 - 2\rho(\frac{x}{\sigma_x})(\frac{y}{\sigma_y}) + (\frac{y}{\sigma_y})^2 = (1-\rho^2)c^2 \tag{3.22}$$

respectively.

Equation 3.22 represents a family of error ellipses centered on the origin of the coordinate system. When $c = 1$, equation 3.22 is the equation of standard error ellipse. The size, shape and orientation of the error ellipse are governed by the distribution parameters σ_x, σ_y and ρ. Several examples demonstrating the effects of different combination of distribution parameters are shown in Figure 3.7.

3.5.2 Line and polygon uncertainty visualization

Unlike point, line uncertainty visualization is quite immature, since the models representing its statistical behavior are immature. Moreover, to determine the line uncertainty model, the line segment is frequently assumed to be straight. The philosophy behind this is that even curvilinear functions can be approximated by straight line

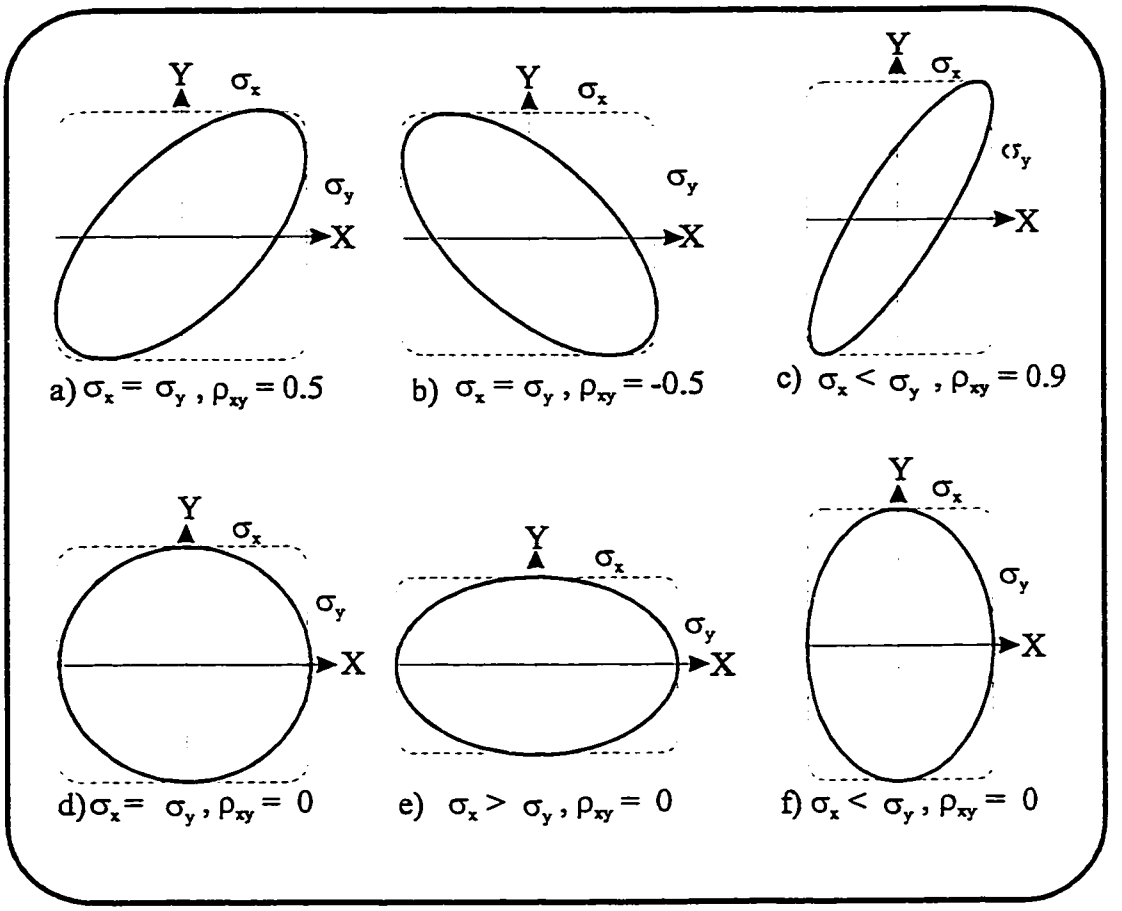


Figure 3.7 Different realization of error ellipses (adopted from Mikhail and Gracie [1981]).

segments. The only available models for line objects are the "epsilon band' model, and the 'error band' model. Their uncertainty visualizations are presented in Figures 3.1 and 3.3. Polygon uncertainty may be structured from the uncertainty of its boundary line segments as represented in Figure 3.8.

3.6 Practical Considerations

The techniques that are highlighted in this chapter can be applied to determine the uncertainty distribution of geospatial objects. The following considerations should be respected when using the methods.

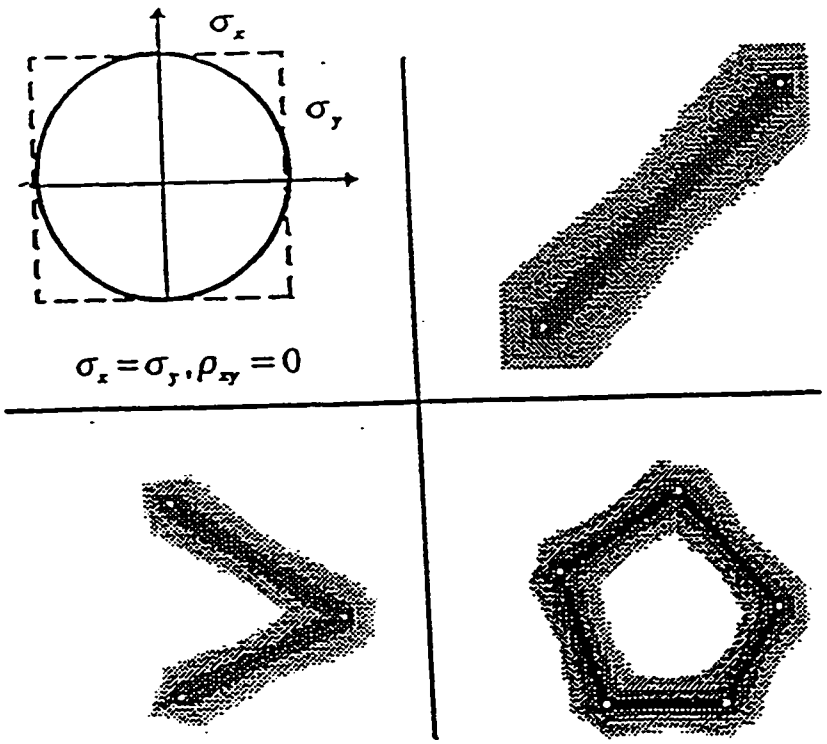


Figure 3.8. The uncertainty visualization of a polygon object (adopted from Shi [1994]).

3.6.1 Incorporation of model error

As mentioned in the previous chapter measurement error is only one source of uncertainty, and a GIS output may be affected by other error sources such as model error. Consequently, not only does the input error aggregate to the output of a GIS operation, but model error does as well. This means that the output of a GIS process may disagree with reality even if the input data are error free. This is due to the fact that the model is an approximation of reality. Although, the uncertainty modeling described so far neglects the contribution of model error, this may be a major source of error in many situations. Thus it should be included in error analysis.

Model error can be included by assigning errors to model coefficients or by adding a residual error term to the model function F. A residual uncertainty occurs because rarely will all the variability in the input data be accounted for by the function.

The effects of model error in empirical functions are more noticeable, because an empirical model has no universal accuracy. For example, an underwater flow function that is developed and calibrated for one area may perform poorly when it is applied to another area.

3.6.2 Scales of data

There are several ways in which geospatial data can be grouped, the most widely used being scales of measurement. Scale of data can be categorized into four classes: nominal, ordinal, interval, and ratio. Nominal data have sufficient information associated with them to classify them into categories, or classes; for example, the type of landuse can be classified as water, forest, or desert. Ordinal data contains sufficient information such that the data can be ranked in ascending or descending order. Interval data have the characteristics that distances between categories are defined as fixed equal size units and that they have no fixed zero value. For the elevation attribute, for example, only distances and not absolute values may be measured. Ratio data have in addition an absolute zero. A value of zero metre of elevation does not mean that it has no elevation. On the other hand, a value of zero mm of rainfall as ratio data indicates no rainfall.

Uncertainty models identified in this chapter are all based on interval/ratio data, since location attributes are of interval/ratio scale. If uncertainty of categorical data are of

concern, other strategies such as use of a confusion matrix should be adopted [Goodchild, 1995].

3.6.3 Storage efficiency

Any algorithm or model in GIS may be evaluated based on its efficiency. The efficiency of a model deteriorates with the volume of data it manipulates. As such when designing any uncertainty model two opposing requirements should be considered: the rigor of the model, and its efficiency. For instance epsilon band which requires less computation time than error band might be considered as a close approximation to the error band model, and be sufficient for some applications. However, GIS users should be able to access both general and efficient models to decide which one best fits to their requirements.

3.7 Chapter Summary

In this chapter, the extent of the research into the methods, modeling and visualization of uncertainty in spatial databases has been reviewed and analyzed. In Section 3.2 formal definitions of space partitioning data models were presented. Partitioning of space is important in uncertainty modeling since it decomposes the problem of complex reality into manageable data models.

Section 3.3 examines the three methods of modeling uncertainty in GIS databases: the analytical method, the Monte Carlo simulation procedure, the experimental approach, and error descriptors. It was emphasized that while analytical methods based on covariance propagation are modest in computer manipulation, they can be employed only when the functional relationships among independent and dependent variables are continuously

differentiable. Moreover, the linearization of the function around the point of linearization must be a valid approximation of the function. Simulation methods may be generalized for all GIS processes, however they are storage and time consuming procedures. The experimental approach is constrained by its cost and time, but can be employed to check the result of other methods, i.e. the concept of check points in photogrammetric processes. Applications of each method in geospatial primitives are also highlighted.

Section 3.5 has been devoted to the analysis of the present method of communicating uncertainty. Statistical terms (text format) currently predominate the uncertainty communication media. It is, however, argued that the visualization method provides a better way of uncertainty presentation.

This chapter identified the following deficiencies in the current line uncertainty models:

(i) they oversimplify the true statistical variation that a line segment may assume,

(ii) they neglect the uncertainty that may arise due to the imperfection of the mathematical model,

(iii) they model straight line segments, and consequently a lack of smoothness is apparent in a map presentation of natural boundaries.

(iv) they do not attempt to reduce the amount of data volume needed for uncertainty determination.

CHAPTER 4

PROPOSED POSITIONAL UNCERTAINTY MODELS OF GEOMETRICAL ENTITIES IN GIS

4.1 Overview

Previous chapters elaborated on the methods by which uncertainty may be modeled, and indicated the deficiencies of the currently used uncertainty models in GIS. It was also argued that a major constraint in the operations of a GIS is the lack of appropriate models of geospatial data uncertainty, and the means by which the uncertainty may be communicated to users. Rigorous uncertainty models and proper communications media are of vital importance in a GIS, because they do not only help GIS users assess the quality of the system products, but also assess if the decisions taken upon the uncertain products are in jeopardy. Indeed, as spatial information is being viewed as a marketable commodity the potential for litigation exists, and failure to deal with uncertainty problems rigorously may well endanger GIS technology. Without proper uncertainty modeling, management of uncertainty in GIS is difficult, if not impossible.

Accordingly, this chapter proposes general uncertainty models of geometrical primitives (particularly straight line and polygon object) in a GIS. The visualization aid is used to communicate the output of the general model to users. This chapter is then structured such that the general uncertainty model of geometric entities, together with their parameters (with particular attention to the observational and model errors), and hypothesis are defined in Section 4.2. Point uncertainty together with the methods to

determine the magnitude of the measurement error for point data are discussed in Section 4.3. Observational error of a straight line segment is determined using the Monte-Carlo simulation method, and described in Section 4.4. Observational errors of line segments can also be calculated analytically using covariance law. The designed strategy for the analytical method is highlighted in Section 4.5. In Section 4.6 model error is investigated and a method for determining the magnitude of model error along the line segment is proposed. The union of the boundary line uncertainties are used to represent the uncertainty of a polygon object in Section 4.7. A logical representation of uncertainty information in GIS databases is proposed in Section 4.8. Having modeled the uncertainty of polygon objects, the solution to the point-in-polygon problem is augmented by probability statements and given in Section 4.9.

4.2 The Proposed General Uncertainty Model of GIS Dataset

The first step in any rigorous uncertainty modeling is to identify the parameters that affect the model, and determine their magnitudes. Knowing all the contributing components in the model helps to determine the components with maximum effects. Once the more important contributing factors in the model are identified, further research on characterizing their behaviors can be accomplished.

A significant portion of geospatial data uncertainty stems from measurement, since all measurements are of limited accuracy [Wolf and Ghilani, 1997]. Measurements are frequently mixed in a mathematical model to estimate the unknown quantities. As a mathematical model is a simplification of a complex process, it introduce uncertainty.

Other important factors that can influence the magnitude of data uncertainty are computational and fuzzy errors. The computational media introduce uncertainty because of their inherent 'round off', and possible 'truncation' processes [Burrough, 1986].

Presenting the real world by objects having crisp boundaries is in error too, as natural objects rarely have clear cut boundaries [Burrough, 1991]. Moreover, the parameters of the general model may have correlation, as such their correlation should also be considered in the uncertainty model. Therefore, a general error model of a GIS product might be presented by the following mathematical function:

$$\sigma^2_{Total} = f(\sigma^2_{Measurement}, \sigma^2_{Model}, \sigma^2_{Computation}, \sigma^2_{Fuzzy}, \ldots) \quad (4.1)$$

where σ^2_{Total} is an indication of the total error for a GIS data.

Since the precision used in GIS exceeds the accuracy of its data the contribution of computational model can be assumed negligible [Goodchild, 1989] in the model. To make the model manageable, and the research conceivable, this chapter focuses on the first two elements of the general model. Moreover, the correlation between the two parameters are assumed negligible, hence the uncertainty model is simplified to, namely:

$$\sigma^2_{Total} = \sigma^2_{Measurement} + \sigma^2_{Model} \quad (4.2).$$

4.3 Determination of Positional Uncertainty of Point Objects

Point data (e.g. power poles or elevation data) may cover a significant portion of any GIS database. Point data can also be used as building blocks for structuring higher dimensional data, 1D lines and 2D polygons. Therefore, the determination of positional

uncertainty of a point is important not only due to being a measure of a point quality, but also because of providing the first step towards modeling the uncertainty of higher dimensional data in a GIS.

Since the majority of GIS data is gathered by the digitization process, the following simple trials were conducted to construct the covariance matrix of the digitized points. Statistical information derived from this trial was averaged to determine a representative covariance matrix (2x2 for a point; or 4x4 for a line) for all the points (or lines). The configuration was made such that a study of possible correlations due to distance and orientation can be carried out. The following steps explain the proposed trials:

(i) A set of precise mathematically definable geometric entities were generated in a GIS system. The entities included a point, a straight line, a triangle, and an octagon (Figure 4.1). Since the entities are mathematically defined their coordinates were used as the 'true' values in the calculation of accuracy [Goodchild et al, 1992].

(ii) The entities were plotted and digitized in point mode at the convenient speed. (Each point is digitized 10 times.)

(iii) Variances in the X and Y direction was computed for each point, and an average computed for all the points digitized. These values fill the diagonal elements of the covariance matrix of digitized points.

(iv) Covariances of X and Y are computed for each point, the numbers are then averaged through all the points. Similar procedures were followed to

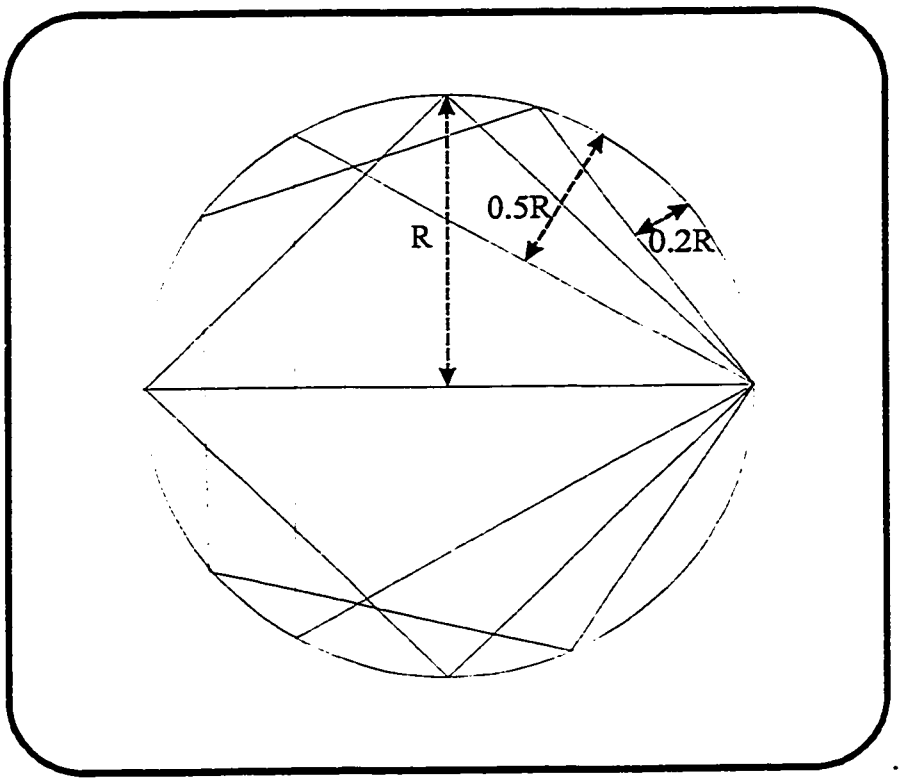


Figure 4.1. Digitized entities, together with the magnitude of the model error

determine the covariance of the line's endpoints. Since no trends were observed in the covariance of the coordinates due to angle and distance for line segments, the covariance of the end point coordinates were also averaged. These elements fill the off diagonal elements of the covariance matrix.

The following 2x2 matrix resulted from the test. The results $\begin{bmatrix} 0.32 & 0.08 \\ 0.08 & 0.35 \end{bmatrix}$ are in square millimetres. The study of possible correlation due to distance and orientation was inconclusive, the reason was attributed to the mode of digitization. It is argued if the stream mode digitization was employed possible correlations might have been encountered [Keefer et al., 1988].

4.4 Simulated Procedure of Determining Observational Uncertainty of Lines

Every measured point is to some degree uncertain, as are the straight line segments generated by connecting the points. Hence, modeling the observational uncertainty of a line segment may be determined by considering (a) the observational uncertainty of the coordinates of endpoints, and (b) the correlation among the points that construct the line (a full covariance matrix should be considered).

A nearly correct map of the region, where the random lines are located with a particular probability, e.g. 39% or 87%, may be derived by the Monte-Carlo simulation method. The coordinates of the endpoints of the line are generated according to their distributions, and to the endpoint correlation. Figure 4.2 illustrates several realizations of such a model. By connecting a large number of endpoints, random positions of the straight lines can be simulated.

Several different shapes and orientations of error ellipses have been examined, once a correlation factor between two endpoints was assigned. As indicated in Figure 4.2.a, when the correlation factor approaches one, the shape of the line variation follows the epsilon band model [Alesheikh, 1997]. When the correlation factor becomes close to zero (no correlation), the uncertainty of the middle point of the line segment gets smaller (Figure 4.2.b). In fact, when the two endpoints are independent, the error band model [Shi, 1994] is reached (Figure 4.2.c). If the correlation factor moves towards negative one, the variance of the middle point gets smaller than error band model (Figures 4.2.d, and 4.2.e).

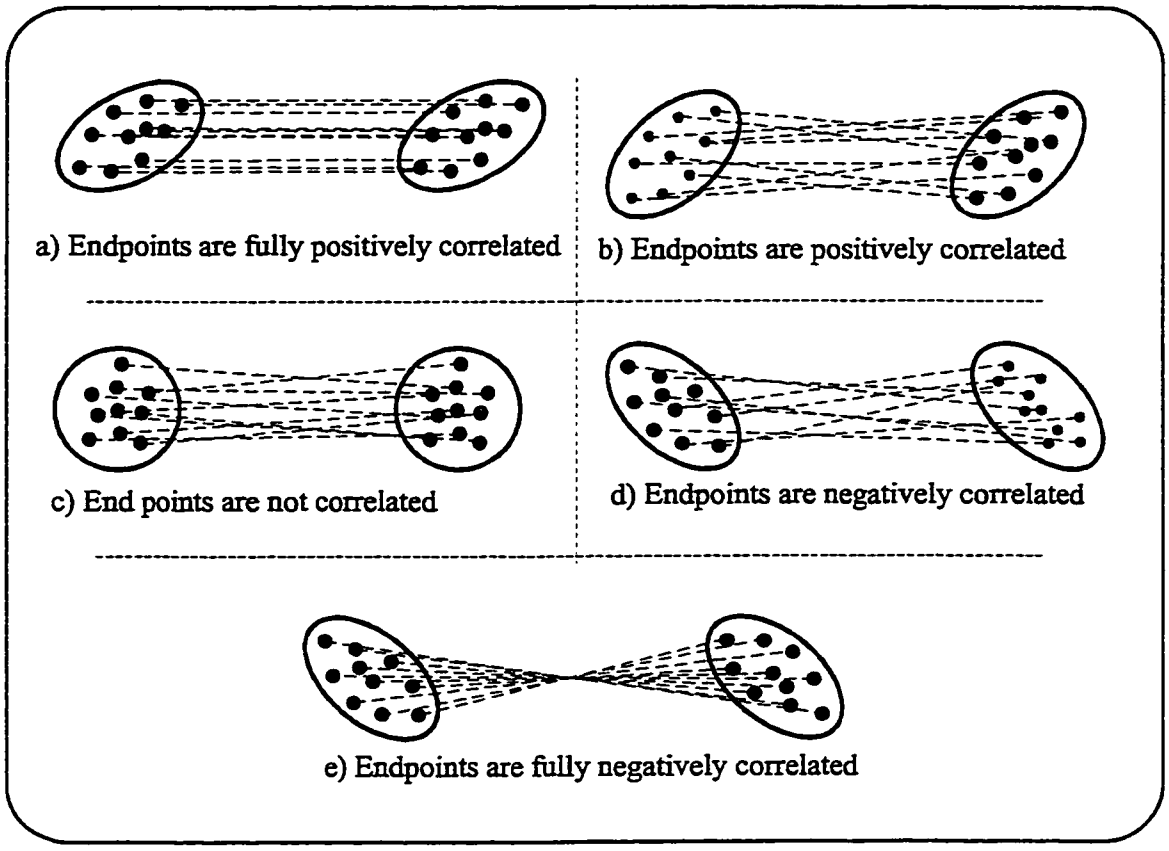


Figure 4.2. Different realizations of line uncertainty models

An immediate conclusion that may be derived from these simulation results is that both the epsilon band and error band models are special representations of the general variations that a line segment may assume.

To visualize the area of constant probability a mesh has been superimposed on top of the random lines. Each simulated line intersects a number of pixels (see Figure 4.3). These intersections are summed up for each cell resulting in a lookup table of intersections which can then be presented in 3D (Figure 4.4.a.). The cells and the relative frequencies show the probability density of the random straight line. Lines of constant probability density can now be constructed in the same way as contour lines in digital

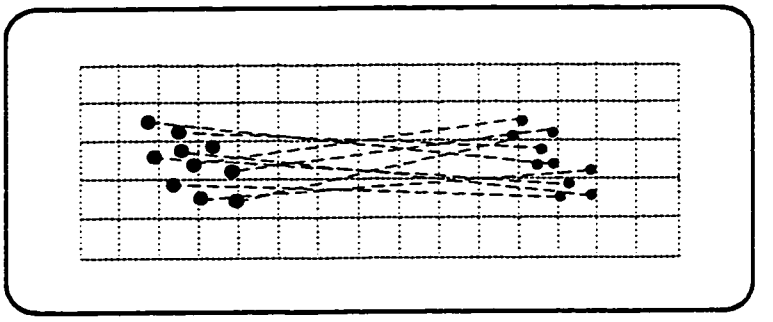


Figure 4.3. Random lines and the intersected pixels

terrain models (Figure 4.4.b, and 4.4.d). The probability associated with the contours of (e.g. 38%) can be determined by the summation of the relative frequency of pixels (f(X, Y)) located inside the contour (Figure 4.4.a). This area presents the confidence interval of a line segment and is proposed as the best presentation of the observational errors of the line segment.

4.5. Analytical Procedure of Determining Observational Uncertainty of Lines

A straight line segment may be defined as a combination of points conforming to a linear function. Hence, the uncertainty of a line segment may be determined as the aggregate of the uncertainty of points constructing the line. Point uncertainty is usually represented by error ellipses that are computed from the variance-covariance matrices of the point coordinates. So, the problem of line uncertainty modeling turns out to be:

(i) the determination of the uncertainty, C_u, of any arbitrary point located along the line segment, based on the statistical information, C_{AB}, of the endpoints coordinates, and

(ii) the computation and visualization of the region that the point ellipses cover.

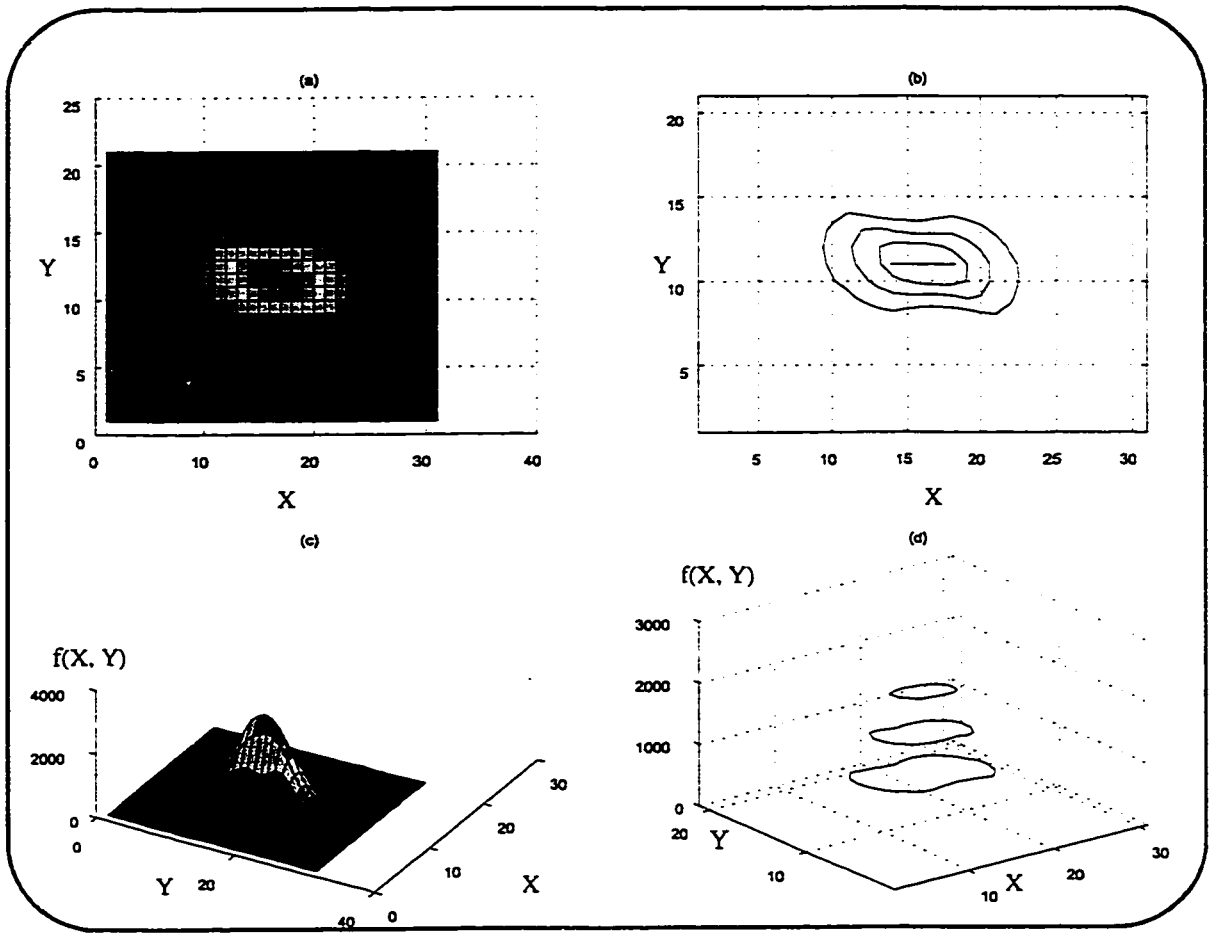


Figure 4.4. The visualization of the line uncertainty using simulation method. a) 2-D view of the Line Density Function (LDF), b) Contour line representation of the LDF, C) 3D view of the LDF, and d) 3D contour of the LDF.

If U is an arbitrary point located along a line segment AB, its coordinates can be defined as (from Equation 3.13):

$$\begin{bmatrix} X_U \\ Y_U \end{bmatrix} = \begin{bmatrix} (1-r)*X_A + r*X_B \\ (1-r)*Y_A + r*Y_B \end{bmatrix} \quad \text{or} \quad \bar{X}_U(\alpha,\beta) = \alpha\bar{X}_A + \beta\bar{X}_B \qquad (4.3)$$

$$\text{where, } r_U = \beta = \frac{D_{AU}}{D_{AB}} = \frac{\sqrt{(X_U - X_A)^2 + (Y_U - Y_A)^2}}{\sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}}, \quad 0 \le r_U \le 1, \text{ and } \alpha = 1-\beta. \quad (4.4).$$

The variance-covariance matrix of U may be derived by applying the error propagation law as follows:

$$C_U = J \; C_{AB} \; J^T \tag{4.5}$$

where J is a Jacobian matrix and C_{AB} is the covariance matrix of the coordinates of the endpoints. Once the error matrix is defined, the error ellipse of the arbitrary points may be represented by:

$$\left(\frac{X - X_U}{\sigma_{X_U}}\right)^2 - 2\rho\left(\frac{X - X_U}{\sigma_{X_U}}\right)\left(\frac{Y - Y_U}{\sigma_{Y_U}}\right) + \left(\frac{Y - Y_U}{\sigma_{Y_U}}\right)^2 = (1 - \rho^2)C^2 \tag{4.6}$$

where: σ_{X_U}, and σ_{Y_U} are the diagonal elements of the covariance matrix of U,

$0 \le \rho = \dfrac{\sigma_{x_U y_U}}{\sigma_{x_U}\sigma_{y_U}} \le 1$ is the correlation factor, and

C is a constant that determines the probability level of the error ellipse (see Chapter 3).

