POSITIONAL UNCERTAINTY ANALYSIS USING DATA UNCERTAINY ENGINE - A CASE STUDY ON AGRICULTURAL LAND PARCELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

İLKSEN URGANCI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
GEODETIC AND GEOGRAPHICAL INFORMATION TECHNOLOGIES

DECEMBER 2009

Approval of the thesis;

POSITIONAL UNCERTAINTY ANALYSIS USING DATA UNCERTAINY ENGINE - A CASE STUDY ON AGRICULTURAL LAND PARCELS

submitted by **İLKSEN URGANCI** in partial fulfillment of the requirements for the degree of Master **of Science in Geodetic and Geographical Information Technologies (GGIT), Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** _____

Assoc. Prof. Dr. Mahmut Onur Karslıoğlu
Head of Department, **Geodetic and Geographical Information Technologies, METU** _____

Assoc. Prof. Dr. S. Zuhal Akyürek
Supervisor, **Civil Engineering Dept., METU** _____

Examining Committee Members:

Prof. Dr. Vedat Toprak
Geological Engineering Dept., METU _____

Assoc. Prof. Dr. S. Zuhal Akyürek
Civil Engineering Dept., METU _____

Prof. Dr. Ali Ünal Şorman
Civil Engineering Dept., METU _____

Assoc. Prof. Dr. Mahmut Onur Karslıoğlu
Civil Engineering Dept., METU _____

Tuncay Küçükpehlivan (M.S.)
Başarsoft A.Ş. _____

Date: 18/12/2009

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: İlksen URGANCI
Signature:

ABSTRACT

POSITIONAL UNCERTAINTY ANALYSIS USING DATA UNCERTAINY ENGINE
A CASE STUDY ON AGRICULTURAL LAND PARCELS

Urgancı, İlksen

M.S., Department of Geodetic and Geographic Information Technologies
Supervisor: Assoc. Prof. Dr. S. Zuhal Akyürek

December 2009, 106 pages

Most of spatial data extraction and updating procedures require digitization of geographical entities from satellite imagery. During digitization, errors are introduced by factors like instrument deficiencies or user errors. In this study positional uncertainty of geographical objects, digitized from high resolution Quickbird satellite imagery, is assessed using Data Uncertainty Engine (DUE). It is a software tool for assessing uncertainties in environmental data; and generating realisations of uncertain data for use in uncertainty propagation analyses. A case study area in Kocaeli, Turkey that mostly includes agricultural land parcels is selected in order to evaluate positional uncertainty and obtain uncertainty boundaries for manually digitized fields. Geostatistical evaluation of discrepancy between reference data and digitized polygons are undertaken to analyse auto and cross correlation structures of errors. This process is utilized in order to estimate error model parameters which are employed in defining an uncertainty model within DUE. Error model parameters obtained from training data, are used to generate simulations for test data. Realisations of data derived via Monte Carlo Simulation using DUE, are evaluated to generate uncertainty boundaries for each object guiding user for further analyses with pre-defined information related to the accuracy of spatial entities. It is also aimed to assess area uncertainties affected by the position of spatial entities. For all different correlation structures and object models, weighted average positional error

for this study is between 2.66 to 2.91 meters. At the end of uncertainty analysis, deformable object model produced the smallest uncertainty bandwidth by modelling cross correlation.

Keywords: Positional Uncertainty, Data Uncertainty Engine, Vector Data, GIS

ÖZ

VERİ BELİRSİZLİK MOTORU KULLANILARAK KONUMSAL BELİRSİZLİK ANALİZİ TARIMSAL ARAZİ PARSELLERİ ÖRNEĞİ

Urgancı, İlksen

Yüksek Lisans, Jeodezi ve Coğrafi Bilgi Teknolojileri E.A.B.D.
Tez Yöneticisi: Doç. Dr. S. Zuhal Akyürek

Aralık 2009, 106 sayfa

Mekansal veri çıkarımı ve güncelleme işlemlerinin çoğunluğu coğrafi varlıkların uydu görüntüsünden sayışlaştırılmasını gerektirmektedir. Sayısallaştırma sırasında ekipman kısıtlamaları ya da insan yanılgıları gibi değişkenlerden kaynaklanan hatalar ortaya çıkmaktadır. Bu çalışmada yüksek çözünürlüklü Quickbird uydu görüntüsünden sayısallaştırılan coğrafi nesnelerin konumsal belirsizliği Veri Belirsizlik Motoru (DUE) kullanılarak değerlendirilmektedir. DUE, mekansal verilerdeki belirsizliğin değerlendirilmesi ve yayılımı analizlerinde kullanılmak üzere belirsiz verilerden simülasyonlar üretmeyi sağlayan bir yazılımdır. Elle sayısallaştırılan arazi sınırlarının konumsal belirsizliğini değerlendirmek ve belirsizlik sınırlarını elde etmek için, Kocaeli'nde çoğunlukla tarımsal arazi parsellerini içeren bir çalışma alanı seçilmiştir. Referans veri ve sayısallaştırılan poligonlar arasındaki farkın coğrafi-istatistiki değerlendirmesi, konumsal hataların oto ve çapraz korelasyon yapıları incelenerek gerçekleştirilmiştir. Bu analiz süreci, DUE içinde tanımlanan belirsizlik modelinin oluşturulmasında kullanılan hata modeli parametrelerini hesaplanmasını sağlamıştır. Eğitme verilerinden elde edilen hata modeli parametreleri test verileri için simülasyonların oluşturulmasında kullanılmıştır. DUE kullanılarak, Monte Carlo Simülasyonu metoduyla türetilen simülasyonlar her bir obje için belirsizlik sınırlarının oluşturulmasında kullanılmıştır. Oluşturulan belirsizlik sınırları ileriki analizler için, mekansal verilerin doğruluğuyla ilgili

kullanıcıya yol gösterici öncül bilgi sağlamaktadır. Bunlara ek olarak, mekansal verilerin pozisyonlarının alansal belirsizlik üzerindeki etkilerinin değerlendirilmesi de amaçlanmıştır. Bu çalışmada, bütün farklı korelasyon ve obje modelleri için, hesaplanan ağırlık ortalama konumsal hata 2.66 ile 2.91 metre arasındadır. Belirsizlik sınır analizleri sonucunda, biçim değiştirebilen obje modeli çapraz korelasyonu modelleyerek en küçük belirsiz bant genişliğini üretmiştir.

Anahtar Kelimeler: Konumsal Belirsizlik, Veri Belirsizlik Motoru, Vektör Veri, CBS

To My Family

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. Zuhal AKYÜREK for her guidance, advice, criticism, encouragement and especially support throughout the preparation of this thesis. I want to thank her for believing in me.

I would like to thank Prof. Dr. Ali Ünal ŞORMAN, Assoc. Prof. Dr. Mahmut Onur KARSLIOĞLU, Prof. Dr. Vedat TOPRAK and Tuncay KÜÇÜKPEHLİVAN for their comments, suggestions and evaluations.

I would like to thank Assoc. Prof. Dr. Gerard HEUVELINK and Dr. James D. BROWN for sharing Data Uncertainty Engine (DUE), and help on DUE.

I would like to express my sincere gratitude to my friends Cem GÜLLÜOĞLU, Aslı ÖZDARICI, Ali Özgün OK and Şeyma ÖZCAN for their suggestions, assistance and encouragement throughout my thesis.

Thanks to my friend Zühal EKŞİ especially, for her support and motivation during the hard times. She always makes me get up when I fall.

Finally I would like to thank to my family for everything we went through this process. They made me the way I am now.

This study is supported by Natural and Applied Sciences BAP fund. Satellite imagery used in this study is purchased by this fund.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVATIONS

GCP: Ground Control Points

GIS: Geographical Information Systems

GPS: Global Positioning Systems

DEM: Digital Elevation Model

DUE: Data Uncertainty Engine

IGF: Indicative Goodness of Fit

Jpdf: Joint Probability Distribution Function

MCS: Monte Carlo Simulation

Mpdf: Marginal Probability Distribution Function

Pdf: Probability Distribution Function

RMSE: Root Mean Squared Error

CHAPTER 1

INTRODUCTION

Geographical Information Systems (GIS) frame a brief abstraction of the real world entities in the form of vector and raster data models associated with attributes. These data models are used to describe objects in GIS; however, these descriptions may involve certain amount of uncertainties caused during extracting environmental variables as objects on computer based systems. These uncertainties can be introduced from the first step of cognition of the real world, to process, analysis and even to decision making (Cheng, 2003).

GISs are commonly utilized as a means of storing and displaying spatial objects and their associated attributes. Hence, a GIS database is a digital representation of the spatial organization of objects and phenomena in the real world. The representations inevitably contain errors arising from the difference between the true value and observed value of a geographical object. So it can easily be said as Heuvelink (1998) stated the problem of spatial data quality is obvious because no map stored in GIS is completely error – free. True state of inputs is not always available to quantify errors in data which can be defined departure from reality. In such cases uncertainty is used as an expression of confidence about our knowledge in data (Heuvelink and Brown, 2007).

When maps that are stored in a GIS database are used as inputs to a GIS operation, the errors in the input will propagate to the output of the operation. According to Cheng (2003), the accuracy of geographical feature and its attribute in GIS may be influenced by vague properties of the spatial objects or uncertainties introduced during the process of data capture. This is because the resulting output is a function of input values, and inaccurate input values automatically affect the computed result (Heuvelink et. al., 1989). Therefore the output may not be

sufficiently reliable for correct conclusions to be drawn from it. Moreover, the error propagation continues when output from one operation is used as input to an ensuing operation. Consequently, when no record is kept of the accuracy of the intermediate results, it becomes extremely difficult to evaluate the true accuracy of the final result (Heuvelink, 1998).

GIS is described by the massive amount of data, sources, and methodologies employed in data production and manipulation. Data in a GIS are often stored in the form of data layers. The data stored in a GIS have been collected in the field, have been classified, interpreted, estimated intuitively and so contain a certain amount of error. Errors also derive from measurement error, from spatial and temporal variation and from mistakes in data entry (Heuvelink, 1998).

Spatial data layers may originate from a variety of sources: digitized maps, points acquired through global positioning devices, and surveyed attributes entered into a software package. Layers may also consist of objects extracted from remotely-sensed imagery.

Within a GIS environment, the proper use of information requires the identification of the uncertainty estimates associated with it. Currently, a major dilemma in evaluating the positional accuracy of a GIS dataset lies in determining the uncertainty of different objects. Positional uncertainty assessments examine error by attempting to predict its propagation from points to lines and to objects within a GIS. The most fundamental geometric elements are points and lines, and an analysis of positional accuracy would logically begin with these simple objects.

Error models for points and the line segments that connect them have been developed, such as the epsilon-band, confidence region, and G-Band models (Shi and Liu 2000). These models are based on an assumed statistical distribution of the error associated with each point, typically a Gaussian (Normal) distribution. Shi and Liu (2000) present a stochastic model of the positional errors in line segments which

assumes that the errors of the endpoints follow two-dimensional (*x*,*y*) normal distributions.

The Data Uncertainty Engine (DUE) which is developed by Brown and Heuvelink (2007) allows uncertainties in model inputs to be described and their impacts propagated through for model predictions. It is a prototype software tool for assessing uncertainties in environmental data, for storing them within a database, and for generating realizations of data to include in an uncertainty propagation analysis (Heuvelink et al., 2007).

Using DUE, the spatial and temporal patterns of uncertainty (autocorrelation), as well as cross-correlations between related inputs, can be incorporated in an uncertainty analysis. Such correlations may greatly influence the outcomes of an uncertainty analysis because models typically respond differently to correlated variability than random errors. DUE also supports the quantification of positional uncertainties in geographic objects, represented as raster maps, time-series or vector outlines (Brown & Heuvelink, 2007).

Objects supported by DUE include spatial vectors, space-time vectors, spatial raster, time-series of raster, simple time-series and objects that are 'constant' in space and time. Attributes supported by DUE include continuous numerical variables, discrete numerical variables and categorical variables.

1.1. Objectives of the Study

Such computer based environments are only the abstraction of real world applications for further research activities used in decision making processes. So that the awareness about the accuracy of used data and executed operations during these processes have crucial effect on outputs.

Here it is tried to increase awareness of using data with positional uncertainty and establish an error- aware decision making processes throughout the use of Data Uncertainty Engine which is developed by Brown & Heuvelink (2007).

Essentially, this study proposes a method to improve the reliability of GIS outputs at real world applications by performing uncertainty analysis in order to identify error aware results in decision making processes

In this thesis, uncertainty boundary estimation with an emphasis on vector polygons which are manually digitized from orthorectified satellite imagery is addressed. By modelling the boundary of the uncertainties of agricultural land parcels, how this information can be used together for spatial analysis containing such datasets is investigated.

Main focus of this study is handling the positional uncertainty in spatial data features within GIS environments, with the help of DUE. The objectives of this study are;

- to analyze estimation and modelling techniques for uncertainty modelling in DUE
- to asses the effectiveness of different object models offered within DUE
- to create uncertainty boundaries of spatial objects digitized from satellite imagery
- to provide uncertainty boundaries for each polygon
- to measure the inclusion and exclusion areas affected by the positional uncertainty in agricultural land parcels

1.2. Thesis Structure

This study is consisted of five chapters. Following this introduction chapter, Chapter 2 concentrates on the sources of uncertainty that are introduced during data capture, concepts and definitions about spatial uncertainty and introduces a review of previous studies on estimation and modelling of positional uncertainty. Also

Chapter 2 gives the basics of mathematical theory behind models and summarizes the models that are used to model the positional uncertainty in DUE.

Chapter 3 explains the methods and data used in the study. Also an introduction about the main parts of Data Uncertainty Engine (DUE) developed by Brown and Heuvelink (2007) and example studies with DUE are given. A flowchart of the methods used in the uncertainty analysis is presented.

Uncertainty analysis and modelling of digitized data and the creation of uncertainty boundaries around vector polygons for case study are explained in Chapter 4. Two sample regions are selected for further analysis to examine different object models (namely, deformable and rigid models) and cross correlation structure between x and y coordinate errors are explained in section 4.4. Chapter 4 also, is a review on results of the uncertainty analysis.

And finally Chapter 5 brings the thesis to a close with a summary and paves the way for future research.

CHAPTER 2

LITERATURE REVIEW

This chapter includes the previous studies carried out about spatial uncertainty in GIS. First sources, concepts and definitions about spatial uncertainty are explained. Then, the previous studies, related to spatial uncertainty estimation and modelling techniques are presented. Some of these methods developed in the previous studies and used in this study are also explained.

2.1. Sources of Uncertainty

Spatial information and data capture procedures contain error inevitably. Spatial measurements can represent the geographical location or extents of a feature. Also a spatial observation can be associated with attribute measurements which specify the characteristics of geographic phenomena. Both the position and attribute contain errors to some degree. An error is defined as a discrepancy between the measured and actual value of a particular attribute for a given entity (Veregin, 1999). Error sources are classified as follows by King's study (2002) when dealing with spatial and attribute measurements.

- Natural - Errors caused by changing conditions in the environment
- Personal - Errors that are created by limitations in the human senses
- Instrumental – Errors which are caused by imperfections in instrument functionality

Natural errors can be caused by the variations in temperature, wind, atmospheric pressure, gravitational fields, and magnetic fields. Human (personal) errors occur from one's inability to perfectly see, perceive, or interpolate observations. Examples of instrumental errors may include uncalibrated parts or lenses in an imaging device,

or non-uniform spacing between divisions on a theodolite or total station instrument (Wolf and Ghilani, 1997).

The types of errors produced by the above mentioned sources can be categorized in three main classes:

- Blunders
- Systematic errors
- Random errors

Blunders are usually large errors such as mistakes in reading or writing observations values or using the wrong datum or projection resulting from the carelessness of the operator. Systematic errors are those that follow some physical law and hence can be predicted. Often systematic errors are removed by deriving corrections based on the physical conditions that created them (e.g., atmospheric interference, solar radiation). Random errors are errors inherent in the nature of measurement, those errors that exist after all blunders and systematic errors have been removed. Random errors can arise from human and instrument imperfection, as well as imperfect corrections. Random errors are impossible to avoid and do not follow any physical laws. Therefore they must be handled according to the mathematical laws of probability and corrected by a series of adjustments (Wolf and Ghilani, 1997)

Uncertainty analyses in GISs mainly focuses on assessment of the random errors that are left when blunder and systematic errors removed by elimination or corrections.

2.2. Spatial Uncertainty

Identifying spatial uncertainty starts with understanding the concepts of spatial accuracy and data quality. Data quality is often used in the context of metadata, and describes the measures and assessments that are intended by data producers to characterize known uncertainties. King (2002) defines accuracy as a measure of how an observation is close to a true value.

The definition of geospatial uncertainty within GIS is a much argued and often unclear subject. When describing spatial uncertainty, a range of comparable terms such as: error, accuracy, precision, vagueness, ambiguity, and reliability have been used almost interchangeably. Vagueness, imprecision, and inaccuracy indicate specific conceptual terms, varying from fuzzy set theory to traditional theories of scientific measurement error, and whether or not it is implied that some true value exists in the real world that can be compared to the value stored in the database (Goodchild, 2007). Today most research undertaken on dealing uncertainty in GIS has concentrated on two main methods: fuzzy set theory to represent vagueness, and probability theory to characterize error (Fisher, 2005).

According to Klir and Yuan (1995) uncertainty is identified as either products of fuzziness or products of ambiguity. Fisher (2005) has developed a taxonomy based on Klir and Yuan's study to represent the relation between geographical feature and definitions different types of uncertainty. Figure 2.1 demonstrates Fisher's different types of uncertainty within spatial information.

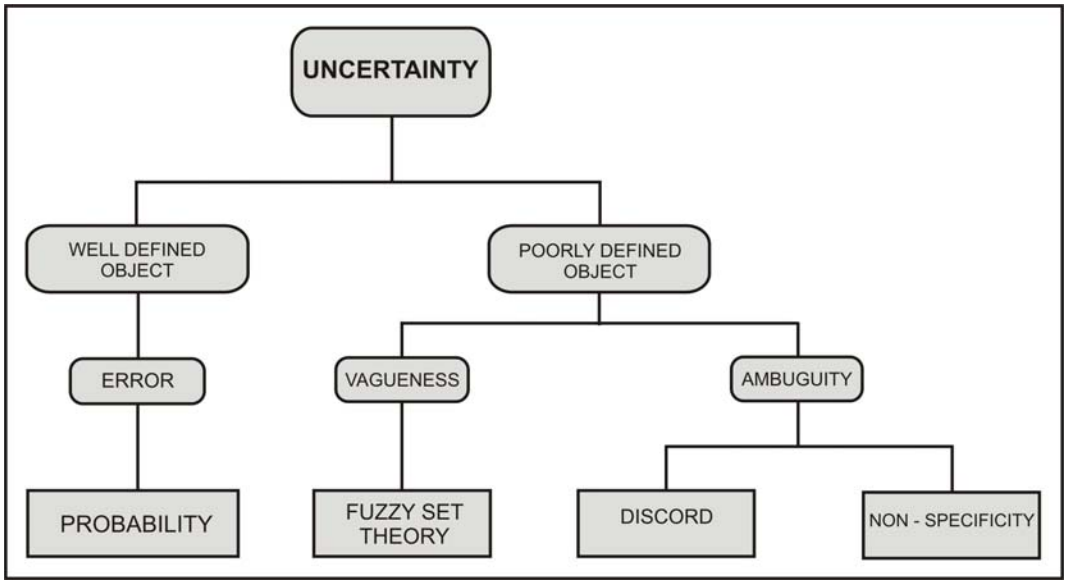

Figure 2.1 Conceptual Model of Uncertainty (taken form Fisher, 2005 and Klir and Yuan, 1995)

To determine which method of uncertainty assessment can be utilized for a particular spatial object, it is crucial to determine whether the class of objects is well defined or poorly defined. If the object is well defined, (e.g., land ownership boundaries), then the uncertainty is caused by errors and is probabilistic. If the object is poorly defined, such as vegetation or soil boundaries, then specific types of uncertainty, vagueness or ambiguity, may be recognized. Fisher (2005) associates vagueness with the poor definition of the class where the object belongs.

Ambiguity is described with instances when doubt about the classification of object because of different perceptions exists. Ambiguity can also be divided in two groups. Discord occurs when an object is clearly defined, yet different perceptions of the classification scheme allow the object to be classified in more than one class. Non-specificity occurs when features have no appropriate class to be assigned to.

As Fisher (2005) and Kurtar (2007) are expressed in their studies line features in other means vector features are well defined geographical objects. In this study spatial uncertainty of well defined objects are evaluated so that in the previous

section much of the discussion focuses on the probability model of spatial uncertainty.

2.3. Models for Assessing Spatial Uncertainty

2.3.1 Probability Models

Focusing mainly on estimation and prediction of the random errors within the system or model, within uncertainty analysis it is assumed that all systematic errors can be corrected. This assumption is mainly utilized for assessing positional uncertainty of spatial entities. So that it tends to be the basis of the many probability-based methods that examine spatial uncertainty (King, 2002). For any measurement of a parameter, there is a probability that it is correctly measured. Methods for determining uncertainty using probability are based on assumptions of standard error theory.

Points, straight line segments, polylines and polygons are main spatial entities in a vector data. Positional uncertainty of a point is usually represented by a circle with a radius of ε. ε is the discrepancy between the true value and the observation. Manual data entry method of digitizing is still a popular method however nowadays automated data entry techniques have been improved. Uncertainty distribution for digitisation is investigated by researches such as Bolstad et al. (1990) found that the digitizing uncertainty is nearly normally distributed.

When probabilistic approaches are considered, distribution functions are the tools for modelling uncertainty in positional uncertainty; like Gaussian distribution, that is commonly used for error distribution. The equation for probability density function of a normal distribution is (Mathworld, 2009):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2} \qquad (2.1)$$

Wolf and Ghilani (1997) proposed an uncertainty model for point features. In this model, uncertainty of a point is modelled based on bivariate normal distribution. In bivariate normal distribution, the dimensions are x and y coordinate pair of the point. Hence uncertainty region of the point is in shape of ellipse (Figure 2.2)

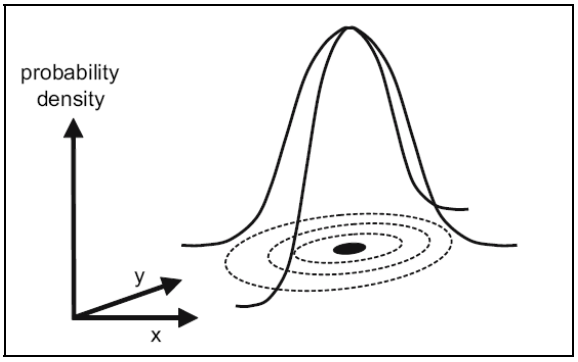

Figure 2.2 Bivariate Probability Distribution of a Point Object's Position (adapted from Heuvelink and Brown, 2007)

The other characteristic of this model is that, a correlation between x and y is taken into consideration and any correlation between these variables cause the ellipse to be rotated towards clockwise direction, if correlation is positive. Otherwise the ellipse is rotated towards counter clockwise direction.

An epsilon error band model is introduced by Perkal (1956, 1966) and developed by Chrisman (1982). The model based on the premise to create an uncertainty band surrounding the line segment. Width of the uncertainty boundary is based on a constant radius around line's most likely true position and called an epsilon band. The quantity epsilon (ε) is derived from the radius of the line's endpoint error circles, assuming a digitization process that yields random coordinate error in a circular normal distribution. The circular normal distribution is two-dimensional (bivariate) and varies normally, meaning that it consists of errors in two directions that are equal and uncorrelated. The main drawback of this model is that it provides no interpretation of error distribution inside the band (Figure 2.3).