Once the error ellipses for arbitrary points along the line segment are defined, the region that these ellipses encompass creates the confidence level of the line. It is this region that represents the rigorous observational uncertainty indicator of a line segment computed analytically. Figure 4.5 illustrates this concept for a line segment.

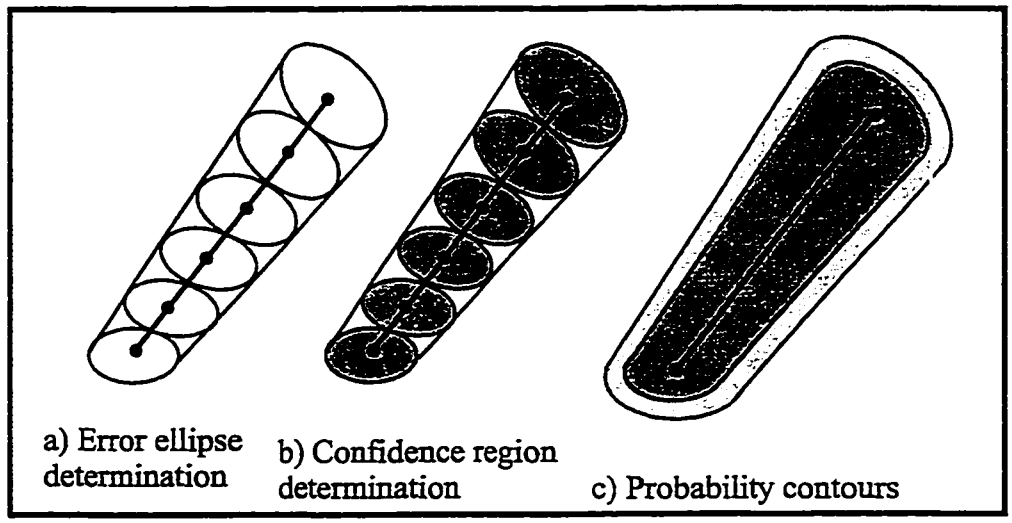


Figure 4.5. Generation of line uncertainty model

An interesting property of this model is that it can easily be extended to determine the uncertainty of a curve, if the curve function y = f(x) is known (for example, in the case of a road design). The only change that occurs in the method is that the distance from two points on the curve should be computed by the arclength equation:

$$D_{AU} = \int_{X_A}^{X_U} \sqrt{1 + (\frac{dy}{dx})^2}\, dx \qquad (4.7).$$

4.6. The Effects of Model Errors in the General Line Uncertainty Model

The previous section addressed the problem of observational uncertainty while assuming the underlying mathematical model, Equation 4.3, is error free. In contrast, this section questions the uncertainty arising from the underlying mathematical model, and determines the effects of such an error in the rigorous uncertainty model.

Model error refers to the deviation between the 'true' function representing a boundary line and its approximated linear function. Model errors in GIS are frequently

caused by:

(i) not knowing the curvilinear function (i.e. where the digitizing operator processes the boundary of a natural feature),

(ii) a simplification assumption (i.e. where the digitizing operator compromises between the magnitude of model error and a simple linear function), and

(iii) a restriction of the software (i.e. where the software cannot represent a nonlinear model at the time of digitization [Davis, 1996] or representation (e.g contours generated in PCI)).

Although the effects of model error may be negligible when line segments represent the boundaries of man-made structures or cadastral boundaries (where the boundaries are straight Figure 4.6.b), they may be of considerable magnitude when the linear function delineates a natural feature (Figure 4.6.c).

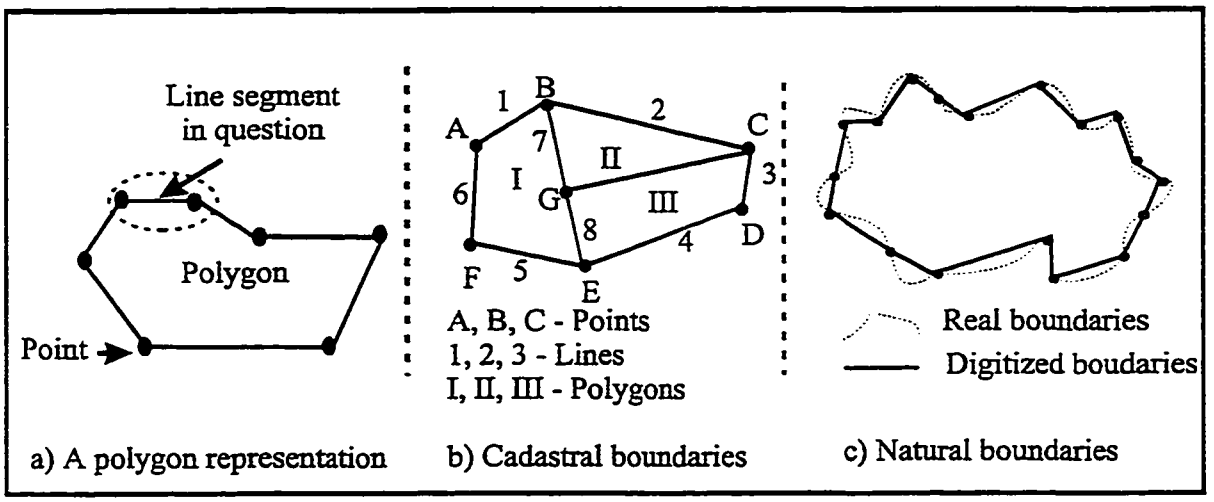


Figure 4.6. The digitized representation of different polygons

Assuming a perfect linear model representing a boundary line causes the uncertainty of the points located along the line to be always less than or equal to the uncertainty of the endpoints. If the model error is under question or the model error is of a considerable magnitude, then its effect should be quantified.

The next section will show that based on the magnitude of the model error the shape of the line uncertainty model may well be approximated by a convex curve rather than a parabola (concave) suggested by the error band model. Figure 4.7 demonstrates the differences between line uncertainty models.

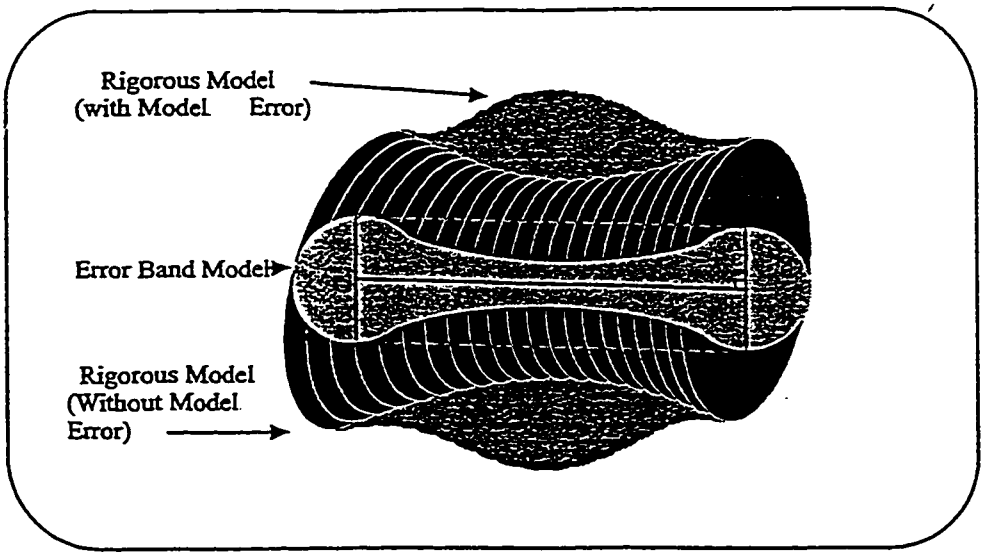


Figure 4.7. Realizations of various line uncertainty models

Functions representing the model error can be determined empirically. As such, they may vary from case to case and from operator to operator. Such an empirical function may be obtained from experience with similar problems, i.e. an operator may determine a suitable model error for his/her operation capability by comparing his/her digitization results with more accurate samples, or from simply knowing something about the

behavior of the problem [Kraus and Mikhail, 1972]. The function depicted in Figure 4.7 is the Gaussian function corresponding to the equation

$$Me\ (d) = \alpha \exp(-\beta d^2) \tag{4.8}$$

that is added to the observational uncertainties. Here, d is the separation distance from the center of gravity of the line, and β are constants that determine the maximum deviation and damping parameter, respectively. For instance, α may be close to zero once the model error is negligible, e.g. in the case of cadastral boundaries, and be higher in the case of digitizing natural feature boundaries. The following sections elaborate on the ways that model error can be quantified.

4.6.1. The magnitude of the model error: empirical method

To determine the variance of the model error along digitized straight line segments, eighteen segments of the boundaries of water bodies are digitized in two different scales. Figure 4.8 presents the area in which the test was accomplished. The figure also identifies a few places that boundary line segments are digitized.

The two versions of the digitized line segments were then compared and the magnitudes of the error at ten points along each line segment were computed. Figure 4.9 shows the two versions of the digitized boundaries for several trials.

Using 18 observations for each point, variances at these points were computed by:

$$\sigma_j^2 = 1/18 \times \sum_{i=1}^{18} V_i^2 \qquad j = 1, 2, \ldots, 10; \tag{4.9}$$

$$\text{where } V_i^2 = (X_i - X)^2 + (Y_i - Y)^2 \tag{4.10}$$

is the minimum distance (in the perpendicular direction to the straight line segments of the Figure 4.9) between the two versions of the digitization. The statistical variances at the ten points were plotted in Figure 4.10.

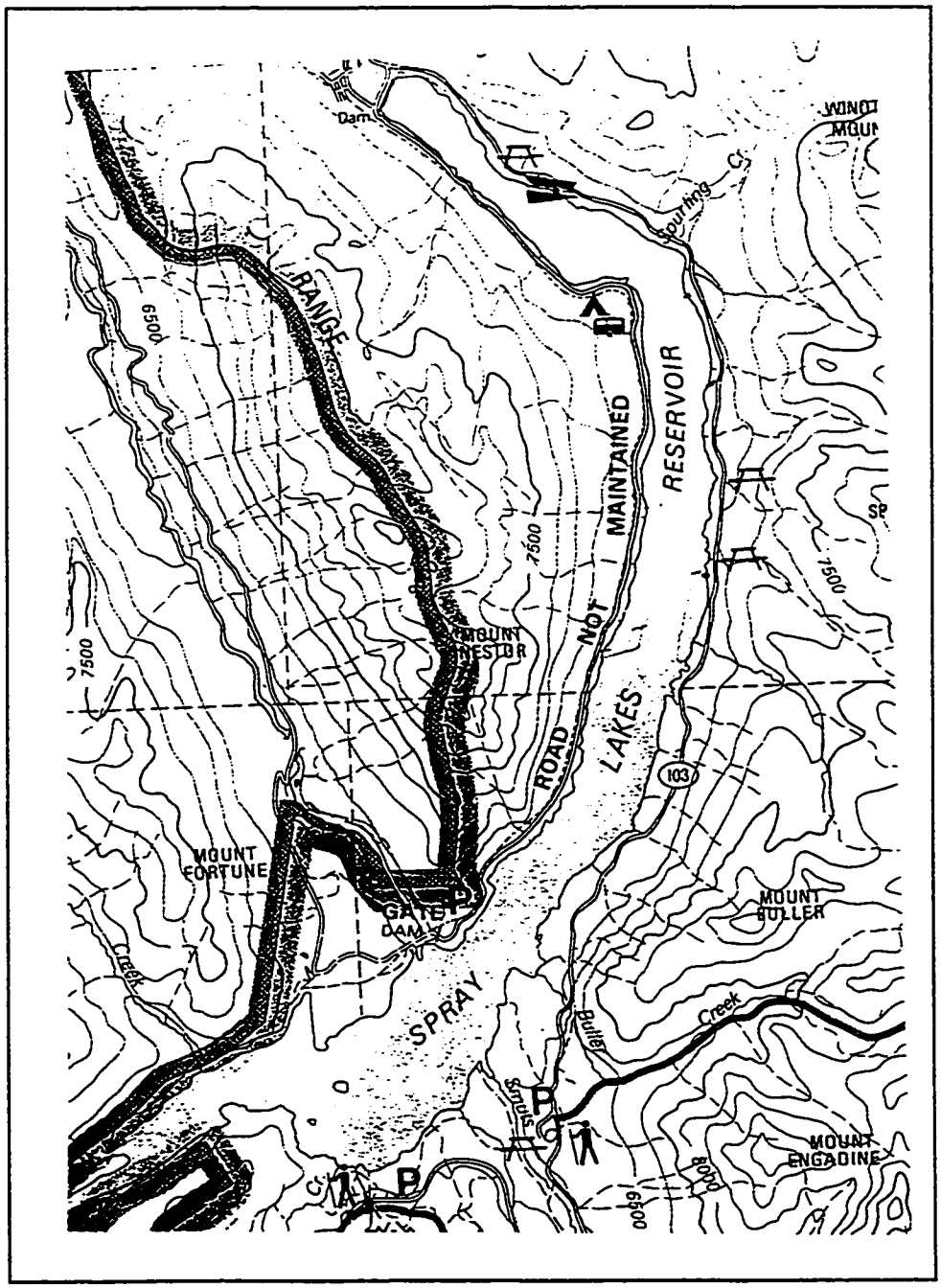


Figure 4.8. The study area for model error determination (Scale 1:100,000)

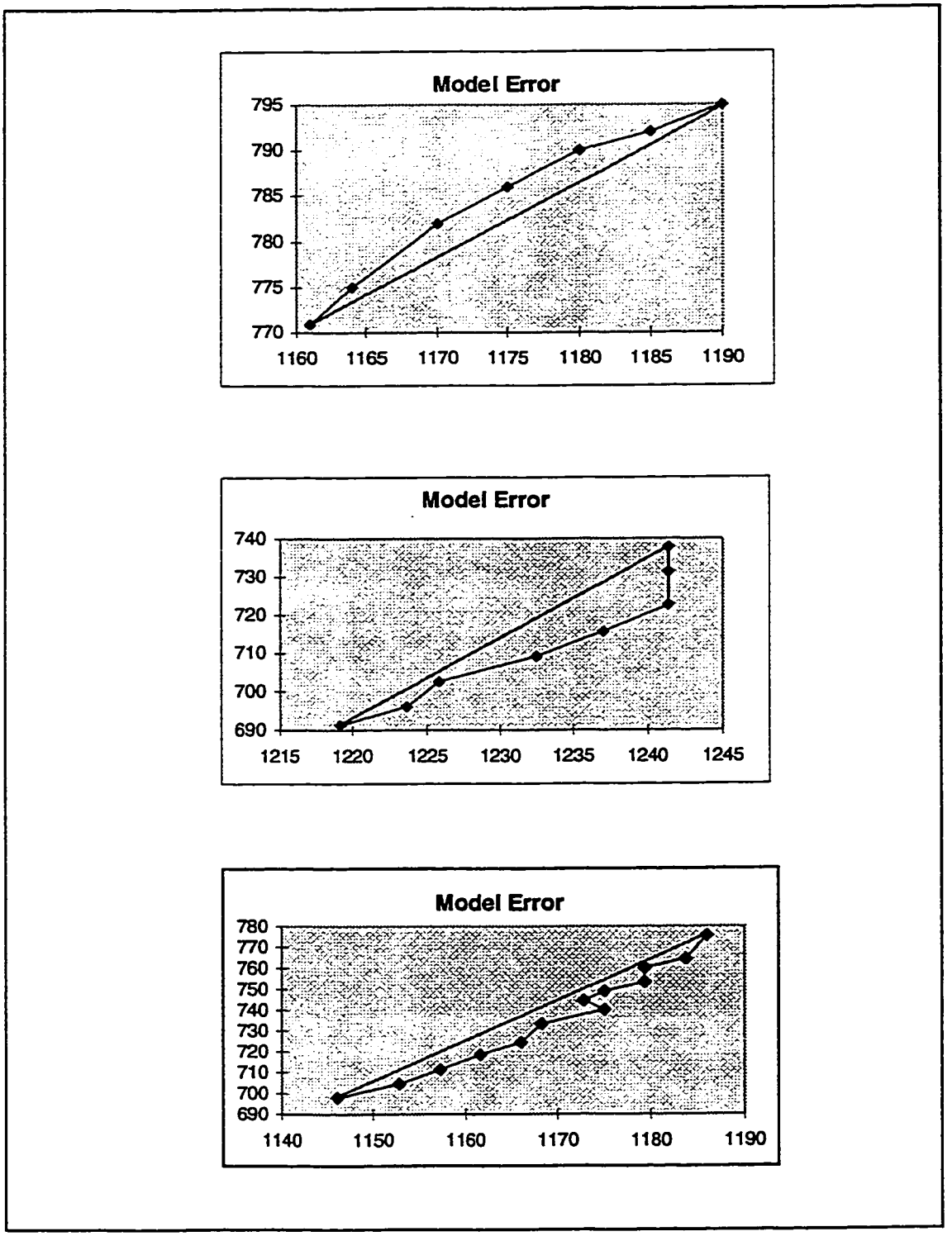


Figure 4.9. Deviations between the two versions of digitized points (assuming some local coordinate systems)

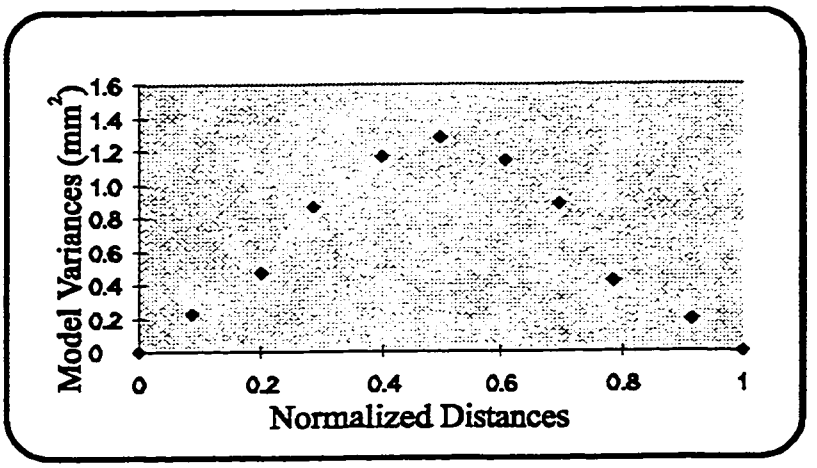


Figure 4.10. The model variances

Any approximation function can be used to estimate the variability of the variances due to model error. A Rayleigh distribution may be more appropriate (Christakos, 1992), however, due to its simplicity a Gaussian function (Equation 4.8) has been used in this thesis. It is worthwhile to mention that the magnitude of the model variance reached about four times the measurement error at the endpoints (1.3 mm^2 in model vs. 0.33 mm^2 in measurement error). This will indicate the importance of the model error in any rigorous modeling. The variance of the model error was then added to the measurement error in perpendicular direction of line segments. This presents the proposed general uncertainty model of line objects which is shown in Figure 4.7.

4.6.2. The magnitude of the model error: Analytical method

Apart from empirical determination of the model error, the effects of it can be analytically quantified once the non-linear function and its approximating straight line segments are known. This is the case when man-made structures such as curved roads are digitized. Because circles constitute the majority of non-linear functions in man-made structures,

the magnitude of the error is quantified for approximating circles. However, the strategy may easily be extended to any other non-linear functions.

As Figure 4.1 demonstrates, the magnitude of the model error decreases once the number of digitized points are increased. The model error is half of the circle's radius, R, when the circle is approximated by two points, and decreases to 0.19R when the circle is estimated by five equidistant points. Once a circle is digitized by n equidistant points, the maximum deviation, ε, can be computed by;

$$\varepsilon = R(1-\cos(\pi / n)) \tag{4.11}$$

This ε can be used instead of α, if Equation 4.8 is applied. If the distance between endpoints, D, is known, the digitization case for instance, the following Lemma can be applied to compute the maximum deviations between a circular curve and its approximated line segment;

Lemma: If the middle ordinate (ε) is the distance from the middle point of the chord to the middle point of the corresponding circular arc (Figure 4.11), its magnitude can be computed from;

$$\varepsilon = OH - OG = R - R\cos(\Delta / 2) = R(1-1/2R\sqrt{4R^2 - D^2}) \tag{4.12}$$

where $D = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}$ is calculated from the coordinates of the digitized points.

A circle (Figure 4.1) is used to determine this individual's model error. To minimize the correlation between measurement error and model error, a circle and its

approximating lines are generated by Arc/Info system (no digitization was carried out). The number of points approximating the circle (R = 5 cm) are then increased from 2 to 20, until the deviation was in the range that was negligible for the experimenter. Using Equation 4.11 the variance of the model error reached was 0.72 mm^2. This constitutes the lower level for the magnitude of model error. As previously mentioned, the value also proves that the magnitude of model error is greater than measurement error for the digitization process. The same circle was then carefully digitized on the computer screen to determine the differences between on screen and hard copy map digitization. Using Equation 4.12 the maximum detected model error was 1.2 mm^2 with an average of 0.83 mm^2 once a normal speed digitization (less than 3 seconds per point) was carried out. The dissimilarity between magnitudes of model uncertainty in the two cases (on screen vs. hard copy) were attributed to the variations in the accuracy of the instruments used.

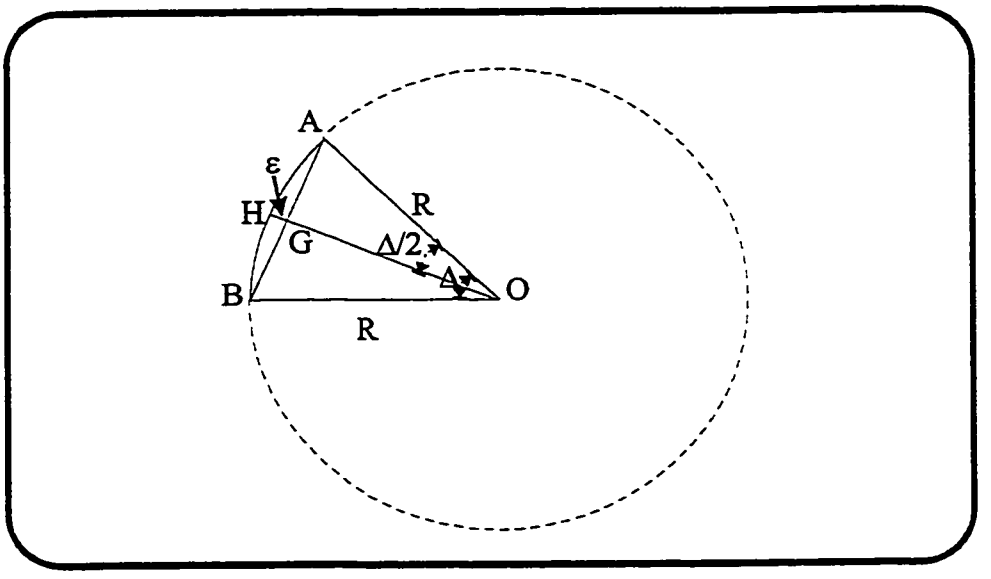


Figure 4.11. The deviation of a straight line segment from a circle

4.7. Polygon Uncertainty Model

A polygon object can be assumed as an areal object defined by its boundary lines. Therefore, polygon uncertainty is closely related to the previously discussed point and line uncertainties. It is important to notice that the uncertainty of a polygon is different from the uncertainty of the boundary of a polygon, since the interior of polygons can have uncertainty too. The interiors of polygons are more homogeneous than the boundaries of polygons, causing the magnitude of uncertainty to be greater on the boundary than interior [Zhou and Lee, 1994]. Hence this research work has focused on the modeling of the boundary uncertainty.

A comprehensive description of the random variations of a polygon may be derived from the covariance matrix of the vertices of the polygon and the model errors of each boundary segment. An indicator of the uncertainty of a polygon object is defined as the combination of the areas that the boundary line uncertainty models cover divided by the polygon area. In other words, the general uncertainty model of a polygon object that is proposed is simply the combination of the area of all of the line uncertainty indicators enclosing the polygon over the polygon's area. In the following, the visualization of the polygon uncertainty will be elaborated using various case studies. Areas covered by various line uncertainty are computed which enables a comparison between different polygon uncertainty models.

4.7.1 Survey lot data sets

The first case study involves a survey lot consisting of seven points, two of which are fixed. The angular and distance measurements are reported in Mepham and Nickerson

[1987]. The coordinate system used is a local coordinate system. The measurements, together with the approximate coordinates of the points, were used in a least-squares adjustment to determine the estimated coordinates of the points and their covariance matrices. The adjustment results were then imported into the general uncertainty model to illustrate the uncertainty of the line segments and the polygon. Figure 4.12 represents the output of the module.

Figure 4.12 clearly illustrates that the union of ellipses creates a region similar to the epsilon band model if model errors are not considered. The reason is attributed to the predominantly positive correlation among endpoint coordinates, which in turn validates the results of the simulation (Figure 4.2.b).

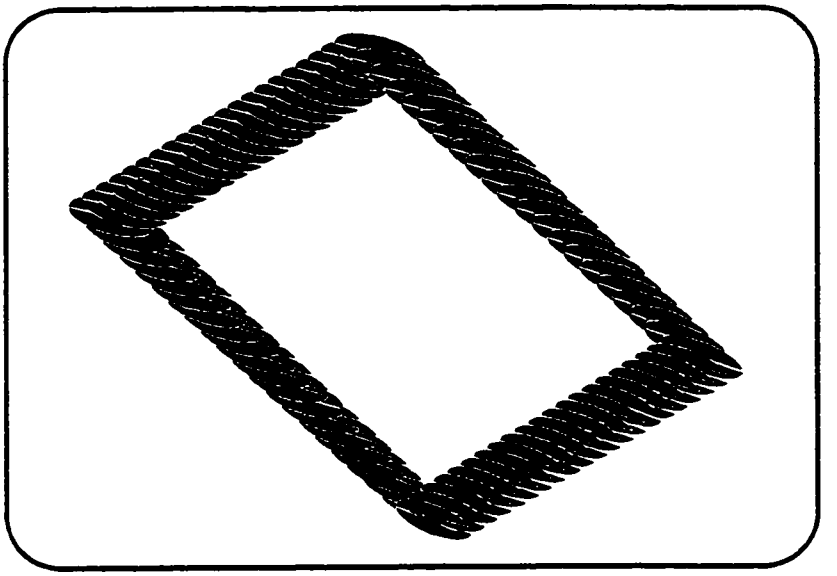


Figure 4.12. Uncertainty visualization of survey lot data sets (Ellipses are enlarged for illustration.)

4.7.2 Simulated data sets

The second case study involves simulation of coordinates of a planar triangle's vertices

together with their covariance matrices. The covariance matrix of the triangle is generated in such a way that the line uncertainty model resembles the error band when the model error is not being considered - negative correlation among end points is predominantly generated. This also supports the results of the simulation procedures (Figure 4.2.d), and consequently the generality of the proposed uncertainty models. The results of the simulated datasets are shown in Figure 4.13.

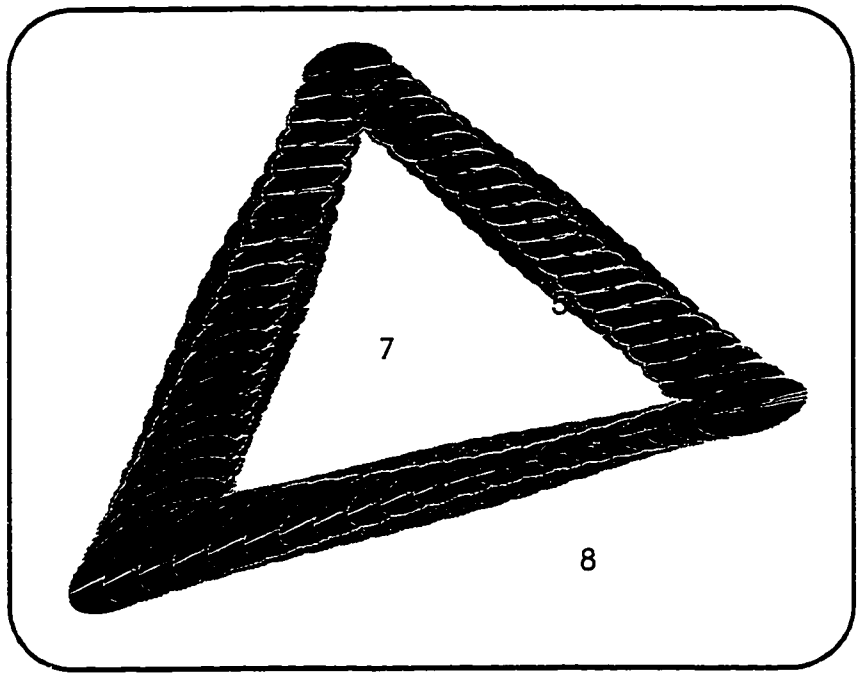


Figure 4.13. An illustration of the uncertainty of the simulated triangle

Figures 4.12, and 4.13 clearly demonstrate the effects of the correlations between endpoint coordinates on the shape of the line uncertainty model.

4.7.3 Comparison of polygon accuracy's using different models

A comparison of the different measures of accuracy of a polygon is given below. The area that is covered by the epsilon band for a line with length D (Figure 3.3) can easily be determined by:

$$A_{ep} = 2D\varepsilon + \pi\varepsilon^2 \qquad (4.13).$$

It can be shown that the area covered by error band model is approximately 0.8 times the area covered by the epsilon band model (Figure 4.2.c):

$$A_{er} = 2D\varepsilon\int_0^1 \sqrt{1-2r+2r^2}\,dr + \pi\varepsilon^2 \approx 1.62D\varepsilon + \pi\varepsilon^2 = A_{ep} - 0.38D\varepsilon \approx 0.8A_{ep} \quad (4.14)$$

The area covered by the proposed general line uncertainty model depends not only on the covariance matrix of the endpoints but also on the function representing the model error. As such a closed formula representing the uncertainty area seems impossible, though the area can be numerically approximated for each case. Table 4.2 shows the difference in the area covered by various line uncertainty models for the two case studies. For the case of the epsilon band and error band models, it is assumed that:

$$\varepsilon = \sigma = 1/4[\sigma^2{}_{xA} + \sigma^2{}_{yA} + \sigma^2{}_{xB} + \sigma^2{}_{yB}]^{0.5} \qquad (4.15).$$

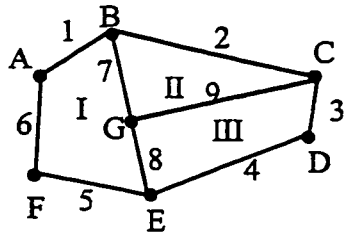
Equation 4.8 is used to approximate the model error for the triangle case study which is characterized by: $\alpha = 1.2$ mm^2 and $\beta = 7$.

Table 4.1. Comparison between areas covered by various line uncertainty measures

	Survey Lot Case (m^2)	Triangle Case (m^2)
Sum of ε-bands A_{ep}	4.155	44.61
Sum of error bands A_{er}	3.324	35.69
Sum of general bands (without model error)	3.983	38.72
Sum of general bands (with model error)		53.37
Area	599.86	6820.8

4.8 Logical Uncertainty Representation in Digital Databases

With the availability of different uncertainty models for geospatial entities, it is suggested that the current data structures be augmented with uncertainty information. Figure 4.13 shows the logical integration of such information in a vector data structure.



A, B, C - Points
1, 2, 3 - Lines
I, II, III - Polygons

Poly #	Line #	Uncertainty(%)
I	1, 7, 8, 5, 6	C_I
II	2, 9, 7	C_{II}
III	3, 4, 8, 9	C_{III}

Polygon-Line List

Line #	From Point	To Point	Uncertainty
1	A	B	C_{AB}, α_{AB}, β_{AB}
2	B	C	C_{BC}, α_{BC}, β_{BC}
3	C	D	C_{CD}, α_{CD}, β_{CD}

Line-Point List

Point #	X Coord.	Y Coord.	Uncertainty
A	X_A	Y_A	C_A
B	X_B	Y_B	C_B
C	X_C	Y_C	C_C

Point Coordinates List

Figure 4.14. Conceptualization of uncertainty inclusion in digital databases

Point uncertainty information is shown as the covariance matrix C_A of that point in the Point-Coordinates list. To visualize the general uncertainty model, information about the correlation between endpoints coordinates and model error are necessary which is given in the Line-Point list by C_{AB}, α_{AB}, β_{AB}. Polygon uncertainty information is considered as the summation of the area covered by uncertainties of boundary line segments over the total area of the polygon, C_I, that should be provided in the Polygon-

Line list. This explicit form of representation facilitates the management of uncertainty, since the manager can assess the quality of the raw data, and aggregate it to determine the product uncertainty.

4.9. The Point-in-Polygon Problem

The point-in-polygon problem has been addressed by several algorithms; one example is the half-line algorithm [Worboys, 1995]. However, the quality of the solution to the point-in-polygon problem has yet to be addressed properly [Leung and Yan, 1997]. A simple computation of the quality of the point-in-polygon solution could, for example, avoid displaying buoys on dry land or rivers outside their floodplain [Chrisman, 1991]. Based on the proposed polygon uncertainty model, the solution to the point-in-polygon problem may be more accurate using probability statements.

In order to enhance the solution to the point-in-polygon problem a two-step procedure is proposed:

(a) determine whether a point is inside, outside, or on the boundary of a polygon, and

(b) determine in which probability region the point falls.

To determine the position of a point with respect to a polygon, the polygon may first be decomposed to several triangles (2-D primitive objects), and then, the position of the point is examined with respect to the triangles. A triangle can be represented in a normalized homogeneous coordinate system by its vertices coordinates: (0, 0, 1), (0, 1, 0), and (1, 0, 0), as shown in Figure 4.12.

Assuming that the coordinates of the points to be examined are (a, b, c), if any coordinate of the point is equal to 1, the point is located on the boundary of the triangle. If

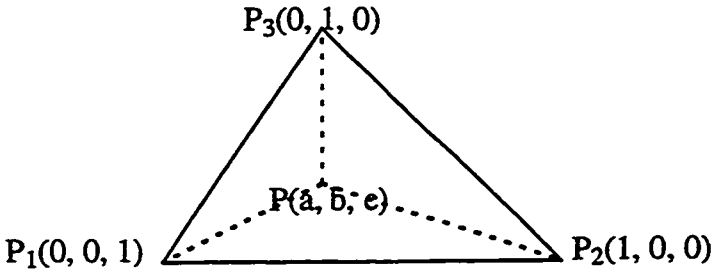


Figure 4.15 Representation of a triangle in a homogeneous coordinate system

any coordinate of the test point is greater than 1 or less than zero, then the point is located outside the triangle. Finally, if all the coordinates of the desired point in the homogeneous coordinate system are less than one, and greater than zero, then the point is located inside the polygon. The above procedure has been used to address the point-in-polygon problem in the case of the simulated triangle and presented in Table 4.2.

Table 4.2. Results of point-in-polygon tests, probability level is fixed(* without model error, ** with model error, *** boundary line)

No.	Epsilon Model	Error Model	General (*)	General (**)	In	Out	Bound (***)
1	*	*	*	*			*
2	*	*	*	*		*	
3	*			*		*	
4	*		*	*	*		
5				*	*		
6	*	*	*	*		*	
7					*		
8						*	

The position of a point inside, on, or outside the triangle $P_1\ P_2\ P_3$ can be determined by:

$$P = P_1 + \lambda(P_2 - P_1) + \mu(P_3 - P_1) = aP_1 + bP_2 + cP_3, \tag{4.16}$$

with $a + b + c = 1$, where

$$\lambda = b = det\ (P - P_1\ \ P_3 - P_1) / det\ (P_2 - P_1\ \ P_3 - P_1),$$

$$\mu = c = det\ (P_2 - P_1\ \ P - P_1) / det\ (P_2 - P_1\ \ P_3 - P_1),$$

$$a = 1 - b - c.$$

given P_1, P_2, and P_3 are non collinear, and det stands for determinant [Blais, 1996b].

To determine which probability level the test point falls in, the coordinates of the point are set in the ellipse equations (Equation 4.6) which create the line uncertainty model. If the equality is valid, then the test point is located on the boundary of the specified probability region. If the left side of the equation is less than the right side, then the test point falls in the ellipses, and consequently within the specified region. If the left side of the Equation 4.6 is greater than the right side, then the test point is located outside the specified probability region.

Table 4.3 clearly shows the different answers to the point-in-polygon problem using various line uncertainty models. The triangle case is examined. In Table 4.3 the locations of the test points are examined based on a fixed probability (0.39, or C=1) level. Similar procedures can be used once the probability levels vary.

4.10 Chapter Summary

At the frontier of the information age, GIS is emerging as a powerful means to handle spatial data effectively and efficiently. However, a constraint in an ever developing GIS is the issue of proper modeling and communicating uncertainties of geospatial data.

This chapter has proposed rigorous uncertainty models for GIS positional data. The major parameters of the model identified and methods to determine the magnitudes of the parameters are supplied. Test studies showed that the magnitude of the model error surpasses the observational uncertainty by a factor of 3. As such it should be considered in the modeling processes.

It is argued that point measurements can be used to build higher-order objects such as line segments and polygons. Following this line of thought, the uncertainty of points were aggregated to realize uncertainties of straight line segments and polygons. The Monte-Carlo simulation and analytical methods were designed and tested to determine the shape of the line uncertainty model. Based on the results, it is concluded that the currently used models, the epsilon band and the error band, are specific realizations that line variations may assume.

Model error is added to the final positional uncertainty to account for the deviation of the true model from approximated ones. To integrate the uncertainty information of positional data, logical representation of uncertainty information in GIS databases was submitted. This structure will be handy for any error propagation procedure to be followed. Having the uncertainty of a polygon object modeled, the solution to the point-in-polygon problem was enhanced by probability statements.

While the emphasis of this chapter was on the modeling (observational and model) uncertainty of straight line segments, uncertainty of the curves remain for further investigation which will be the theme of the next chapter.

CHAPTER 5

UNCERTAINTY MODELING AND DATA REDUCTION OF CURVILINEAR BOUNDARIES

5.1 Overview

Geospatial Information Systems (GISs) are proving successful in reducing practical problems of organizing and integrating spatial data and for carrying out complex spatial analysis. As GIS users become more familiar with the technology and as GIS is accorded a more sophisticated role through incorporation into decision support systems, the quality of the results is more frequently questioned. The quality of GIS results is hardly achievable without rigorous modeling of their spatial data constituents: points, lines and polygons. Analytical modeling of geospatial data, in turn, requires a functional representation of the data.

In the previous chapter, the uncertainty of the linear objects was modeled based on the implicit assumption that endpoints are connected by a straight line. Though the assumption is valid for most man-made structures, the boundaries of natural features are frequently not so constrained. One method to get around this problem is to divide the boundary into several small segments and use the uncertainty models devised in Chapter 4. The jagged representation of the boundary lines, however, lacks the aesthetical presentation cartographers require. Moreover, the required data volumes for such a representation are huge, hence, the method hampers the efficiency that computer specialists look for. An intelligent way to represent natural boundaries might be with the use of non-linear functions.

This chapter aims to identify statistically-best functions representing natural boundaries, and to model their associated uncertainties. In doing so, data reduction and smoothness of boundaries are of prime concern. As such, this chapter is structured so that the problem of model identification of the boundary object is addressed first. Efficient ways of identifying the model are determined. Afterwards, the uncertainties of the best-fit model to the digitized lines are examined. The chapter will conclude with some practical considerations and a brief summary.

5.2 Model Identification Strategy

Model identification problems are often encountered in practice, because any data processing and quality analysis of experimental results can rarely be carried out without first deciding on an appropriate model. Though model identification can be accomplished for several purposes, such as enhancing the reliability of results [Blais, 1991a], it is mainly used to reduce the volume of data and consequently to improve the computational efficiency of presenting object's boundaries in this study. Moreover, having boundary lines presented as mathematical functions enables the study of the function uncertainty.

To approximate such an optimal function, three questions should be answered:

(i) what is the optimal norm for the approximation?

(ii) what are the appropriate base functions?

(iii) do all the estimated parameters of the model significantly contribute to the accuracy of the results?

For example, given data for regression, an error norm (such as the quadratic norm)

has to be decided first, and then the base functions for an algebraic or other regression should be selected. In the curve and surface fitting, the selection of the unknown parameters to be included in the model very often has to be identified too.

Blais [1991a] aptly elaborated on the selection of optimal norm in model identification procedures. Due to its computational simplicity [Vanicek and Krakiwsky, 1986] the quadratic norm is used in this thesis.

Selecting the family of the approximating functions (e.g. algebraic, trigonometric, or exponential polynomials) should be made using external information, such as the appearance of the true function or previous trials. Therefore, it varies from one application to another.

Deciding on the order of the approximating function and the significance of the involved parameters can be carried out using statistical testing procedures. But, it is now recognized that classical statistical test procedures that are too closely and implicitly related to specific models (e.g. the normal function) are often inappropriate to identify the optimum order of a model [Sakamoto et al., 1986]. It is also well known that in selecting optimal degrees or orders for problems such as regression models, the maximum likelihood principle fails in the sense that minimizing the error variance (implied by a Gaussian assumption) simply implies the highest degree or order possible for the available data.

Akaike [1971 and 1973] showed an information interpretation of the likelihood and pointed out that an objective evaluation of the goodness of fit of the assumed models becomes possible by extending the concept of likelihood. The criterion is now well

known as AIC (Akaike Information Criterion). The introduction of this objective criterion is very important since it enables the selection of optimal order and significant parameters in a candidate mathematical family.

5.3 AIC

An objective evaluation of the goodness of fit of an approximating model can be made by the log likelihood function. The function identifies the model as the best if the expected log likelihood is the largest among the candidate family. The maximum log likelihood is shown to be a biased estimator, however [Akaike, 1974]. This tendency is more noticeable for models with a larger number of free parameters. In other words, if the model with the largest maximum log likelihood is chosen, a model with an unnecessarily large number of free parameters is likely to be selected.

Through examination of the relationship between the biased and the number of free parameters of a model, Akaike [1973] claimed that

(maximum log likelihood of a model) - (number of free parameters of the model) (5.1)

is an unbiased estimator. Therefore, the AIC criterion has been defined as

AIC = -2(maximum log likelihood of a model) +

2(number of free parameters of the model) (5.2)

A model which minimizes the AIC (minimum AIC estimate, MAICE) is regarded to be the most appropriate model. Equation 5.2 implies that when there are several models whose values of maximum likelihood are about the same level, the one with the smallest

number of free parameters should be selected.

The originally proposed AIC, that is:

$$AIC = \log \hat{\sigma}^2 + (2m + 3) / n \tag{5.3}$$

for an approximating model with m parameters and n observations, has been shown to require corrections to the penalty term for different applications [Huang, 1990]. That is:

$$AIC_H = \log \hat{\sigma}^2 + (m / n) \log n \qquad \text{for general models,} \tag{5.4}$$

$$AIC_H = \log \hat{\sigma}^2 + 2(m/n) \log n \qquad \text{for unstable models} \tag{5.5}$$

$$AIC_H = \log \hat{\sigma}^2 + m \qquad \text{for explosive models} \tag{5.6}$$

where *log* $\hat{\sigma}^2$ denotes the log likelihood of the model fit and where

$$\hat{\sigma}^2 = E(y - a_o - a_1 x - a_2 x^2 - \ldots - a_m x^m)^2 \tag{5.7}$$

is the model variance in the above equations. Blais [1991a] showed that the previous three criteria result in the same order of an autoregressive model in practical applications, supporting the independence of AIC from translations and scaling.

5.4 Polynomial Regression Model and its AIC

Since the boundary delineating two natural features is frequently represented by an algebraic polynomial [Gong and Chen, 1992], and due to its simplicity, in this section, the polynomial regression model is developed. AIC will then be derived for such a model and

used as a tool for determining the optimum order of the model. In the polynomial regression, a common approach is to fit a function of the form:

$$y = a_0 + a_1x + a_2x^2 + \ldots a_mx^m + \varepsilon \qquad (5.8)$$

to data such as those derived from digitization [Alesheikh, 1997], where ε is an independent normal random variable that follows $N(0, \sigma^2)$. The integer m is the order of the polynomial regression model. This model determines the variation of y as the sum of the polynomial in the deterministic variable x and the random error ε.

Fitting a polynomial to the data is equivalent to estimating a probability distribution for the variable y as a function of the variable x. The Appendix demonstrates the required algebraic manipulation for calculating the AIC of such a polynomial. The AIC of a regression model (Equation 5.8) has the form:

$$AIC_m = \log(\hat{\sigma}^2) + 2(m+2)/n \qquad (5.9).$$

Figure 5.1 depicts simulated data (using Equation 5.8 with $a_3 = 1$ and other a_i zero) together with several fitted algebraic polynomials. The deviations between the lower order polynomials are much greater than the higher order ones, showing that increasing the order of the complex models will not necessarily contribute significantly to improvement of its error variance. To determine which order constructs the statistically-best fit function, the AIC criterion (Equation 5.9) was applied.

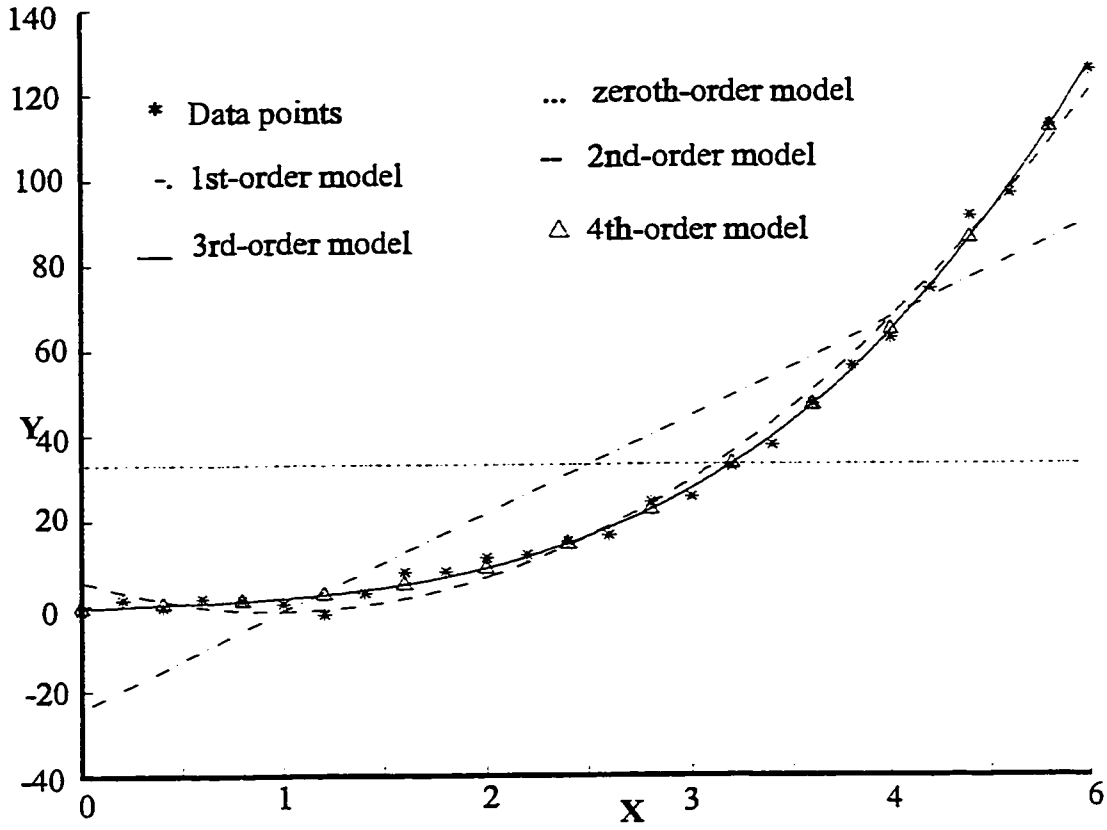


Figure 5.1. Data points and their fitted algebraic polynomials

Table 5.1 demonstrates the resulting AICs together with the parameters of the model. As expected, the variance of the model (the third column) decreases when the order of the

Table 5.1. The values of AICs for several algebraic polynomials

Case No.	Order	$\hat{\sigma}^2$	AIC	AIC_H	a_0	a_1	a_2	a_3	a_4
1	0	1478.00	133.2	3.15	33.00	----	----	----	----
2	1	254.08	115.3	2.43	-24.40	22.8	----	----	----
3	2	8.63	79.1	1.21	5.60	-14.7	7.50	----	----
4	3	0.27	42.1	-0.40	-0.14	0.5	-0.26	1.03	----
5	4	0.254	43.3	-0.38	-0.62	2.4	-1.28	1.21	-0.01

polynomial increases. However, the improvement of the variance slows down when the model reaches higher order polynomials. To compare different AIC criteria, the values of Equation 5.4 were also computed and presented in Figure 5.2.b. Figures 5.2.a and 5.2.b clearly demonstrate that both criteria pinpoint a similar model, in this case a third order polynomial, meaning that AICs are independent of the translation and scaling. The statistically best function in this case is then:

$$y = -0.1415 + 0.5396x - 0.2603x^2 + 1.0348x^3 \tag{5.10}$$

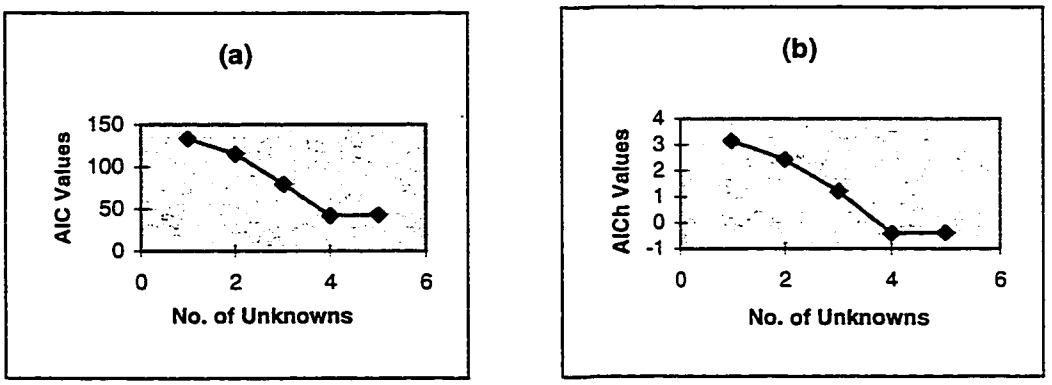


Figure 5.2. AIC values of polynomial regression. (a) Using Equation 5.7 (b) Using Equation 5.4

5.5 The Multiple Regression Model and its AIC

The proceeding approximation procedure estimates the unknown model by a polynomial of order zero first, and goes on by increasing the order of the model until the minimum AIC is reached (See Table 5.1). The model that lays the minimum AIC is then selected as the statistically best function. For instance, to identify a function of the form

$$y = a_3x^3 + \varepsilon \tag{5.11}$$

it starts with $y = a_0$, then $y = a_0 + a_1x$ then $y = a_0 + a_1x + a_2x^2$, and so on to reach a minimum AIC where that model will be chosen. Consequently the method is unable to

identify the desired Equation (5.11) in its original form. Moreover, because of the need to store all the model's parameters, a_0, a_1, a_2, ..., it may be categorized as an inefficient algorithm. As this example shows, the stored parameters may not contribute significantly to the information contents of the model. An estimated model with an unnecessarily large number of independent variables may be difficult to handle too [Burrough, 1986].

To identify the most significant independent variables of the model and to reduce the storage inefficiency of the method, it is proposed that the AIC of all possible combinations of the independent variables be computed and the model with minimum AIC be selected.

To accommodate for such a modification, Equation 5.8 can easily be changed to:

$$y = a_o + a_1x + a_2x_2 + \ldots a_mx_m + \varepsilon \qquad (5.12)$$

in which m independent variables are assumed. Following similar algebraic manipulation as described in Appendix, the AIC for Equation 5.12 can be derived as:

$$AIC(m) = \log(\hat{\sigma}_{(m)}^2) + 2(m+2)/n \qquad (5.13)$$

where $\log(\hat{\sigma}_{(m)}^2)$ denotes the error variance of the model fit due to m independent variables. This method has been applied in several trials. Table 5.2 presents the results of such a test for the data used in Figure 5.1.

Table 5.2 clearly demonstrates how well the method can detect the desired function. The storage requirements of this method is one-fourth of that needed by using Equation 5.10. Compared to Equation 5.10, the statistically best function determined by

Table 5.2 AICs of possible combinations of independent variables

No.	Variables	No. of Unknowns	Residual Variance	AIC	a_0	a_1	a_2	a_3
1	1, x_1, x_2, x_3	4	0.274	42.110	-0.142	0.540	-0.260	1.035
2	1, x_1, x_2	3	8.632	79.090	5.570	-14.680	7.501	-----
3	1, x_1, x_3	3	0.271	39.998	0.075	0.006	-----	1.001
4	1, x_2, x_3	3	0.270	39.981	0.112	-----	-0.016	1.006
5	x_1, x_2, x_3	3	0.264	39.692	----	0.332	-0.179	1.026
6	1, x_1	2	254.080	115.280	-24.432	22.823	----	----
7	1, x_2	2	43.400	95.330	-7.843	----	4.760	----
8	1, x_3	2	0.260	37.520	0.082	----	----	1.001
9	x_1, x_2	2	12.610	81.370	-----	-10.314	6.787	-----
10	x_1, x_3	2	0.260	37.560	-----	0.047	-----	1.000
11	x_2, x_3	2	0.263	37.650	-----	-----	0.017	0.999
12	1	1	1477.860	133.160	33.000	-----	-----	-----
13	x_1	1	408.040	118.630	-----	15.637	-----	-----
14	x_2	1	70.760	98.850	-----	-----	4.258	-----
15	**x_3**	**1**	**0.253**	**35.230**	**-----**	**-----**	**-----**	**1.002**

multiregression model has the form:

$$y = 1.0024x^3 \qquad (5.14).$$

In reality, the data shown in Figures 5.1 and 5.2 were obtained by assuming the following model as the true structure:

$$y_i = x_i^3 + \varepsilon_i \tag{5.15}$$

where ε_i's are normal random numbers with zero mean and variance 0.25.

5.6 The Piecewise Polynomial Approximation

The previous sections of this chapter were concerned about the approximation of the boundary features by the use of algebraic polynomials. While this method of approximation is appropriate for many circumstances, its use is restricted by the oscillatory nature of high degree polynomials, and because a fluctuation over a small portion of the interval can induce large fluctuations over the entire range restricts their use [Burden et al., 1978]. Moreover, the higher order of the approximated polynomial may results in an unstable solution [Burrough, 1986].

An alternative approach that is suggested is to divide the interval into a collection of segments and to construct an approximating polynomial on each section. Approximating the model by this type of functions is referred to as piecewise polynomial approximation. The orders of the piecewise polynomial may vary to accommodate some constraints on the approximating model. The following is a general expression for a piecewise polynomial function:

$$p(x) = p_i(x) \qquad x_i \le x \le x_{i+1} \qquad i = 0, 1, \ldots, k\text{-}1 \tag{5.16}$$

$$p_i^{(j)}(x) = p_{i+1}^{(j)}(x_i) \qquad j = 0, 1, \ldots, r\text{-}1; \qquad i = 0, 1, \ldots, k\text{-}1 \tag{5.17}$$

The points $x_1, \ldots, x_{k-1}$ divide the interval into k subintervals. The functions $p_i(x)$ are polynomials similar to Equation 5.8. Equation 5.17 governs the order, r, of the continuity

constraints at the breakpoints. The most widely used piecewise polynomial for approximation purposes is the cubic spline.

5.6.1 Cubic spline approximation

Cubic splines are piecewise polynomials that are constrained to preserve smoothness at the nodes between the subintervals or segments. The curve defined by a cubic polynomial can pass exactly through four points, but in order to fit a longer sequence it is necessary to use a succession of polynomial segments. To ensure that there are no abrupt changes in slope or curvature between successive segments, the polynomial function is not fitted to four points, but only to two. This allows the use of additional constraints which will determine that the resulting spline has continuous first derivatives between segments (the slope of the line will be similar on both sides of the endpoints) and continuous second derivatives (the rate of change in the slope of the line will not change across the endpoints). To fully identify the function for each segment, the coordinates of the points together with the slope of tangent lines at the points are required. Moreover, end conditions that determine the shape of the line in the first and last segment should be specified. Various end conditions may be selected depending on the desired shape of the line at the end points. Relaxed or natural end conditions [Davis, 1986] that do not require specifying tangent vectors at the end points are used in this chapter.

5.6.2 Breakpoint determination

A natural spline normally passes through all the data points, leaving no room for errors. Applied in this way, the procedure partitions the n digitized boundary points into n-1 segments which may then be presented by n-1 spline segments. If carried out, the method

burns the storage dramatically. It is proposed to relax the error restrictions to some limit, T, to compromise for storage load. If so, the defined threshold can be used to determine

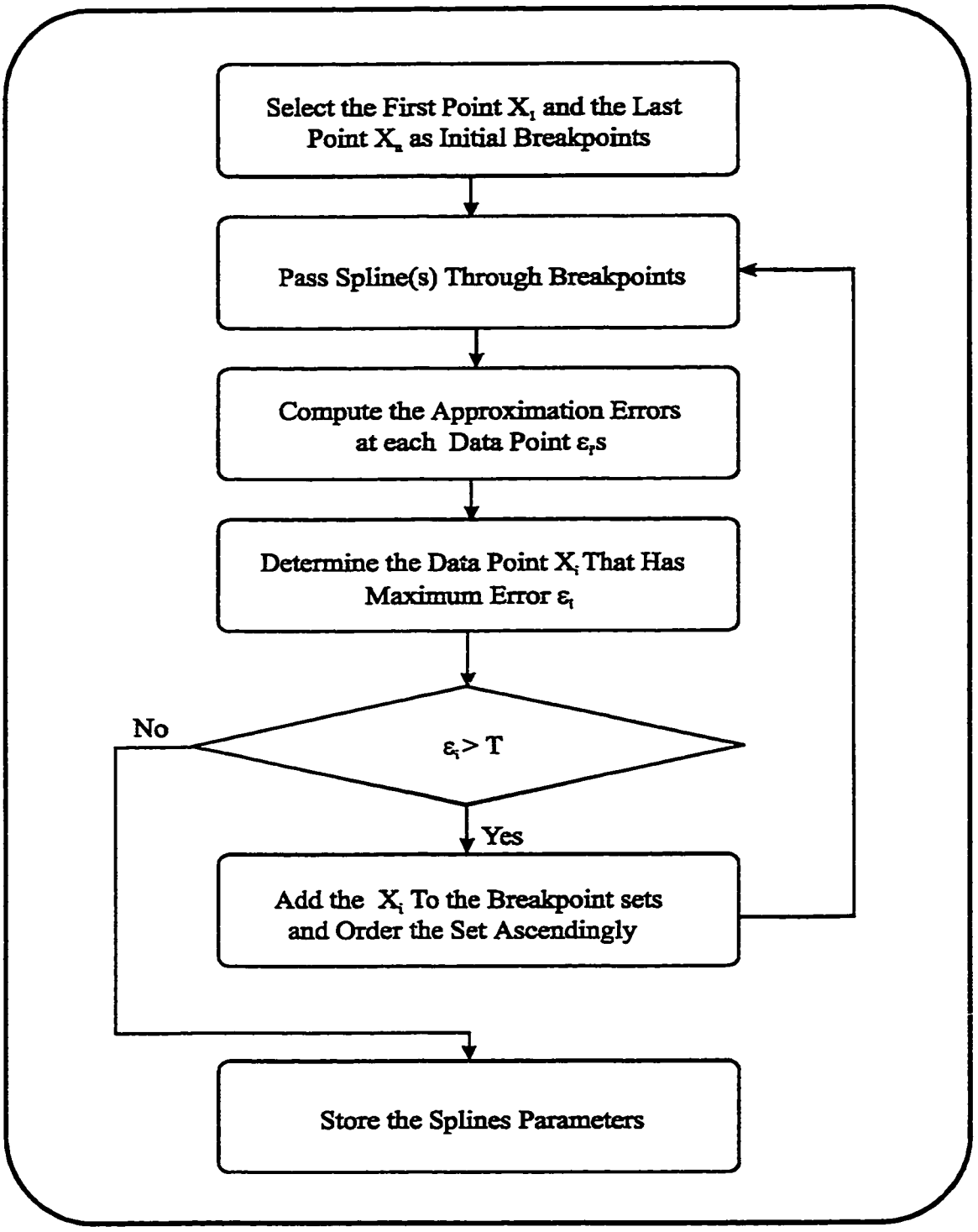


Figure 5.3 Algorithms for determining the breakpoints

the position of the breakpoint. Depends on the magnitude of the threshold, the amount of data reduction varies. The above flowchart presents the proposed algorithm for specifying the location of breakpoints.