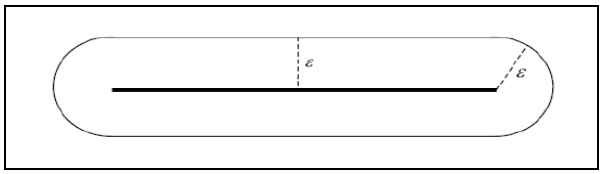

Figure 2.3 Chrisman's Epsilon Model (taken from Chrisman, 1982)

Caspary and Scheuring (1992) improved Chrisman's band model based on Dutton experiment (Dutton, 1992) which generates realizations of line segments using Monte Carlo Simulation between two uncertain end points. Dutton finds out that the error is the greatest on the measured point and least in the halfway between them. The experiment showed that there is a need to narrower band towards the interior of the line segment (Figure 2.4).

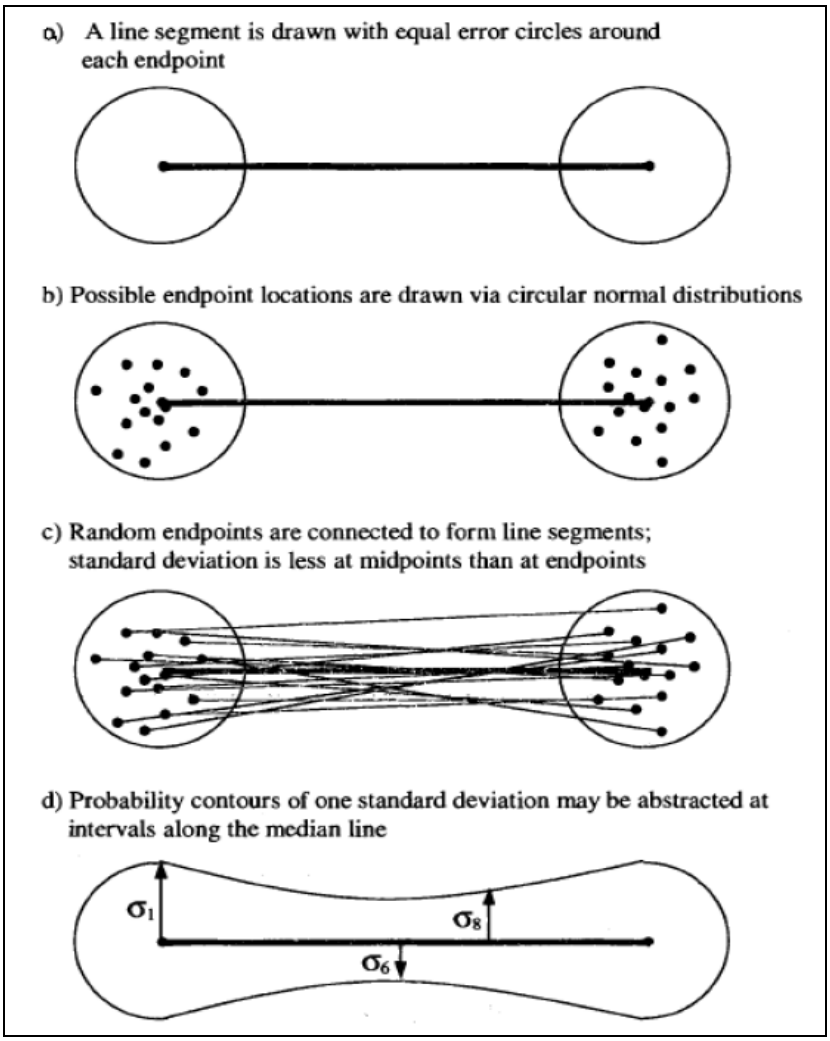

Figure 2.4 Dutton's Experiment Simulating Line Segments (taken from Dutton, 1992)

After Dutton experiment, Caspary and Scheuring refine the fixed radius epsilon band and they proposed a new error band which has a narrower width error band toward the interior of the line segment. This model also uses circular normal distribution as Chrisman's band model. Like Chrisman (1982) and Dutton (1992) they use equal endpoint coordinate errors following a circular normal distribution. Monte Carlo simulation and error propagation are used to derive the error band which becomes narrower towards the midpoint. (Figure 2.5)

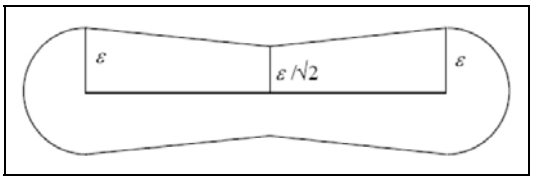

Figure 2.5 Caspary and Scheuring Error Band Model (taken from Caspary and Scheuring, 1992)

This study is important for positional accuracy of spatial objects because it utilizes the error propagation law to derive errors along the line and determines an error band by the error circle boundaries of all points along the line rather than error of the end points strictly. Also this model represents a straight line as four – dimensional random vector (four coordinates of two endpoints) (King, 2002).

Shi and Tempflie (1994) expand the work of Caspary and Scheuring's study as taking it one step further for modelling positional errors in line features. As in the previous model, this model also makes the assumption that errors fits on a normal distribution and uncorrelated. Also the model assumes that positional error distribution of a random point on the line segment is dependent on the errors and distribution of the endpoints. They define the probability distribution function and confidence region of a line segment with equal variances and covariances which indicates independent endpoint errors. A boundary that is formed by computing the probability distribution of a point in a direction to perpendicular to the line indicates how the point's position can vary from its true or mean position (King, 2002). Figure 2.6 of Shi and Tempfli's (1994) graphic displays a line segments' probability distribution and how it can deviate from its expected or mean location.

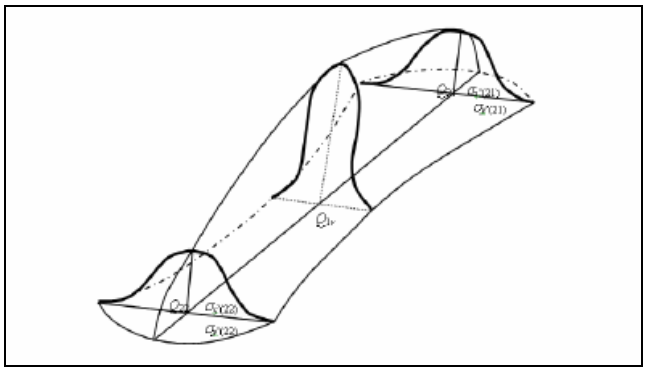

Figure 2.6 Probability density function of a line (taken from Shi and Tempfli, 1994)

Goodchild and Hunter (1997) modified the epsilon band model for assessing positional uncertainty of a digitised line feature. The width of the epsilon band was estimated by calculating the proportion of the measured location of the line feature falling on or inside the epsilon band. If the value of the proportion is equal to or larger than the predefined tolerance (e.g. 0.95), the width of the epsilon band is a measure of the positional uncertainty of the line feature.

Shi and Liu (2000) further developed the studies of Caspary and Scheuring (1992) and Shi and Tempfli (1994) by creating a more general model of the error band, called the G – Band model. The main improvement of the G – band model is that it accounts for the correlations between two endpoints, a condition that is not handled by the previous error band models. Endpoint errors are assumed to follow two-dimensional normal distribution but unlike from the previous studies errors of the endpoints can be different from each other.

Figure 2.7(a) illustrates the more general case of the G-band, which allows each endpoint to have varying errors in both dimensions. In this case, the endpoint errors in the two directions are correlated and of varying magnitudes. Figure 2.7 (b) and (d) displays errors with the same magnitude in the points on the line segments, but at (b) the errors are cross correlated between x and y coordinate pairs of each point, in (d) there is no indication of a cross correlation structure. While Caspary and Scheuring's (1992) error band relies on the assumption of directional independence, this condition applies only special occasions of the G-band. Figure 2.7 (c) illustrates

a special case of the G-band, when endpoint errors are independent and equal. Under these conditions, the G-band reduces to the error band models of Caspary and Scheuring.

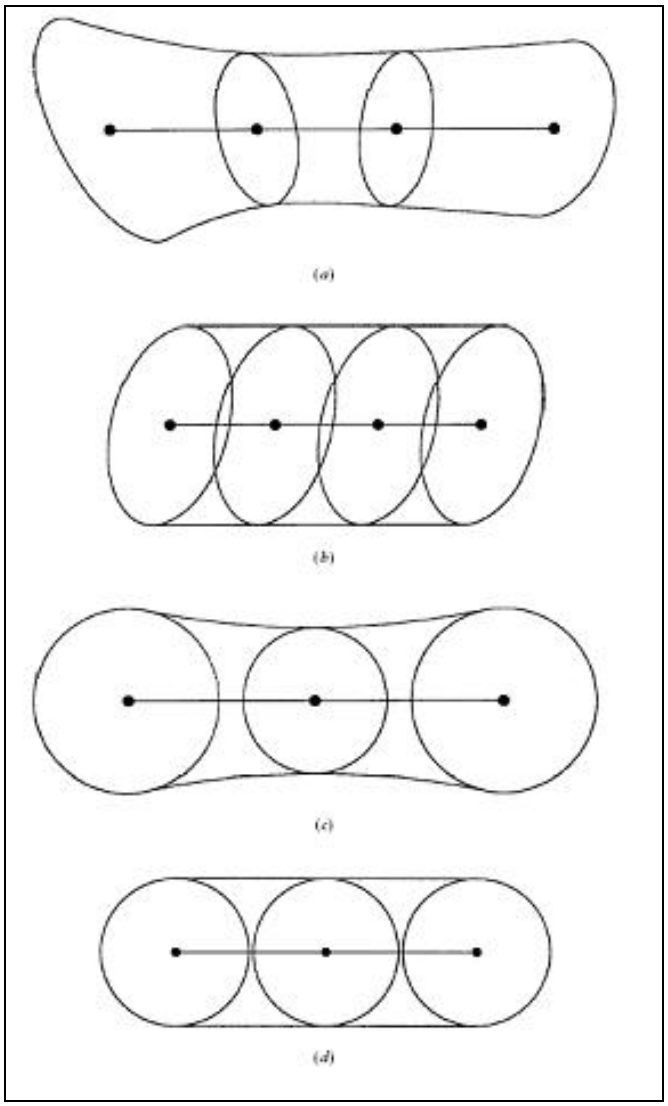

Figure 2.7 G Band of Line Segments with Different Statistical Characteristics
(Adapted from Shi and Liu, 2000)

Kurtar (2006) studied on the G-Band model to develop it for non-linear functional curves. Models for arc, arc string, cubic spline and clothoid are proposed. The proposed uncertainty models for arc geometry is managed using three coordinates,

arc string and cubic spline are based on G-Band model. However the uncertainty models for arc by centre point and clothoid are modelled using epsilon ellipse model. Kurtar (2006) propose epsilon band model for arc by centre point and clothoid because they are created with a single geographic coordinate together with other scalar variables. Hence, the band models of these functional curves use a unique epsilon ellipse for all of random points on the arc line segments.

Yarkınoğlu (2007) worked on a road network geometry consisted of line segments. A new uncertainty model is proposed for network implementation based on the Propagated Error Band model developed by Leung and Yan (1998). This model and G-Band are utilized in this study. Study is carried out in two parts; first models are applied for each line segment of the selected network separately and on the second one uncertainty that propagated form the consecutive segment are utilized considering the positional correlation.

Unlike previous studies Leung and Yan (1998) developed an error model for points, lines and also polygons. Errors in vertices points of lines and polygons are used as the basis for constructing the model. Two main assumption are taken into account in this model (1) all the errors in vertices points have the same standard deviation, (2) the distribution of any point of an entity displays a circular normal distribution.

Zhang and Kirby (2000) examined in their study that “the possibilities by which spatial correlation may be usefully explored in the handling of positional errors in vector data”. An empirical study using photogrammetric data for a suburb – containing line and polygon features of buildings, roads and natural geographic objects such as lakes – displayed that geostatistics can be applied to analyse vector data and photogrammetric data in particular without major complications. Also the study confirmed that spatial correlation should be incorporated in the analysis of positional errors. Also they proposed that geostatistical methods can be used to explore spatial correlation in attribute errors.

Heuvelink et al. (2007) developed a probabilistic framework for representing and simulating uncertain environmental variables both for positional and attribute uncertainty. They categorize the spatial uncertainty in two groups of attribute and positional uncertainty and create taxonomy for both these groups. Uncertainties are associated with estimated probability distribution function (pdf). Then Brown and Heuvelink (2007) developed a software tool called as Data Uncertainty Engine (DUE) using this probabilistic framework study. DUE can deal with both attribute and positional uncertainties. In this study for the assessment of positional uncertainty DUE is utilized so that framework of DUE will be explained in section 2.4 of this chapter.

Kiiveri (1997) suggest a statistical model which assumes that effect of all sources of error give a ‘smooth or rubber sheet distortion’ on true map. Each error source is not modelled separately so that smoothness preserves continuous lines and topology is guaranteed in resulting simulations. Transmitting the uncertainty between set operations such as; intersection and unions are also evaluated within this study. Also uncertainty in length and area of fence lines, roads and property boundaries are modelled.

Bogaert et al. (2005) developed a general framework for error assessment of area measurements of planar polygonal surfaces with application to agriculture. Two cases are evaluated such as; correlated and independent measurements carried out with Global Positioning Systems (GPS) devices. Geostatistical analysis is based on time series data obtained via GPS measurements. An operator made the measurements trough walking along the border line of a field. At the end of the study it is observed that error in area measurements are linked to both the operator's speed and the acquisition rate of GPS device.

2.3.2. Fundamental Mathematical Expression of Uncertain Spatial Features

This chapter aims at describing the positional uncertainty of the spatial features mathematically.

2.3.2.1. Probability Density Function of a Point

Location of a point feature is usually captured by techniques such as, ground survey, photogrammetric or remote sensing survey, map digitising or scanning, or a combination of these. A point in GIS is mostly obtained by digitising a map produced using through the photogrammetric method. Its positional uncertainty is then caused by control, photography, aero-triangulation, orientation, compilation, drafting, printing and digitisation (Shi, 1994) and can be estimated by error propagation law.

Let $f_P(Px, Py)$ denote the probability density function (pdf) of a point *P*. This is the probability that any point (Px, Py) is located at the true location of *P*. Let $X_P = (P_X, P_Y)$, $\mu_P = (\mu_{Px}, \mu_{Py})$ where μ_{Px} and μ_{Py} are the expected (or mean) values of Px and Py respectively, $\sum_P = \begin{pmatrix} \sigma_{Px,Px} & \sigma_{Px,Py} \\ \sigma_{Py,Px} & \sigma_{Py,Py} \end{pmatrix}$ which is covariance matrix of P. If the uncertainty of *P* follows a bivariate normal distribution, $f_P(Px, Py)$ will be

$$f_P(Px, Py) = \frac{1}{(2\pi)\sqrt{|\sum_P|}} e^{(-(X_P-\mu_P)(X_P-\mu_P)^T/2\sum_P)} \qquad (2.2.)$$

where Z^T is the transpose of any vector *Z* (Cheng , 2003).

2.3.2.2. Probability Density Function of a Line or Polygon

Positional uncertainty of a line or polygon is difficult to model statistically when compared to a point feature. Because an infinite number of points located on lines and line segments that form a polygon have correlated. Cheng (2003) stated in her

study that if individual positional uncertainties of all the points on a line are identically and uniformly distributed, a probability density function (pdf) of the uncertainty of the whole line would be like a three-dimensional box. However, the normality of the random uncertainty of a digitised point is usually made for ease of computations (Bolstad et al., 1990; Dutton, 1992; Caspary and Scheduring, 1992). In statistics, it is hard to find a distribution that fits the whole line, because any location on the line follows bivariate normal distribution. Similar situation is also valid when defining a pdf of a polygon.

Cheng et al. (2003) studied the pdfs of the individual points rather than the pdf of the whole spatial feature. These pdfs are in fact different from the pdf of the whole spatial feature; however they can be used to derive multivariate pdf of all the points of the geographical object.

Cheng at al. (2003) define this pdf as follows;

Let $h_A\left(Ax_1, Ay_1, Ax_2, Ay_2, ..., Ax_{N_A}, Ay_{N_A}\right)$ denote the pdf of all the points of a spatial feature *A*. If the positional uncertainty of these points follows a multivariate normal distribution it will become;

$$\frac{1}{\left((2\pi)\sqrt{\left|\sum_A\right|}\right)} e^{-\frac{1}{2\sum_A}(X_A-\mu_A)(X_A-\mu_A)^T} \tag{2.3}$$

where the following represents;

$$\sum_A = \begin{pmatrix} \sigma_{Ax_1,Ax_1} & \sigma_{Ax_1,Ay_1} & & \sigma_{Ax_1,Ax_{N_A}} & \sigma_{Ax_1,Ay_{N_A}} \\ \sigma_{Ay_1,Ax_1} & \sigma_{Ay_1,Ay_1} & & \sigma_{Ay_1,Ax_{N_A}} & \sigma_{Ay_1,Ay_{N_A}} \\ & & \ddots & & \\ \sigma_{Ax_{N_A},Ax_1} & \sigma_{Ax_{N_A},Ay_1} & & \sigma_{Ax_{N_A},Ax_{N_A}} & \sigma_{Ax_{N_A},Ay_{N_A}} \\ \sigma_{Ay_{N_A},Ax_1} & \sigma_{Ay_{N_A},Ay_1} & & \sigma_{Ay_{N_A},Ax_{N_A}} & \sigma_{Ay_{N_A},Ay_{N_A}} \end{pmatrix}$$

$X_A = \left(Ax_1, Ay_1, Ax_2, Ay_2, ..., Ax_{N_A}, Ay_{N_A}\right)$ and $\mu_A = \left(\mu_{Ax_1}, \mu_{Ay_1}, \mu_{Ax_2}, \mu_{Ay_2}, ..., \mu_{Ax_{N_A}}, \mu_{Ay_{N_A}}\right)$

2.4. Probabilistic Framework of DUE

Heuvelink et al. (2007) developed taxonomy for uncertain objects and models for each object to define uncertainty model. In this framework, objects are represented as abstractions of real entities that include features with observed boundaries, such as buildings, trees, and land ownership boundaries. These boundaries contain positional information, such as absolute coordinates in space and time or relative distances between locations. If the coordinates or distances are uncertain, the boundaries contain positional uncertainty.

2.4.1. Taxonomy of Uncertain Objects

Heuvelink et al. (2007) classified the uncertain geographical objects "by their primitive parts and by the types of movement they support under uncertainty" in order to describe positional uncertainty. These classes are as follows by their definition;

- Objects that are single points (point objects);
- Objects that comprise multiple points whose relative position in space-time (internal geometry) cannot change under uncertainty (rigid objects);
- Objects that comprise multiple points whose relative position in space-time can vary under uncertainty (deformable objects).

In contrast to rigid and deformable objects, the positional uncertainty of a point object always leads to a unitary shift in the object's position. Rigid and deformable objects may comprise groups of isolated points, such as the 'trees' in a 'orchard' or groups of interconnected points, such as a 'main road centerline' or a time series of 'water levels', and closed lines or polygons (in 2D or 3D), such as, 'buildings' or 'lakes'. However, the positional uncertainty of a rigid or deformable object is always characterised by the uncertainties of its individual points (Brown and Heuvelink, 2006)

The positional uncertainty of a rigid object leads to - as Brown and Heuvelink (2006) defined - a “unitary shift in the object’s position (translation) and/or an angular shift (rotation)” of the object for any given outcome of the pdf, because the primitive nodes are perfectly correlated. By other means, positional uncertainty cannot alter the topology of a rigid object. In contrast, the topology of a deformable object may be altered by positional uncertainty, because the uncertainties in its primitive points are partially or completely independent of each other.

2.4.2. Models Used for Positional Uncertainty

Methods for defining positional uncertainty in geographic entities include partial and full application of probability theory to vector data. These methods are explained in section 2.2. Heuvelink et al. (2007) developed a general probability method based on the previous studies for point, rigid and deformable objects.

A point object contain at least two uncertain coordinates namely *x* and *y* in 2D Cartesian space. Location of each coordinate is uncertain so that they can be represented by random variables X and Y with a marginal (cumulative) probability distribution function (mpdf) F_X and F_Y.

$$F_X(x) = \Pr ob(X \le x) \text{ and } F_Y(y) = \Pr ob(Y \le y) \tag{2.4}$$

where x and y are real numbers. Marginal distributions may be defined for each coordinate of an uncertain point object. When the errors in *x* and *y* directions are correlated, a multivariate or joint pdf (jpdf) is required:

$$F_{XY}(x, y) = \Pr ob(X \le x \text{ and } Y \le y) \tag{2.5}$$

When the uncertain coordinates are independent, the jpdf is the product of the two mpdfs.

The random variables X and Y will typically have a mean (expected value) $E[X] = \mu_x$ and a standard deviation $\sqrt{E\left[\sum(X-\mu_X)^2/n-1\right]} = \sigma_X$. As a measure of central tendency, the mean provides information about positional bias and standard deviation as measure of dispersion provides information about the average departure of X and Y from their means.

Rigid objects are consisted of multiple points whose internal angles and distances cannot change under uncertainty. The positional uncertainty of a rigid object is characterized by movement of a chosen single point whether a translation of the point and/or a rotation around the point. The point itself may be a primitive point of the rigid object or a reference point associated with the object (e.g. its centroid) (Brown and Heuvelink, 2007; Heuvelink et al., 2007). So that a joint probability distribution function is required for the positional coordinates of this reference point (x and y).

A spatial object is treated as deformable if its component vertices can move with a degree of independence. Description of a positional uncertainty for a deformable object composed of *n* vertices requires a 2*n* dimensional joint probability distribution function where this jpdf contains mpdfs for the coordinates of the individual vertices, together with all the auto and cross correlations between them (Brown and Heuvelink, 2007; Heuvelink et al., 2007).

$$F_{X_1Y_1,X_2Y_2,.....,X_nY_n}(x_1,y_1,x_2,y_2,...,x_n,y_n) = \Pr ob(X_1 \le x_1, Y_1 \le y_1, \\ ...., X_n \le x_n, Y_n \le y_{n1}) \quad (2.6)$$

Heuvelink et al. (2007) stated that obtaining equation (2.6) as the product of n jpdfs specified in equation (2.5) is not practical because data collection and pre-processing will introduce statistical dependencies between points. For example, GPS surveys, georeferencing of remote sensing data, and manual digitizing will all introduce positive correlations between positional uncertainties. Considering these

conditions, DUE proposes a technique that is more practical in the estimation of statistical dependence with pdfs.

The positional uncertainties of objects may be statistically dependent in space and time, and between coordinate dimensions. If the uncertainties are statistically dependent, these dependencies must be estimated alongside the mpdfs. In practice, few parametric shapes are available to describe the jpdf whose mpdfs are statistically dependent. For continuous numerical variables, a joint normal distribution is often assumed. Given this assumption, the jpdf comprises a vector of means and a covariance matrix. The covariance matrix contains the variance of each mpdf along the diagonal and the covariance of each pair of mpdfs that comprise the jpdf elsewhere. Using expert judgement, the uncertainty may be assumed ‘second-order stationary’, whereby the associated pdf has a variance that is constant and for which the covariances depend only on the distance between locations (Heuvelink, 1998). In that case, the covariances may be estimated from a simple function (semi variogram, cross-variogram), which can be fitted directly to a sample of observed errors at control points (Goovaerts, 1997).