At the first step, all the data points are approximated by one linear spline passing through the two end points. The error at each data point is then computed. Error is defined as the distance between the data point (X_i, Y_i) and the approximation function (a line in this case). The magnitude of the errors at the data points in the first run is computed by:

$$\varepsilon_i = \pm(y_i - a_0 - a_1 x_i) / (1 + a_1^2)^{0.5} \qquad i = 2, 3, \ldots, n\text{-}1 \qquad (5.18)$$

where the sign is chosen so that the distance is nonnegative, and a_0, a_1 are the parameters of the linear spline.

If the maximum error ($\max(\varepsilon_i)$) is less than the magnitude of the threshold, the algorithm stops and the parameters of the line together with the position of the endpoints are saved. Otherwise, the point with maximum error is selected as a new breakpoint, a quadratic spline is fitted to the three nodes, and the error at each data point is computed. Similar to the first run, if the maximum error gets greater than the assigned tolerance, the point with maximum error is selected as a new breakpoint. Cubic splines are then fitted to the four breakpoints. To determine the error, the distance function (Equation 5.19) is minimized (making its first derivative with respect to X equal to zero and solving for X):

$$Z(x) = \varepsilon^2 = (X - X_0)^2 + (P - Y_0)^2 \qquad (5.19)$$

where, $P = a_0 + a_1X + a_2X^2 + a_3X^3$ (5.20).

Once X is computed, it is used in Equation 5.20 to determine the corresponding Y coordinates on the spline. The distance between the points (X, Y) and (X_i, Y_i) is the required error at i-th point. It is worth mentioning that at the second run, Equation 5.20 will be a quadratic polynomial.

In summary the algorithm starts with a linear spline passing through two end points and then continues with a quadratic spline using three points. Once the number of the breakpoints gets larger than three, cubic splines are used to pass through all the breakpoints. If the maximum error is greater than the specified threshold, the point with maximum error is added to the breakpoint sets. The algorithm recursively runs till the maximum error gets smaller than the threshold, where spline parameters together with the position of breakpoints are stored.

5.7 Uncertainty Modeling of Curvilinear Boundaries

To estimate the level of uncertainties of curvilinear boundary features, two types of models are proposed. They can be categorized as simulation and simplified analytically based models.

5.7.1 Simulation based uncertainty model

In this procedure, the estimated regression models are assumed to be the 'true' boundary section. The realizations y_i for several values of x (e.g. $x_i = i/100$, i = 0, 1, 2, ... 100) were obtained using the Monte Carlo method (See Section 3.6). Proper random components are then added to the values of the y_i's. Polynomials of orders similar to the 'true' function

were then fitted to the generated points. Figure 5.4 presents the results of such a simulation model for Equation 5.14.

To determine the probability regions around the approximated line, the procedures outlined in Section 4.4 can be applied.

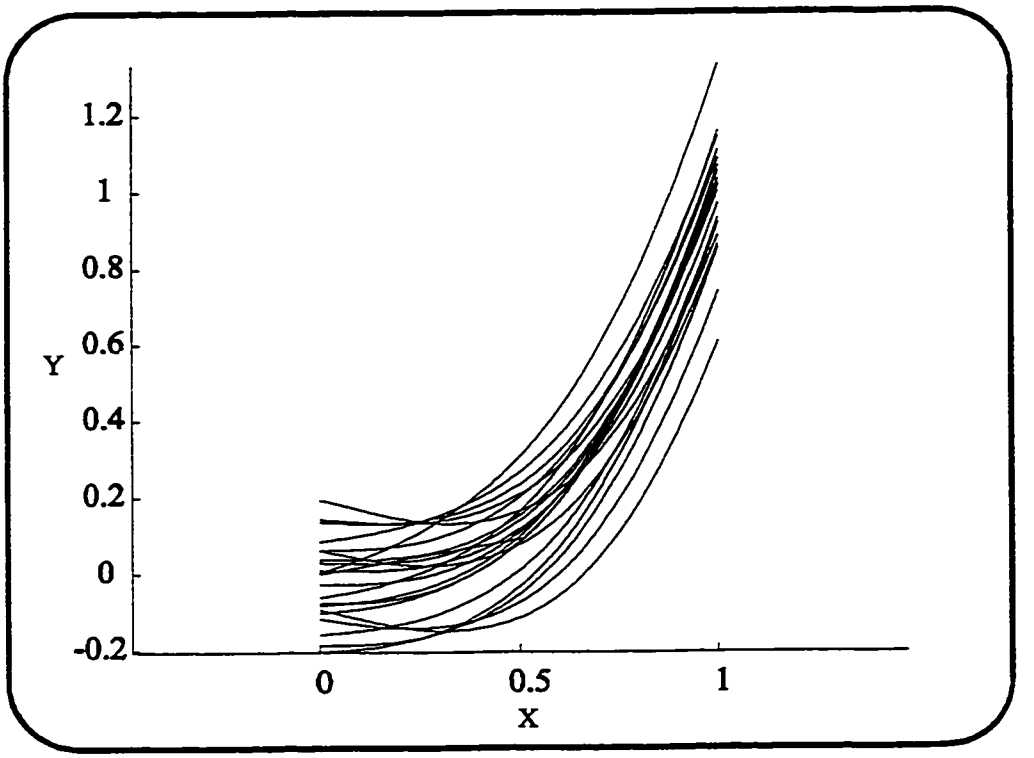


Figure 5.4 A realization of the Monte Carlo approach for Equation 5.14

Although the model provides uncertainty indicators and is easily extended to all processes, it is computationally intensive and the selection of an appropriate probability density function for the data values may be questionable (See Chapter 3).

5.7.2 Simplified uncertainty model

Sections 4.5 and 4.6 highlighted the procedures for determining rigorous analytical

uncertainty models of linear objects. Similar methods can be implemented for determining the uncertainty of curvilinear objects. However, a simplified method has been suggested for the cases when storage and simplicity are of primary concern. In the proposed method, a confidence level is selected first and is used to create a buffer around the curvilinear boundary. The buffer zone constructs a region so that the true line is located inside the threshold with the specified probability level. Figure 5.5 depicts two regions around the fitted model.

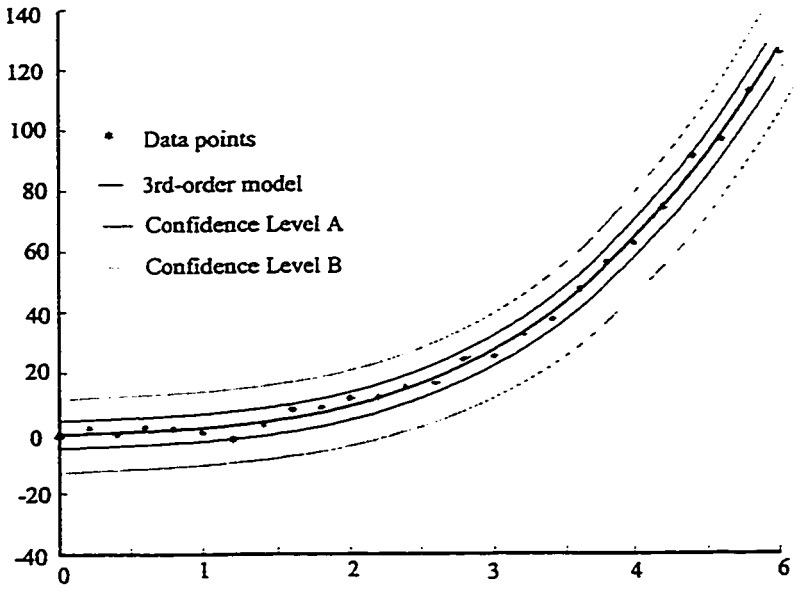


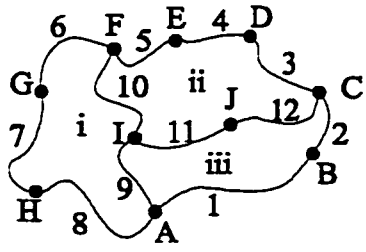
Figure 5.5. Regression model with the visualization of approximation error

If the observational error is assumed to be constant along the function, then a similar shape will result.

5.8 Logical Spline Representation in GIS Database

Once the parameters of the spline functions, and their approximating uncertainty have been identified, they should be stored in the database. It is recommended that the

following modification be implemented in the logical data model presented in Section 4.8 (see Figure 5.6).



A, B, C - Breakpoints/Nodes
1, 2, 3 - Splines
i, ii, iii - Polygons

Poly #	Spline #	Uncertainty (%)
i	1, 7, 8, 5, 6	C_I
ii	2, 9, 7	C_{II}
iii	3, 4, 8, 9	C_{III}

Polygon-Spline List

Spline #	From Point	To Point	Spline Para.	Uncer.
1	A	B	$a^1_0, a^1_1, a^1_2, a^1_3$	ε^1
2	B	C	$a^2_0, a^2_1, a^2_2, a^2_3$	ε^2
3	C	D	$a^3_0, a^3_1, a^3_2, a^3_3$	ε^3

Spline-Point List

Point #	X Coord.	Y Coord.	Uncertainty
A	X_A	Y_A	C_A
B	X_B	Y_B	C_B
C	X_C	Y_C	C_C

Point Coordinate List

Figure 5.6. A logical representation of geospatial data and their uncertainty

In the above figure, most data reduction occurs in the Point-Coordinate table because some of the digitized points will be deleted and only those that construct the breakpoints will remain. The approximated functions are uniquely identified by their parameters, which is stored in the Spline-Parameter field of the Spline-Point table. If the simplified uncertainty model is applied, the uncertainty column will be affected too. Assuming a constant width throughout the approximation, the uncertainty column of the Spline-Point table can be removed. Instead, a tag can then be used to indicate the uncertainty of the approximated function. If the uncertainty of the approximated function varies from one spline to another one, the uncertainty column of the Spline-Point should remain as

indicated in Figure 5.6. The magnitude of the data reduction varies from one case to another. The following case studies determine the data reduction ratio in each case. Regardless of the data reduction effect, the appearance of the boundary lines are much closer to reality than the jagged representation of straight line segments.

5.9 Case Studies

Several case studies have been analyzed to examine the methods suggested for polynomial regression, model identification, and spline approximation and their uncertainties. Two studies will be highlighted here.

5.9.1 Matlab data sets

The first case study uses the data sets that resemble the shape of Matlab's Titanium data [Matlab 1992]. Since the variation of the data were claimed to be difficult to model, it will be used to illustrate the rigor of the proposed methodologies. It is worth noting that the units of the data sets are assumed similar in the perpendicular directions, since the shape of the data sets is the important aspect. Original data sets recorded a property of titanium measured as a function of temperature.

5.9.1.1 Polynomial regression

Figure 5.7 presents the data points together with their fitted algebraic polynomials. As the figure shows, adding to the order of the polynomial by one does not substantially reduce the errors magnitudes when the regression polynomials reach higher order. Moreover, the greater the order of the polynomial, the more resonance the shape of the model took.

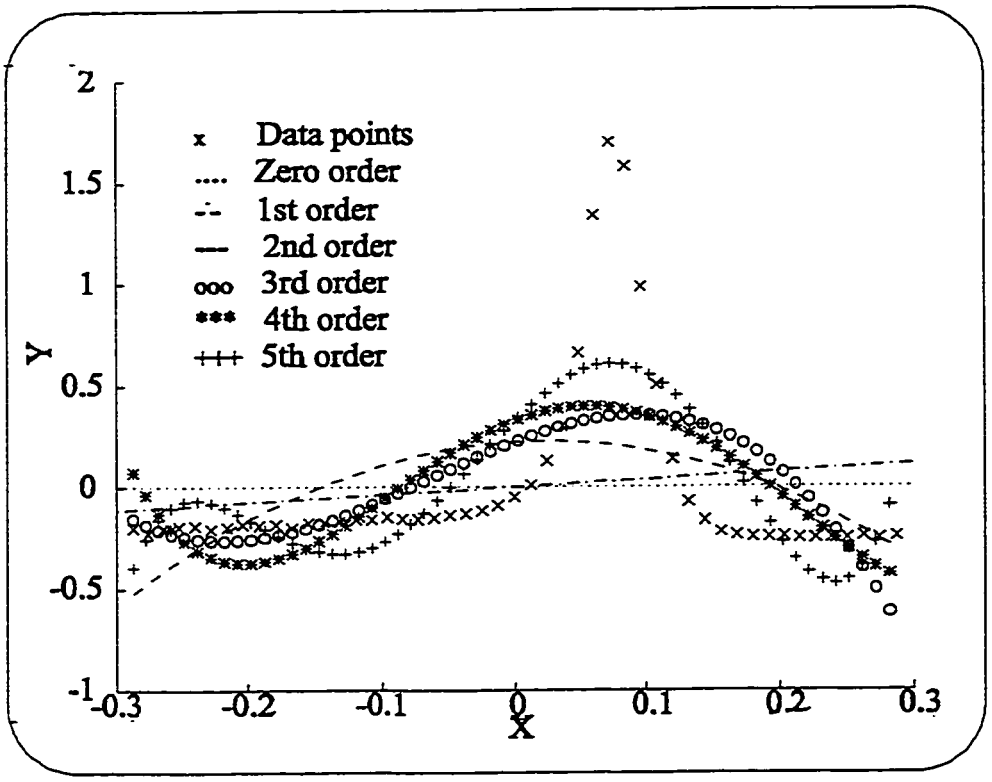


Figure 5.7 Titanium data sets together with its fitted algebraic polynomials

Table 5.3 presents the results of the test for identifying the Minimum AIC Estimate (MAICE). Since the values of the AICs wave from one model to the other, the selection

Table 5.3. The values of AICs and the estimated parameters of fitted polynomials

Case No.	Order	$\hat{\sigma}^2$	AIC	a_0	a_1	a_2	a_3	a_4	a_5
1	0	0.1406	50.36	0.805	—	—	—	—	—
2	1	0.1351	51.51	0.0	0.378	—	—	—	—
3	2	0.1171	50.47	0.221	0.378	-7.70	—	—	—
4	3	0.1022	49.57	-.142	0.54	-.26	1.034	—	—
5	4	0.0983	50.75	-.617	2.409	-1.28	1.21	-0.01	—
6	5	0.0736	46.59	0.326	6.586	-19.9	-263	165.7	2323

of the best model gets difficult. Table 5.3 also shows the parameters of the fitted models, together with the model variances.

Blais [1991a] argued that in cases such as this, the selected mathematical model may not represent the data variation well, hence reformulating of the problem using different base functions might be in order.

5.9.1.2 Spline approximation

To approximate the Titanium data sets with splines and determine the data storage efficiency, procedures highlighted in Section 5.6.2 were followed. It was assumed that the maximum error should not exceed, T = 0.1. Figure 5.8 represents the results of such a trial. It was found that using seven segments can satisfy the threshold requirement.

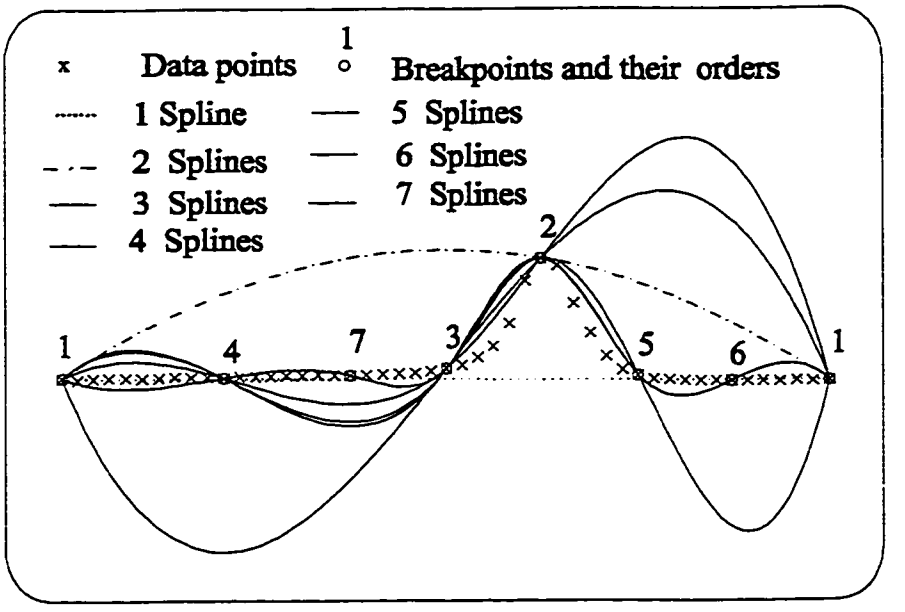


Figure 5.8. Spline approximation to Titanium data sets

This means a saving of about 84% of the storage required for the Point-Coordinate table of Figure 4.11. Splines needed only eight breakpoints, whereas straight lines required 49 points, all of which must be stored in the table. Though the addition of the spline parameters in the Spline-Point table may, at first glance, seem a burden on storage, the number of cubic polynomials required to present an arc is much less than what is required by straight line segmentation. This fact accounts for a saving of 66% on the Spline-Point table of Figure 5.6. Titanium data sets were represented by 48 straight line segments (and hence by 48*3=144 data elements), while only seven cubic spline segments (7x7=49 data elements) were needed to represent the same data set. In total, a saving of about 75% (1-73 / 291) in data storage is observed in this case. Moreover, the functions are smooth as cartographers used to represent the natural feature boundaries.

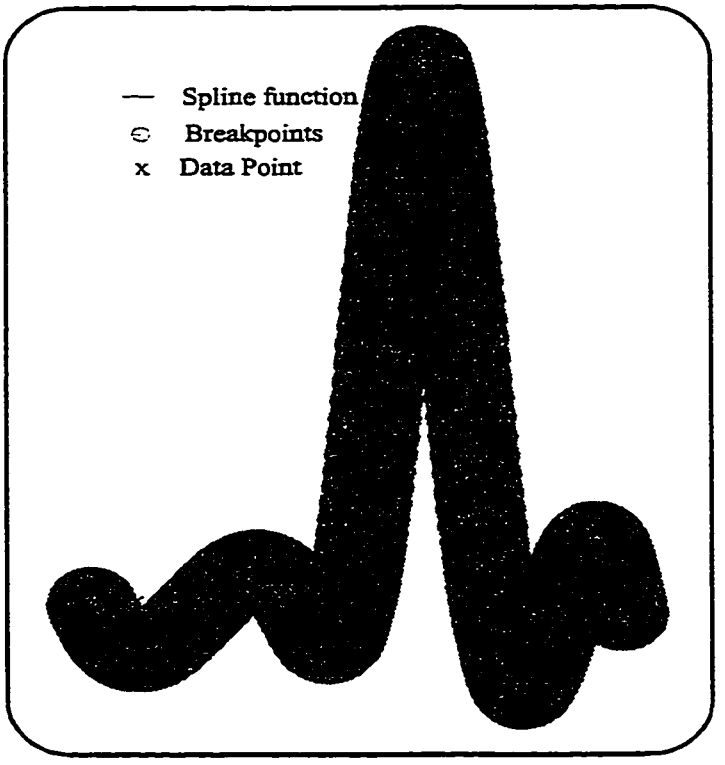


Figure 5.9 Spline approximation and its simplified uncertainty region

To show the uncertainty of the approximation function, a buffer zone of specific width (i.e. confidence level) was created along the spline as shown in Figure 5.9.

5.9.2 Digitized boundary data set

The second case study involved the data sets gathered through on screen digitization of an scanned image. Several algebraic polynomials have been used to approximate the data. Since polynomials are the best way to approximate a smooth function locally, the degree of the approximating polynomial has to be chosen unacceptably large when the function approximates a large interval such as an arc, as shown in Figure 5.10.

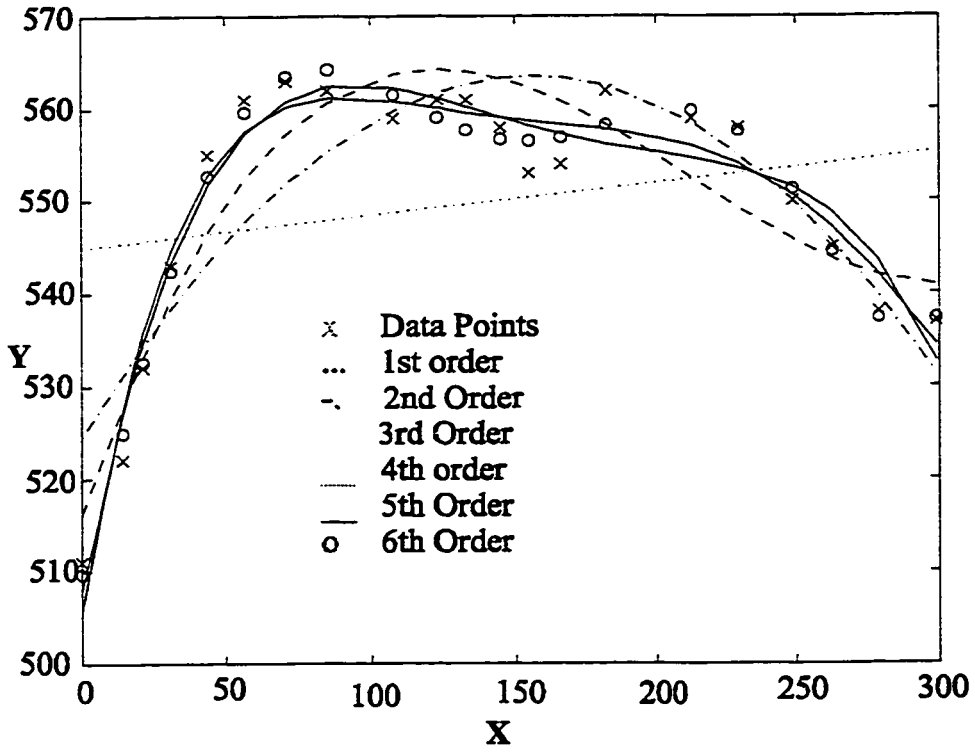


Figure 5.10. Digitized data points together with their fitted polynomials

Applying the 'principle of parsimony' in the model identification, AICs of the fitted models were computed and are presented in Table 5.4. Since the AIC values are oscillating, deciding on the best-fit model may be problematic. Table 5. 4 shows the

parameters of the fitted model and the accuracy of the model fit. It is clear that the accuracy of the model gets better as the orders of the polynomials increase, but the deviations between their variances get smaller, once the polynomial order gets larger.

Table 5.4 Parameters of the algebraic polynomials together with their AIC values

No.	Order	$\hat{\sigma}^2$	AIC	a_0	a_1	a_2	a_3	a_4	a_5
1	0	198.8	90.03	549.7	----	----	----	----	----
2	1	188.6	91.55	544.9	0.035	----	----	----	----
3	2	49.94	81.43	524.5	0.498	-0.002	----	----	----
4	3	30.42	78.91	516.2	0.905	-0.005	8.1e-6	----	----
5	4	12.28	72.64	507.9	1.567	-0.016	6.3e-5	-9.2e-8	----
6	5	10.84	75.50	505.8	1.853	-0.023	1.3e4	-3.6e-7	3.6e-10

Based on the assumption that the value of threshold should be equal to eight, T = 8 pixels, the breakpoints have been identified. Figure 5.11 demonstrates the fitted spline

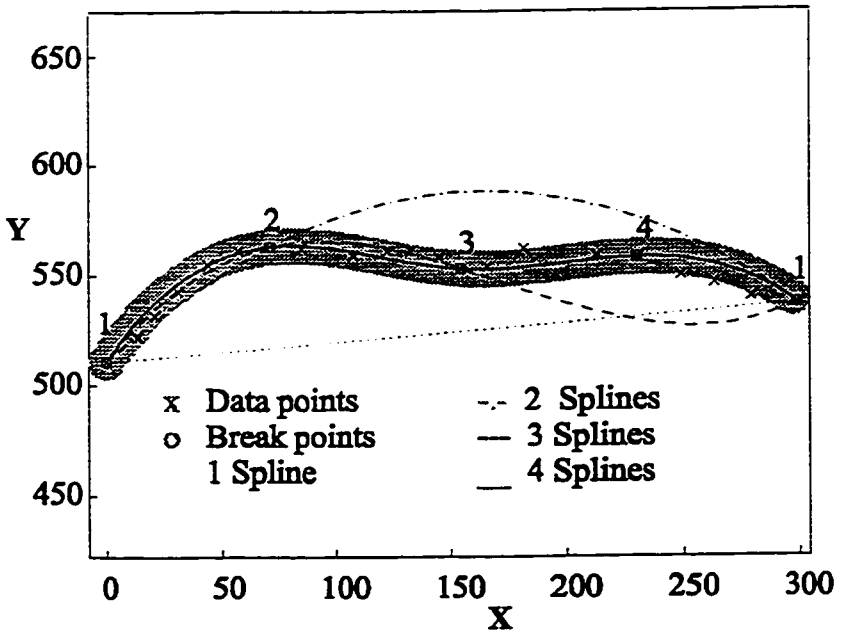


Figure 5.11 Spline approximation and its simplified uncertainty region

together with its uncertainty region. The number of breakpoints, as shown in Figure 5.11, are 5. The Spline-Point list requires only 4*7=28 data elements to fully represent the spline. Compare to Line-Point list of Figure 4.14, which needs 20*3=60 data elements, it reduces the volume of data by a factor of 53%. In total the magnitude of saving for this case study reaches the 65% level.

5.10 Practical Considerations

For determining the statistically-best function, AIC should be carefully used. Since the values of AIC may not decrease for the few first orders as expected. Without a clearly defined minimum, an alternative model formulation should be considered.

AIC is shown to be an effective mathematical tool in identifying the best model within one family of models. In the case study, it was assumed that the boundary can be represented by a polynomial regression. AIC was then used to identify the optimum order of the polynomial. The AIC cannot identify the optimum model among different mathematical models. Therefore, in the case study, AIC would probably not identify an optimal trigonometric or exponential model. In other words, information about the type of the mathematical model should come from the nature of the problem; AIC will identify the optimum model within one category of mathematical models. Moreover, an appropriate coordinate system should be assumed in the use of algebraic polynomials.

5. 11 Chapter Summary

Linear objects cover a large proportion of any GIS databases. As such their uncertainty is of vital significant since they provide criteria for the assessment of fitness for use. The

quality of the linear objects may not be determined if the functions representing the objects are not identified first.

Due to its simplicity, algebraic polynomials were selected for approximating boundary lines. AIC was then used to identify the best order and the most significant parameters of the polynomial. Simulation results indicated that a saving of about 75% could be achieved if statistically non-significant parameters of the regression were ignored.

Algebraic polynomials are best for approximating a function locally. To approximate a sequence of large data points, such as an arc, a polynomial of high order is often required, implying possibly large oscillations between data points

Piecewise polynomials were used to approximate data sets of relatively large extent. A logical data model have been proposed for keeping the parameters of the approximated function. The average results of the tests showed that a data reduction of about 70% may be achieved through the use of the proposed spline approximations. It is also indicated that the magnitude of the data reduction directly related to the amount of error tolerance an operator considers acceptable.

Two types of uncertainty models have been proposed for the approximated function; simulation, and simplified analytically based models. The simplified model reduced the data storage significantly, though it may not be categorized as rigorous model.

CHAPTER 6

A PROPOSED DECISION SUPPORT STRATEGY IN GIS

6.1 Overview

So far, this thesis has evaluated the sources of errors, the forms these errors may take, and the theories that may be used to handle the errors. This thesis has also reviewed and proposed some of the means by which errors may be measured, modeled and communicated. However, even if the uncertainties of GIS products are modeled and presented properly, the question of "how to handle the uncertainty" still remains [Hunter 1993].

Accordingly, the objective of this chapter is to present a strategy for handling uncertainty in GIS as a decision support system. The chapter begins with an introduction to the proposed strategy, and examines its three key elements. This includes the uncertainty assessment, communication to users, handling through reduction and absorption. Different data transfer standard formats are also discussed. The application of the strategy in a case study is tested and the results evaluated.

For implementation purposes, an uncertainty management prototype is designed to test the concepts and methods developed in this thesis. The prototype is used to deal with uncertainty of different spatial objects, such as points, straight line segments, curvilinear segments, and polygons. The chapter ends with a brief description of the prototypes implemented. The software package together with its Graphic User Interface (GUI) have been prepared to facilitate the application of the strategy.

6.2 A Decision Support Strategy

The proposed decision support strategy is presented in Figure 6.1. The strategy is comprised of the key stages of problem identification, evaluation of alternative methods, and implementation of the adopted solution to simplify a decision.