CHAPTER 3

DATA AND METHODOLOGY

3.1. Study Area

Study area is located within the boundaries of Kocaeli Province in northwest Anatolia. Area contains agricultural fields and villages scattered around 30 km². Geographical extent of the study area is upper left 40° 47’ 42.25”N 30° 3’ 21.7”E, lower right 40° 46’ 16.8”N 30° 5’ 17.5”E. Elevation within the selected region ranges from 5 metres to 410 metres (Figure 3.1).

This area is selected due to agricultural field density and low variation at the altitude which is also corrected by orthorectification. When compared to urban areas and buildings, agricultural fields are located exactly on terrain surface. In remotely sensed images height of geographical objects may cause positional bias in terms of X and Y coordinates of feature. Therefore the errors that may be occurred depending on the heights of building due to the perspective distortion during image acquisition are removed or alleviated. Another characteristic of agricultural field is its boundaries are static. For instance lakes or shorelines are affected from seasonal fluctuations during the year and do not have static boundaries. However agricultural land parcel boundaries are static and can be defined with cadastral studies. The availability of the cadastral data is another reason in selecting this area for the case study.

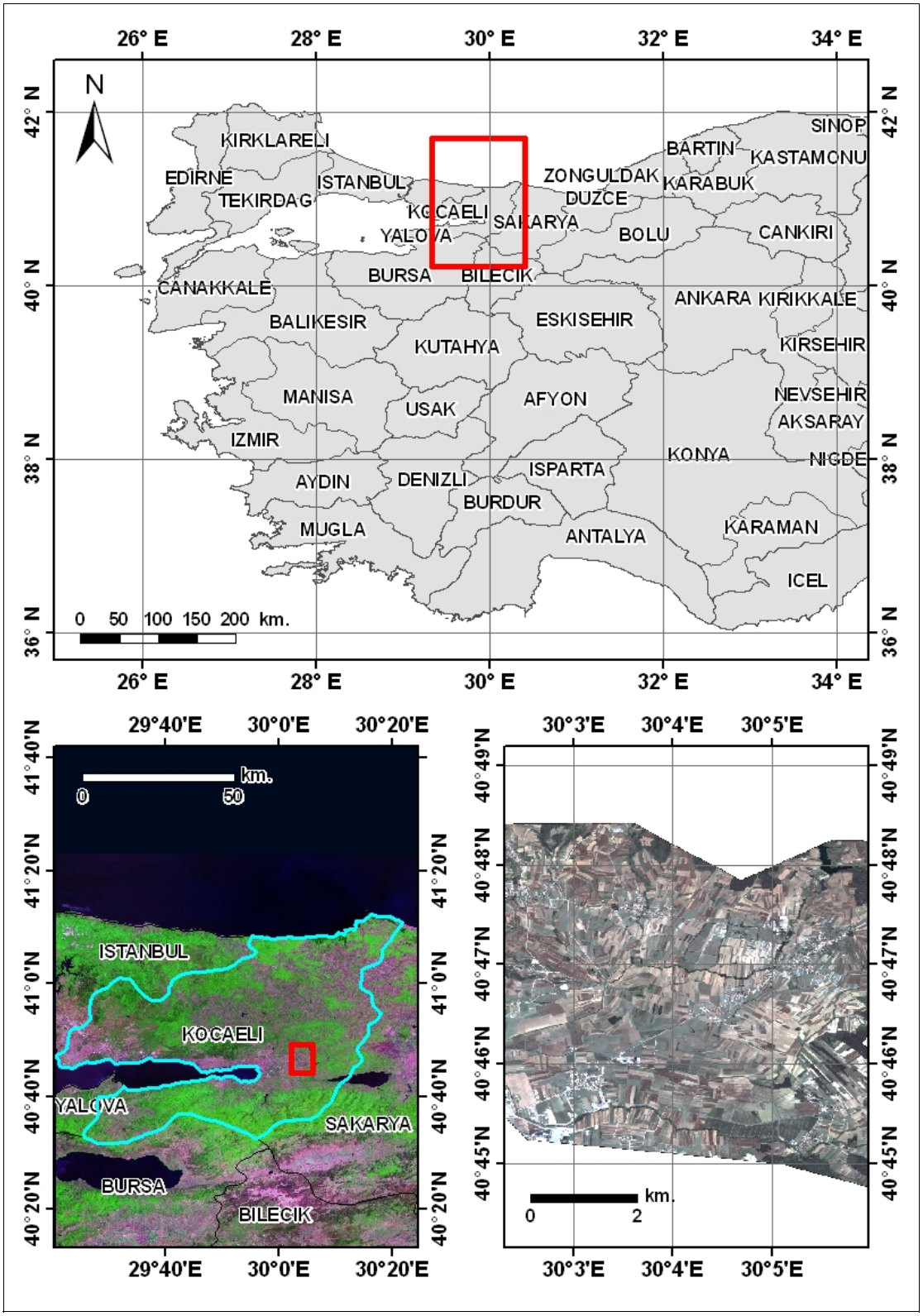

Figure 3.1 Location of the Study Area

3.2. Data Description

Data that constitute the base information for the study is obtained from Municipality of Kocaeli. They consist of 1:5000 scaled digital base map series. These map sheets were produced from aerial colour photographs acquired in 2005 with 35 cm. / pixel nominal spatial resolution via photogrammetric production. Planimetric accuracy of this production is known to be about ±35 cm. (RMSE) on the ground. 1:5000 scaled digital base maps contain detailed information about structure of land as parcel boundaries, road network, fences, man – made structures etc (Figure 3.2).

Second data used in thesis is satellite imagery of study area. Two images of the same region acquired at different dates ordered for the study. Satellite imagery is obtained with the fund provided from Scientific Research Project (BAP) under the supervision of Natural and Applied Sciences Institute. Pre – processing stage is explained in orthorectification section. One of these images is used as basis for digitizing agricultural land parcels. Also digitizing process is explained in digitizing section.

Agricultural field boundaries that are used as basis for all analysis are extracted from 1:5000 scaled digital base maps as vector dataset. These data used as reference and assumed to be the true positions of field boundaries. And the accuracy of this cadastral data is analysis is at most 1 meter on the ground.

All raster and vector dataset (e.g. satellite imagery, cadastral boundaries, 1:5.000 scaled base maps, simulations) utilized in this study are referenced with the following datum and projection domain; ITRF96 (International Terrestrial Reference System 1996) Datum, GRS80 (Geodetic Reference System 1980) Ellipsoid, 3º Gauss Kruger (aka Transverse Mercator) Projection.

Figure 3.2 1:5.000 scaled GCP reference dataset. Map Sheet G-24-D-03-A

3.3. Orhorectification

A very high resolution optical satellite imagery acquired on 16th of June 2007 was utilized in this study in order to digitize 221 polygon vector dataset. Metadata of this image, acquired by Quickbird satellite platform, is given in Table 3.1. Image can be previewed in Figure 3.3.

Table 3.1 Metadata of Quickbird Image

Catalog ID	1010010005B06700
Satellite Platform / Operator	Quickbird / Digital Globe US
Date / Time of Acquisition	**16.06.2007** / 09:15
Processing Level / Type	2A / Standard Ortho Ready
Scan Mode & Direction / Number of Looks	Full Swath & Forward / 1 (Stereo mode off)
Number of Bands / Band Combination	4 / Red Green Blue (RGB) + Near Infrared (nIR)
Radiometric Resolution	11 bits (2048 shades of gray) per band
Resampling Kernel / Nominal Spatial Resolution	Pansharpen / 60 cm. per pixel
Mean Ground Sampling Distance	0,705 m.
Mean Sun Azimuth	147,8 °
Mean Sun Elevation	70,3 °
Mean Satellite Elevation	65,0 °
Mean in Track View Angle	23,4 °
Mean Cross Track View Angle	- 0,8 °
Mean Off Nadir View Angle	23,5 °

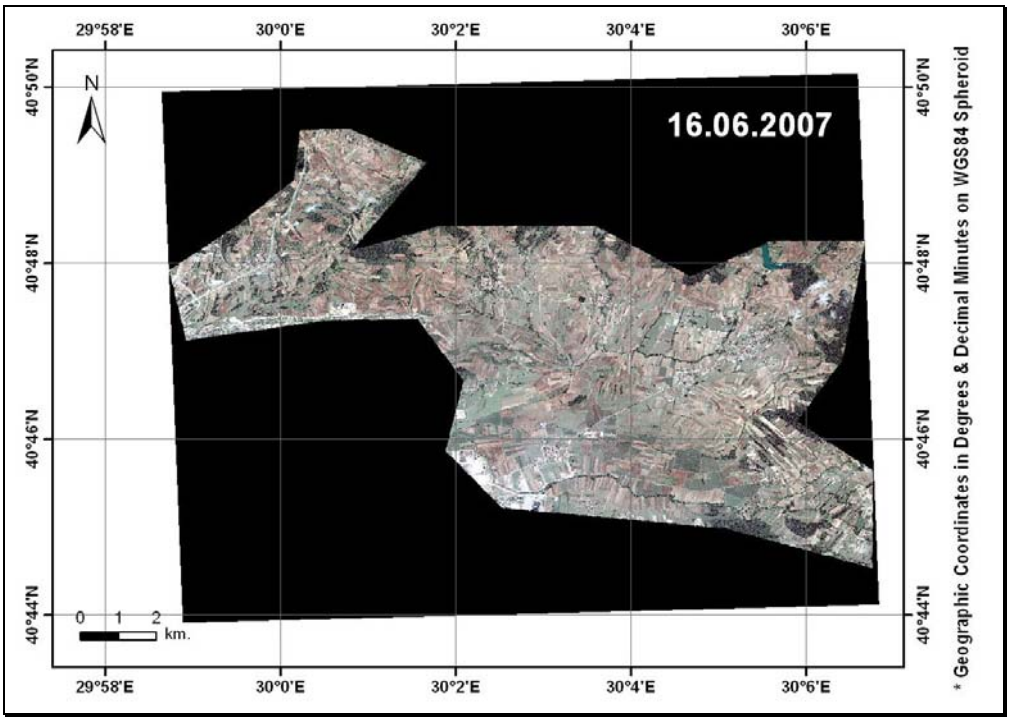

Figure 3.3 Previews of Quickbird Image

Quickbird image is processed under PCI Geomatica v10.0.3 and ortho-rectification is performed in OrthoEngine via Toutin's (high resolution) rigorous physical model which fully utilizes ephemeris data.

For year 2007 image, having the catalog id of 1010010005B06700, 50 ground control points (GCPs) are selected from 1:5.000 scaled digital base map series (Figure 3.2). These map sheets were produced from aerial colour photographs acquired in 2005 with 35 cm. / pixel nominal spatial resolution via photogrammetric production. Planimetric accuracy of this production is known to be about ±35 cm. (RMSE) on the ground.

Besides, with reference to the same dataset having contour intervals of 2,5 m., digital elevation model (DEM) of the study area having 1,20 m. / pixel spatial resolution was produced with an accuracy of ±48 cm. (RMSE) on the vertical axis (Figure 3.4). DEM is produced using universal kriging surface interpolator algorithm via second order trend surface in ArcGIS v. 9.3 Geostatistical Analyst.

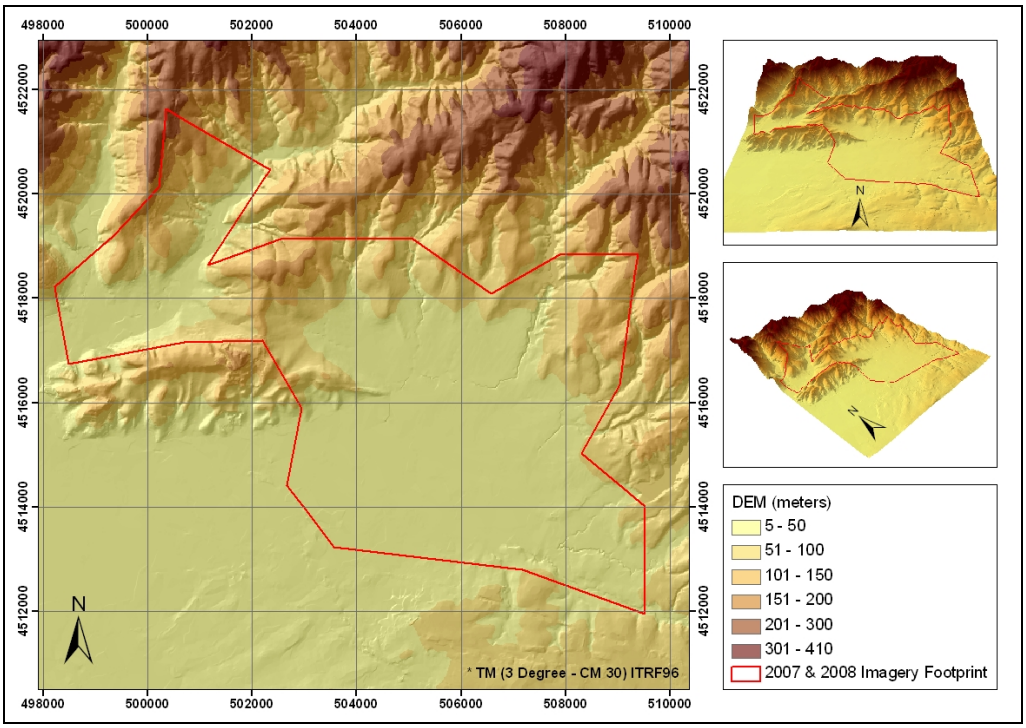

Figure 3.4 DEM of the study area produced from 1:5.000 scaled map sheets.

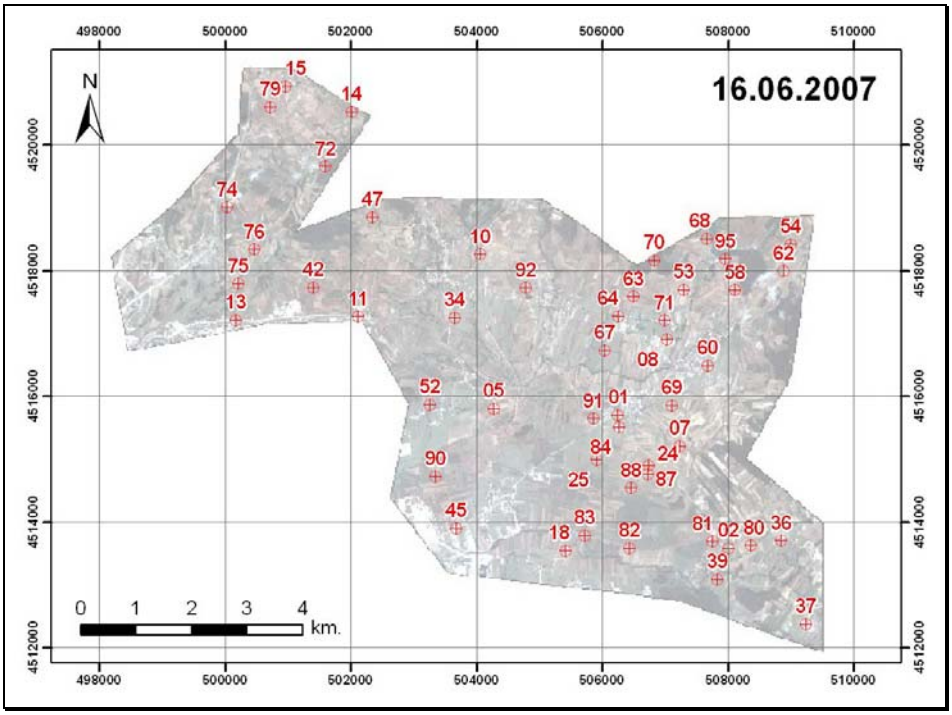

Figure 3.5 Distribution of GCPs

Utilizing Toutin's (high resolution) rigorous physical model, bundled in OrthoEngine v10.0.3, achieved planimetric accuracy for ortho-rectified image is ±0,98 pixel (RMSE) for 2007 image. Detailed results are given in Table 3.2. Full residual report, including complete list of GCPs is given in Appendix A.

Table 3.2 Calculated error values for ortho-rectification via rigorous physical model

	ID 1010010005B06700 **Date 16.06.2007**
Number of GCPs	50
GCP RMSE X axis	0,72
GCP RMSE Y axis	0,67
GCP RMSE X & Y	**0,98**

3.4. Digitizing of Test Data

To undertake uncertainty analysis, experimental test data is generated using digitising. Digitising is performed using ESRI ArcGIS v.9.3 from orto - rectified satellite imagery acquired in year 2007. Spatial features are constituted of agricultural land parcel boundaries. All digitizing process is performed at 1:1000 scale on a display with 1028x1024 pixel resolution in ArcGIS v.9.3 for eliminating any variation, which can be caused by resolution and scale variation. All 700 polygon boundaries digitized in clockwise direction starting from lower left corner of parcels. While generating geographical features via digitising auto - complete polygon option was selected from topology tasks to avoid gaps between adjacent parcel boundaries. Both reference data and data generated from satellite imagery via digitising can be seen in Figure 3.6. In order to ease further analysis, each of the 700 polygons is registered with a unique feature ID which is the same both in reference and digitized vector dataset.

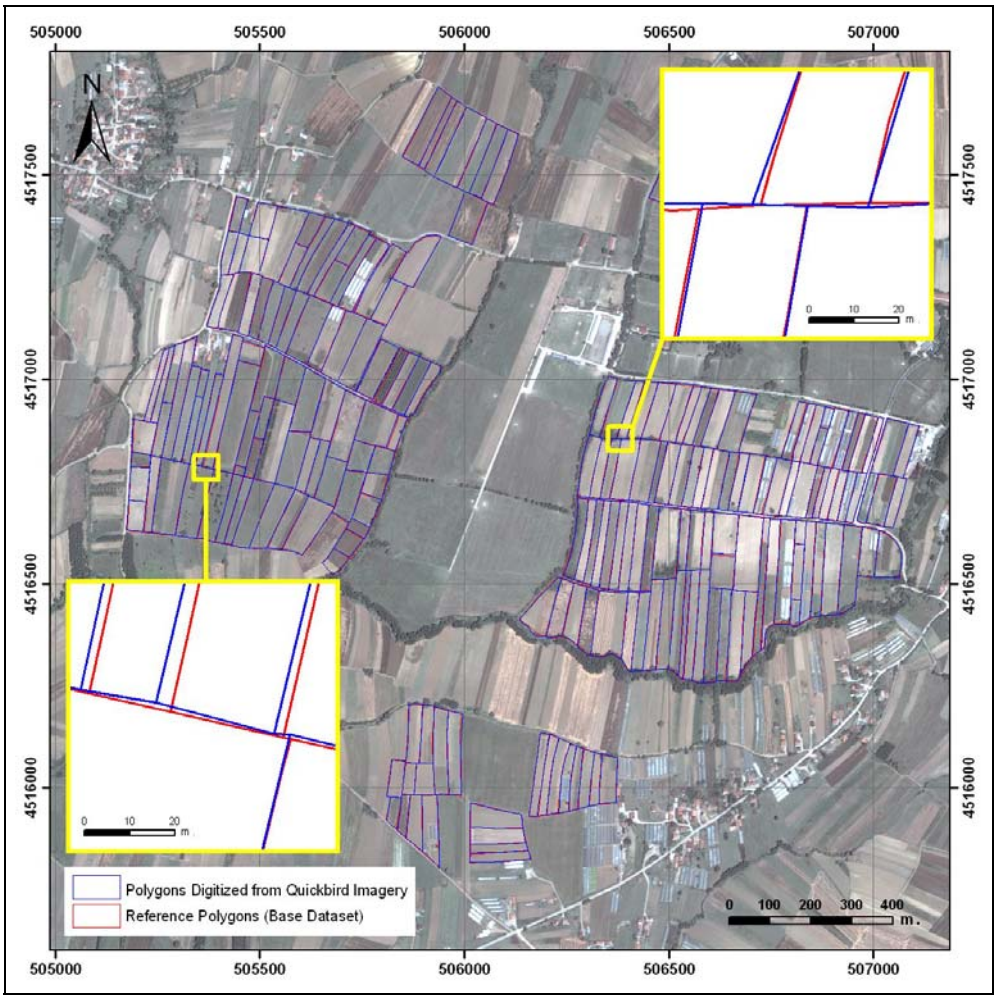

Figure 3.6 Polygons Digitized from Satellite Imagery and Reference Polygons

3.5 Data Uncertainty Engine (DUE)

Data Uncertainty Engine (DUE) is a free software that aids the user in defining probability distributions for uncertain spatial objects and draws random samples using Monte Carlo Simulation (MCS) from these distributions. DUE is developed by by James D. Brown and Gerard B.M. Heuvelink (2007) using java programming language It runs on the Java TM Runtime Environment (JRE) version 5.0 (1.5.0) or higher.

DUE enables to describe spatial and temporal patterns (time – series data) of uncertainty which can be called auto correlation and also cross – correlations available in the related inputs that can be included in uncertainty model. Determination of a probability distribution function (pdf) in DUE for the positional uncertainty of 2D spatial vectors includes correlations within and between coordinates. Such correlations may greatly influence the outcomes of an uncertainty analysis because models typically respond differently to correlated variability than random errors. DUE also supports the quantification of positional uncertainties in geographic objects, presented as raster maps, time-series or vector outlines.

Objects that comprise multiple points, such as lines and polygons, may be assumed "rigid" under uncertainty, where all internal coordinates move identically, or "deformable", whereby each internal point can move separately. The uncertainty of a rigid object is completely specified by a translation and/or rotation of that object about a single point. Examples of rigid object might include buildings whose boundaries are theoretically rigid or field whose boundaries are treated as rigid for simplicity. In contrast, the uncertainty of a "deformable" object requires the marginal uncertainties to be defined at all internal points, together with any relationships between them (Heuvelink et al.,2007)(Figure 3.7).

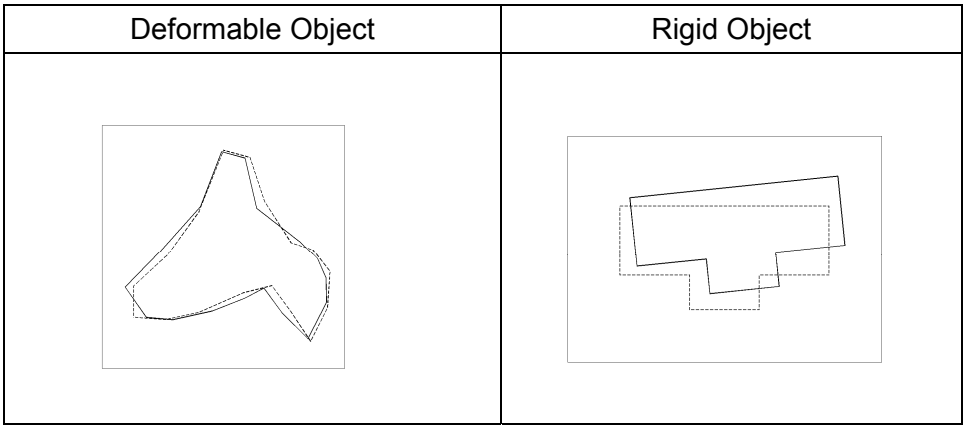

Figure 3.7 Examples of Deformable and Rigid Object

Both attribute and positional uncertainties can be evaluated using probability distribution functions. Objects supported by DUE include spatial vectors, space-time vectors, spatial rasters, time-series of rasters, simple time-series and objects that are constant in space and time. Attributes supported by DUE include continuous numerical variables, discrete numerical variables and categorical variables (Brown & Heuvelink, 2007).

An uncertainty assessment with DUE is separated into five stages. These stages are presented as 'tabbed windows' in DUE. Also the workflow of DUE is shown in Figure 3.8.

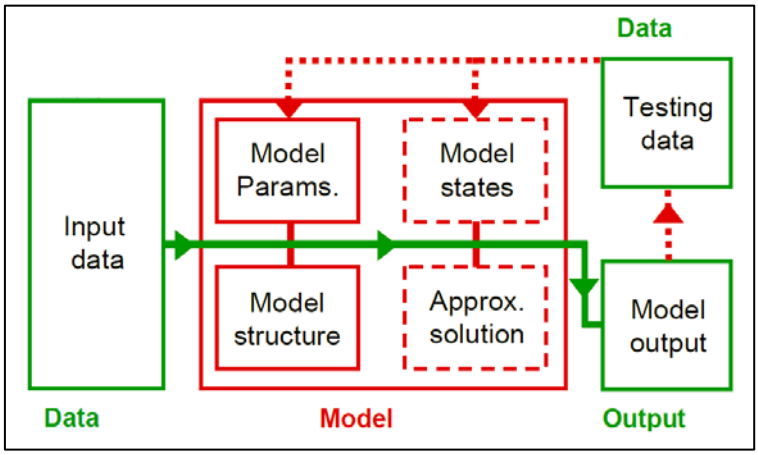

Figure 3.8 Workflow of DUE (Adapted from Brown and Heuvelink, 2007)

1. Input Window - Loading and saving data

Supported Data Files in DUE are as follows.

- ESRI Shapefile for spatial vector datasets (e.g. points,lines,polygons)
- ASCII Raster for 2D raster data (.asc)
- ASCII file for simple time series data (.tsd)
- Also searching, retrieving and saving pdfs for time series in a DUE – enabled Oracle – ArcSDE database.

A snapshot of Input window is presented in Figure 3.9.

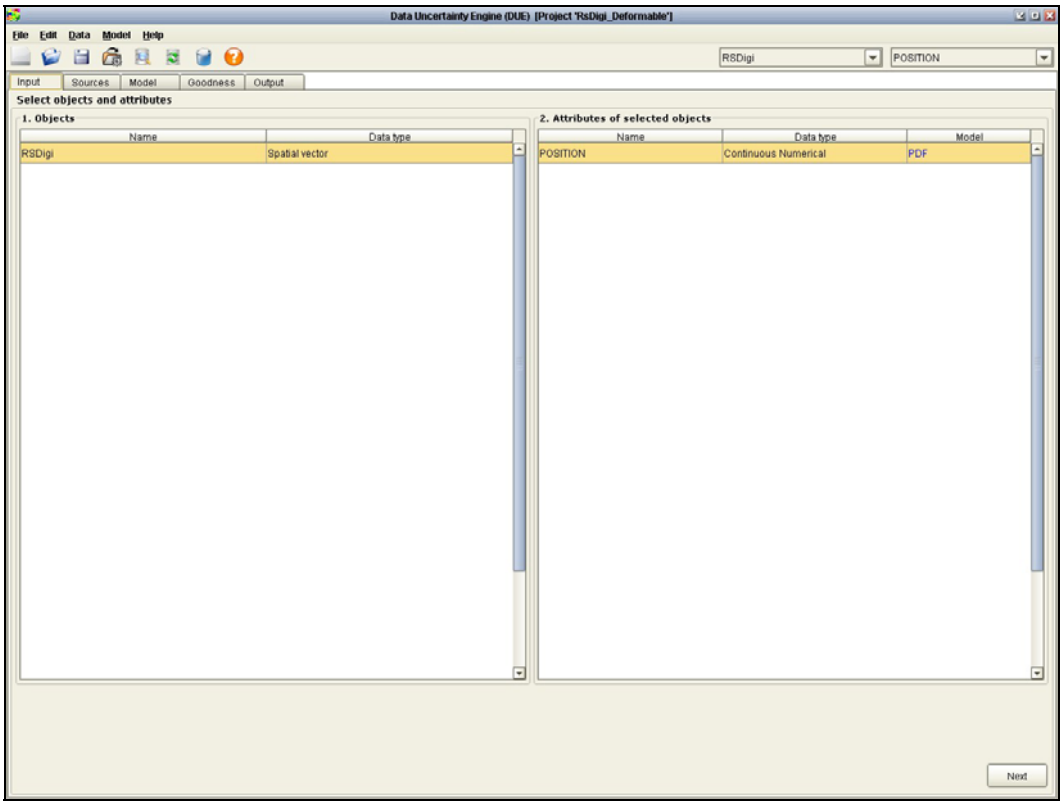

Figure 3.9 Input window of DUE

2. Sources Window - Identifying the causes or sources of uncertainty

A library of sources included in the software as an example, sources of uncertainty generally classified as follows;

- Instrumental Accuracy
- Sampling Design
- Sample Representativeness
- Statistical Modeling
- Classification Accuracy
- Class Definition

3. Model Window - Defining an uncertainty model

Once data is imported into DUE, an uncertainty model can be defined for the objects and attributes selected in the opening tab. In the first window of the “Model” pane, an uncertainty model structure is chosen for the selected objects and attributes.

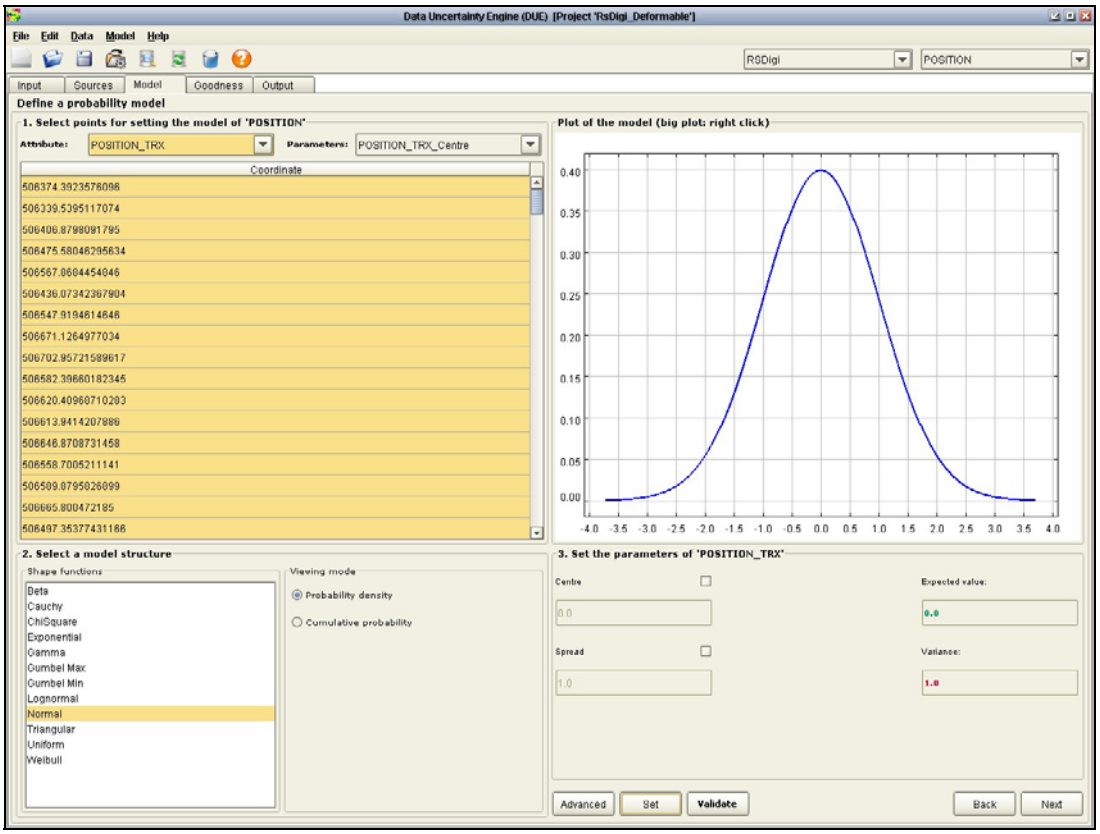

Figure 3.10 Model Window of DUE

If sample data are available, it can be loaded. In the absence of sample data, an uncertainty model is defined through expert judgement. If the uncertainties are assumed spatially correlated, then a correlation model must be defined in the following model window. (Figure 3.10)

4. Goodness Window - Reflecting on the quality of the model

When constructing uncertainty model in DUE stages 2 and 4 (describing the sources of uncertainty and assessing goodness) can be skipped.

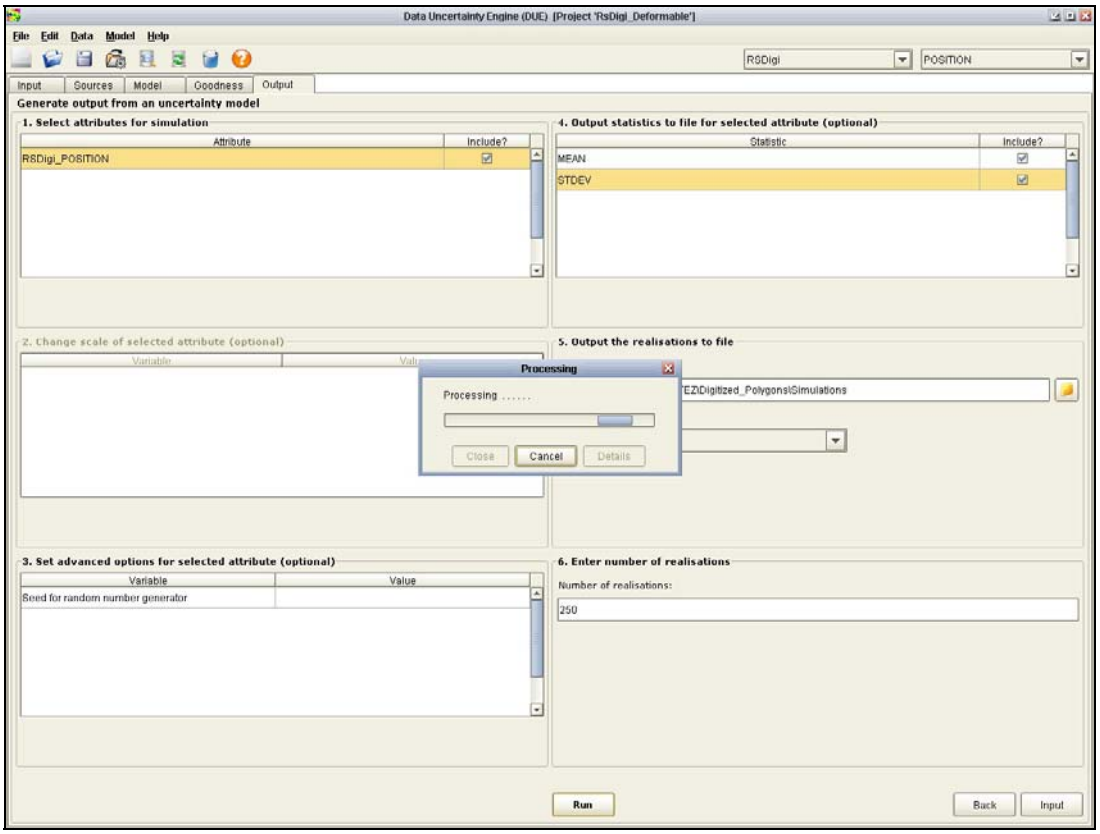

Figure 3.11 Output Window for DUE

5. Output Window - Simulating from an uncertainty model

Once complete, an uncertainty model is used to generate realisations of the uncertain objects and attributes in the output window (Figure 3.11). In order to simulate from an uncertainty model, the output scale, the number of realisations and the location for writing data must be specified

3.5.1 Worked Examples with DUE

DUE has the functionality to evaluate both attribute and positional errors in spatial data. But this study focuses on positional uncertainty. Following examples display the use of DUE within spatial projects.

The Dutch Ministry of Housing, Spatial Planning and the Environment initiated a project that aims to provide tools for handling uncertainty in spatial planning. Positional uncertainty in delineated breeding bird areas was evaluated as one of the aspects of the study. Heuvelink (2007) stated that these areas are considered to have exact boundaries however in reality there is no certain knowledge about where these areas begin and end. Vullings et al. (2007) explains the aim of the project to analyse how positional uncertainty of these areas would affect the outcome of the spatial planning process.

The shape files of breeding bird areas were loaded into DUE and a statistical model (probability distribution function) of uncertain position of the polygons was constructed. Parameters of standard deviations in the x and y coordinates and correlation between neighbouring vertices of each polygon were defined. Standard deviations in the x and y directions were assumed to be equal. The correlation was assumed to depend only on the distance between points, where a Gaussian-shaped semivariogram was employed. Then DUE was used to simulate 100 possible realisations of the delineated breeding bird areas. Many of the simulated polygons were topologically corrupt so that they were replaced by new simulations. After running the spatial planning process, it turned out that the positional uncertainty about the breeding bird areas marginally affected the final plan. (Heuvelink 2007, Vullings et al., 2007)

Cultivation of fields is executed using GPS driven farming vehicles in Netherlands such that precision farming requires that field boundaries are measured with cm level accuracy. To avoid losses caused by unharvested crops or wasted inputs de Bruin et al. (2008) presented a general error propagation method using DUE.

Propagation of positional error measurements that are carried out with three different scenarios is investigated. Spatial dependence structure is based on a statistical model that compromises of temporal correlation in positional measurements.

A potato field of 15 ha is used as case study and error model was parameterised on measurements' scenarios presenting;

(1) Manual digitization of field boundary
(2) Real Time Kinematic GPS survey
(3) Differential GPS- based field checks

Methods used in Bogaert et al. (2005) (explained in Chapter 2, page 18) are adopted in this study with extension to Data Uncertainty Engine.

14 corner points of the polygon are modelled in scenario 1 and 2; however in scenario 3 of differential GPS number of vertices were 1258. Temporal autocorrelation and cross correlation structure were modelled using semivariogram analysis. For each model 250 simulations of the case area is created using DUE.

Inclusion and exclusion areas for each of 250 simulations are computed. Then mean and standard deviation of these 250 realisations are calculated to observe the performance of scenarios. Best results are obtained with Real Time Kinematic GPSs.

CHAPTER 4

UNCERTAINTY ANALYSIS

In order to analyze positional uncertainty using Data Uncertainty Engine (DUE), measures of spatial correlations, semivariograms of positional displacements both in X and Y coordinates and cross – semivariogram of X and Y displacements are needed to be constructed. Modelling of semivariograms is performed to define spatial autocorrelation and cross correlation structures in positional errors. These model parameters are used as inputs for running simulations from digitized features using joint normal distribution. These model parameters are also used to construct a valid variance – covariance structure for uncertainty models used in DUE.

Realisations of possible positional displacements generated by DUE give possible locations of digitized dataset within pre–defined uncertainty error model. These realisations allow user to define the uncertainty boundaries for uncertain dataset.

To run simulations it is necessary to compute semivariograms and then to fit suitable semivariogram models. This chapter explains steps of uncertainty analysis from data preparation to running simulations.

Three different sample applications are carried out with DUE on different locations of the same dataset. In the first example a training set of 221 polygons are selected to identify error model for a larger set of 479 polygons out of a total number 700 polygonal boundaries. In the second and third example sample regions of 95 and 30 polygonal boundaries are selected respectively for analysis. Both in these two examples 5% of population are selected as training data for error model parameter estimation.

4.1. Data Preparation

To provide initial estimation of positional errors of digitized land parcel boundaries from Quickbird imagery, reference data produced from cadastral data are overlaid with vector data digitized from satellite imagery. In order to obtain main vertices of polygons, line simplification algorithm in ArcGIS Desktop v9.3 ArcToolBox is used to eliminate redundant point's information in polygon features. Every polygon is registered with a unique feature ID; also vertices are labelled with unique ID indicating which points belong to which polygon. Point IDs are given starting from the lower left corner of the polygon and then following a clockwise direction. Vertices of digitized polygons, which are converted to 506 unique points, can be seen in Figure 4.1.

It is possible to match polygon vertices across two data layers and to compute displacements for individual points. Matching points both in reference data and digitized polygons' X and Y coordinates are extracted from each other to provide displacements of individual points in X and Y coordinates.

Initial information about positional errors in X and Y coordinates is acquired via descriptive statistics measures and histograms. Table 4.1 represents descriptive statistics both in positional displacements and absolute values of displacements.

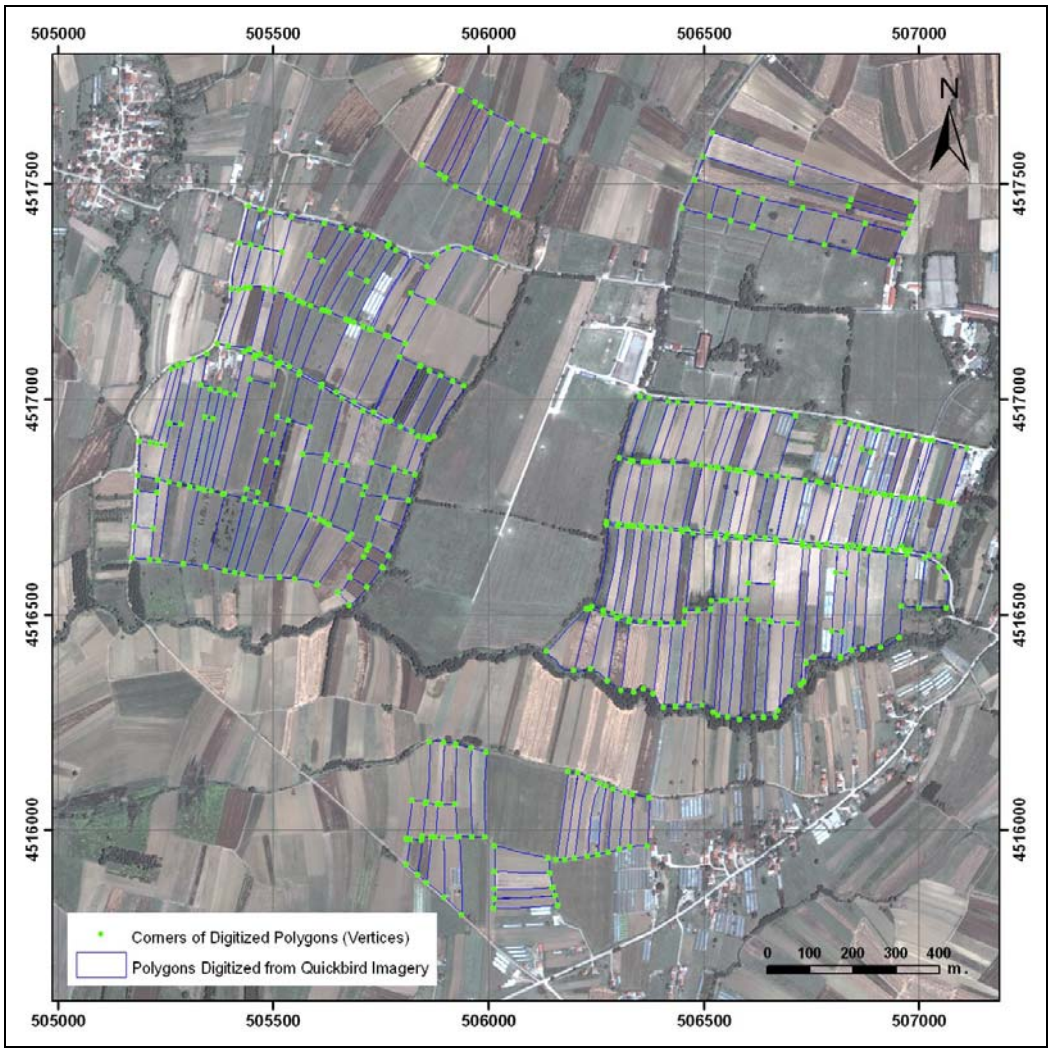

Figure 4.1 Polygon Vertices Points

Table 4.1 Descriptive Statistics of Displacements in Coordinates

	ΔX	ΔY	Abs(ΔX)	Abs(ΔY)
N	497	497	497	497
Mean (m)	**-.05944**	**.21696**	**.86467**	**.88424**
Standard Deviation (m)	**1.156449**	**1.274988**	**.769248**	**.943032**
Variance (m)	**1.337**	**1.626**	**.592**	**.889**
Skewness	.087	1.469	1.572	3.225
Kurtosis	1.343	7.648	2.835	18.265
Range (m)	**8.120**	**12.414**	**4.060**	**8.884**
Minimum (m)	-4.052	-3.526	.008	.004
Maximum (m)	4.068	8.888	4.068	8.888

In DUE, errors are assumed to follow a normal distribution. Also semivariogram construction compromises this assumption; positional errors represent normal distributions.

Histograms for displacements both in X and Y axis are calculated before semivariogram construction to check normality assumption. Figure 4.2 is prepared using SPSS v15.0. Negative and positive values on histograms indicate direction in positional errors. When histogram of positional errors displays normal distributions for both X and Y, absolute values show skewed distribution since all errors are converted to positive values. It can be interpreted from the histograms that majority of the positional errors in both X and Y axes are less than 1 meter.

Displacements in X and Y coordinates are re-organised from ArcGIS “shp” shape file format to GSlib’s text based “dat” file format in order to construct semivariograms, standardised semivariograms and correlograms in the geostatistical software package GSlib, which is a free geostatistical software.

The data input files for GSlib need to be in a specific format common to many geostatistical software packages. Each data file starts with a header line containing a descriptive title. Next follows a line with the number of variables. The following set of lines contains the variable names, one per line. Next are the actual values, with a new column for each observation, and the values separated by tabs or spaces. The last line in the file should be a blank line.

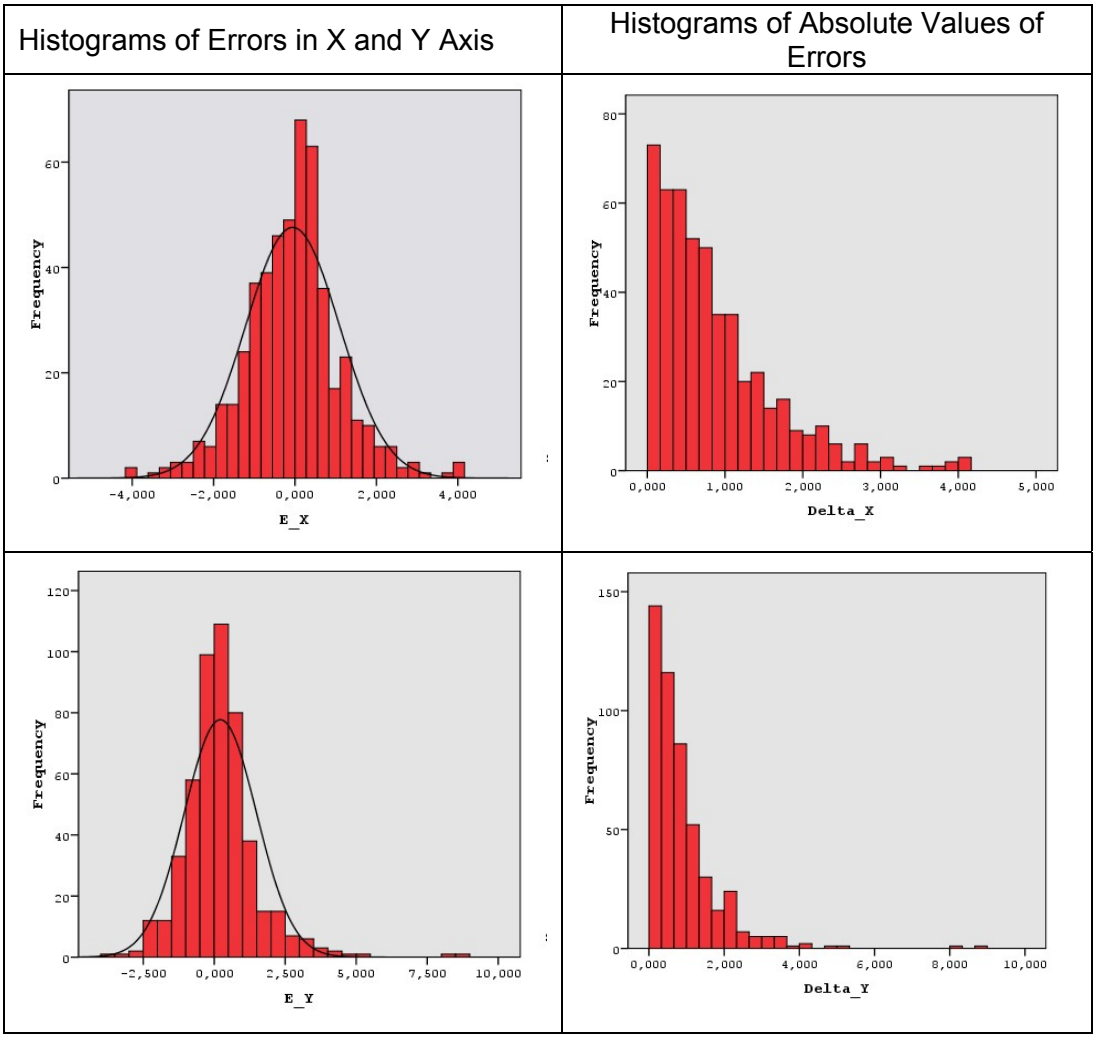

Figure 4.2 Histograms of Errors in X and Y Axis

4.2. Parameter Estimation

Practical use of probability distribution functions usually involves a combination of choosing a parametric shape and estimation. Auto correlations, cross correlation of uncertainty in the positions of vertices and the random deviations of polygons require description of statistically dependent joint probability distribution functions.