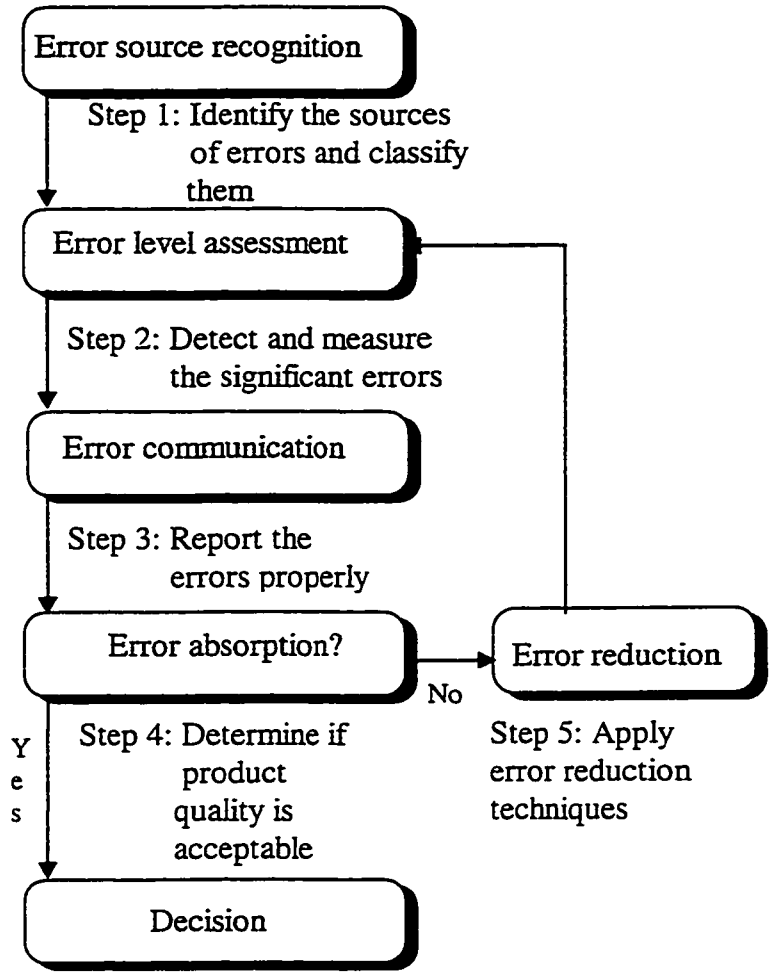


Figure 6.1 The proposed decision support strategy

It is clear that the way in which a GIS is used varies between users with different skills and responsibilities, similar to the effect of different types of decision and application areas. The proposed strategy not only works when the decision maker knows

the required quality indicators in advance, but also when the needed accuracy has been decided on through pre-analysis.

While the concept is relatively new to GIS, a derivation of the strategy has already been used in geodetic surveying where, for instance, proposed locations of survey stations and a priori observations together with their estimated precision, are input into a network adjustment program. The output of the program will be the estimate of the unknowns and their respective precision. Thus, if a priori information of the network configuration and their measurements do not give the required precision, alternative locations and measurements are investigated until satisfactory results are achieved. The difference between the pre-analysis approach and the proposed strategy is that the strategy is applicable for any type of data element (such as spatial, and other attribute) quality. If the quality of the selected data elements does not satisfy the application requirements, methods of uncertainty reduction or uncertainty absorption must then be considered.

6.2.1 Error source recognition and error classification

The first level of the strategy addresses the basic problem of identifying the sources of errors and classifying their forms in spatial data. The significance of this step is the recognition that every GIS application imposes a specific level of accuracy in associated data elements. Recognizing the most important source of error allows data reduction efforts to focus on the specific cause of the uncertainty. Chapter 2 clearly indicated that errors in GIS products originated in the data, system, model, operation and user.

The significant forms of uncertainty for a specific application should also be identified, since particular data reduction strategies are required for various types of uncertainty. It has already been shown that different applications will place different priorities on the various uncertainty forms (refer to Chapter 2). While positional errors are the focus of this thesis and more generally the focus of Geomatics methodology, other attribute errors are a major concern in other applications such as human geography, or perhaps the combination of the two may be significant. In addition other forms of uncertainty such as logical consistency and time require more attention in applications where topological or temporal information is paramount.

6.2.2 Error level assessment

Once the significant aspects of the errors have been identified for a particular application, the magnitude of the uncertainty should be measured and modeled. The objective of this step is to compute an uncertainty indicator. The error descriptor facilitates the assessment of 'fitness for use'. Various methods have been proposed for assessing spatial error [Alesheikh, 1997], and this thesis contributes significantly to the endeavour.

When considering a specific form of uncertainty such as positional uncertainty, the accuracy needed for one application may be drastically different from one GIS user to another. For example, a Geomatics engineer monitoring dam movement needs very accurate positional data (e.g. millimetre level) while the use of such accurate data for estimation of timber land revenue is a waste of time and money. At the other extreme, the GIS industry has witnessed many cases where some locations " ... seemed suddenly to be 2 or 3 kilometres out into the sea" due to inappropriate and inaccurate data usage

[Blakemore 1985, p.346]. Attention to error level assessment early in project is essential in reducing the risks of making inappropriate data modeling assumptions.

It is important to notice that the uncertainty modeling and uncertainty analysis offer much more than only the computation of the output error. An error analysis can also determine how significantly each individual input affects the output error. This information is quite useful, since it allows users to determine how much the quality of the output improves, given a reduction of error in a particular input. This indeed facilitates the cost-benefit analysis. Thus the improvement foreseen due to any uncertainty reduction procedures such as intensified sampling can be weighed against extra sampling costs.

6.2.3 Error communication

Ideally, the error indicator computed in the previous section should be understood by users regardless of their educational background. The existing metadata can facilitate this understanding. Several tests (psychological, aesthetical) may be conducted to determine the best medium for presenting uncertainty. Methods of communicating uncertainty vary from one application to the another [Fisher, 1994]. Since most people are familiar with static visualization (i.e. conventional maps) to convey a model of the Earth, this thesis practiced such a method in communicating the uncertainty model. However, animation, holographic, and sonic uncertainty representations are attracting researchers [Pang et al. 1996].

The concept of reporting the quality of a product is not new to Geomaticians, with some of the earliest forms being positional accuracy statements found on hardcopy topographic maps. However, it is just recently that GIS users recognized the need for

standardized data quality representation [Plunkett and McKenna, 1996]. To represent the fitness of data for a specific application, some measures of accuracy of the data (metadata) should be provided.

As for digital spatial data, the first example of quality reporting was perhaps implemented by Statistics Canada over a decade ago to inform users about the accuracy of the data that they were using [Lundin et al. 1989]. Although during the first era of digital data transfer the quality report was not mandatory, it is fortunate that, this situation is now rapidly changing. In the United States, considerable time and effort have been devoted to the development of the US Spatial Data Transfer Standards [Fegeas et al. 1992]. Other countries, such as Australia, are poised to adopt the standard, albeit with local modifications. Regarding data accuracy and quality, the standard has specific provisions for the following components: data lineage, positional accuracy, attribute accuracy, logical consistency, and completeness. The objective here is that this information will constitute a better data quality report that gives sufficient information to make users able to assess the appropriateness of data.

Canada has long been active in the development of transfer standards such as the Map Data Interchange format (MDIF) and the Map and Chart Data Interchange Format (MACDIF). Transfer standards in Finland are being developed and special considerations have been made to record time attributes which will help users assess the currency of data [Rainio 1991]. The Hungarians have accepted the US Standards and see the two main problems as being (i) the evaluation of data quality, and (ii) the transfer of qualitative information [Divenyi, 1991]. South Africa's National Exchange Standard (NES) has

followed the US practice of data quality reporting, but only in text format. Finally, the NATO Digital Geographic Information Exchange Standard (DIGEST) embodies the same data quality components as the US standards but contains additional parameters to ensure security and privacy [Smith, 1991].

6.2.4. Error absorption

The next stage of the decision strategy deals with deciding whether product quality satisfies the requirements for the task at hand, and recognizing what can and cannot be modified. Indeed, if the quality is found to be acceptable, then the issue of uncertainty has effectively been dealt with. However if the quality is not acceptable, then further action is required. Regardless of the efforts put on the error reduction techniques, it should be realized that there will always remain some residual uncertainty which users must decide to either absorb (accept) if they wish to use the data, or reject. This is simply because all of the measurements are of limited accuracy and no model will ever perfectly reflect the real world. The amount of uncertainty absorbed can be considered to be the risk associated with using the data or product [Miller, 1992].

6.2.5 Error reduction

The purpose of this stage is to reduce the magnitude of uncertainty to an acceptable level. If the uncertainty of the product does not meet the requirements, then error reduction techniques should be applied. With regards to parcel-based systems, Bedard [1987] recognized that actions such as field checking of observations, strengthening geodetic control networks, defining and standardizing technical procedures, mandatory registration of all rights in land, and improved professional training all contribute to confirming "...

the precision and crispness in the description and location in space and time of a spatial entity" [Bedard, 1987, p. 181]. Other provisions such as the use of more rigorous data models, collecting more data, sampling with higher frequency, improving the spatial and temporal resolution, and improving procedures for model calibration may be taken to improve the accuracy of the product. However, it should be pointed out that more accurate data requires greater investment in effort and time.

If uncertainty reduction is required, all the contributing factors should be considered. Improving the accuracy of the parameter with the highest error contribution may not be the best solution. This is due to the fact that some input data may be much more expensive to acquire than the others [Agumya and Hunter, 1996]. However, in most cases it is reasonable to struggle for a balance of errors [Heuvelink, 1993]. When the error has a marginal effect on the output, then there is little to gain from collecting it more accurately. For instance, if the data gathered through a digitization process that is more sensitive to the scale of map and less so to the digitizing device, then it is more important to find a larger scale map than buying a more accurate digitizer.

The digitization example draws attention to the fact that the balance of errors must also include model error. It is unwise to spend much effort in collecting data if it is thrown away by using a poor computational model. On the other extreme, a simple computational model may be as good as a complex model if the latter needs lots of data that cannot be accurately acquired.

6.3 Case Study

In this section, an application of the proposed strategy, subject to the use of the proposed uncertainty models, is investigated. It is argued that for the application in question, the use of error estimates (determined in Chapter 4) and current computer capabilities, can provide a manager with effective means to handle the uncertainty of the product.

The problem is to determine the area of a polygon together with its accuracy. This issue can be of critical concern to managers in applications such as computing the timber land revenue in planning a recreational area. From a decision-maker's perspective it is vital to be aware of the uncertainty of GIS products so that proper action may be taken to reduce and absorb the risks associated with decisions based on a particular methodology and product.

The dataset for this test digitized from paper topographic maps covers an area of approximately 245 hectares of the Barrier Lake Provincial Recreation Area of Alberta, Canada, at the scales of 1:100,000 and 1:50,000. The area is enclosed by Highways 40, 68 and a contour representing 1440m as shown in Figure 6.2. What follows presents different decision-based approaches that can be selected to address the uncertainty issues in determining the area of the polygon.

6.3.1 Uncertainty ignorance

This solution, that is unfortunately used by most untrained users, ignores uncertainty totally – i.e. the 'do nothing' option. The manager simply generates the polygon from the digitized points, computes the enclosed area, and calculates the timber revenue. This method makes no attempts to provide any accuracy indicators of the product and hardly

provides an acceptable solution.

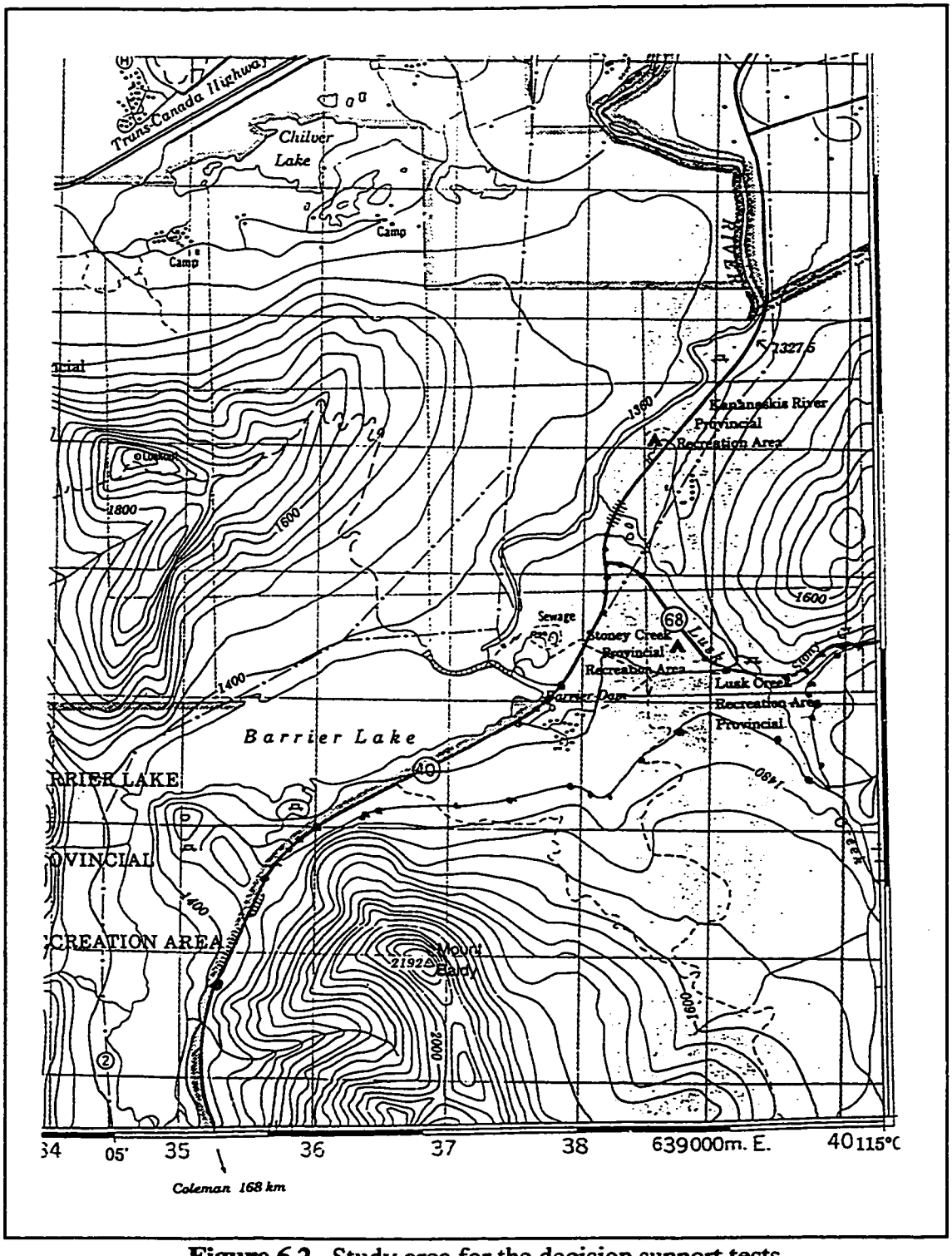


Figure 6.2. Study area for the decision support tests

6.3.2. Epsilon band

A skilled user may attempt to provide an uncertainty indicator of the boundary lines by drawing epsilon bands around them and determines the uncertainty of the polygon by dividing the area covered by the band over the polygon's area. The idea is to assume that the true positions of the boundary lines are lying somewhere within the bands with a specific probability. Figure 6.3 presents the results of such an assumption.

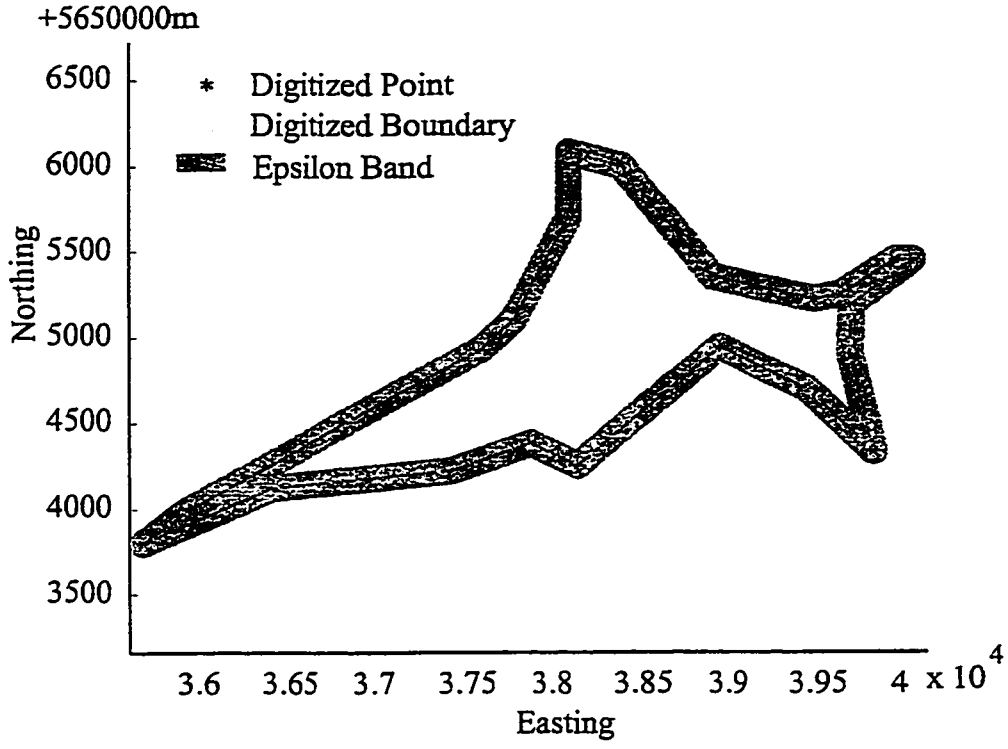


Figure 6.3. Use of Epsilon Band to portray polygon uncertainty

While the idea of the epsilon band has been evaluated in Chapter 3, it should be understood that the epsilon band is an error descriptor and does not satisfy the requirements of an error model. This is because it does not provide a means of generating various randomized versions that satisfy all the requirements of the feature

concerned – in this case, boundary lines. Therefore, it is considered that epsilon bands are inappropriate for displaying polygon uncertainty; however, portrayal of uncertainty by epsilon bands is better than not portraying uncertainty at all. The uncertainty of the polygon may then be computed, as shown in Table 6.1.

6.3.3 Error band

An improved approach that a manager may follow is to create an error band around all of the linear features, determine the area covered by such a model, and compute the polygon uncertainty. Figure 6.4 presents the results of such an assumption.

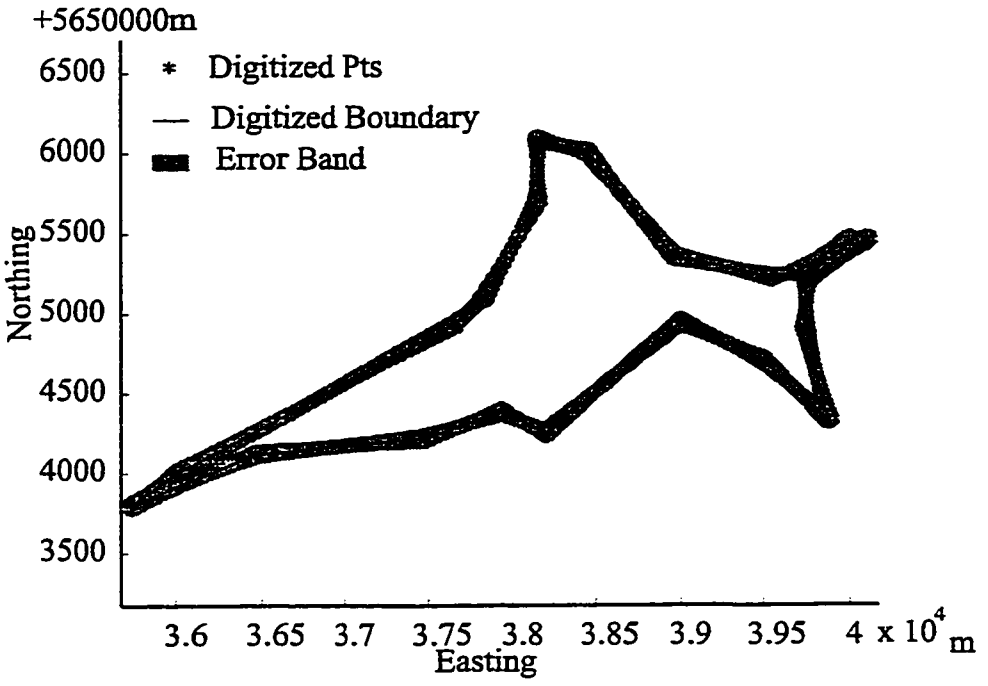


Figure 6.4. Use of Error Band to portray the polygon uncertainty

Though supported by analytical method, the amount of simplifications made to derive such a model makes the suitability of this uncertainty indicator minimal. A thorough

evaluation of this uncertainty model was made in Chapter 3. The uncertainty of the polygon is presented in Table 6.1.

6.3.4. The general uncertainty model

If the manager decides to determine the uncertainty in calculating the area of the polygon completely, he or she should use the general uncertainty models proposed in Chapter 4 of this thesis. Figure 6.5 represents the visualization of the polygon uncertainty.

The appropriateness and generality of this model have been evaluated in previous chapters. The area covered by different error models together with their uncertainty indicators are presented in Table 6.1. Information provided in Chapter 4 was used to determine the area, each model covers.

If the general model is applied the total uncertainty of the area is about 60 hectares. Assuming a revenue of $10000 per hectare, the total revenue may bear a risk of $600,000. It is worth noting that a smaller uncertainty indicator in the following table does not necessarily mean a better model since the model does not respect the general line variations.

Table 6.1 Uncertainty area and error indicator of test polygon

	Polygon (He)	Polygon Uncertainty
Area covered by E-bands A_E	33.4	13%
Area covered by error bands A_B	26.7	10.6%
Area covered by general models	60.12	24%
Area	250.595	

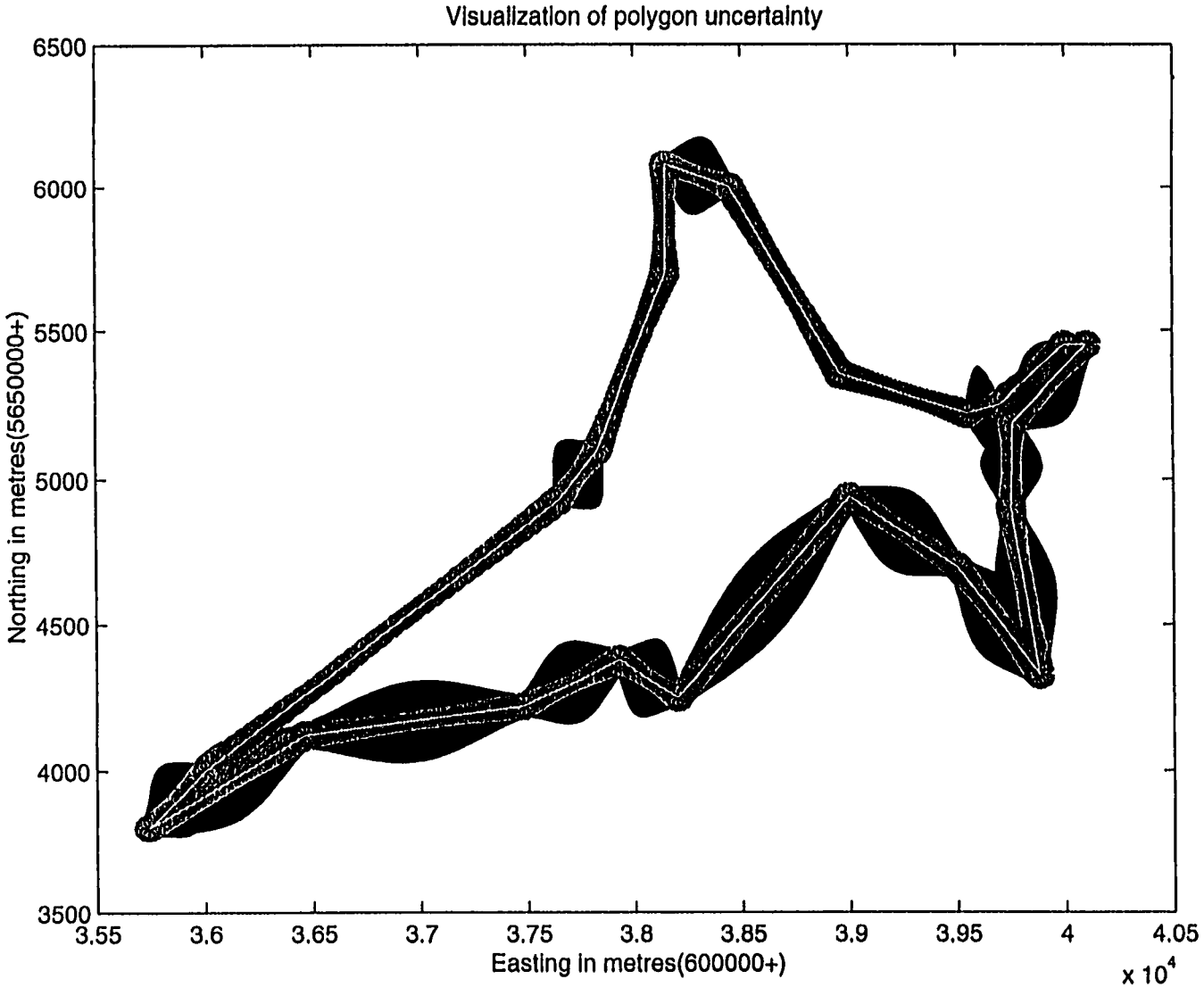


Figure 6.5. Use of general model to portray the polygon uncertainty

6.3.5 Uncertainty management

Once a manager receives and visualizes the uncertainty associated with the polygon, the important question remains as to how the information shall be applied in practice. From a decision-strategy point of view, the users must select between reducing the uncertainty in the product, or else absorbing it. For instance if the manager determines that the accuracy does not fit to the application requirements, he or she should use an error reduction strategy. The most common form of error reduction would be to recollect and reprocess digitized data from a larger scale map, e.g. 1:50,000. Figure 6.6 shows the visualization of the polygon uncertainty using the larger scale map.

The test shows if such a larger scale map is used the uncertain area reduces to 29 hectares which may mean a risk of $290,000 for the decision to be made - a considerable improvement from the previous indicator.

Once the uncertainty of the area is satisfactorily reduced, the user should absorb the remaining error. This comes from the fact that all of the data are of limited accuracy, and there always remains a slight chance of uncertainty in data due to modeling and measurement of reality.

6.4 Prototype Development

Ultimately, the success of any research rests on whether it can be effectively used by people, integrated with other systems, and improved by other researchers. However, the research community has sometimes been guilty of viewing the system as being devoid of the human factor. Recent history can provide salutary lessons by showing examples of

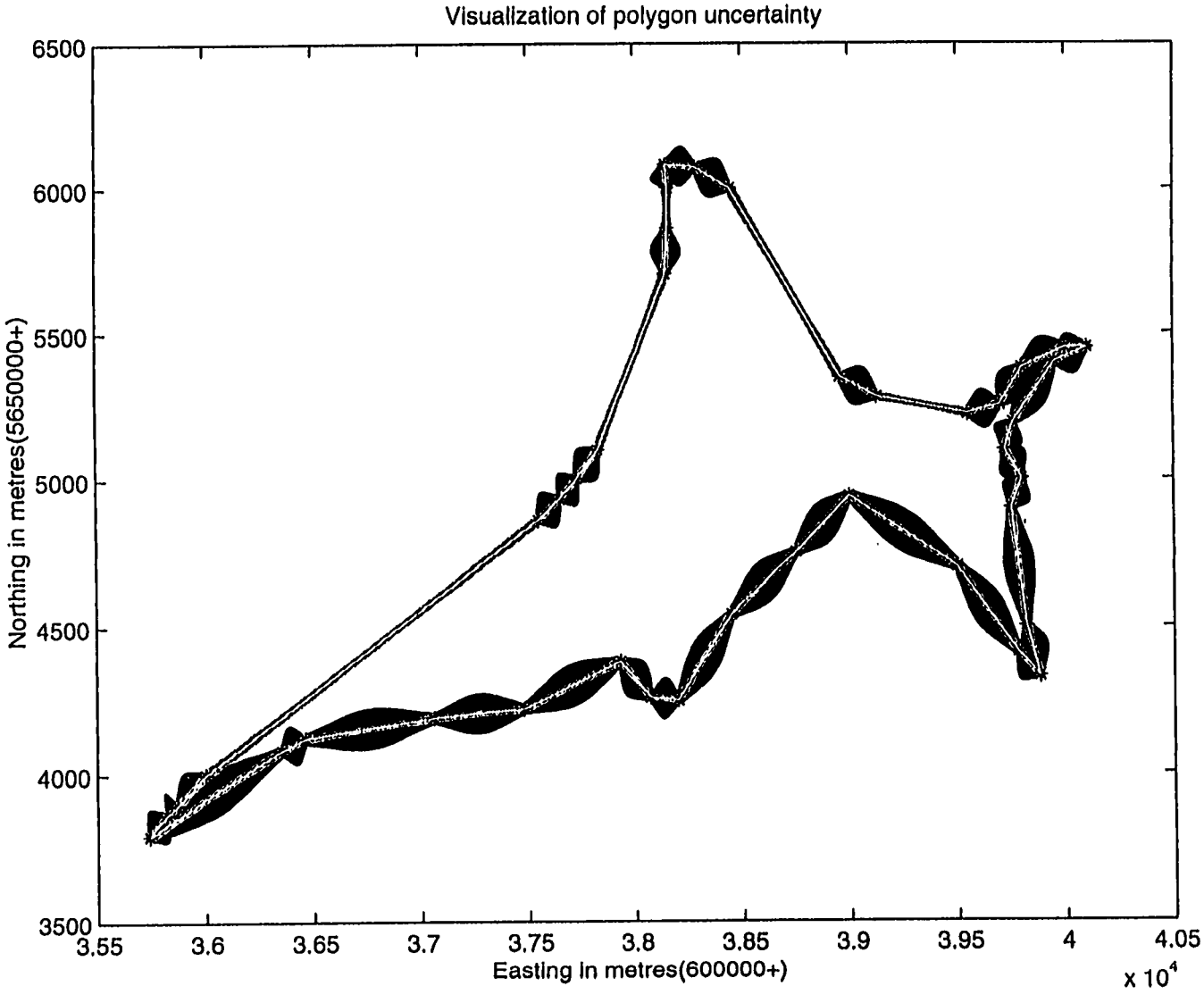


Figure 6.6. Visualization of polygon uncertainty using large scal map

what can happen if ease of use of the research products have not been fully account for in system design, system interface, and its implementation [Preece et al. 1994].

To alleviate these problems a prototype software package has been developed. The package aims at applying the concepts developed in this thesis. A Graphical User Interface (GUI) was also designed. The interface opens a gateway to the collection of software modules for modeling and visualization of uncertainties of different spatial objects, such as points, straight line segments, curvilinear segments, and polygons. The GUI to the prototype software package is shown in Figure 6.7. The interface was written in Matlab script and implemented in Matlab 4.2.c version.

Several interactions are allowed between the user and the software programs. They are: Push buttons, Command entry, Popup menus, and Editable text. A complete description of these interactions can be found in Matlab Reference Guide [Matlab, 1992]. The software can be integrated into any commercial GIS software as an uncertainty layer, or can be stand alone as the data gathered through GIS operations (e.g. digitization) might be directly introduced to the system. What follows are brief descriptions of some of the algorithms and the results of the algorithm executions. The software can be used interactively where the required data are entered through keyboard or mouse. This mode may be used for instructional purposes or when the data volume are not huge. The software makes necessary provisions to read the required data from a text file. This mode may be used if the software are integrated with currently used digitizing programs.

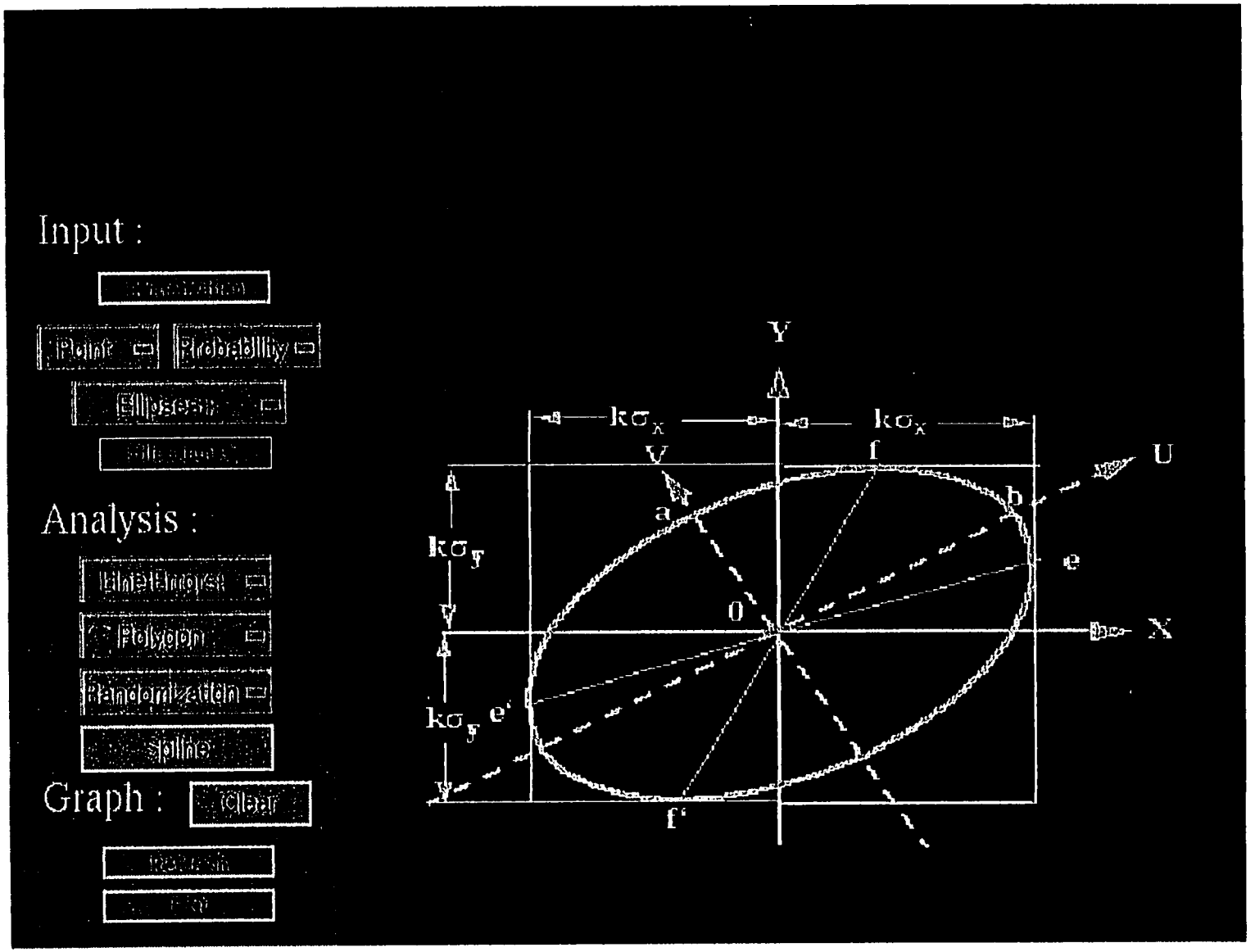


Figure 6.7. The interface for uncertainty management

6.4.1 Point uncertainty visualization

A point is the simplest form of geometric entities. Its uncertainty in 2D can be modeled and visualized by its coordinates, its covariance matrix (2x2), and a tag (C) which indicates the probability associated with the model. The software provides two options for entering these elements: interactive and reading from a text file. The following algorithm that uses Equation 3.20 is applied to display the point uncertainty. Figure 6.8 presents the results of executing the algorithm.

Point Uncertainty Visualization:

Algorithm 6.1 (PtUnVis.m)

Given the position of a point (μ_x, μ_y), its covariance matrix $C_x = \begin{bmatrix} \sigma^2_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2_y \end{bmatrix}$

and the probability level C to construct the point uncertainty visualization and return the points on the boundary of the ellipses.

function [x, y_1] = PtUnVis(μ_x, μ_y, C_x, C)

Rho = σ_{xy} /(σ_x σ_y);

x = - C * σ_x : 0.001 : C * σ_x;

Temp = sqrt(Rho^2 * x^2 - (1 - Rho^2) * C^2);

y_1 = Rho * x +Temp;

y_2 = Rho * x - Temp;

plot (x + μ_x , y_1 + μ_y , x + μ_x , y_2 + μ_y); % Draw the ellipse

fill (x + μ_x , y_1 + μ_y , 'g'); % Fill the ellipse with

fill (x + μ_x , y_2 + μ_y , 'g'); end % green color

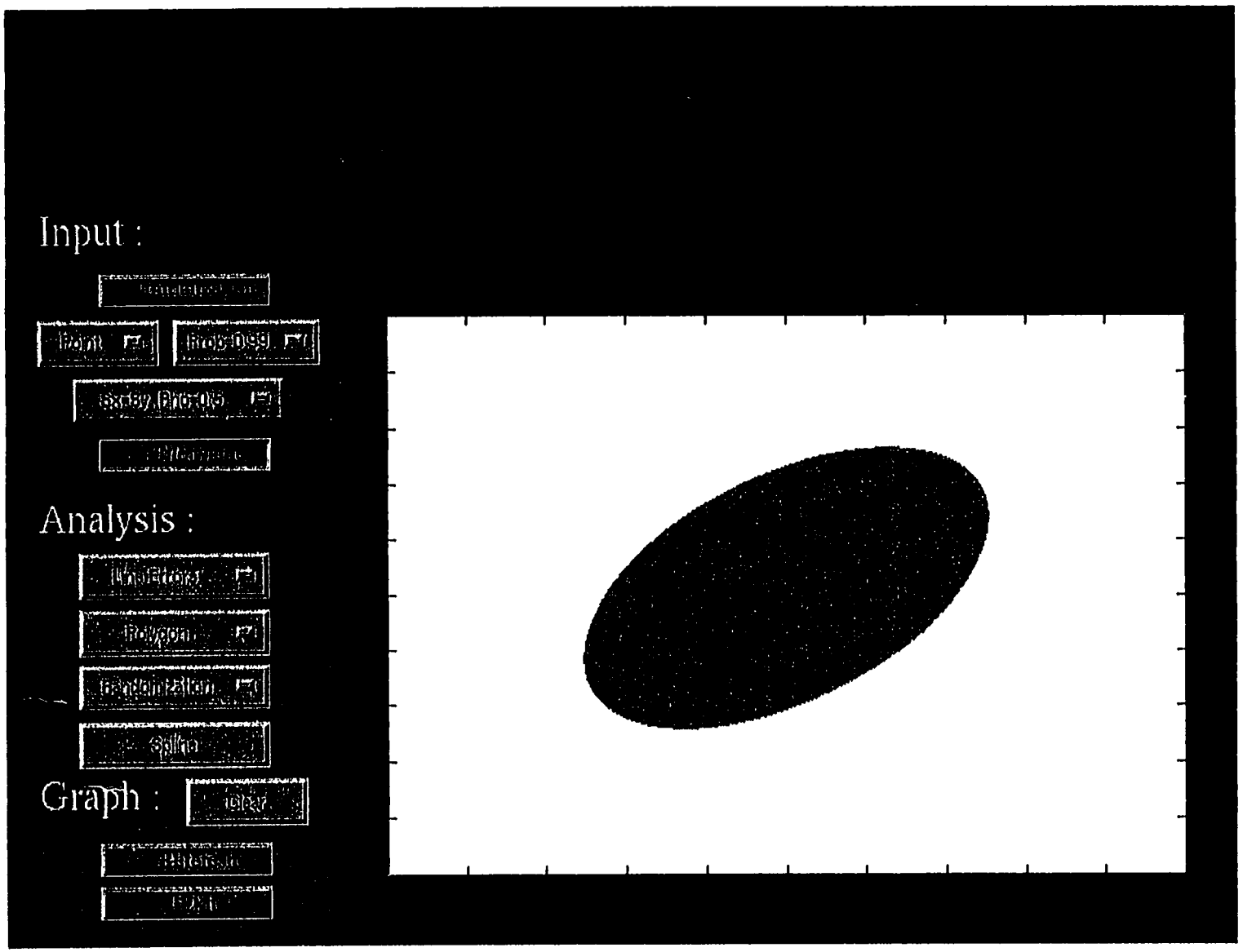


Figure 6.8. A realization of point uncertainty

6.4.2 Line uncertainty visualization: Measurement error

The observational uncertainty of a straight-line segment is defined as the envelope of the uncertainty of points along that line. As such Equations 4.3 and 4.5 are computed for several points (k) along the line and displayed on the screen using the Algorithm 6.2. The results of the execution of the algorithm is shown in Figure 6.9. The GUI allows the algorithm to be executed in both interactive and batch using a text file.

Visualization of Measurement Line Uncertainty:

Algorithm 6.2 (VisMeLU.m)

Given the coordinates of the two endpoints $\{(X_A, Y_A), (X_B, Y_B)\}$, their 4x4 covariance matrix (COV_{AB}), and the probability level C to construct the observational uncertainty of straight line segment

```
function [] = VisObLU(X_A, Y_A, X_B, Y_B, COV_AB, C)
J = zero(2, 4)                          %Initialize Jacobian matrix
for r = 0:1/k:1;
    X_u = (1 - r) X_A + r X_B ;
    Y_u = (1 - r) Y_A + r Y_B ;
    J(1, 1) = 1 - r;  J(1, 3) = r;      %Fill the Jacobian matrix
    J(2, 2) = 1 - r;  J(2, 4) = r;
    C_u = J COVAB J';
    PtUnVis(X_u, Y_u, C_u, C); end      %Call Algorithm 6.1
plot(X_A, Y_A, 'r*', X_B, Y_B, 'r*');
plot((X_A, X_B), (Y_A, Y_B), 'r-');     % Draw a line between endpoints
end
```

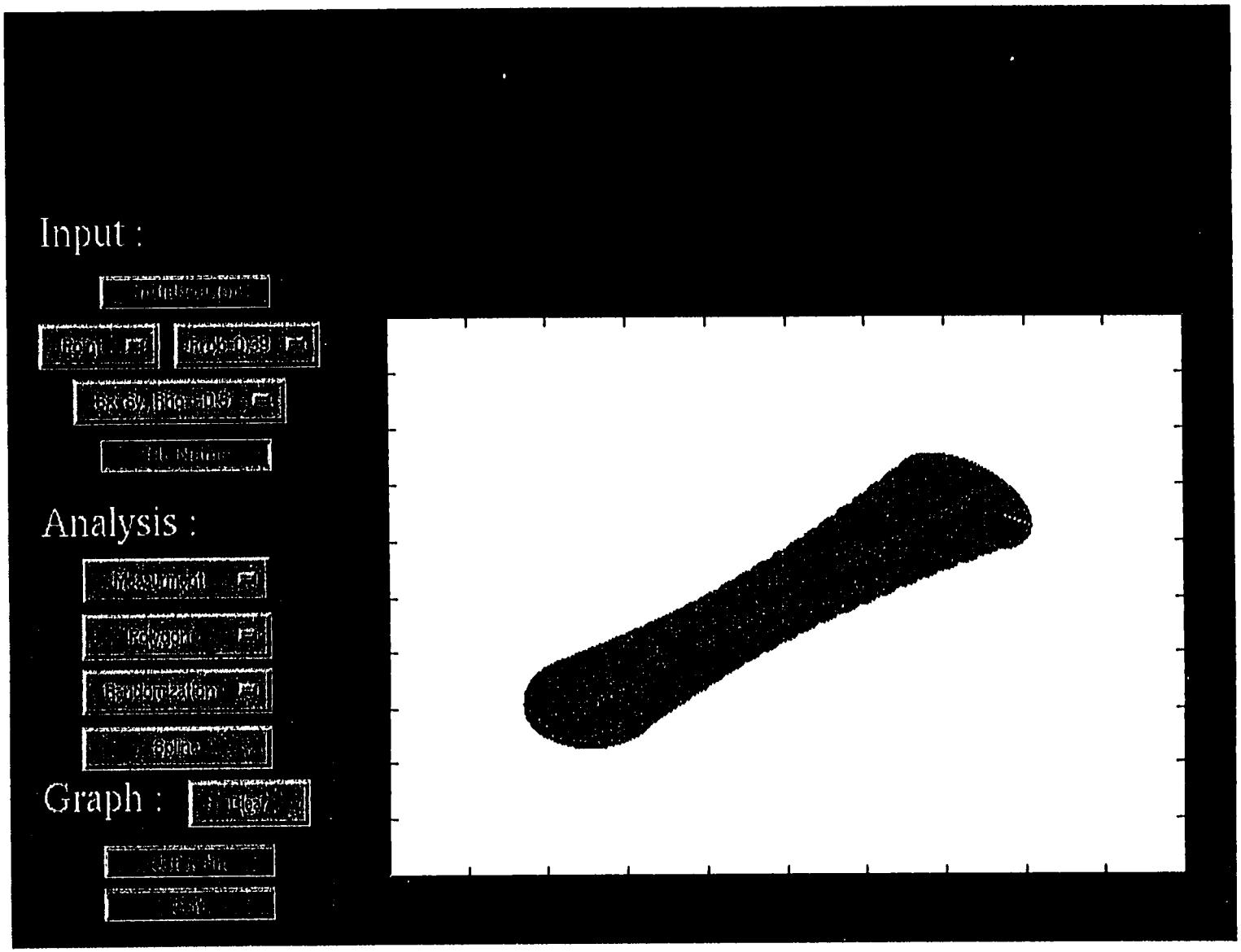


Figure 6.9. A realization of the measurment uncertainty of a line segment

6.4.3 Line uncertainty visualization: Measurement plus model errors

A straight-line segment may be affected by both measurement and model errors. If so, the magnitude of the observational error for each point along the line segment should be added to its model error. Model error is approximated by Equation 4.8 and is added to the measurement error in perpendicular to the line segment. Algorithm 6.3 represents the pseudo-code for visualizing such an uncertainty model. Figure 6.10 shows the results of executing the algorithm.

Visualization of Measurement and Model Line Uncertainty:
Algorithm 6.3 (ViMeMoLU.m)

Given the coordinates of the two endpoints $\{(X_A, Y_A), (X_B, Y_B)\}$, their 4x4 covariance matrix (COV_{AB}), the probability level C and the parameters of model error (α, β) to construct the observational and model uncertainty of a straight line segment

```
function [] = ViObMoLU(X_A, Y_A, X_B, Y_B, COV_AB, C, α, β)
m = (Y_B - Y_A) / (X_B - X_A);          % Line slope
Alfa = atan(m);  Rmalfa = [cos(Alfa)  sin(Alfa); -sin(Alfa)  cos(Alfa)];
Ralfa =  [cos(Alfa)  -sin(Alfa); sin(Alfa)  cos(Alfa)];
b = Y_B - m X_B;                        % Line intercept
J = zero(2, 4)                          %Initialize Jacobean matrix
Dist = 0;    AllXmax = [];  AllYmax = [];  AllXmin = [];  AllYmin = [];
for r = 0:1/k:1;
    X_u = (1 - r) X_A + r X_B ;
    Y_u = (1 - r) Y_A + r Y_B ;
```

```
        J(1, 1) = 1 - r;  J(1, 3) = r;              %Fill the Jacobian matrix
        J(2, 2) = 1 - r;  J(2, 4) = r;
        Cu = J COVAB J';
        [x , yl] = PtUnVis(Xu, Yu, Cu, C);     %Call Algorithem 6.1
        for i = 1: length(x),
%Determine the position of the points with maximum distance from a line
            temp = abs((m * x(i) - yl(i) ) / SQRT(m^2 + 1));
            if (temp > Dist)
                Dist = temp;
                Xmax = x + Xu ;
                Ymax = yl + Yu ;
            end
        end
%Add the model error in perpendicular direction to the line segment
        [Xmaxt; Ymaxt] = Rmalfa * [Xmax; Ymax];
% Rotate the coordinates (the direction of the line coincides with the x).
        Ymaxt = Ymaxt + α*exp(-β(r * k - k/2)^2);
        [Xmax; Ymax] = Ralfa * [Xmaxt; Ymaxt];   %Rotate back to original
        Xmin = 2x - Xmax;    Ymin = 2yl - Ymax;   %system
        AllXmax = [AllXmax; Xmax];    AllYmax = [AllYmax; Ymax];
        AllXmin = [AllXmin; Xmin];          AllYmin = [AllYmin; Ymin];
end
fill(AllXmax, AllYmin, 'b');
fill (AllXmin, AllYmin, 'b');
end
```

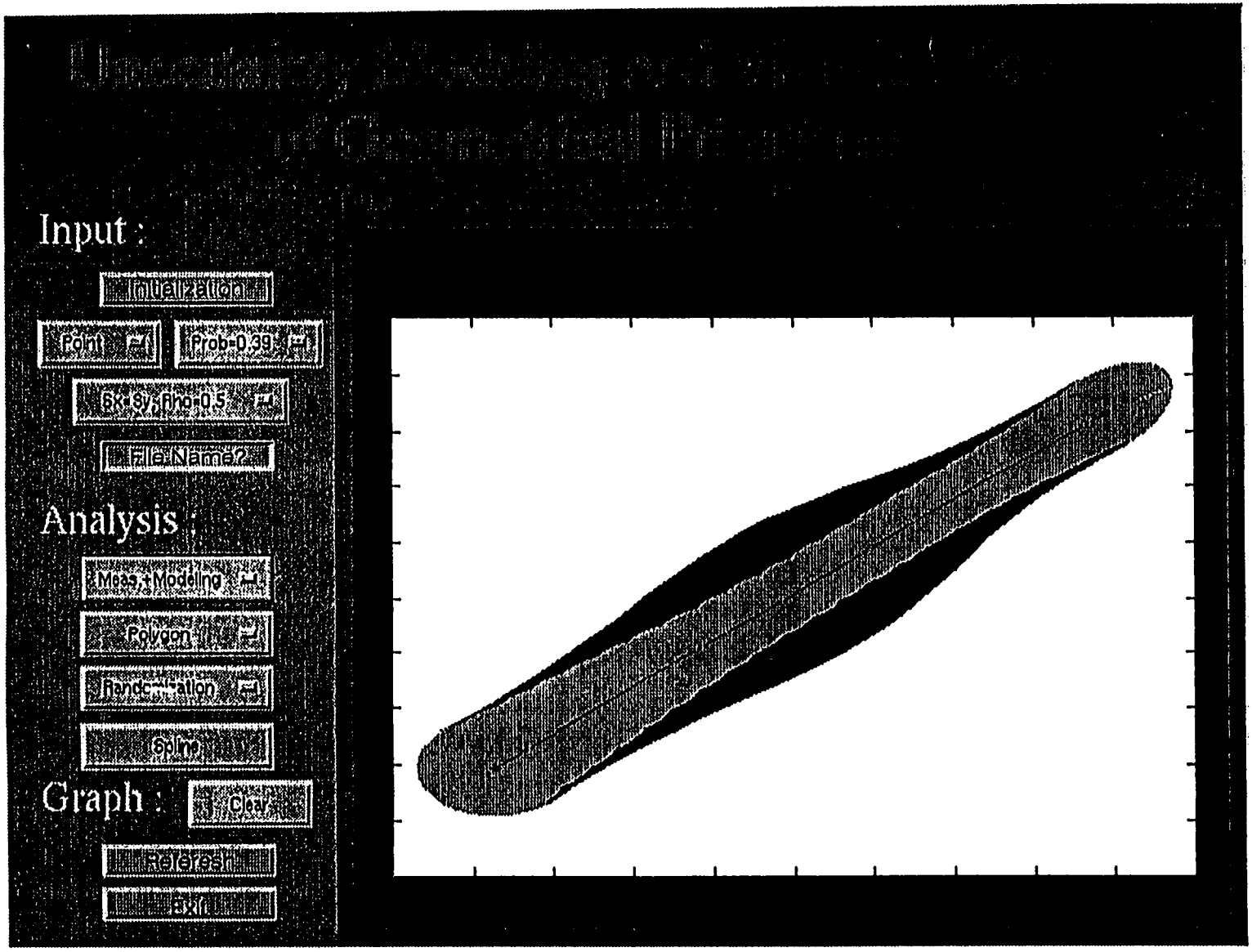


Figure 6.10. The line uncertainty visualization (measurements and model errors).

6.4.4 Line uncertainty visualization: Simulation method

A straight-line segment may be visualized as the connection of its endpoints. If so, the uncertainty of a line segment may also be visualized by connecting randomly generated endpoints. The locational uncertainty of the endpoints may be computed using Equation 3.20. Algorithm 6.4 applies such a concept. It first generates several endpoints that are constrained to Equation 3.20, and then connects the endpoints. To visualize the locational uncertainty of the line, a dense mesh may be superimposed on the random lines and the frequency of the intersected pixels displayed in color shades (See Figure 4.4). Figure 6.11 presents the results of running the algorithm.

Visualization of Observational Line Uncertainty:
Simulation method
Algorithm 6.4 (ViObLUS.m)

Given the coordinates of the two endpoints $\{(X_A, Y_A), (X_B, Y_B)\}$, their covariance matrix (C_{AB}), and the probability level C to construct the observational uncertainty of straight line segment

```
function [] = ViObMoLUS(X_A, Y_A, X_B, Y_B, C_AB, C)
Rho1 = C_AB(1, 2) / (C_AB(1, 1) * C_AB(2, 2)) ;
Rho2 = C_AB(3,4) / (C_AB(3, 3) * C_AB(4, 4)) ;
for i = 1: 10,
    XY1 = randn(1, 2) + [X_A, Y_A];
    Temp1 = (XY1(1, 1) - X_A) / C_AB(1, 1);
    Temp2 = (XY1(1, 2) - Y_A) / C_AB(2, 2));
    while (Temp1^2 - 2*Rho1 Temp1*Temp2 + Temp2^2 > (1 - Rho1^2)*C^2),
```

```
        XYl = randn(1, 2) + [X_A, Y_A];
    end,
    XYr = randn(1, 2) + [X_B, Y_B];
    Temp1 = (XYr(1, 1) - X_B) / C_AB(3, 3);
    Temp2 = (XYr(1, 2) - Y_A) / C_AB(4, 4));
    while (Temp1^2 - 2*Rho2 Temp1*Temp2 + Temp2^2 > (1 - Rho2^2)*C^2),
        XYr = randn(1, 2) + [X_A, Y_A];
    end
end
plot(XYl, XYr, 'r-');
end,
```

6.4.5 Polygon uncertainty visualization

The uncertainty of a polygon object is defined as the area that the surrounding line uncertainty models covers over the area of the polygon. The preceding line uncertainty algorithms can be used to visualize the uncertainty of a polygon object. Equation 3.15 may be used to compute the area of the polygon. To compute the area that uncertainty of each line covers; a) the line uncertainty model is computed, b) points on the boundary of the line uncertainty model that have the maximum perpendicular distance from the line are selected, and c) the area of the polygon that these vertices construct is computed. Algorithm 6.5 follows this concept. Figure 6.12 shows the output of the algorithm. The information for computing the polygon uncertainty model can be entered interactively or read from a text file.

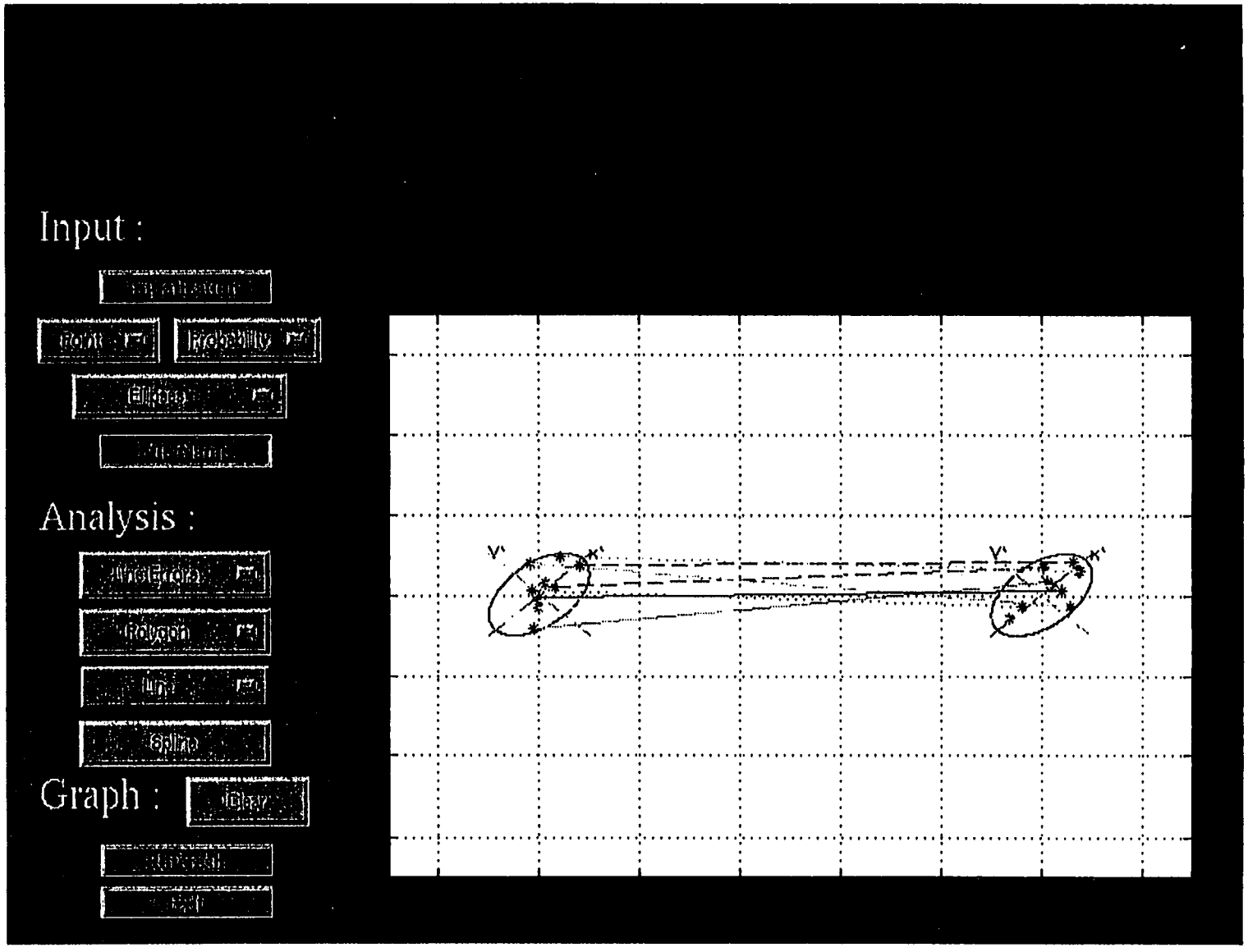


Figure 6.11. Results of running Algorithm 6.4

Visualization of Polygon Uncertainty:

Algorithm 6.5 (ViPoUn.m)

Given the coordinates of polygon vertices $X = [X_1, X_2, \ldots, X_n]$, $Y = [Y_1, Y_2, \ldots, Y_n]$, their Covariance matrix C_{n*n}, the probability level C, a series of tags indicating the line numbers that have model uncertainty tag = $[T_1, T_2, \ldots, T_m]$, and the parameters of model error (α, β) to construct the uncertainty of the polygon objects and return the position of points on the error ellipses that have maximum distance from the lines

```
function [Xendsp1, Yendsp1] = ViPoUn(X, Y, C_{n*n}, C, T)
LineArea = 0;  ntmp = length(X);
for nt = 2 : ntmp
    Xa = []; Ya = []; Xendsp1 = []; Yendsp1 = [];
    Xmax = []; Ymax = []; Xmin = [];  Ymin = [];
%Find the points that have the maximum distance from the line
    m = (Y(nt) - Y(nt -1)) / (X(nt) - X(nt - 1));
    J = zeros (2, 4);
    for r = 0.0:1/k:1;
        Xu = (1 - r) * X(nt -1) + r * X(nt);
        Yu = (1- r) * Y(nt -1) + r * Y(nt);
        J(1, 1) = 1 - r;  J(1, 3) = r;  J(2, 2) = 1 - r; J(2, 4) =  r;
        JsigmaXY = J*C(((2*nt - 3)): (2*nt), ((2*nt - 3)):(2*nt));
        sigmaU = JsigmaXY*J';
        [x , y1] = PtUnVis(Xu, Yu, Cu, C);   %Call Algorithm 6.1
%Determine the position of the points with maximum distance from the line
        for i = 1: length(x),
            temp = abs((m * x(i) - y1(i) ) / SQRT(m^2 + 1));
```

```
            if (temp > Dist)
                Dist = temp;
                Xmax = x + Xu ;
                Ymax = y1 + Yu ;
            end
        end
        for ntt = 1, length(tag)
            if (nt-1 == tag(ntt))     % If needed add model error
%Add the model error in perpendicular direction to the line segment
                [Xmaxt; Ymaxt] = Rmalfa * [Xmax; Ymax];
                Ymaxt = Ymaxt + α*exp(-β(r * k - k/2)^2);
                [Xmax; Ymax] = Ralfa * [Xmaxt; Ymaxt];
            end,
        end,
        Xmin = 2x - Xmax;    Ymin = 2y1 - Ymax;
        AllXmax = [AllXmax; Xmax];    AllYmax = [AllYmax; Ymax];
        AllXmin = [AllXmin; Xmin];          AllYmin = [AllYmin; Ymin];
    end,
end,
%Order the vertices
for i = 1: length(Xmin),
    Xmt(i) = Xmin(length(Xmin)+1-i);
    Ymt(i) = Ymin(length(Ymin)+1-i);
end,
Ya = [Ymax;Ymt'];
Xa = [Xmax;Xmt'];
%Determine the area of the line uncertainty models
```

```
    Xendsp1 = [Xa(1);Xa];
    Yendsp1 = [Ya(length(Ya)); Ya; Ya(1)];
    sum = 0;
    for j = 2: length(Xendsp1),
        sum = sum +  (Yendsp1(j+1) - Yendsp1(j-1)) * Xendsp1(j);
    end
    lineArea = lineArea + abs(sum/ 2)  ;
end
% Determine the area of the polygon
Xendsp2 = [Xends(1);Xends(1:length(Xends)-1)];
Yendsp2 = [Yends(length(Yends)-1); Yends(1:length(Yends)-1); Yends(1)];
sum = 0;
for j = 2: length(Xendsp2),
    sum = sum +  (Yendsp2(j+1) - Yendsp2(j-1)) * Xendsp2(j);
end
polArea = abs(sum/ 2);
polUnc = 100 * lineArea/polArea;
title(['The polygon Uncertainty is: ', num2str(polUnc), '%']);
end
```

6.4.6 Uncertainty of points with respect to a polygon

The uncertainty of a point with respect to a polygon may be computed by determining in which probability level the point is located. Algorithm 6.6 applies this concept. The algorithm first determines the region of constant probability by its boundary vertices and then decides whether the point is located inside, on, or outside the boundary of the region. Figure 6.13 shows the results of the execution of the algorithm for a simple polygon.