Using the familiar assumption of second order stationarity (Isaaks and Srivastava, 1989, Goovaerts 1997), covariances and cross-covariances can be estimated from

semivariograms that are fitted to samples of observed reference positions (Heuvelink et al, 2007).

Semivariograms of the deviations are computed using geostaistical software package GSlib.

4.2.1. Semivariogram

A semivariogram is a geostatistical technique which can be used to examine the spatial continuity of a regionalized variable and how this continuity changes as a function of distance and direction.

The correlation structure of the positioning can be made by estimating semivariogram of displacements both in X and Y coordinates. The basic idea is to look at points separated by lag distance *h* and to compare measurement values. (Kanevski & Maignan, 2004)

$$\gamma_{ij}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (Z_i(x) - Z_i(x+h))^2 \qquad (4.1)$$

Computation of a variogram involves plotting the relationship between the semivariance $\gamma_{ij}(h)$ and the lag distance (*h*). (Equation 4.1) The semivariance can be defined as one half of the variance of the difference between points separated by a lag distance (*h)*, such that: where $\gamma_{ij}(h)$ is the semivariance, $Z_i(x)$ is the measured sample value at point *i*, $Z_i(x+h)$ is the measured sample value at point *i*+*h* where *h* is the lag distance, and *N* is the number of observations for a particular separation or lag distance. The lag distance is defined as the distance separating two points within a dataset in a specific direction (Isaaks and Srivastava, 1989).

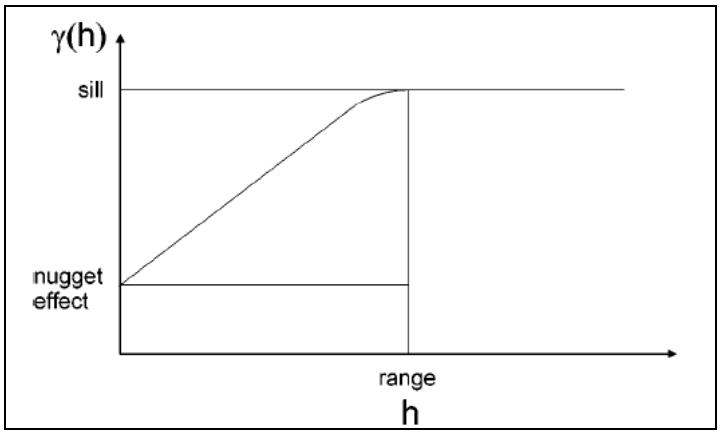

Figure 4.3 Parameters of a semivariogram

When the semivariance is plotted against the lag distance, certain features are displayed on the resulting semivariogram (Figure 4.3). These features include the sill, range and nugget effect.

4.2.2. Semivariogram Construction

The formula involves calculating the difference squared between the values of the paired locations with an increment of *h* lag distances. For semivariogram construction parameters of lag distance, lag tolerance and number of lags are needed. Rule of thumb indicates that the multiplication of the lag distance with the number of lags should be about largest distance among all points, in this case it is about 1800 meters. Actual largest distance is 1940 meters but above 1800 meters semivariance values starts to diminish which indicates no significant correlation structure.

Selection of lag distance and number of lags requires a trial-error procedure of different values. Optimum values representing the auto and cross correlations are selected as follows. Lag distance is selected between points is about 4.5 metres and lag tolerance is selected as 2 meters. Number of lags is selected as 400. Resulting semivariogram graph for positional errors in X coordinate can be seen in Figure 4.4.

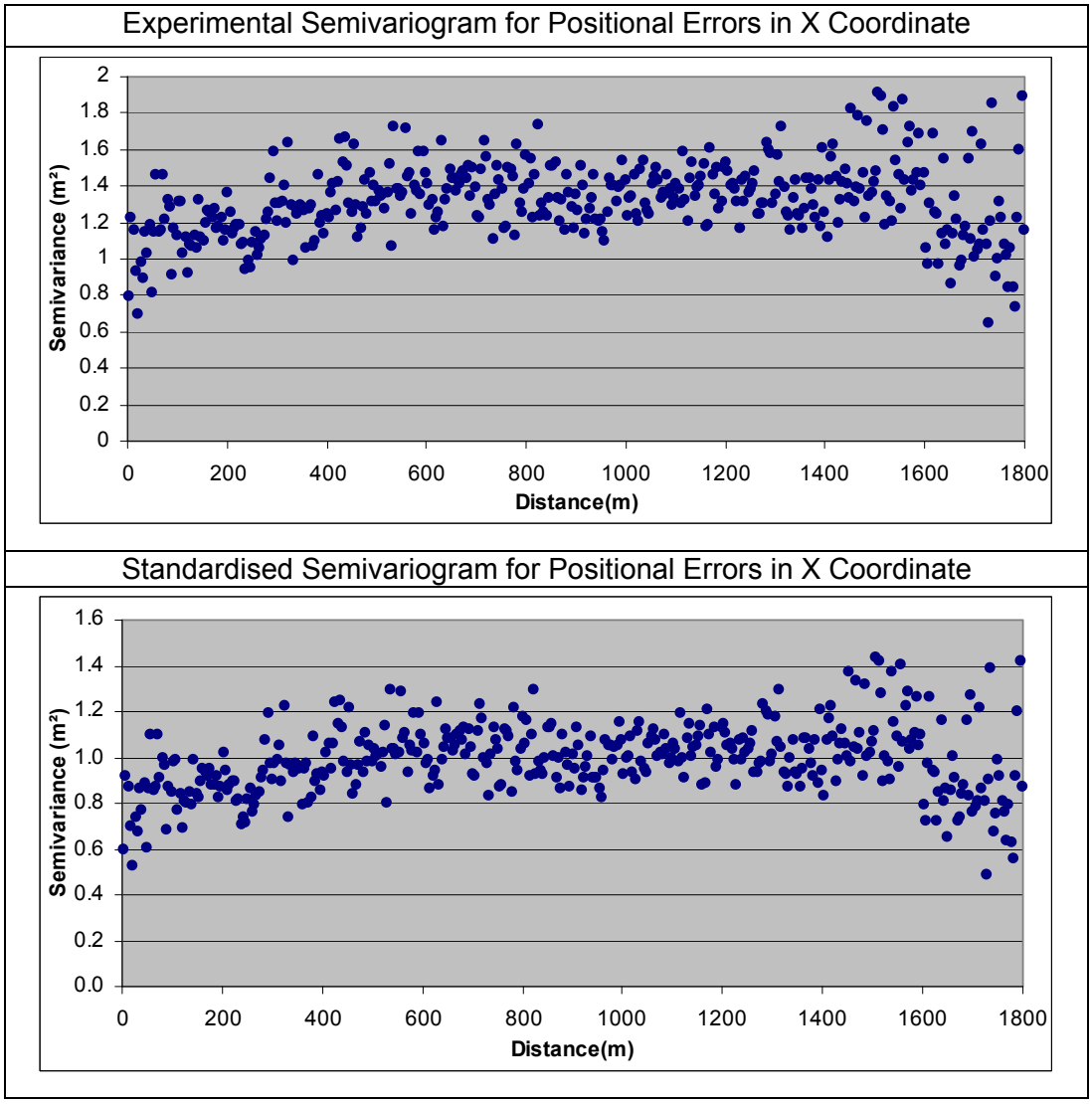

Figure 4.4 Semivariograms for X Coordinate Errors

However semivariogram models included in DUE uses sill that is standardised at value 1 so that semivariogram values standardised by diving each semivariogram value by its lag variance. Both semivariograms and standardised semivariograms for X and Y coordinate errors are presented in Figure 4.4 and 4.5 respectively. Model fitting is performed by using standardised semivariograms.

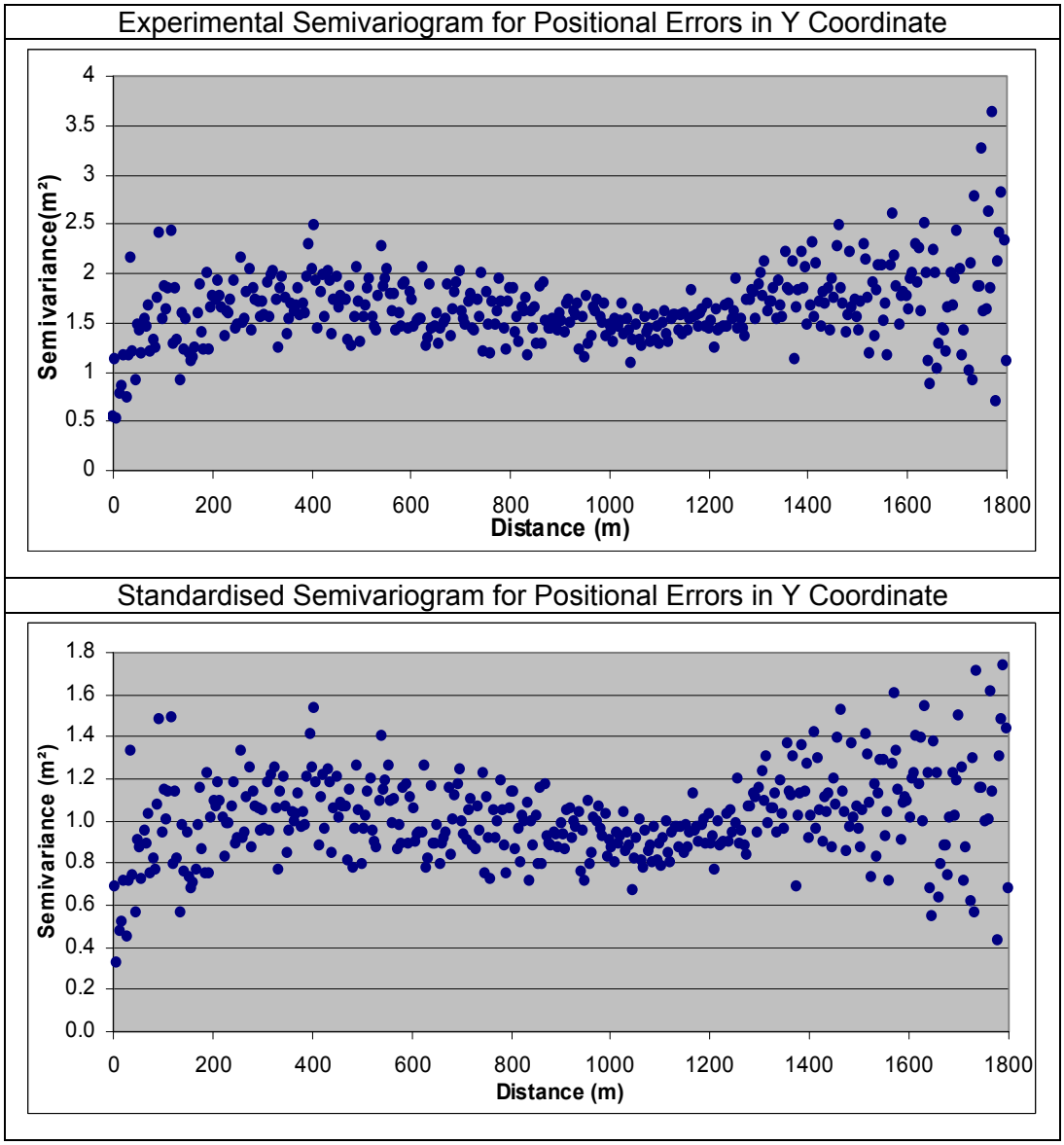

Figure 4.5 Semivariograms for Y Coordinate Errors

In order to evaluate whether positional errors in X and Y coordinates have any affect on another, cross correlation between errors should be investigated. To evaluate cross correlation a cross – semivariogram is constructed with the aid of GSlib. Also in DUE cross correlation can be defined in uncertainty model calculation. Calculated cross semivariogram can be seen in Figure 4.6.

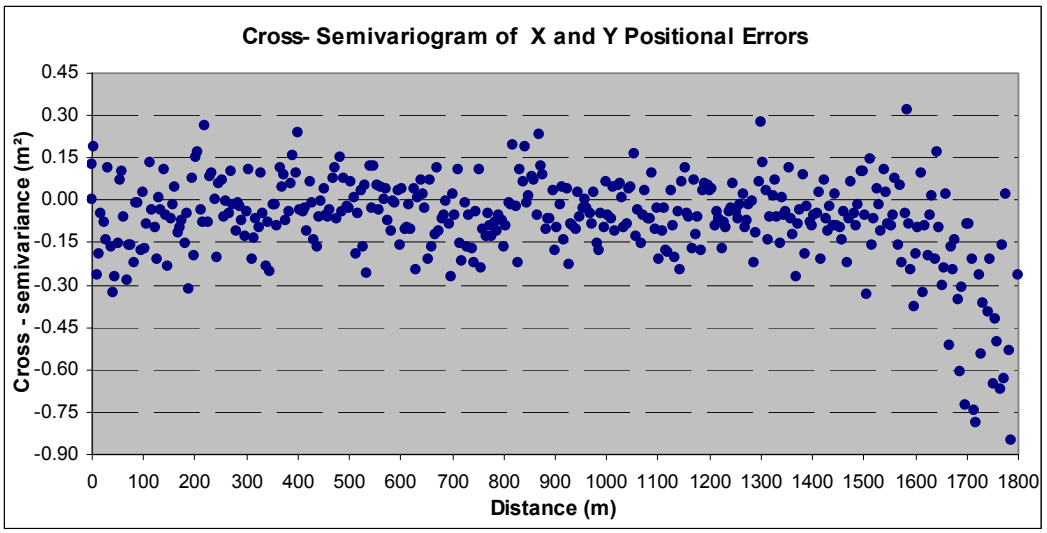

Figure 4.6 Cross - Semivariogram for X and Y Coordinate Errors

4.2.3. Semivariogram Modelling & Model Fitting

A unique principle of variography is that a number of theoretical variogram models, either alone or in combination, can be used in order to capture a statistically quantifiable portion of the spatial variability in the dataset. The model chosen will depend on several factors including whether the semivariogram reaches a sill value, and the behaviour of the semivariogram at the origin. Most commonly used models to describe the variability are exponential, spherical and gaussian models. All three semivariogram models are displayed with equivalent practical range in the below Figure 4.7.

Model fitting requires comparing different alternatives. Inspection of standardized semivariograms suggests using exponential, Gaussian and spherical semivariogram models.

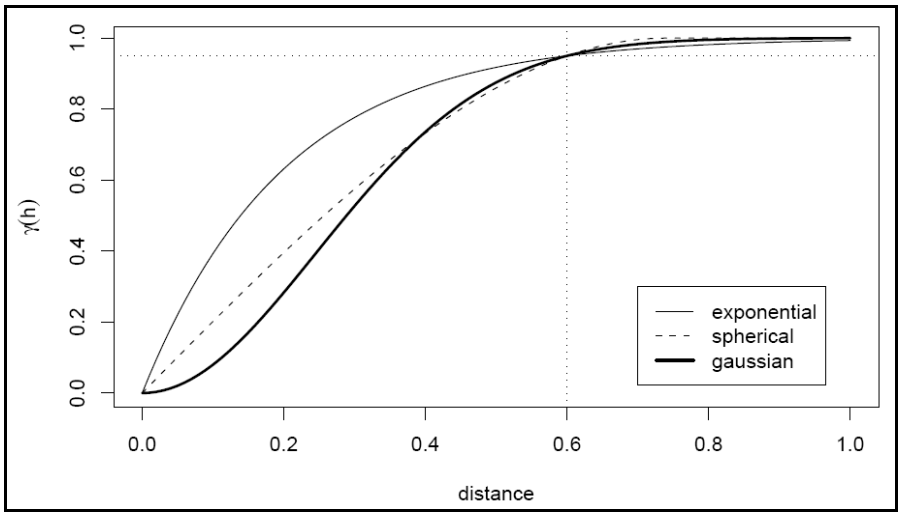

Figure 4.7 Three Generic Semivariogram Models

Following equations is used to define and construct semivariogram models for experimental semivariograms above. Then these models are evaluated using three different quality indicators to decide which model gives the best estimate for the data.

Exponential Model

$$\gamma(h) = \begin{cases} 0 \ , & h = 0 \\ c_0 + c.\left(1 - \exp\left(\dfrac{-h}{a}\right)\right) , & h \neq 0 \end{cases} \tag{4.2}$$

Spherical Model

$$\gamma(h) = \begin{cases} c_0 + c.\left(\dfrac{3h}{2a} - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3\right) , & if \ \ h \leq a \\ c_0 + c \qquad\qquad\qquad , & if \ \ h > a \end{cases} \tag{4.3}$$

Gaussian Model

$$\gamma(h) = c_0 + c.\left(1 - \exp\left(\frac{-h^2}{a^2}\right)\right) \tag{4.4}$$

c_0 indicates nugget, c is the partial sill, addition of nugget and partial sill gives the sill. *h* is the lag distance on the *i-th* lag, and *a* is the range of correlation.

Fitted models for semivariogram of positional errors in X are shown in Figure 4.8 where semivariogram model parameters respectively in Table 4.2 and Table 4.3. The region in the blue box is expressed to show slight differences in models and how models close to each other. There are slight differences in the sill and nugget parameters of these three models however differences in range of the correlations can be seen in Figure 4.8.

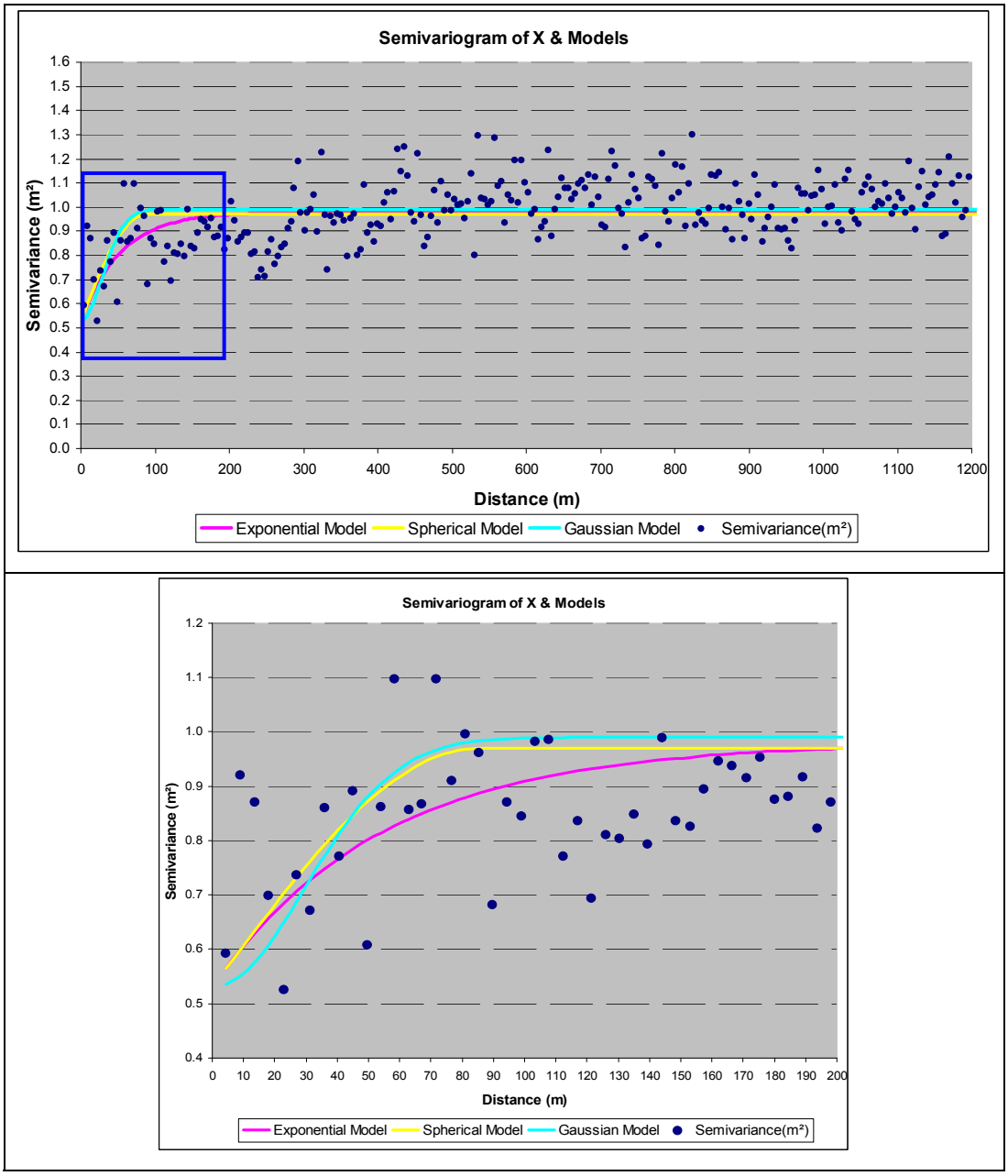

Figure 4.8 Fitted Semivariogram Models to Positional Errors in X axis

Fitted models for semivariogram of positional errors in Y are shown in Figure 4.9. where semivariogram model parameters respectively in the table. The region in the blue box is zoomed to show slight differences in models and how models close to each other in this data, too. The sill and nugget parameters are nearly the same in

all three models however differences in range of the correlations can also be seen in Figure 4.9.