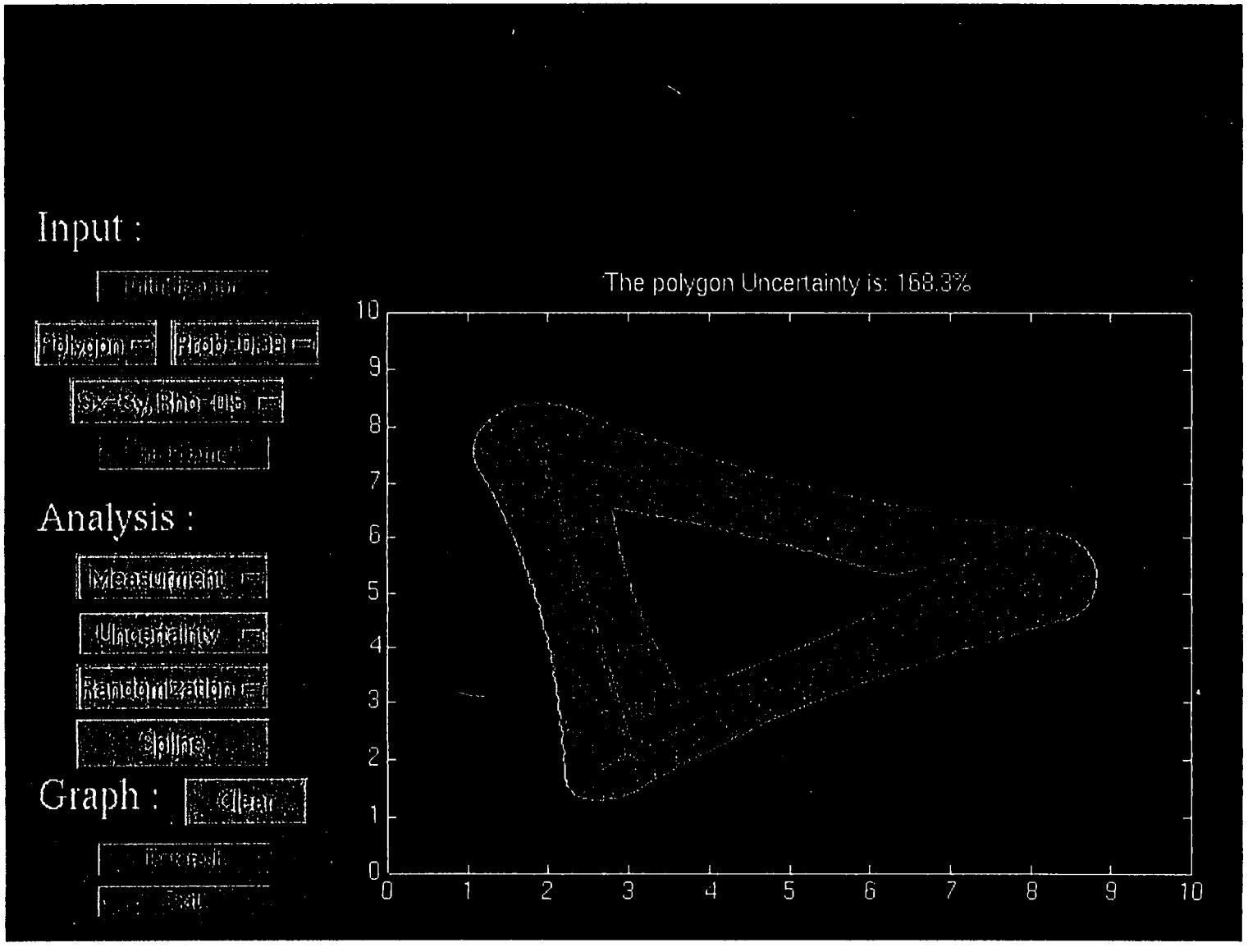


Figure 6.12. A visualization of polygon uncertainty and its value

Point Uncertainty with Respect to a Polygon:

Algorithm 6.6 (PoUnRPo.m)

Given the coordinates of a polygon vertices $X = [X_1, X_2, \ldots, X_n]$, $Y = [Y_1, Y_2, \ldots, Y_n]$, their Covariance matrix C_{n*n}, the probability level C, a series of tags indicating the line numbers that have model uncertainty $tag = [T_1, T_2, \ldots, T_m]$, the parameters of model error (α, β), and the coordinates of the test point (xpoint, ypoint) to determine whether the point is located out, or on the specified probability region.

```
function [] = PoUnRPo(X, Y, C_n*n, C, T, Xp, Yp)
[Xends, Yends] = ViPoUn(X, Y, C_n*n, C, T);  %Call algorithm 6.5
n = length(Xends) - 1;
x = Xends(1: n);
y = Yends(1:n);
plot (xpoint, ypoint, 'w+');
inpgon = 0;     smalld = 0.1;
for i=1:n-1
    if (abs(xpoint - x(i)) < smalld & abs(ypoint - y(i)) < smalld)
                                                % Point is on boundary
        title(['The selected point is on the specified confidence region']);
        return;
    end
    x1 = x(i);  y1 = y(i);  x2 = x(i+1);  y2 = y(i+1);
    if (y2 < y1)
        x1 = x(i+1);    y1 = y(i+1);  x2 = x(i);  y2 = y(i);
    end
    if (ypoint > y1 & ypoint <= y2)
```

```
    dx = x2 - x1;
    if (abs(dx) > smalld)
        xint = x1 + (ypoint - y1) * dx / (y2 - y1);
        if(xpoint - xint < 0.0)
            inpgon = 1 - inpgon;
        else
            if((abs(xpoint - xint)) < smalld)
                title(['The selected point is on the specified
                confidence region']);
                return;
            end
        end
    else
    if (xpoint - x1 < 0.0)
        inpgon = 1 - inpgon;
    else
        if(abs(xpoint - x1) < smalld)
            title(['The selected point is on the specified confidence region']);
            return;
        end
    end
end
else
    if(abs(ypoint - y1) < smalld & abs(ypoint - y2) < smalld)
        if (xpoint >= x1 & xpoint <= x2)
            title(['The selected point is on the specified confidence region']);
```

```
                return;
            end,
            if (xpoint >= x2 & xpoint <= x1)
                title(['The selected point is on the specified confidence region']);
                return;
            end
        end
    end
end
if inpgon == 2
    title(['The selected point is on the specified confidence region']);
    return;
elseif inpgon == 1
    title(['The selected point is on the specified confidence region']);
    return;
elseif inpgon == 0
    title(['The selected point is outside the specified confidence region']);
    return;
end
```

6.4.7 Data reduction through spline approximation

Splines can be applied to reduce the data volume and to smoothen the boundary of the geospatial features. Algorithm 6.7 applies the procedure highlighted in Section 5.6.2 recursively. It first approximate all the data points with a straight line segment connecting the first and the last points. It then computes the maximum error. If the magnitude of the maximum error is greater than the specified threshold it considers the point with maximum

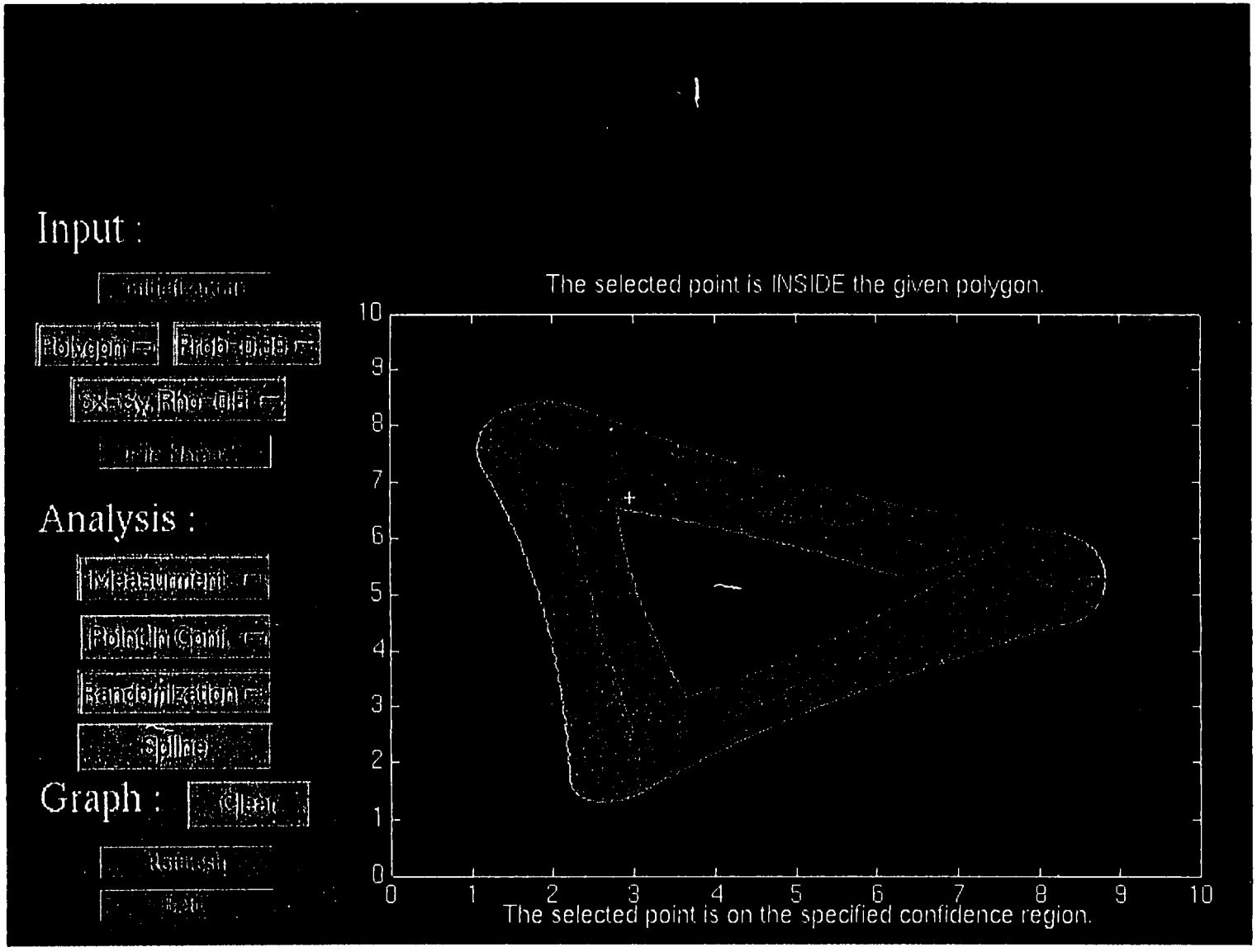


Figure 6.13. The uncertainty of a point with respect to a polygon

error as the new nodes. The algorithm continues till the error gets less than the predefined threshold. Figure 6.14 shows the results of the algorithm execution for some digitized points.

Data Reduction Through Spline Approximation:

Algorithm 6.7 (DaReTSA.m)

Given the coordinates of digitized points Xd = [X_1, X_2, ..., X_n], Yd = [Y_1, Y_2, ..., Y_n], and the value of the threshold T to construct the spline approximation and return the break points location, the coefficients of the spline, the number of break points, and the order of the spline;

```
function [breaks, coefs, l, k] = (Xd, Yd, T)
step 1;
X = [X1, Xn]; Y = [Y1, Yn];
%Connect the two points;
m = (Yn - Y1 ) /( Xn - X1);     %the slope of the line
b = Yn  - m Xn;                 %the intercept of the line
%Determine the approximation error
    Dist = T;
    for k = 2: length(Xd) -1
        error = abs((m * Xk - Yk)/(sqrt(m^2 + 1));
        if (error > Dist)
            Dist = error;   Xtmp = Xk;    Ytmp = Yk;
        end
    end
    if (Dist < T)
```

breaks = X; coefs = [m, b]; l = length(X); k = 1;

else

X = sort([X; Xtmp]) ; Y = sort([Y; Ytmp]);

step 2;

$l_0 = 1;\ m_0 = 0;\ z_0 = 0;$

for j = 2: length(X) - 1;

$$\alpha_j = 3(Y_{j+1}(X_j - X_{j-1}) - Y_j(X_{j+1} - X_{j-1}) + Y_{j-1}(X_{j+1} - Xj)) /$$

$$((X_{j+1} - Xj)(X_j - X_{j-1}))$$

$$l_j = 2(X_{j+1} - X_{j-1}) - (X_j - X_{j-1})m_{j-1}$$

$$m_j = (X_{j+1} - X_j) / l_j$$

$$z_j = (\alpha_j - (X_j - X_{j-1})z_{j-1}) / l_j$$

end,

$l_n = 1, \quad z_n = 0;\ c_n = z_n$

for j = length(X) - 1:-1:1,

$$c_j = z_j - m_j c_{j+1};$$

$$b_j = (Y_{j+1} - Y_j) / (X_{j+1} - X_j) - (X_{j+1} - X_j)(c_{j+1} + 2c_j) / 3;$$

$$d_j = (c_{j+1} - c_j) / 3(X_{j+1} - X_j)$$

end

%The procedure is complete for this iteration. The natural cubic spline

%in each interval is $S_j(x) = Y_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$

%Determine the approximation error

Dist = T;

for k = 2: length(Xd) -1

for l = 1:j,

error = min(Dist[(X_k, Y_k), S_j];

```
        end
        if (error > Dist)
            Dist = error;   Xtmp = Xk;    Ytmp = Yk;
        end
    end
    if (Dist > T)
        X = sort([X; Xtmp]) ; Y = sort([Y; Ytmp]);
        Go to step 2;
    end
    breaks = X;    coefs = [Yj, bj, cj, dj];      l = length(X);  k = 3;
end
```

6.5 Chapter Summary

In this chapter, a decision-based strategy for managing uncertainty in geospatial databases has been presented. Using this strategy, users should consider the significant forms of uncertainty likely to be present in their product, the ways that those uncertainties may be modeled and measured, and how they can be communicated. Finally, the users must compare the uncertainty present in the product with what they need. If it happens that the magnitude of uncertainty is unacceptable, the uncertainty must be reduced. However, no matter how precise the uncertainty reduction is carried out, there always remains a slight amount of uncertainty that should be accepted by users. The case study presented was a common one - determining the area of a polygon under the presence of uncertainty. The successful handling of uncertainty in the case study can be extended to other applications.

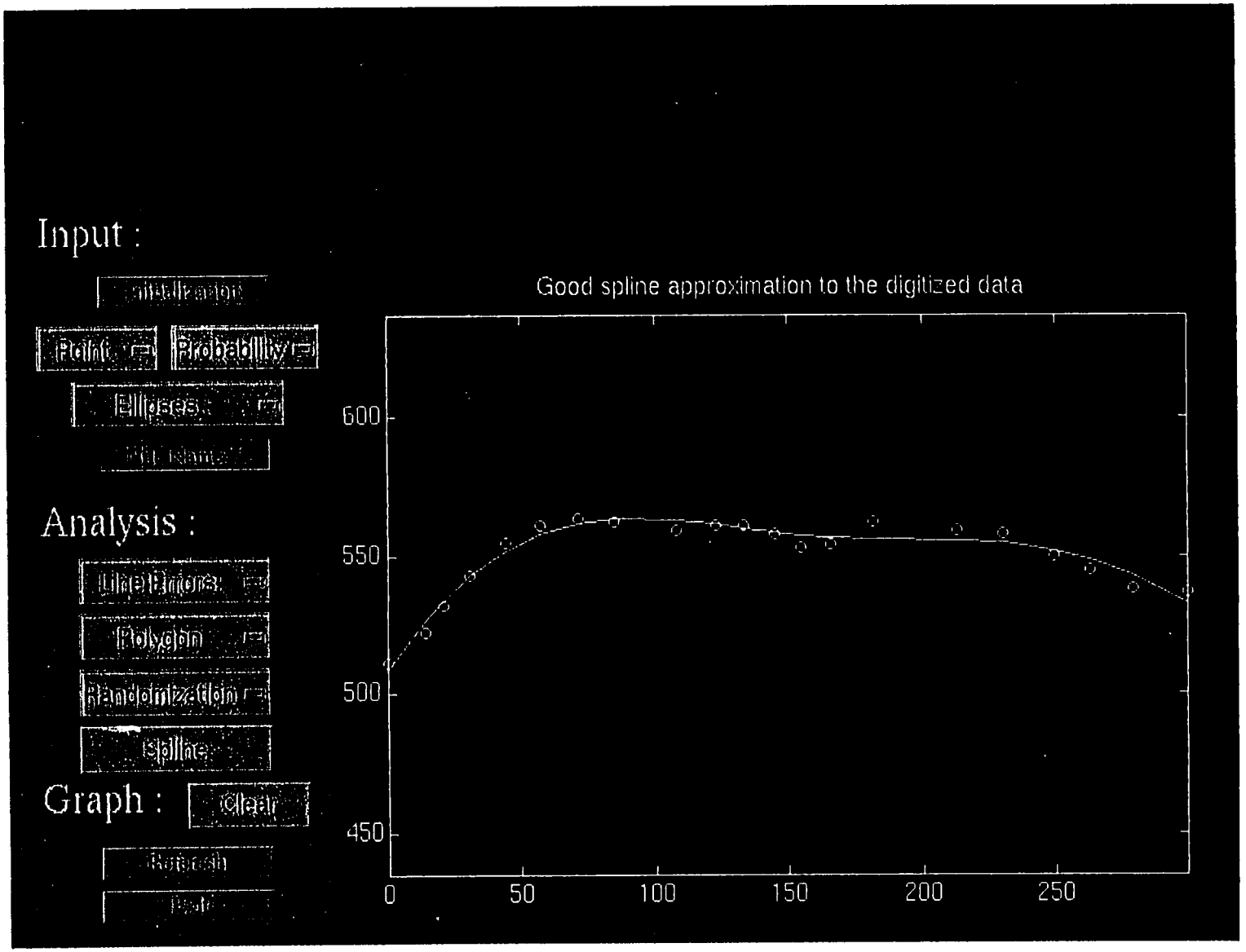


Figure 6.14. A visualization of spline approximation

For the implementation purposes, a prototype software package controlled by a GUI has been designed and tested. The GUI allows several easy interactions between the user and the computer.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

7.1 Overview

This research study explored uncertainty management in object-based Geospatial Information Systems, a topic of considerable concern to users of GISs throughout the world. This investigation included evaluating uncertainty related theories, assessing available uncertainty models, improving the uncertainty models of geometric primitives, incorporating model error into modeling processes, identifying the statistically best function representing curvilinear boundaries, and analyzing methods to reduce the data volume in GIS databases. Furthermore, a strategy was proposed for handling uncertainty in spatial databases. The proposed models and design methods for uncertainty management were implemented in software packages and successfully tested.

In summary, the research clearly has both practical and theoretical significance for users and managers since it not only presents a conceptual framework for uncertainty management, but also provides required implementation considerations. Moreover, the research has an added significance for spatial database system designers interested in incorporating uncertainty reporting in their own systems, and to those people concerned with modeling and communicating the impact of data uncertainty on spatial decision making. The following conclusions can be drawn from this dissertation:

7.2 Conclusions

- Handling uncertainty in GIS is of vital importance if the GIS technology is to continue advancing at the current rate. In spite of little research in uncertainty management in GIS, the revolution of the digital era, the assessment of fitness of data for their intended use, the mandatory spatial data transfer standards, the advancement in science, and the protection of GIS growth, forces the GIS discipline to further expand the study of uncertainty in GIS.

- The effect of error taxonomy in uncertainty handling in GIS is essential. The proposed three-part taxonomy of uncertainties which synthesized the dimensions of the uncertainty problem, and distinguished between error sources and the forms that they might take is intended to facilitate the management of uncertainty.

- Among several theories that have been considered to handle uncertainty, geostatistics and probability theory can be used to manage those uncertainties caused by random components. Fuzzy set theory may be applied to handle vague concepts, e.g. linguistic variables, while evidential theory may be of use to manage the uncertainty due to information incompleteness. It is argued that a single theory that can handle all types of uncertainty is yet to come.

- Different methods exist in GIS uncertainty modeling. Four classes of methods have been identified and formulated. They are: analytical, Monte Carlo, empirical, and error descriptors. It is elaborated that the analytical method is just an approximation technique. The Monte Carlo method requires several iterative operations and hence,

the method may not be efficient in practice. Empirical method is based on comparing the object with its 'true' value, making it time and money consuming. Error descriptors cannot simulate the possible locations of the spatial objects; therefore, they lack the conditions of being statistical error models. It is concluded that the selection of the best method depends on the application, time, and cost.

- Relatively few models or procedures exist by which errors may be quantitatively assessed. Even those models that were expected to identify the behavior of uncertainty lack generality. Communicating uncertainty to GIS users should be improved by implementing new technologies such as visual or sonic aids.

- Current uncertainty models of straight line segments are enhanced by considering all the significant factors: measurement and model errors. The theoretical comparison among the proposed line uncertainty model and the previous models show that the former is more general, and as such can handle general variations that a line segment may take. It is also shown that the error band and epsilon band model are special cases of the proposed model.

- The empirical comparisons between the proposed line uncertainty model and the epsilon band model revealed that the magnitude of uncertainty of linear objects have been underestimated by the later model. The difference is attributed to the magnitude of the model error and correlation among the endpoint coordinates. The value of the underestimation was magnified once the comparison made between the proposed model and the error band model.

- Using the proposed line uncertainty models, an uncertainty model for polygon objects is proposed. Using the polygon uncertainty model, the solution to the point-in-polygon problem is enhanced by probability statements.

- Due to their simplicity, algebraic polynomials are proposed to represent boundary lines. AIC is used to parsimoniously select the significant model parameters. Numerical results reveal that the application of the method for a large interval causes unwanted resonance. As the order of the polynomials gets large, numerical instability may occur which needs extra care.

- The proposed recursive procedure of parameter selection can make the data reduction level reach 75%. However it requires much computational efforts.

- Data reduction and the smooth appearance of curvilinear boundaries are achieved once they were approximated by nonlinear splines, and represented by their parameters in the GIS databases. Numerical results show that in average 70% storage efficiency is observed once spline parameters stored instead of the coordinates of points. It is argued that the amount of data reduction directly depend upon the presumed error threshold. The proposed data structure for saving the uncertainty information associated to geometrical primitives uses the current relational model. As such its application to off-the-shelf GIS is simple.

- A general strategy for handling uncertainty in GIS is formulated. The recursive methodology aims at reducing the magnitude of error until the predefined uncertainty level is reached. The developed prototype software package assists the

implementation of the proposed strategy. The gateway to the software packages was established through a GUI. By using the software, users are able to visualize the uncertainty associated with each geometric primitive. Moreover, due to its user friendliness it can be used to train lay users about the inherent uncertainty of GIS data.

7.3 Contributions

This thesis significantly contributed to the management of uncertainty in object-based GIS. The following summarizes the contributions:

- The thesis has proposed a framework for uncertainty management. The components of the framework have been identified and their integration has been investigated. A software package with GUI has been developed to assist the implementation of the proposed uncertainty strategy.
- In this thesis, positional uncertainty is analyzed, modeled and managed for vector data elements. As such, it represents the local variation of the geospatial primitives: points, lines and polygons.
- Given feature information with uncertainty measures (e.g. point positions with their covariance matrices) the research modeled uncertainty propagation to derive the proposed line uncertainty models. Model errors were also included in the proposed generalization of the usual line uncertainty representations.
- The uncertainty of the line segments was then used to develop the corresponding uncertainty of the polygon boundaries. The models not only provided the uncertainty

distribution of spatial objects, but also made it possible to efficiently monitor uncertainty propagation of spatial analysis on those spatial data (e.g. point-in-polygon problem). The models were applied in several case studies, which have demonstrated their generality.

- Instead of saving point coordinates to represent linear objects, this thesis proposes to store the parameters of mathematical functions. Due to their simplicity, algebraic polynomials are used to approximate the functions. Because of unbiased property of AIC in model identification, it is then proposed to identify the best statistical model. Uncertainty of the curvilinear boundary was then modeled using simulation and analytical techniques.

- The proposed multivariate linear regression identified the most significant parameters of the polynomial. The proposed strategy in determining the significant parameters was based on iteration and exhibited 75% savings in data volume compared to the polynomial regression.

- Uncertainty modeling and management problems in GIS are considered as closely related matters rather than independent issues. A framework to systematically deal with uncertainty management questions in GIS was established. The framework provided a new systematic way to deal with uncertainty management in GIS.

7.4 Recommendations

Future research is recommended in the following areas:

i) How is the uncertainty of geospatial objects related to the resolution (spatial, spectral, and temporal) of the primary data sources?

ii) What is the relationship between spatial database uncertainty and its impact upon decision making and risk analysis? A method should be devised to determine acceptable levels of uncertainty in spatial information by analyzing the risks associated with decisions based upon use of that information.

iii) How is temporal and logical uncertainty to be handled in decision based systems? This thesis examined the positional uncertainty in GIS. However, the uncertainty caused by logical inconsistency (topological errors) is not considered in the proposed strategy. Furthermore, assuming data to be static over a long period of time is likely erroneous.

iv) What uncertainty communication methods are most effective for different combinations of users, applications, and type of decisions? It is shown that visual techniques offer fresh alternatives to the reporting and understanding of the problem, and new methods using multimedia, aural senses and virtual reality are already appearing. However, there are still problems in determining which techniques are best suited to particular combinations of user skills, applications, and the decisions to be made.

v) The prototype software package prepared during the course of this study, has been written in Matlab script. It would be desirable to have the programs rewritten in a

current high level language such as C++. Such development would facilitate the integration of the current GIS systems with error handling capability.

REFERENCES

Abler, R.F. (1987). "The national science foundation national center for geographical information and analysis." *International Journal of Geographical Information Systems*, No. 1, pp. 303-326.

Agumya, A. and G.H. Hunter (1996). "Assessing fitness for use of spatial information: Information usage and decision uncertainty." *Proceedings of GIS/LIS '96,* Denver, Colorado, pp. 349.360.

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE Transactions in Automatic Control.* Vol. Ac-19, pp. 716-723.

Akaike, H. (1973). "Information theory and an extension of the Maximum Likelihood Principle." *Second International Symposium on Information Theory*, edited by B.N. Petrov and F. Csaki, pp. 267-281, Akademia Kiado, Budapest.

Akaike, H. (1971). *Determination of the Number of Factors by an Extended Maximum Likelihood Principle.* Research memorandum 44, Institute of Statistical Mathematics, Tokyo.

Alai, J. (1993). *Spatial Uncertainty in a GIS.* M.Sc.E. Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Alesheikh, A.A. (1997). "Uncertainty modeling of line and polygon objects in GIS." *Fourth International Conference on Civil Engineering*, Tehran, Iran.

Alesheikh, A.A. and R. Li (1996). "Rigorous uncertainty models of line and polygon objects in GIS." *Proceedings of GIS/LIS '96,* Denver, Colorado, pp. 906-920.

Alesheikh, A.A. and R. Li (1994). "A review of boundary uncertainty in GIS." *Proceedings of CPGIS '94 Conference.* Calgary, Alberta, Canada.

Alesheikh, A.A. (1993). *Generating an Orthoimage Mosaic Using Hybrid GIS*. Masters Thesis. The University of New Brunswick, Fredericton, N.B., Canada.

Alexander, S. and N. Waters (1996). "Decision support methods for resolving land-use conflicts." *Proceedings of GIS/LIS '96,* Denver, Colorado, pp. 1151-1168.

Aronoff, S. (1989). *Geographical Information Systems: A Management Perspective*, WDL Publications, Ottawa, Ontario.

Beard, M.K. (1989). "Use error: The neglected error component." *Proceedings of the Ninth International Symposium on Computer-Assisted Cartography (Auto-Carto 9)*, Baltimore, Maryland, pp. 808-817.

Beard, M.K., B.P. Buttenfield, and S. Clapham (1991). Visualizing the Quality of Spatial Information: Scientific Report of the Specialist Meeting. NCGIA, Technical Report 91-26. Santa Barbara, California.

Bedard, Y. (1987). "Uncertainties in land information systems databases." *Proceedings of the ACSM-ASPRS Auto-Carto 8 Conference*, Baltimore, Maryland, pp. 175-184.

Bertin, J. (1985). *Graphical Semiology*. University of Wisconsin Press. Madison, Wisconsin.

Blackemore, M. (1985). "High or low resolution? Conflicts of accuracy, cost, quality and application in computer Mapping." *Computers and Geosciences*, Vol. 11, No. 3, pp. 345-348.

Blackemore, M. (1984). "Generalization and Error in Spatial Databases." *Cartographica*, Vol. 21, No. 2, pp. 131-139.

Blais, J.A.R. (1998). Personal communication.

Blais, J.A.R. (1996). *Design and implementation of land information systems.* Lecture materials, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Blais, J.A.R. (1991a). "On some model identification strategies using information theory." *Manuscripta Geodetica*, Vol. 16, No. 5, pp. 326-332.

Blais, J.A.R. (1991b). "On some practical applications of information theory in Geomatics." *Canadian Institute of Surveying and Mapping (CISM) Journal*, Vol. 45, No. 2. Pp. 239-247.

Blais, J.A.R. and M. Boulianne (1988). "Comparative analysis of information measures for digital image processing." *International Archive of Photogrammetry and Remote Sensing*, Vol. 27, Part B8, Commission 3, Kyoto, pp. 34-44.

Blais, J.A.R. (1988). *Estimation and Spectral Analysis*. The University of Calgary Press. Calgary, Canada.

Blais, J.A.R. (1987). "Information theory and optimal estimation." *Manuscripta Geodetica*, Vol. 12, No. 4, pp. 238-244.

Burden, R.L., Faires, J.D., and A.C. Reynolds (1978). Numerical Analysis. Prindle, Weber & Schmidt, Incorporated. U.S.A.

Bureau of the Budget (1947). *National Map Accuracy Standards*. US Government Printing Office, Washington, DC.

Burrough, P.A. (1989). "Fuzzy mathematical methods for soil survey and land evaluation." *Journal of Soil Science*, Vol. 40. pp. 477-492.

Burrough, P.A. (1986). *Principles of Geographic Information Systems for Land Resources Assessment*, Claredon Press, Oxford.

Calkins, H.W. and N.J. Obermeyer (1991). "Taxonomy for surveying the use and value of geographic information." *International Journal of Geographic Information Systems*, Vol. 5, No. 3, pp. 341-351.

Canada Map Office (1977). *A Guide to the accuracy of maps*. Technical Report. Ministry of Supply and Services.

Carver, S. (1991). "Adding error handling functionality to the GIS toolkit." *Proceedings of the Second European Conference on GIS (EGIS '91)*, Brussels, Belgium, pp. 187-196.

Caspary, W. and R. Scheuring (1992). "Error-bands as measures of geometrical accuracy." *Proceedings of the Third European Conference on GIS (EGIS '92)*, Munich, Germany, pp. 227-233.

Chapman, M.A., A. Alesheikh, and H.A. Karimi (1997). "Error modeling and management for data in GIS." Proceedings of the CoastGIS '97. Aberdeen, Scotland.

Chrisman, N. (1997). *Exploring Geographic Information Systems*. John Wiley & Sons, New York.

Chrisman, N. (1991). "The error component in spatial data." Eds D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographic Information Systems: Principles and Applications* (Longman: London) Vol. 1, pp. 165-174.

Chrisman, N.R. (1989). "Error in categorical maps: Testing versus simulation." *Proceedings of the Ninth International Symposium on Computer-Assissted Cartography (Auto-Carto 9)*, Baltimore, Maryland, pp. 521-529.

Chrisman, N. (1982). *Methods of Spatial Analysis Based on Error in Categorical Maps*. Ph.D. Thesis, University of Bristol, UK.

Christakos, G. (1992). Random Field Models in Earth Sciences. Academic Press Inc., San Diego, California, USA.

Cohen, P.R. (1985). *Heuristic Reasoning with Uncertainty: An Artificial Intelligent Approach.* Pitman Advanced Publishing Program, London, England.

Cressie, N. (1991). *Statistics for Spatial Data.* John Wiley Publications, New York.

Csillag, F. (1991). "Visualization of data quality." Eds. M.K. Beard, B.P. Buttenfield & S.B. Clapham, *Scientific Report for the Specialist Meeting on NCGIA Research Initiative 7: Visualization of Spatial Data Quality*, National Centre for Geographic Information & Analysis Technical Paper 91-26, University of California, Santa Barbara, pp. C-44-C49.

Dangermond, J. (1983). "A classification of software components commonly used in geographic information systems." *Design and Implementation of Computer-based Geographic Information Systems*, International Geographical Union, Amherst, New York, pp. 23-57.

Davis, B. [1996]. *GIS: A Visual Approach.* Word Press, New York.

Davis, F.W. and D.S. Simonett (1991). "GIS and remote sensing." Eds D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographic Information Systems: Principles and Applications* (Longman: London) Vol. 1, pp. 191-213.

Davis, J.C. (1986). *Statistics and Data Analysis in Geology.* John Wiley & Sons. New York.

Davis, R.E., F.S. Foote, Anderson, J.M., and E.M. Mikhail (1981). Surveying, Theory and Practice. McGraw-Hill Inc. New York, U.S.A.

DCDSTF (1988). "The proposed standard for digital cartographic data." *The American Cartographer*, Vol. 11. Pp. 9-140.

DeFranti, T.A., Brown, M.D., and B.H. McCormick (1989). "Visualization: Expanding the scientific and engineering research opportunities." *Computer*, Vol. 22, No. 6, pp. 27-38.

Divenyi, P. (1991). "Standardization efforts in Hungary." Ed. H. Moellering, *Spatial Database transfer Standards*: Current International Status, Elsevier, New York, pp. 111-122.

Dougherty, E.R. (1990). *Probability and Statistics for the Engineering, Computing, and Physical Sciences*. Prentice Hall, Englewood Cliffs, New Jersey.

Drummond, J. (1995). "Positional Accuracy." In Elements of Spatial Data Quality. Edited by Guptill, S.C. and Morrison, J.L., Elsevier Science Ltd, New York, U.S.A. pp. 31-58.

Dunn, R., Harrison, A.R. and White, J.C. (1990). "Positional accuracy and measurement error in digital databases of land use: An empirical study." *International Journal of Geographic Information Systems*, Vol. 4, No. 4, pp. 385-398.

Dutton, G. (1992). "Handling positional uncertainty in spatial databases." *Proceedings of the 5th International Symposium on Spatial Data Handling*, Charleston, South Carolina, Vol. 2 pp. 460-469.

Edwards, G. and K.E. Lowell (1996). "Modeling uncertainty in photointerpreted boundaries." *Photogrammetric Engineering and Remote Sensing*, Vol. 62, No. 4, pp. 337-391.

Epstein, E.F. (1991). "Legal aspect of GIS." Eds D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographic Information Systems: Principles and Applications* (Longman: London) Vol. 1, pp. 489-502.

Epstein, E.F. and H. Roitman (1987). "Liability for information." *Proceedings of 1987 URISA Annual Conference*, Fort Lauderdale, Florida, Vol. 4, pp. 115-125.

Fegeas, R.G., Cascio, J.L. and R.A. Lazar (1992). "An Overview of FIPS 173, the spatial data transfer standard." *Cartography and Geographic Information Systems*, Vol. 19, No. 5. pp. 278-293.

Fisher, P. (1994). "Animation and Sounds for the Visualization of Uncertain Spatial Information." *In Visuallization in Geographical Information Systems*. Edited by: Hearnshaw, H.M., and D.J. Unwin. John Wiley & Sons, New York.

Foley, J.D., Dam A.V., Feiner, S.K., and J.F. Hughes (1990). *Computer Graphics, Principles and Practices*. Addison-Wesley Publishing Company, Ontario.

Gong, P. and J. Chen (1992). "Boundary uncertainties in digitized maps 1: Some possible determination methods." *Proceedings of GIS/LIS '92*. Pp. 274-281.

Goodchild, M. and D.A. Quattrochi (1997). "Scale, multiscaling, remote sensing, and GIS." In *Scale in Remote Sensing and GIS*. Edited by: Quattrochi, D.A. and M.F. Goodchild, Lewis Publisher, New York, pp. 1-11.

Goodchild, M.F. (1995). "Attribute Accuracy". in Elements of Spatial Data Quality Edited by: Guptill, S.C. and J.L. Morrison. Elsevier Publication. New York.

Goodchild, M., L. Chih-Chang,and Y. Leung (1994a). "Visualizing Fuzzy Maps". *In Visuallization in Geographical Information Systems*. Edited by:Hearnshaw, H.M., and D.J. Unwin. John Wiley & Sons, New York. pp. 158-167.

Goodchild, M., B. Buttenfield and J. Wood (1994b). "Introduction to Visualizing Data Quality." *In Visuallization in Geographical Information Systems*. Edited by:Hearnshaw, H.M., and D.J. Unwin. John Wiley & Sons, New York. pp. 141-149.

Goodchild, M.F., B. Parks, and L. Staeyert (1993). *Environmental Modeling with GIS*. Oxford University press, New York

Goodchild, M.F. (1992a). "Geographic information science." *International Journal of Geographical Information Systems*, Vol. 6, No. 1, pp. 31-45.

Goodchild, M.F. (1992b). *Final Report, Research Initiative 1: Accuracy of Spatial Databases*, National Centre for Geographic Information and Analysis, University of California, Santa Barbara.

Goodchild, M.F. and S. Gopal (1989). *Accuracy of Spatial Databases*. Taylor and Francis.

Goodchild, M.F. and W. Min-Hua (1989). "Modeling errors for remotely sensed data input to GIS. "*Proceedings of the Ninth International Symposium on Computer-Assisted Cartography (Auto Carto 9)*, Baltimore, Maryland, pp. 503-537.

Goodchild, M.F. (1989). "Modeling Errors in objects and fields." Eds. M. Goodchild and S. Gopal, *Accuracy of Spatial Databases*, Taylor and Francis, London, pp. 107-115.

Goodchild, M.F. and O. Dubuc (1987). "A model of error for choropleth maps, with applications to geographic information systems." *Proceedings of the eigth international symposium on Computer-Assissted Cartography (Auto-Carto 8)*, Baltimore, Marylland, pp. 165-174.

Guptill, S.C. (1991). "Spatial data exchange and standardization." Eds D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographic Information Systems: Principles and Applications* (Longman: London) Vol. 1, pp. 515-530.

Heuvelink, G. (1993). *Error Propagation in Quantitative Spatial Modeling*. Ph.D. Thesis, Universiteit Utrecht, The Netherlands.

Heuvelink, G., P.A. Burrough and A. Stein (1989). "Propagation of error in spatial modeling with GIS". International Journal of GIS. Vol. 3. Pp. 303-332.

Hopkins, L.D. (1977). " Methods for generating land suitability maps: A comparative Evaluation." *Journal of the American Institute of Planners*, Vol. 43, No. 4, pp. 386-400.

Huang, D. (1990). "Selecting order for general autoregressive models by minimum description length." *Journal of Time Series Analysis*, Vol. 11, No. 2, pp. 107-113.

Hudson, D. (1988). "Some comments on data quality in a GIS." *Proceedings of the ACSM-ASPRS Annual Convention*, St. Louis, Missouri, Vol. 2, pp. 203-210.

Hunter, G.J. (1996). "Accuracy and GIS - Why worry about it?" *GIS Asia Pacific*, Vol. 2, No. 2, pp.16-17.

Hunter, G.J. and M.F. Goodchild (1995). "Dealing with errors in spatial databases: A simple case study." *Photogrammetric Engineering and Remote Sensing*, Vol. 61, No. 5, pp. 529-537.

Hunter, G.J., M.F. Goodchild and M. Robey (1994). " A toolbox for assessing uncertainty in spatial databases." *Proceedings of the AURISA '94 Conference*. Sydney, Australia, pp. 367-379.

Hunter, G.J. (1993). *Handling Uncertainty in Spatial Databases*. Ph.D. Thesis, Department of surveying and land information, The University of Melbourne, Melbourne, Australia.

Jackson, M.J. and P.A. Woodsford (1991). "GIS data capture hardware and software." Eds D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographic Information Systems: Principles and Applications* (Longman: London) Vol. 1, pp. 239-249.

Karimi, H.A. and D. Hwang (1996). "Towards managing model uncertainty in GISs: An algorithm for uncertainty analysis of air quality advection models." *Geomatica*, Vol. 50, No. 3, pp. 251-259.

Kaufmann, A. (1975). *Introduction to the Theory of Fuzzy Subsets.* Academic Press, Inc.

Keefer, B.L., Smith, J.L. and T.G. Gregoire (1988). "Simulating manual digitizing error with statistical models." *Proceedings of GIS/LIS 88 Conference*, San Antonio, Texas, Vol. 2, pp. 475-483.

Kitchen, T.R. (1989). "Parcel mapping in New York State: Is it absolutely necessary to have a high degree of absolute accuracy?" *Proceedings of the URISA '89 Conference*, Boston, Massachusetts, Vol. 1, pp. 90-99.

Krakiwsky, E.J. (1992). *The Method of Least Squares: A Synthesis of Advances.* Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Kraus, K., and E.M. Mikhail (1972). "Linear least squares interpolation." Photogrammetric Engineering, Vol. 38. No. 10, pp. 1016-1029.

Lam, S. (1992). *Uncertainty Management Using Fuzzy Sets in a Commercial Vector GIS.*, M.Sc.E. Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Larouche, C. (1995). *Automation of Photogrammetric Operations Using Advanced Image Matching Techniques.* Ph.D. Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Laurini, R. and D. Thompson (1992). *Fundamentals of Spatial Information Systems.* Academic Press, New York.

Leung, Y. and J. Yan (1997). "Point-in-Polygon analysis under certainty and uncertainty." Geoinformatica, vol 1. Pp. 93-114.

Leung, Y. (1988). *Spatial Analysis and Planning Under Imprecision.* Elsevier Science Publishers B.V. North Holland.

Lichti, D. (1998). "*Least Squares Adjustment.*" Department of Geomatics Engineering, Lecture Notes, University of Calgary, Calgary, Alberta, Canada.

Lundin, B., Yan, J. and J.P. Parker (1989). "Data quality reporting methods for digital geographic products at Statistics Canada." *Proceedings of the 1989 CISM National Conference on GIS*, Ottawa, Canada, pp. 236-251.

Maguire, D.J., M.F. Goodchild and D.W. Rhind (1991). *Geographical Information Systems; Principles and applications*. John Wiley & Sons. Inc. New York. USA.

Masry, S.E. and Y.C. Lee (1988). *An Introduction to Digital Mapping* . Lecture Notes No. 56, Department of Geodesy and Geomatics Engineering, The University of New Brunswick, Fredericton, N.B., Canada.

Matlab (1992). *Spline Toolbox User's Guide*. The MathWork Inc. MA.

Mepham, M.P. and B.G. Nickerson (1987). " Preanalysis." In Papers for the CISM Adjustment and Analysis Seminar. Edited by E.J. Krakiwsky. pp. 150-181.

Mikhail, E.M. and G. Gracie (1981). *Analysis and Adjustment of Survey Measurements*. Van Nostrand Reinhold Company, New York.

Mikhail, E.M. (1976). *Observations and Least Squares*. Dun-Donnely Publisher, New York

Miller, R.G. (1992). *Modeling and Managing Uncertainty in G.I.S.*, M.Sc.E. Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Miller, R.G., Karimi, H.A. and M. Feuchwagner (1989). "Uncertainty and its management in geographic information systems." *Proceedings of the CISM National Conference on GIS*, Ottawa, Canada, pp. 252-260.

NIST (1992). *Spatial Data Transfer Standard (FIPS 173)*, National Institute of Standards and Technology, US Department of Commerce, Washington, D.C.

Obermeyer, N.J. (1989). "A systematic approach to the taxonomy of geographic information use." *Proceedings of the GIS/LIS '89 Conference*, Orlando, Florida, Vol. 2, pp. 421-429.

Openshaw, S. (1989). "Learning to live with errors in spatial databases" Eds. M. Goodchild and S. Gopal, *Accuracy of Spatial Databases*, Taylor and Francis, London, pp. 263-276.

Pang, A.T., C.M. Wittenbrink and S.K. Lodha (1996). "Approaches to uncertainty visualization." *Http://www.cse.uscs.edu/research/slrg/uncertainty.html.*

Perkal, J. (1966). *On the Length of Empirical Curves*. Discussion Paper Number 10. Michigan Inter-University community of Mathematical Geography.

Plunkett, G. and J. McKenna (1996). "Metadata requirement for GIS: An international perspective." 1996 Canadian Geomatics CD-ROM. Paper No. 91.

Preece, J., Y. Rogers, H. Sharp, D. Benyon, S. Holland and T. Carey (1994). *Human-Computer Interaction*, Addison-Wisely, Reading, MA.

Quattrochi, D.A. and M.F. Goodchild (1997). *Scale in Remote Sensing and GIS*. Lewis Publishers, New York.

Raymond, E.W. (1989). "Accuracy considerations in the use of parcel maps A source document for a GIS." *Proceedings of the URISA '89 Conference*, Boston, Vol. 1, pp. 82-88.

Rainio, A. (1991). "The standardization of geographical data interchange in Finland." Ed. H. Moelling, *Spatial Database Transfer Standards: Current International Status*, Elsevier, New York, pp. 75-84.

Sakamoto, Y., Ishiguro, M. and G. Kitagawa (1986). *Akaike Information Criterion Statistics*. KTK Scientific Publishers, Tokyo, Japan.

Seddon, B. (1971). *Introduction to Biogeography*. Harper and Row, New York, NY., pp. 220.

Shaffer, G. (1976). *The Mathematical Theory of Evidence*. Princton university Press.

Shannon, C.E. and W. Weaver (1959). *The Mathematical Theory of Communication*. University of Illinois Press. Urbana, Illinois.

Shi, W. (1994). *Modeling Positional and Thematic Uncertainties in Integration of Remote Sensing and Geographic Information Systems*. Ph.D. Thesis, ITC, The Netherlands.

Sinton, D. (1978). "The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study." In: Dutton, G. (ed.) *Harvard Papers on Geographic Information Systems*, Vol. 6, Addison-Wesley, Reading, MA.

Smith, I, (1991). "Digital geographic information working group: Exchange Standards." Ed. H. Moellering, *Spatial Data transfer Standards: Current International Status*, Elsevier, New York, pp. 223-235.

Smith, T.R. and Y. Jiang (1991). "Knowledge-based approach in GIS." Eds D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographic Information Systems: Principles and Applications* (Longman: London) Vol. 1, pp. 413-425.

Sookman, B.B. (1989). "Liability of information providers." *Proceedings of URISA '89 Conference*, Boston, Massachussetts, Vol. 4, pp. 256-276.

Star, J.L., J.E. Estes and F. Davis (1991). "Improved integration of remote sensing and Geographic Information Systems: A background to NCGIA Initiative 12." Photogrammetric Engineering and Remote Sensing. Vol. 57, No. 6. Pp. 643-645.

UCGIS (1994). "GIS education priorities." Http://www.ucgis.org.

Vanicek, P. (1992). "*Advanced Least Squares Adjustment.*" Lecture Notes. Department of Geodesy and Geomatics Engineering, The University of New Brunswick, Fredericton, N.B., Canada.

Vanicek, P. and E.J. Krakiwsky (1986). *Geodesy: The Concept*, 2nd Rev. ed., North-Holland, Amsterdam.

Veregin, H. (1989a). "Error modeling for the map overlay operation." Eds. M. Goodchild and S. Gopal, *Accuracy of Spatial Databases*, Taylor and Francis, London, pp. 3-18.

Veregin, H. (1989b). *A Taxonomy of Error in Spatial Databases*. National Centre for Geographic Information and Analysis Technical Paper 89-12, University of California, Santa Barbara, 115 pp.

Wolf, P.R. and C.D. Ghilani (1997). *Adjustment Computations: Statistics and Least Squares in Surveying and GIS*. John Wiley & Sons. New York.

Worboys, M.F. (1995). *GIS, A Computing Perspective*. Taylor & Francis Inc. London, UK.

Zadeh, L.A., (1956). "Fuzzy sets." *Information and Control*, Vol. 8, pp. 338-353.

Zadeh, L.A. (1995). "Discussion: Probability theory and fuzzy logic are complementary rather than competitive." Technometrics. Vol. 37, No. 3, pp. 271-276.

Zhou, F. and Y.C. Lee (1994). "Polygon uncertainty modeling and representation." Proceedings of Canadian Conference on GIS, June 6-10, Ottawa, Vil. 1. Pp. 185-191.

Zhou, F. (1995a). *Uncertainty Management for Object-Based Geographic Information Systems*, Ph.D. Thesis, Department of Geodesy and Geomatics Engineering, The University of New Brunswick, Fredericton, N.B., Canada.

Zhou, W. (1995b). *Large Object Support Using an Object Oriented Approach in Spatial Information Systems*. Ph.D. Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

Zhou, J. (1996). *Modeling and Visualization of Positional Uncertainty in GIS and Its Application to VISAT*. Masters Thesis, Department of Geomatics Engineering, The University of Calgary, Calgary, Alberta, Canada.

APPENDIX

DERIVATION OF AIC FOR ALGEBRAIC POLYNOMIAL

In this section the maximum likelihood of the algebraic polynomial is derived first, and AIC will then be presented for the polynomial [Sakamoto et al. 1986].

Polynomial Regression Model

Let us assume the polynomial regression model is of the form

$$y = a_o + a_1 x_i + a_2 x_i^2 + \ldots a_m x_i^m + \varepsilon_i$$

where ε_i is independent normal random variable that follows $N(0, \sigma^2)$. The integer m is called the order of the polynomial regression model. This model determines the variation of y as the sum of the polynomial in the deterministic variable x and the random error ε.

The polynomial regression model of order m is a conditional distribution model of which, given the independent variable x, the distribution of the dependent variable y is a normal distribution with the mean $a_0 + a_1 x + a_2 x^2 + \ldots a_m x^m$ and the variance σ^2 i.e.,

$$f(y \mid a_0, \ldots a_m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y - a_o - a_1 x - \ldots - a_m x^m)^2\}$$

Let us denote the polynomial model of order m by MODEL(m). Fitting a polynomial to the data is equivalent to estimating a probability distribution for the variable y as a function of the variable x.

The Likelihood of a Polynomial Regression Model

Given the probability density function of y that follows MODEL(m) as and a set of n data points $\{ (x_i, y_i); i = 1, \ldots, n\}$, the likelihood is given by

$$L(y_1, \ldots y_n \mid a_o, \ldots a_m, \sigma^2) = \pi_{i=1}^{n} f(y_i \mid a_o, \ldots a_m, \sigma^2)$$

$$= (\frac{1}{2\pi\sigma^2})^{\frac{\pi}{2}} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a_o - a_1 x_i - \ldots - a_m x^m{}_i)^2\}.$$

The values of $a_0, a_1, \ldots, a_m$ and σ^2 that maximize the likelihood constitute the maximum likelihood of the regression coefficient and the residual variance of the polynomial regression model, MODEL(m), respectively.

Maximum Likelihood Estimates of the Parameters

By taking the natural logarithm of the likelihood function; the log likelihood can be obtained by:

$$l(y \mid a_o, \ldots a_m, \sigma^2) = \log L(y_1, \ldots y_n \mid a_o, \ldots a_m, \sigma^2)$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a_o - \ldots - a_m x_i^m)^2.$$

To obtain the maximum likelihood estimates, values of the parameters that maximize the log likelihood must be found. It can be proven that the log likelihood is maximized with respect to $a_0, a_1, \ldots, a_m$ when

$$S = \sum_{i=1}^{n}(y_i - a_0 - \ldots - a_m x_i^m)^2.$$

is minimized. Thus in the case of the polynomial model fitting, the maximum likelihood is equivalent to the ordinary least squares methods. S will be maximized when

$$\left\{\begin{array}{l} \dfrac{\partial S}{\partial a_0} = -2\sum_{i=1}^{n}(y_i - a_0 - \ldots - a_m x_i^m) = 0 \\ \dfrac{\partial S}{\partial a_1} = -2\sum_{i=1}^{n} x_i(y_i - a_0 - \ldots - a_m x_i^m) = 0 \\ \ldots\ldots \\ \dfrac{\partial S}{\partial a_m} = -2\sum_{i=1}^{n} x_i^m(y_i - a_0 - \ldots - a_m x_i^m) = 0 \end{array}\right\}$$

Thus the maximum likelihood estimates $\hat{a}_0,\ldots,\hat{a}_m$ are obtained by solving the above system of linear equations (normal equations). The necessary condition that σ^2 maximizes the log likelihood is

$$\frac{\partial l}{\partial \sigma^2}\Big|_{\sigma^2=\hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(y_i - \hat{a}_0 - \ldots - \hat{a}_m x_i^m)^2 = 0$$

After algebraic manipulations, the maximum likelihood estimate of the residual variance is

$$\hat{\sigma}^2 = \frac{1}{n}\left\{\sum_{i=1}^{n} y_i^2 - \sum_{i=0}^{m}\hat{a}_i \sum_{j=1}^{n} x_j^i y_i\right\}.$$

$$l(y \mid a_o, \ldots\ a_m, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log d(m) - \frac{n}{2}.$$

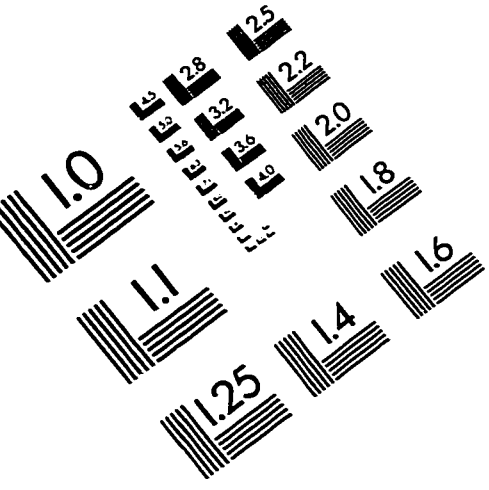
MODEL(m) has m + 2 free parameters, namely the regression coefficients a_0, ... , a_m and the variance σ^2 . Therefore, on substituting k = m+ 2 the AIC of MODEL(m) can be found as (Sakomoto et al., 1986)

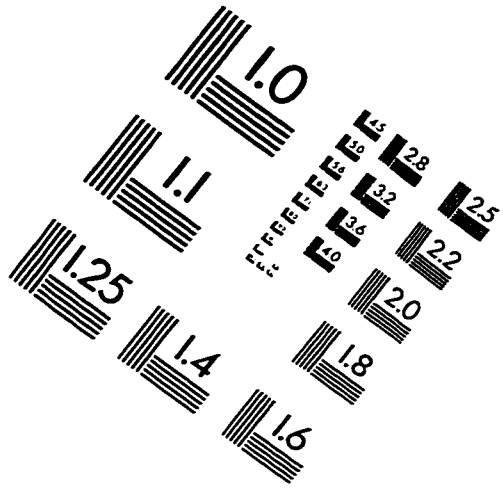
$$AIC(m) = n\log 2\pi + n\log d(m) + n + 2(m+2).$$

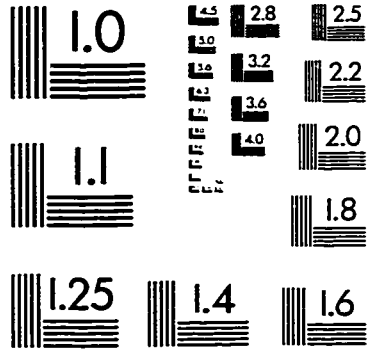
Because AIC is independent of any scale or translation changes, the above formula can be devided by n and subtracted from a constant. The AIC of MODEL(m) is the parsimoniously found as

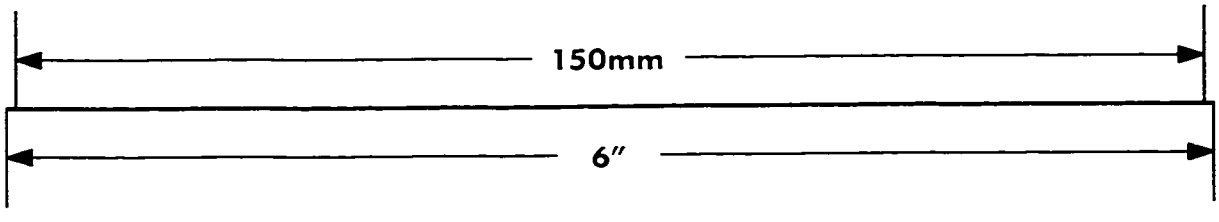
$$AIC(m) = \log d(m) + 2(m+2) / n.$$

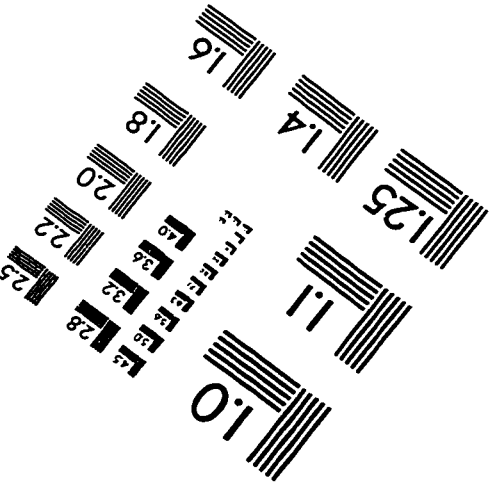
IMAGE EVALUATION
TEST TARGET (QA-3)

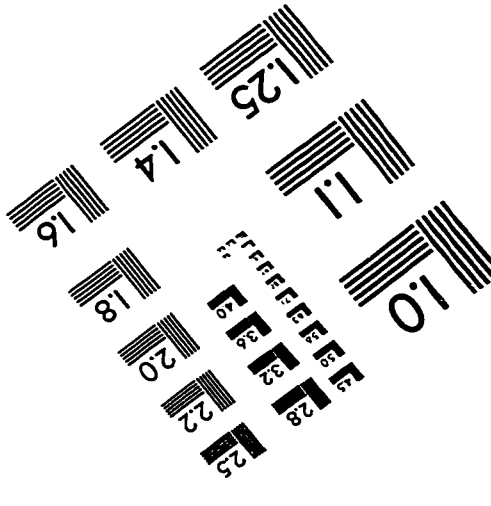












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989