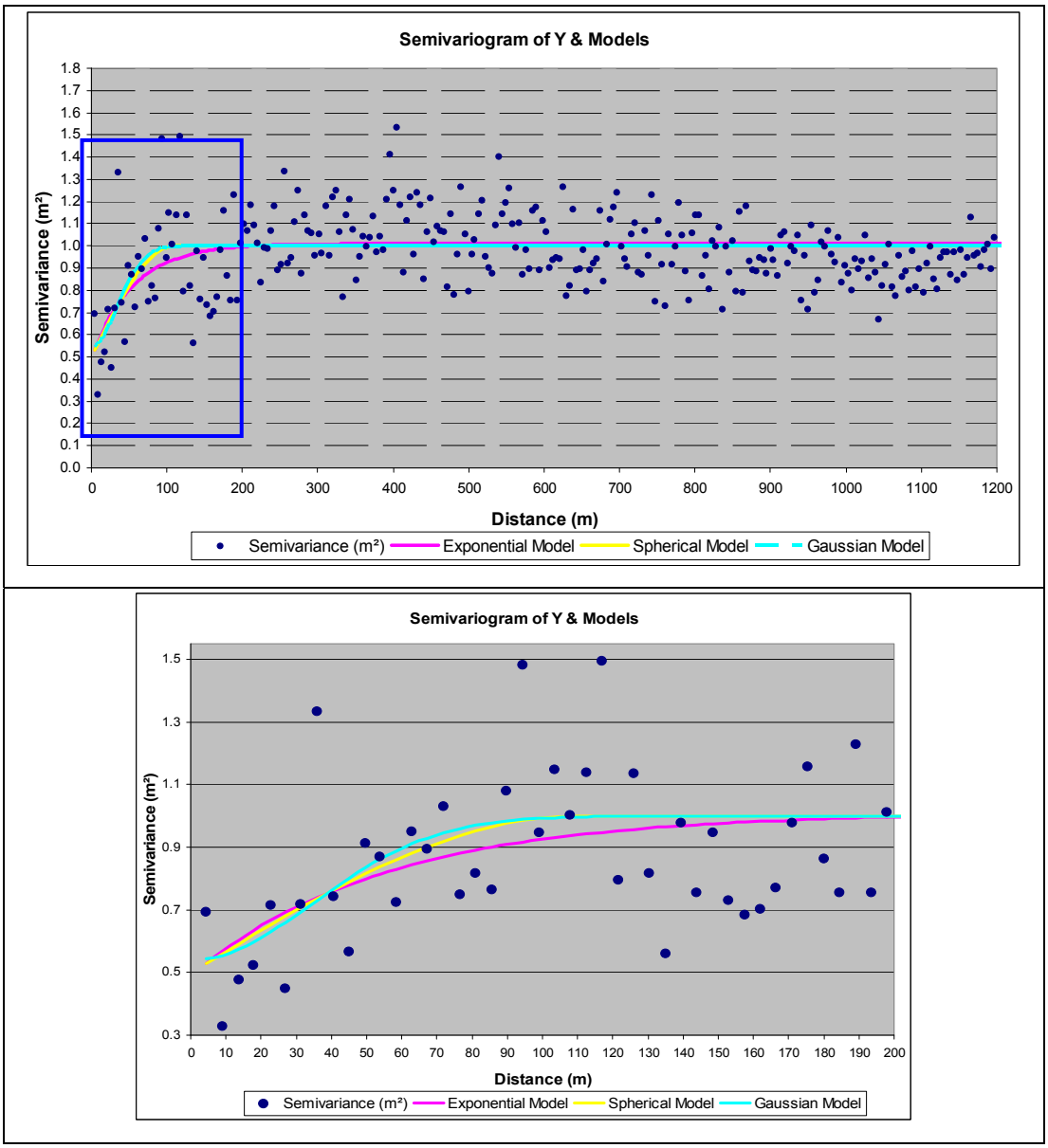

Figure 4.9 Fitted Semivariogram Models to Positional Errors in Y axis

4.2.4. Model Fitting Indicators

During model fitting parameters and models are controlled via three different quality indicators commonly used in variogram fitting. Usually they are giving comparable results but they can differ depending on the function to be fitted.

4.2.4.1. Indicative Goodness of Fit

The Indicative Goodness of Fit (IGF) measure (Kanevski & Maignan, 2004) is based on a least squares estimator and is defined as:

$$IGF = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=0}^{n(k)} \frac{N(h_i)\left(\sum_{i=0}^{n(k)} N(h_i)\right)^{-1}}{h_i\left(h_{\max}(k)\right)^{-1}}\left[\frac{\gamma(h_i)-\hat{\gamma}(h_i)}{\sigma^2}\right]^2 \tag{4.5}$$

where $\gamma(h_i)$ is the value of semivariogram at the i-th lag, $\hat{\gamma}(h_i)$ is the semivariogram model value for the lag h_i, *N* is the number of directional semivariograms, $N(h_i)$ is the number of pairs for the lag h_i, $h_{\max}(k)$ is the maximum lag distance for the *k-th* direction, σ^2 is the variance of the data for the variogram. This goodness of fit measure is standardized so that values for different variograms, using different models can be compared. A value closer to zero indicates a better fit.

4.2.4.2. Residuals Sum of Squares

In statistics, the residual sum of squares (RSS) is the sum of squared errors. It is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data (Wikipedia, 2009).

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{4.6}$$

RSS Calculation for Semivariogram Models

$$RSS = \sum_{i=1}^{n} (\gamma(h_i) - \hat{\gamma}(h_i))^2 \quad (4.7)$$

4.2.4.3. Cressie's Indicator

Cressie's indicator is a weighted RSS (WSS) where the weight hi given to each lag h_k is usually taken proportional to the number of N (h_k) of data pairs that contribute to the estimated semivariogram.

The implicit assumption is that the reliability of an experimental semivariogram value increases with statistical mass. An alternative that gives more weight to the first lags consists of dividing the number of data pairs by the squared model value: $\frac{N(h_i)}{\hat{\gamma}(h_i)^2}$

(Cressie , 1985 in Goovaerts 1997)

$$I_C = \sum \frac{N(h_i)}{\hat{\gamma}(h_i)^2} [\gamma(h_i) - \hat{\gamma}(h_i)]^2 \quad (4.8)$$

Corresponding quality indicator results to models and parameters used in semivariogram model fitting step are seen in Tables 4.2 and 4.3 representing models for errors in X and Y respectively. Smallest values for indicators are represented in bold characters. But as followed from Table 4.1 models represent very close fitting results to each other as previously assumed with reference to graphs (Figure 4.8 & 4.9). It is assumed in this study that all distributions display an exponential spatial correlation model in terms of positional errors. But to observe how different models respond to uncertainty in the output, all three models are used for simulations for generating comparable results in evaluation.

Table 4.2 Model Parameters and Indicators for X Errors

		Exponential Model	Spherical Model	Gaussian Model
Model Parameters	**Nugget**	0.53	0.53	0.53
	Sill	0.98	0.97	0.99
	Range	162	85	72
Quality Indicators	**I.G.F**	**0.035506**	0.037212	0.039967
	RSS	**8.4217**	8.7522	8.5556
	Cressie	3758.51219	4032.70433	**3647.47663**

Table 4.3 Model Parameters and Indicators for Y Errors

		Exponential Model	Spherical Model	Gaussian Model
Model Parameters	**Nugget**	0.49	0.5	0.54
	Sill	1.01	1	1
	Range	165	110	85
Quality Indicators	**I.G.F**	0.047693	0.047608	**0.047319**
	RSS	**19.9955**	20.1433	20.1491
	Cressie	6579.215	6675.73	**6678.076**

4.3. Simulation of Digitized Field Boundaries

Simulations of digitized agricultural parcel boundaries are performed in two phases. First stage is analysis of training data (221 polygons) using all semivariogram models' (Exponential, Spherical and Gaussian models) parameters. In the first phase, it is also compared that how different models affect the resulting uncertainty boundaries and distributions of generated simulations around both reference and digitized data.

Second phase focuses on test data that consists of 479 field boundaries out of 700. These test data are simulated with respect to best fitting (Table 4.2 & 4.3) exponential model's parameters in uncertainty model, then 1000 realisations are performed per each polygon. Distribution of training and test data can be viewed in Figure 4.10.

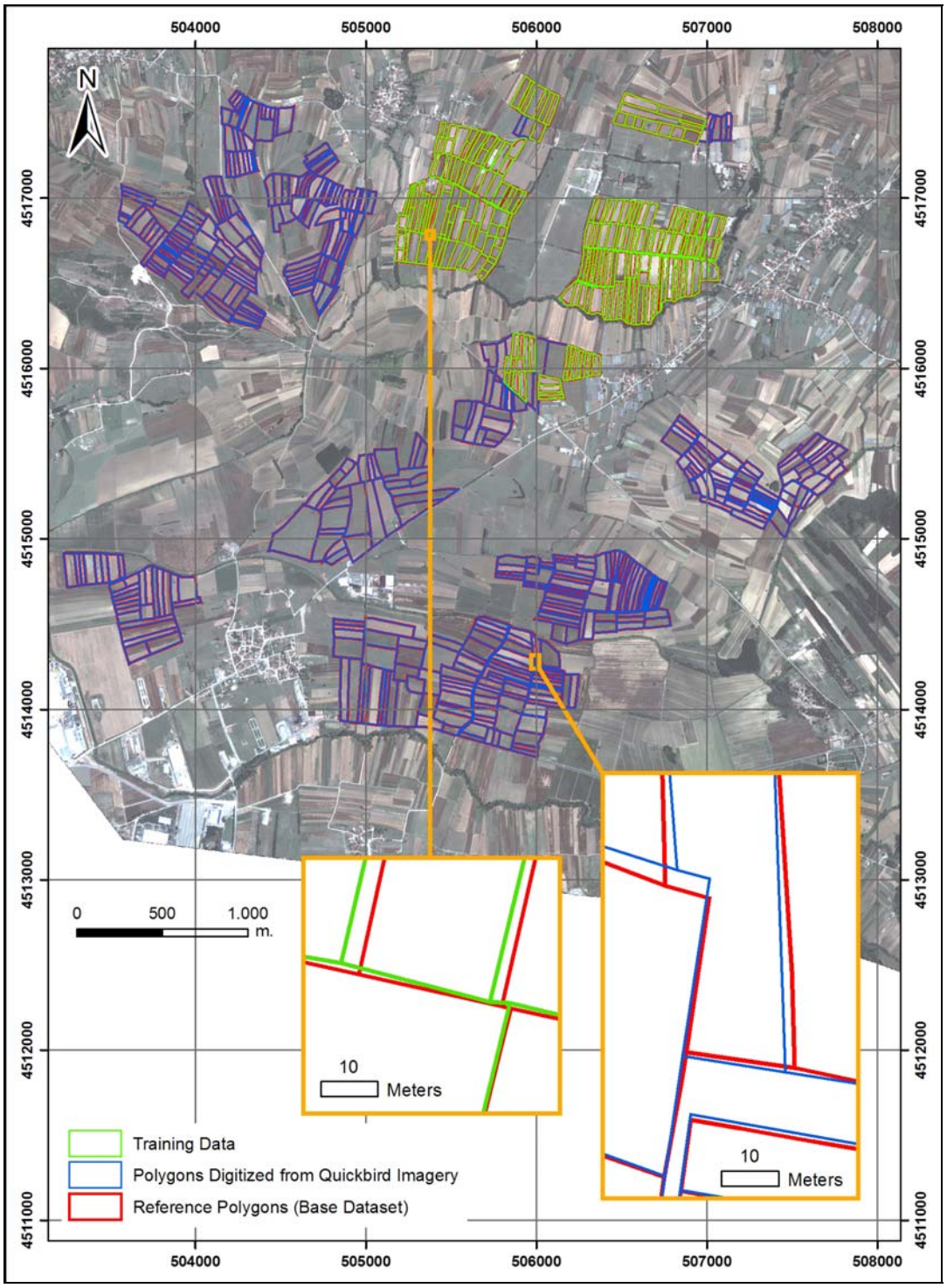

Figure 4.10 Distributions of Training and Test Data

4.3.1. Simulation of Training Data

The Data Uncertainty Engine (DUE) v3.1 (Brown and Heuvelink, 2007) is used for generating 250 realisations of error model with parameterisations representing each model described before. Number of simulations is determined as 250 with reference to the study of de Bruin et al. (2008), in which they generated the same number of simulations with DUE in order to identify errors arising from positional measurements carried by GPS. The field boundaries are classified as rigid objects containing multiple vertices whose relative positions do not change under uncertainty (Brown and Heuvelink, 2007). Examples of rigid objects might include buildings whose boundaries are theoretically rigid or fields whose boundaries are treated as rigid for simplicity (de Bruin et. al, 2008). The positional uncertainty of a rigid object can be characterised by a translation and possibly rotation of the object in DUE.

The model parameters are entered as “expert judgement” on the model page of DUE 3.1. The standard deviations or spreads of normally distributed errors σ_X and σ_Y defined as 1.2 metres and 1.3 metres respectively (Table 4.1). Normal distribution curves, regarding the coordinate errors, are centred on the object coordinates in case of $\mu_X = \mu_Y = 0$, otherwise an offset is added to model bias. Semivariograms are modelled before they were employed in DUE to define the dependence models. In the case of cross – correlation between X and Y errors, linear model of co - regionalization is used to ensure a valid bivariate covariance structure (Heuvelink, 2007). Negative cross – correlations are not currently supported in DUE. In this study, positional errors in X and Y coordinates display very small magnitude of negative cross correlation, around -0,075 (Figure 4.6). Thus, it is not possible to model the cross – correlations structure with respect to the defined uncertainty model. In the output window, number of simulations is defined as 250 and output file type is selected as ArcGIS shapefile format for further analysis on simulations.

Simulation results regarding the rigid object model with exponential function can be seen in Figure 4.11.

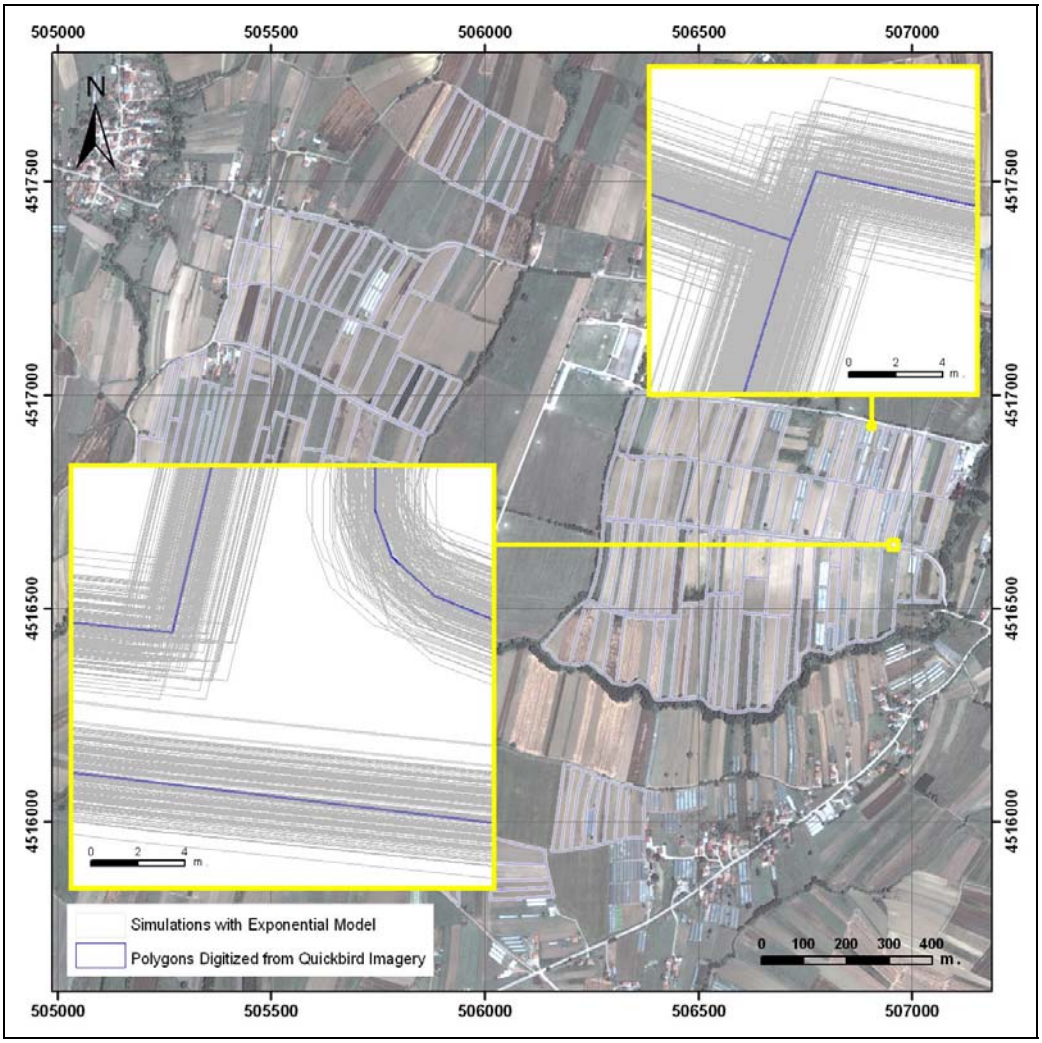

Figure 4.11 Simulations Derived from Uncertainty Model

Initially, all shapefiles, each containing 250 simulations per polygon, are merged into a single shapefile containing 55250 simulations, pertaining 221 polygons. In order to determine the exterior simulation boundaries for each polygon, 55250 simulations are dissolved with respect to their parent polygon id. By means of this “union” operation, exterior simulation boundaries are extracted as polygons. To determine the interior simulation boundary for each polygon, “intersection” of 250 simulations,

each simulation set corresponding to one parent polygon, is computed and inner simulation boundary is extracted. By repeating this procedure 221 times, all inner simulation boundaries are extracted. Both inner and outer simulation boundaries are converted into polylines, merged into a new shapefile and dissolved as multipart polyline objects representing both inner and outer simulation boundaries, resembling a donut shape. Results are presented in Figure 4.12. It can be interpreted that in the absence of cross – correlation structure uncertainty boundaries showed an epsilon error band model structure. If cross – correlation could be modelled within uncertainty model, uncertainty bands may result in G - band structure which displays narrower band width towards the midpoints.

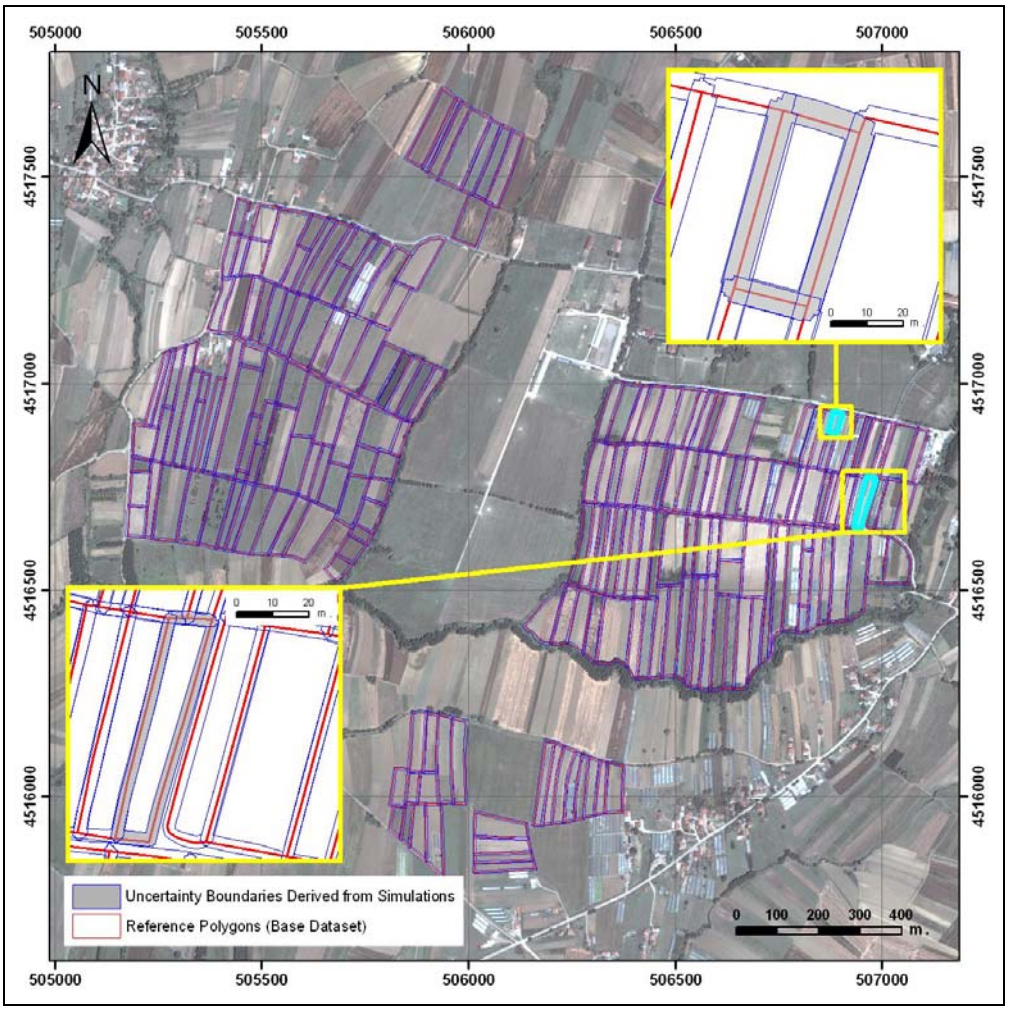

Figure 4.12 Uncertainty Boundaries Derived from Simulations

In order to explore the spatial distribution of simulations from digitized vector dataset, digitized polygons and reference data polygons are converted into polyline objects and multiple buffers, each increasing by 50 cm, are applied in each direction; in and out. A buffer distance of 50 cm. is selected because digitized dataset to which simulations are referencing is digitized from the Quickbird satellite imagery having a nominal resolution of 60 cm/pixel. Knowing the smallest discriminable image unit corresponds 60 cm, multiple buffer intervals are selected as 50 cm for computational easiness. Once multiple buffer rings are created and simulation polygons are converted into polylines, a spatial query is performed for each ring at that buffer distance. As a result of the following query; “Number of simulation polyline features which are completely covered by the relevant buffer ring” gives the spatial distribution. To give better understanding about the spatial distribution of simulations, a histogram is constructed with respect to buffer distances (Figure 4.13).

Both histograms in Figure 4.13 confirm the joint normal distribution of uncertainty model that generated simulations. Distribution of simulations around digitized data represent the farthest simulation that falls at a distance of 6 meters, however the frequency of simulated polygons are accumulated within 1 to 3 meters distance from digitized data. However, simulated polygons are distributed within 10 meters distance around reference data and accumulation takes place between 2 to 4 meter distances.

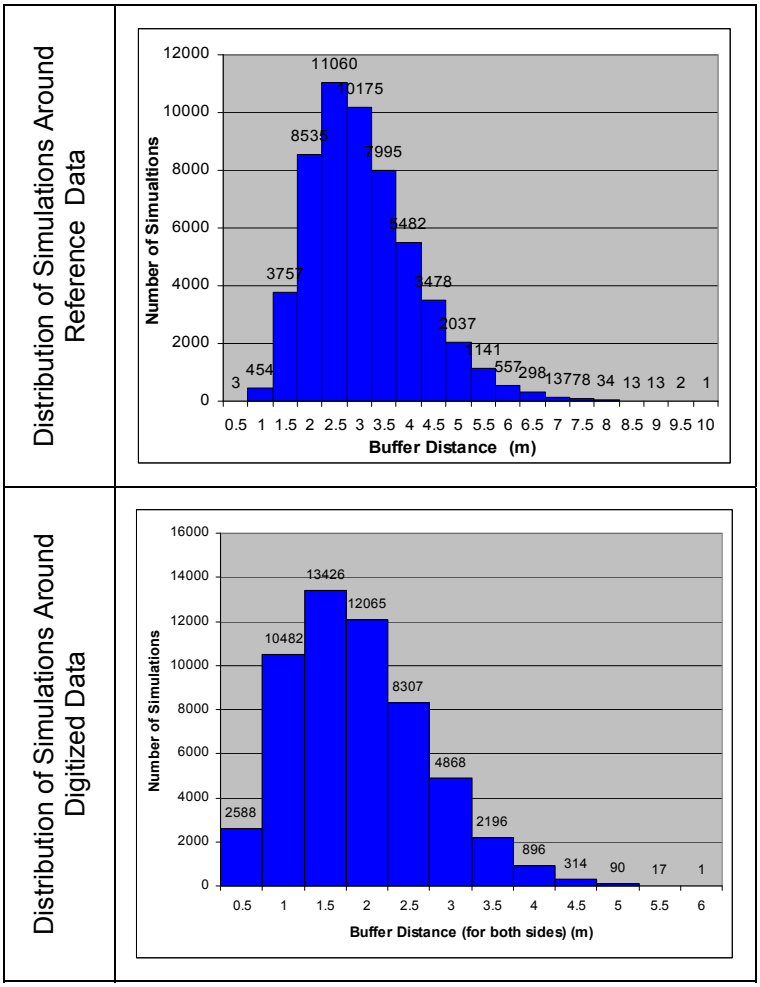

Figure 4.13 Numbers of Simulations (Exponential Model) Included in Buffers

When percentage of distributions around digitized and reference data are calculated, it is observed that 84.83% of the simulations fall within 2.5 meters around digitized data (Table 4.4). But in this study, departure of simulations from true value is essential. From all the realisations derived, with 85.90% probability, digitized vectors contain an error of 4 meters whether they are on the left or right hand side of the reference data (Table 4.5).

Table 4.4 Percentage of Exponential Model Simulations' Distributions around Digitized Data

Buffer Distance (m)*	Number of Simulations	Percentage	Cumulative Percentage
0.5	2588	4.684%	4.684%
1	10482	18.972%	23.66%
1.5	13426	24.300%	47.96%
2	12065	21.837%	69.79%
2.5	***8307***	***15.035%***	***84.83%***
3	4868	8.811%	93.64%
3.5	2196	3.975%	97.61%
4	896	1.622%	99.24%
4.5	314	0.568%	99.80%
5	90	0.163%	99.97%
5.5	17	0.031%	100.00%
6	1	0.002%	100.00%

*Buffer Distance indicates buffer ring distance to both sides (left and right side of the polygon), thus 0.5 meters to both sides of the reference data covers an area of 1 meter around the line.

Table 4.5 Percentage of Exponential Model Simulations' Distributions around Reference Data

Buffer Distance (m)	Number of Simulations	Percentage	Cumulative Percentage
0.5	3	0.005%	0.005%
1	454	0.822%	0.827%
1.5	3757	6.800%	7.627%
2	8535	15.448%	23.075%
2.5	11060	20.018%	43.093%
3	10175	18.416%	61.510%
3.5	7995	14.471%	75.980%
4	***5482***	***9.922%***	***85.902%***
4.5	3478	6.295%	92.197%
5	2037	3.687%	95.884%
5.5	1141	2.065%	97.949%
6	557	1.008%	98.957%
6.5	298	0.539%	99.497%
7	137	0.248%	99.745%
7.5	78	0.141%	99.886%
8	34	0.062%	99.948%
8.5	13	0.024%	99.971%
9	13	0.024%	99.995%
9.5	2	0.004%	99.998%
10	1	0.002%	100.000%

4.3.1.1. Comparison of Uncertainty Model Results

All three semivariogram models are used to simulate positional uncertainty. The parameterised error models are entered as expert judgement on the model window in DUE. All the same procedure applied to training data is executed over each model's resulting realisations and distributions are calculated to provide comparable results. Figure 4.14 show the distribution of simulations around reference data and digitized data, it can be followed from the figure that gaussian and spherical models' parameters results create similar distributions compared to exponential model.

The range of distributions around reference data is 10 meters in all model results. Exponential models' resulting range around digitized data 6 meters, but spherical and gaussian model the display same range of 6.5 meters.

Also the percentage and cumulative percentage of distributions are calculated for both reference data and digitized data. (Table 4.6 and Table 4.7) The semivariogram models' parameter values and goodness of fit results in modelling stage presented similar results, so that in the uncertainty analysis simulations derived from these parameters shows slight differences. According to the different model results, from this point forward only the best fitting model parameters will be used to generate simulations from error model.

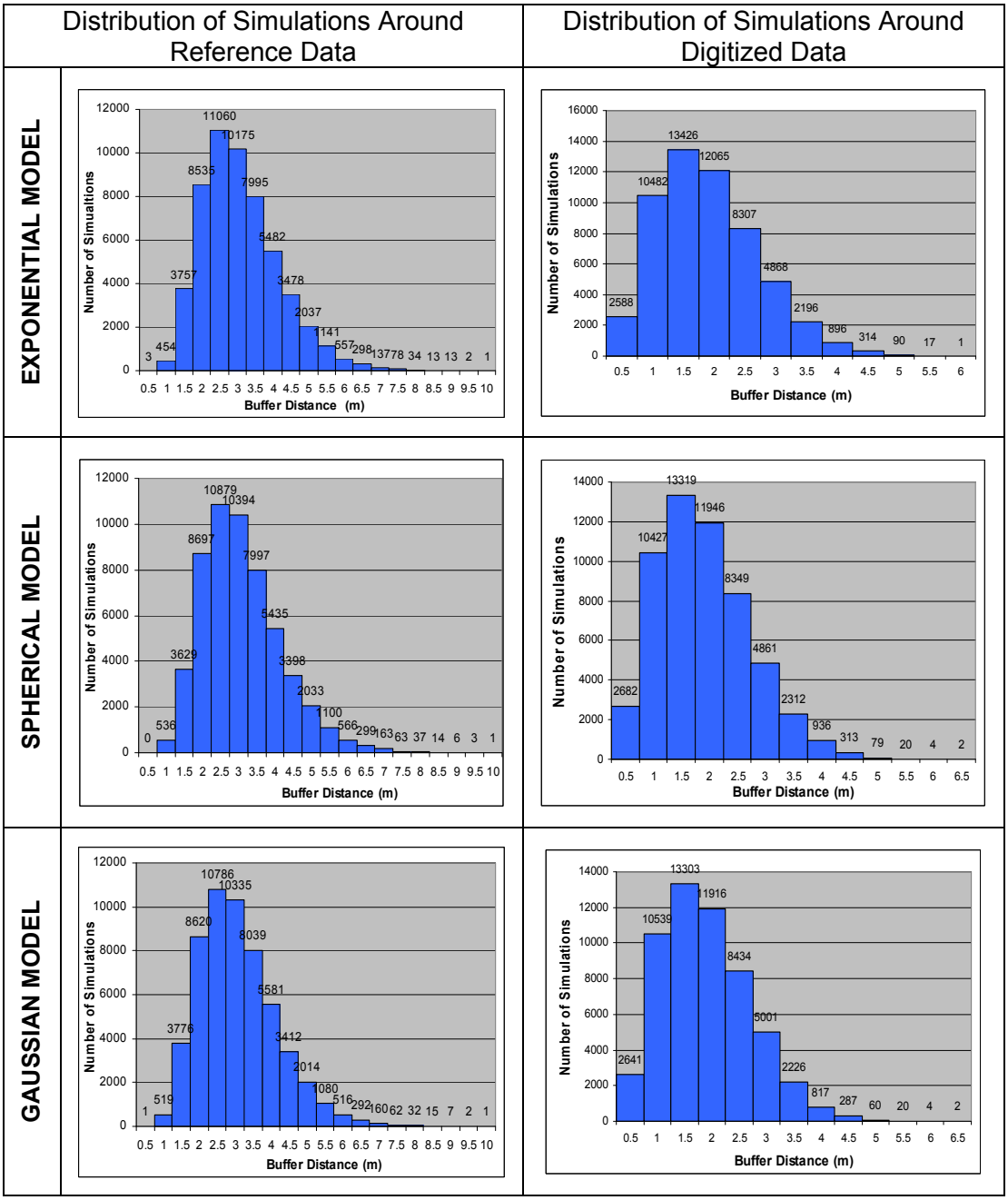

Figure 4.14 Exponential – Spherical and Gaussian Models' Simulation Distributions

Table 4.6. Percentage of Distribution around Digitized Data

Buffer Distance(m)	Number of Simulations			Percentage			Cumulative Percentage		
	Exponential	Gaussian	Spherical	Exponential	Gaussian	Spherical	Exponential	Gaussian	Spherical
0.5	2588	2641	2682	4.68%	4.78%	4.85%	4.68%	4.78%	4.85%
1	10482	10539	10427	18.97%	19.08%	18.87%	23.66%	23.86%	23.73%
1.5	13426	13303	13319	24.30%	24.08%	24.11%	47.96%	47.93%	47.83%
2	12065	11916	11946	21.84%	21.57%	21.62%	69.79%	69.50%	69.46%
2.5	**8307**	**8434**	**8349**	**15.04%**	**15.27%**	**15.11%**	**84.83%**	**84.77%**	**84.57%**
3	4868	5001	4861	8.81%	9.05%	8.80%	93.64%	93.82%	93.36%
3.5	2196	2226	2312	3.97%	4.03%	4.18%	97.61%	97.85%	97.55%
4	896	817	936	1.62%	1.48%	1.69%	99.24%	99.32%	99.24%
4.5	314	287	313	0.57%	0.52%	0.57%	99.80%	99.84%	99.81%
5	90	60	79	0.16%	0.11%	0.14%	99.97%	99.95%	99.95%
5.5	17	20	20	0.03%	0.04%	0.04%	100.00%	99.99%	99.99%
6	1	4	4	0.00%	0.01%	0.01%	100.00%	100.00%	100.00%
6.5	0	2	2	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%

Table 4.7. Percentage of Distribution around Reference Data

Buffer Distance(m)	Number of Simulations			Percentage			Cumulative Percentage		
	Exponential	Gaussian	Spherical	Exponential	Gaussian	Spherical	Exponential	Gaussian	Spherical
0.5	3	1	0	0.01%	0.00%	0.00%	0.01%	0.00%	0.00%
1	454	519	536	0.82%	0.94%	0.97%	0.83%	0.94%	0.97%
1.5	3757	3776	3629	6.80%	6.83%	6.57%	7.63%	7.78%	7.54%
2	8535	8620	8697	15.45%	15.60%	15.74%	23.08%	23.38%	23.28%
2.5	11060	10786	10879	20.02%	19.52%	19.69%	43.09%	42.90%	42.97%
3	10175	10335	10394	18.42%	18.71%	18.81%	61.51%	61.61%	61.78%
3.5	7995	8039	7997	14.47%	14.55%	14.47%	75.98%	76.16%	76.26%
4	**5482**	**5581**	**5435**	**9.92%**	**10.10%**	**9.84%**	**85.90%**	**86.26%**	**86.09%**
4.5	3478	3412	3398	6.30%	6.18%	6.15%	92.20%	92.43%	92.24%
5	2037	2014	2033	3.69%	3.65%	3.68%	95.88%	96.08%	95.92%
5.5	1141	1080	1100	2.07%	1.95%	1.99%	97.95%	98.03%	97.91%
6	557	516	566	1.01%	0.93%	1.02%	98.96%	98.97%	98.94%
6.5	298	292	299	0.54%	0.53%	0.54%	99.50%	99.50%	99.48%
7	137	160	163	0.25%	0.29%	0.30%	99.74%	99.78%	99.78%
7.5	78	62	63	0.14%	0.11%	0.11%	99.89%	99.90%	99.89%
8	34	32	37	0.06%	0.06%	0.07%	99.95%	99.95%	99.96%
8.5	13	15	14	0.02%	0.03%	0.03%	99.97%	99.98%	99.98%
9	13	7	6	0.02%	0.01%	0.01%	99.99%	99.99%	99.99%
9.5	2	2	3	0.00%	0.00%	0.01%	100.00%	100.00%	100.00%
10	1	1	1	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%

4.3.2. Simulation of Test Data

Test data, containing 479 polygons out of 700, are evaluated in uncertainty analysis. Data Uncertainty Engine (DUE) is used for generating 1000 realisations of the error model with parameterisations.

Model parameters are inserted as “expert judgement” on the model page of DUE 3.1. Standard deviations or spreads of normally distributed errors, σ_X and σ_Y, are defined as 1.2 metres and 1.3 metres respectively from training data results (Table 4.1). In case of $\mu_X = \mu_Y = 0$, if coordinate errors exhibit normal distribution, they are centred on the coordinates of objects; otherwise an offset is added to model bias. To define spatial dependence model, semivariograms are modelled before they were employed in DUE. Exponential semivariogram model parameters that are drawn out from training data are used for describing uncertainty model. These model parameters vary in range of 162 meters for X error autocorrelations and 165 meters for Y autocorrelations. In DUE’s simulations output window number of simulations is defined as 1000 per each polygon and output file type is selected as ArcGIS shapefile format as before.

Resulting 479,000 simulations are evaluated with respect to the same procedure explained in section 4.3.1 and distribution of simulations around reference data is displayed below in Figure 4.15 and Table 4.8. Compared to training set, it is observed that distribution of simulations display a more dispersed range of 12.5 meters. In training data, 4 meters distance contain 85.902 % of all simulations, but in test data 4 meters of distance cover up 89.87 % of all realisations.

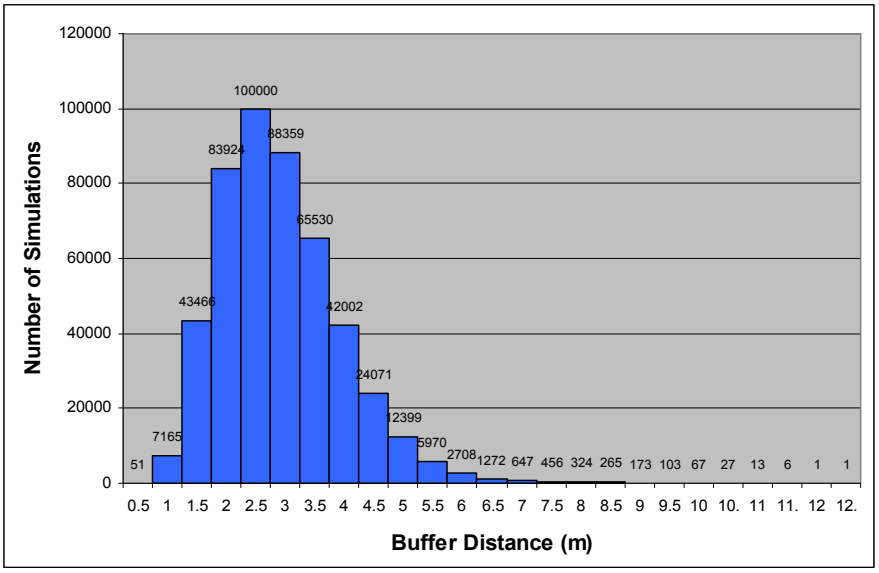

Figure 4.15 Distribution of Simulations around Reference Data

Table 4.8.Percentage of Distribution for Test Data

Buffer Distance(m)	Number of Simulations	Percentage	Cumulative Percentage
0.5	51	0.0106%	0.0106%
1	7165	1.4958%	1.5065%
1.5	43466	9.0743%	10.5808%
2	83924	17.5207%	28.1015%
2.5	100000	20.8768%	48.9783%
3	88359	18.4466%	67.4248%
3.5	65530	13.6806%	81.1054%
4	***42002***	***8.7687%***	***89.8741%***
4.5	24071	5.0253%	94.8994%
5	12399	2.5885%	97.4879%
5.5	5970	1.2463%	98.7342%
6	2708	0.5653%	99.2996%
6.5	1272	0.2656%	99.5651%
7	647	0.1351%	99.7002%
7.5	456	0.0952%	99.7954%
8	324	0.0676%	99.8630%
8.5	265	0.0553%	99.9184%
9	173	0.0361%	99.9545%
9.5	103	0.0215%	99.9760%
10	67	0.0140%	99.9900%
10.5	27	0.0056%	99.9956%
11	13	0.0027%	99.9983%
11.5	6	0.0013%	99.9996%
12	1	0.0002%	99.9998%
12.5	1	0.0002%	100.0000%

4.4. Positional Uncertainty Analysis of Sample Areas

Two sample areas are selected for further analysis in order to cover up the whole functionality of DUE. Effects of cross correlation structure on model results are analyzed in the first sample region for 95 polygons. Second example, comprising from 30 parcel boundaries, is carried with two scenarios, first is rigid object model and second is deformable object model with cross correlation.

In the previous analysis, involving 221 training and 479 test parcel boundaries, four main corner vertices of 221 vector polygons are used as samples to provide positional errors and autocorrelation structure. On these sample regions a different method is applied for sampling. Initially 5 % of population (total number of parcel boundaries in each region) is sampled. Spatial random sampling is performed via ArcGIS Desktop v9.3 Geostatistical Analyst to provide sample polygons. Afterwards, these parcel boundaries are traced with snapping on reference data and digitized data with different intervals. Finally, vertices of the line segments are converted to point features and used for point based displacement calculation.

4.4.1. Analysis of Cross – Correlated Areas

First the sample region contains a number of 95 agricultural land parcel boundaries are selected from the previous 700 polygonal vector dataset. Selected region and region's location in the whole dataset can be viewed in Figure 4.16.

5% of these 95 polygons are sampled to generate vertices for evaluation of error modelling and uncertainty model parameter estimation. Sample parcel boundaries are traced with snapping option in order to preserve exact positions of reference and digitized data. Tracing operation is carried with 5 meters of intervals starting from the lower left corner of polygons. This is carried out for all polygons both in reference and digitized dataset.

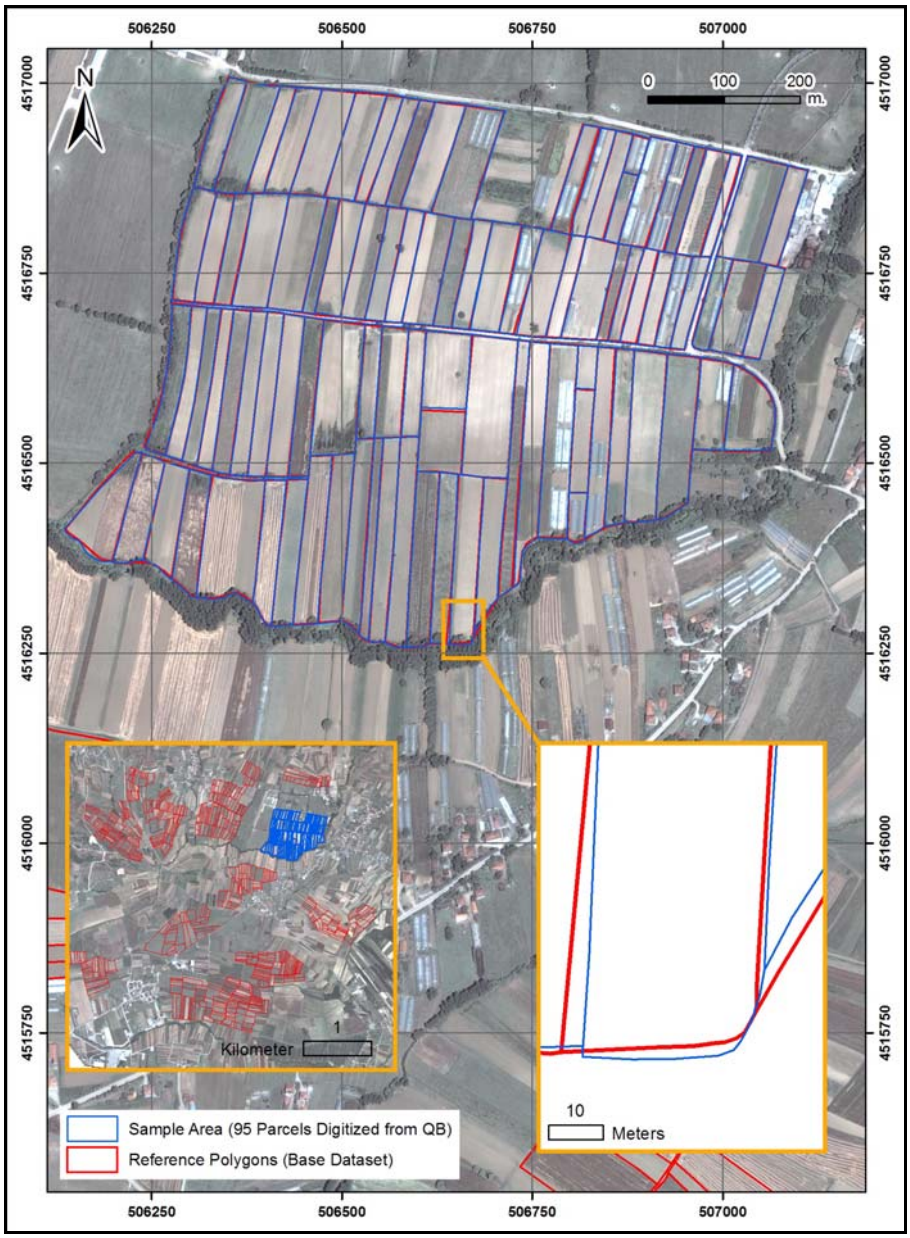

Figure 4.16 Sample Region of 95 Parcel Boundaries

For all sampled polygons, vertices for each 5 meters line segment are converted to point features. A unique ID is given to all points in both data for matching these points in further analysis. Preliminary assessment of positional errors is carried out with the discrepancy of points in reference and digitized data. Selected polygons and sample vertices can be seen on Figure 4.17.

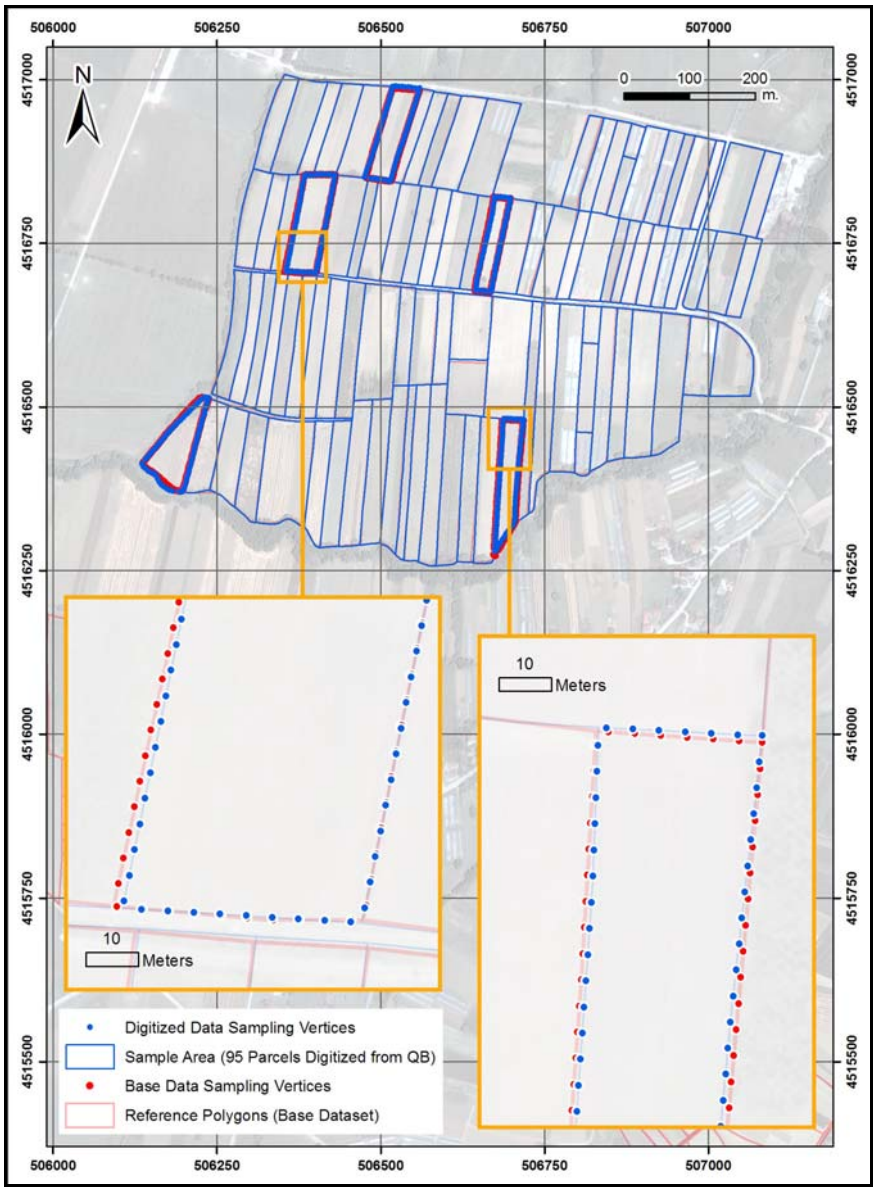

Figure 4.17 Sampled Polygons and Vertices for 95 Polygon Sample Region

Total number of 387 points from 5 polygons is included in calculation. Absolute values are calculated for displacements because minus or plus signs only indicates the direction of errors in positions. Main focus on this study is the absolute distance variation.

Statistical measures for positional errors on X and Y axis are given in Table 4.9.

Table 4.9 Descriptive statistics of Displacements in Coordinates for 95 Polygon Sample Region

	Errors in X Axis	Errors in Y Axis
Standard Deviation	0.85	0.94
Mean	0.98	0.97
Maximum Value	4.8	7.2

Calculated absolute valued positional errors in X and Y coordinates are converted to ".dat" file format in order to compute standardised semivariograms and cross semivariogram. Geostatistical software package GSlib is used to construct standardised semivariograms.

Since sample points are located within 5 meters interval, optimum lag distance parameter for all semivariograms is selected as 5 meters while constructing semivariograms. Optimum number of lags representing the autocorrelation structure is found as 24 for errors on X and Y axis. However, number of lags for cross – correlation structure is 20. All three semivariograms are displayed in Figure 4.18.

For semivariogram modelling, different model alternatives (e.g. Gaussian, exponential, spherical) are evaluated and finally spherical model is selected as the best fitting model for the following semivariograms given in Figure 4.18. Fitted semivariogram models are controlled by the most powerful quality indicator – indicative goodness of fit (IGF). These semivariogram model parameters are further employed to construct uncertainty model within DUE.

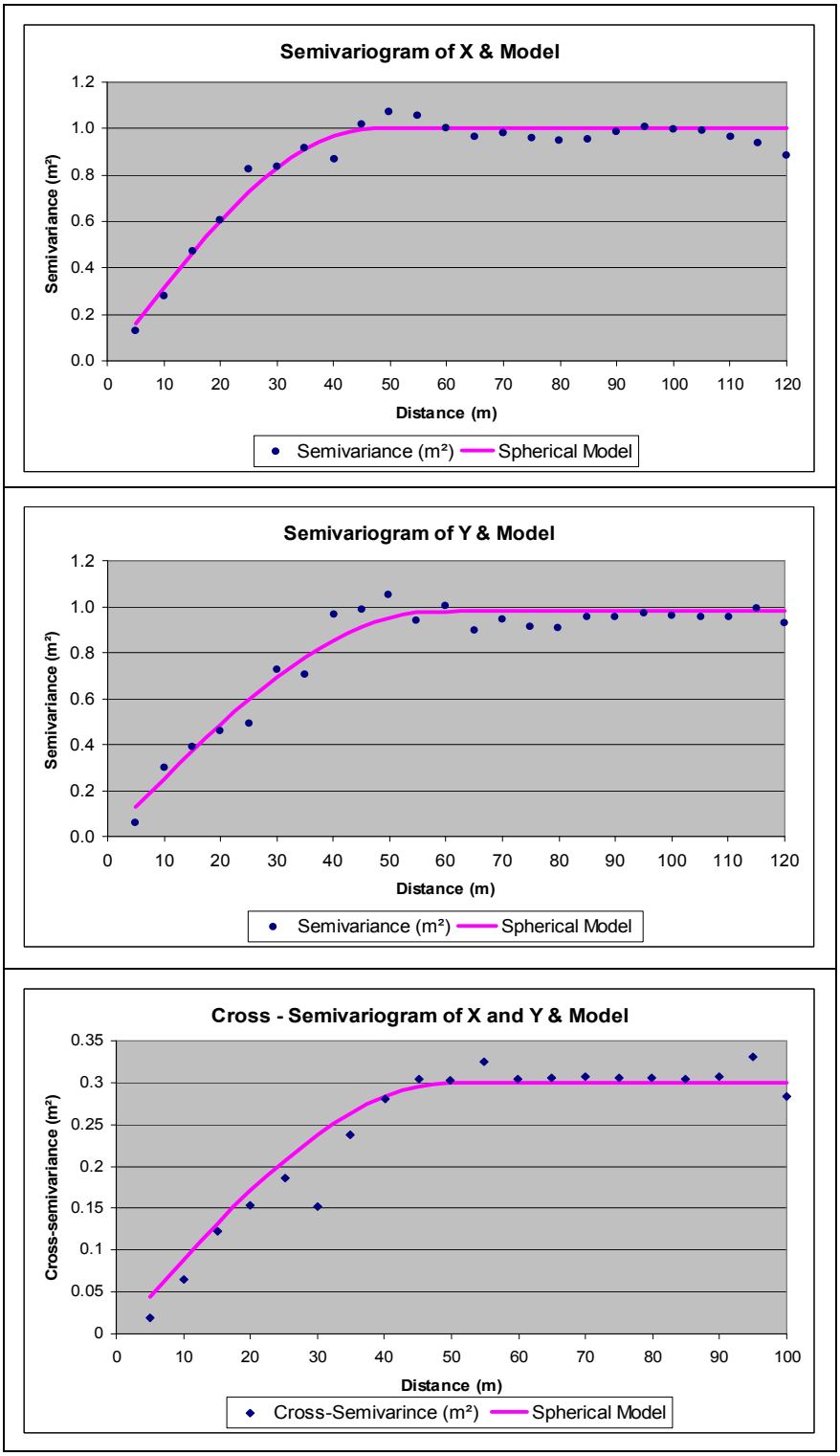

Figure 4.18 Semivariograms for Positional Errors and Best Fitting Models

IGF results and parameters regarding all semivariograms are shown in Table 4.10.

Table 4.10 Model Parameters and IGF Results for 95 Polygon Sample Region

Model Parameters	X Model	Y Model	XY Cross Model
	Spherical	Spherical	Spherical
Nugget	0	0	0.7
Sill	1	0.98	**0.3**
Range	**46.8**	**58**	**50**
I.G.F	0.007337	0.013826	0.01205

250 realisations for each polygon are created with the error model parameters described in Table 4.10 using DUE. Parcel boundaries are defined via rigid object model with type of movement as translation.

Parameters are entered as “expert judgement” on the modelling tab of DUE 3.1. Standard deviations or spreads of normally distributed errors, σ_X and σ_Y, are both defined as 1 metre (Table 4.9). Normal distribution curves, regarding the coordinate errors, are centred on the object coordinates in case of $\mu_X = \mu_Y = 0$, otherwise an offset is added to model bias. Semivariograms are modelled before they were employed in DUE to define the dependence models. Auto-correlation and cross-correlation structures are defined by using semivariogram parameters in the second modelling window of DUE 3.1. Range parameters for X and Y are entered as 47 and 58 meters respectively. For auto correlation structure, sill is considered to be constant on DUE and taken as 1. With reference to this sill value, all semivariograms are standardised to achieve 1 meter of sill. For this test case, since cross – correlation displays a positive structure, it can be modelled within DUE. Parameters of range and sill are employed as 50 meters and 0.3 respectively.

Simulations are generated within DUE in ESRI shapefile file format as polygon feature. Similar procedures explained before are repeated to obtain the uncertainty boundaries around data. Totally, 23750 realisations are evaluated to achieve the distribution structure around reference data.

The uncertain boundaries are displayed in Figure 4.19.

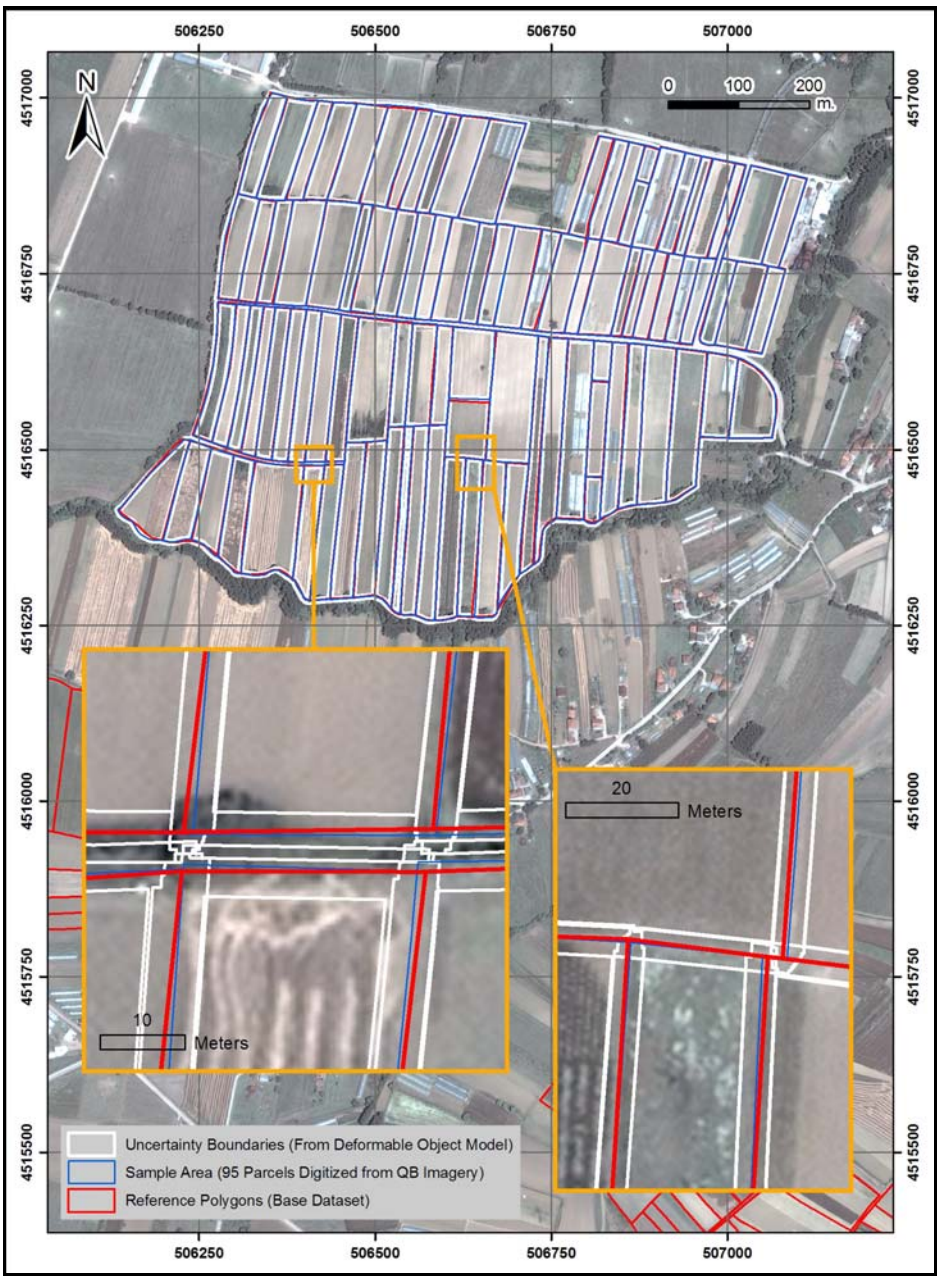

Figure 4.19 Resulting Uncertainty Boundaries for 95 Polygons.

Distribution of simulations around reference polygons are given in Figure 4.19 and percentage of this spatial dispersion is displayed in Table 4.11.

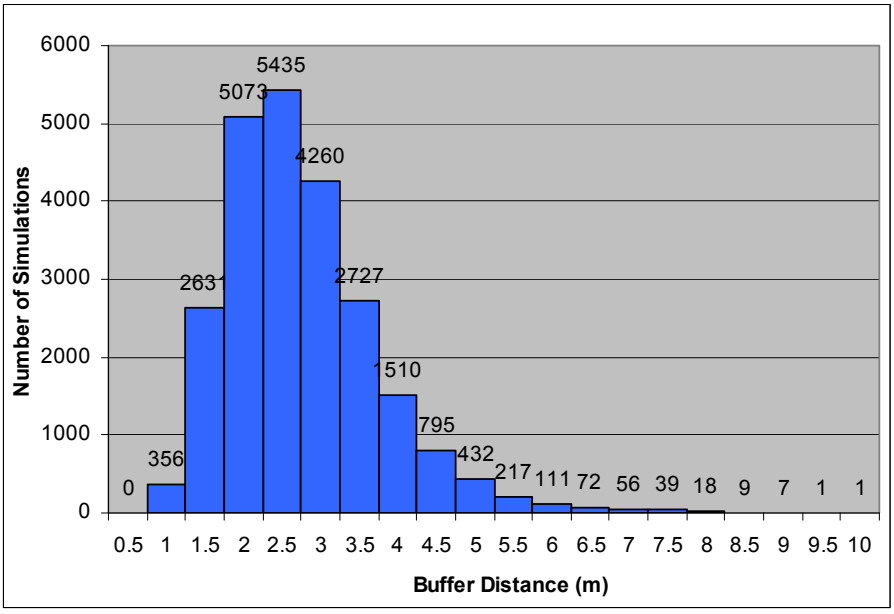

Figure 4.20 Distribution Graphic of Simulations.

Table 4.11 Percentage and Cumulative Percentage of Distributions

Buffer Distance (m)	Number of Simulations	Percentage	Cumulative Percentage
0.5	0	0.000%	0.000%
1	356	1.499%	1.499%
1.5	2631	11.078%	12.577%
2	5073	21.360%	33.937%
2.5	5435	22.884%	56.821%
3	4260	17.937%	74.758%
3.5	**2727**	**11.482%**	**86.240%**
4	**1510**	**6.358%**	**92.598%**
4.5	795	3.347%	95.945%
5	432	1.819%	97.764%
5.5	217	0.914%	98.678%
6	111	0.467%	99.145%
6.5	72	0.303%	99.448%
7	56	0.236%	99.684%
7.5	39	0.164%	99.848%
8	18	0.076%	99.924%
8.5	9	0.038%	99.962%
9	7	0.029%	99.992%
9.5	1	0.004%	99.996%
10	1	0.004%	100.000%

When compared with the previous model, which ignores cross – correlation, nearly 85 % of simulations dispersed within 4 metres. However if cross – correlation structure is considered in the model, about 86% of simulations can be covered within a range of 3.5 metres.

4.4.2. Analysis with Different Object Models

For the second case area, a number of 30 agricultural land parcels are sub-sampled from the vector dataset containing 700 parcels. Selected region and the location of this sub-region within the whole dataset can be viewed in Figure 4.21.

In this example, two scenarios for manual digitization of vector data are considered; “rigid object model” scenario and “deformable object model” scenario. A spatial object is considered as deformable if its vertices can move separately (Brown and Heuvelink , 2007).

In order to generate vertices for the evaluation of error modelling and uncertainty model parameter estimation, 5% of these 30 polygons are sub-sampled. Sample parcel boundaries are traced with snapping option in order to preserve the exact positions of reference and digitized data. Tracing operation is carried with 2 meters intervals, starting from the lower left corner of the polygon in clockwise direction and the same procedure repeated for two polygons in the reference and digitized data.

Vertices of line segments, each having 2 meters length, are converted to point features. A unique ID is given to all points in both data for matching these points in further analysis. Initial assessment of positional errors is carried out with the discrepancy of points in reference and digitized data.

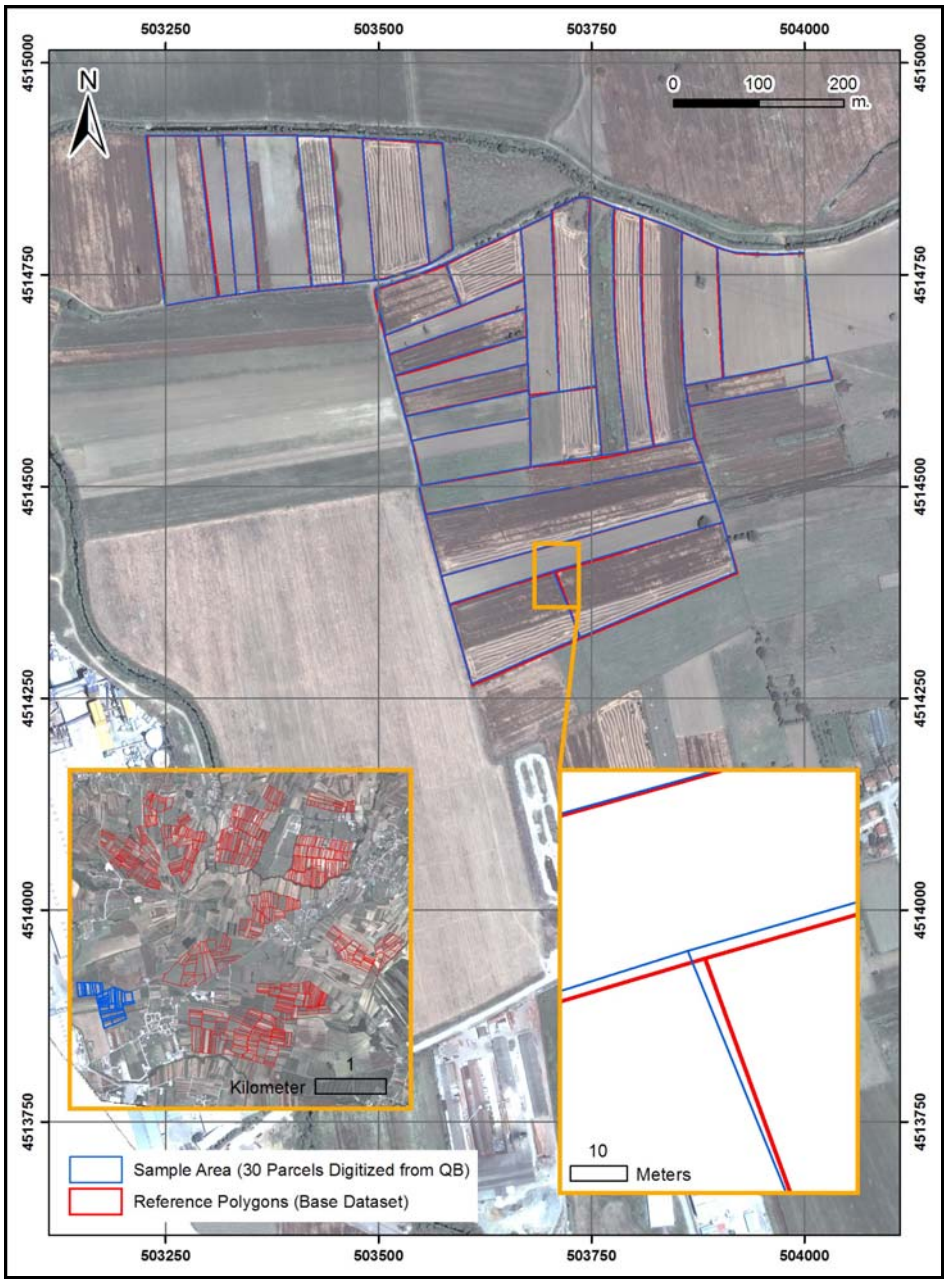

Figure 4.21 Sample Region of 30 Parcel Boundaries

Total number of 409 points derived from 2 polygons is included in calculation. Absolute values are calculated for displacements since either minus or plus sign only indicates the direction of positional error.

Selected polygons and sample vertices can be seen on Figure 4.22.

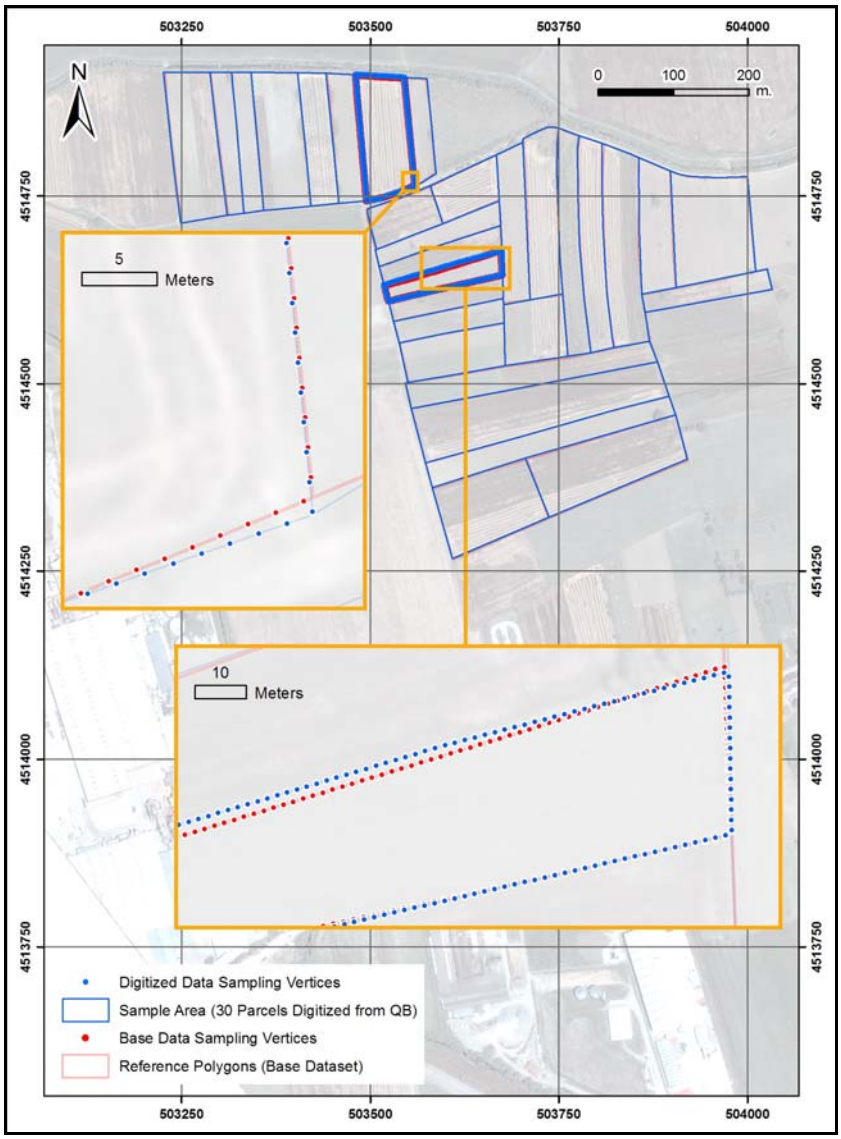

Figure 4.22 Sampled Polygons and Vertices for 95 Polygon Sample Region

Statistical measures regarding positional errors on X and Y axis are given in the following Table 4.12.

Table 4.12 Descriptive statistics of Displacements in Coordinates for 30Polygon Sample Region

	Errors in X Axis	Errors in Y Axis
Standard Deviation	0.78	0.83
Mean	0.86	0.84
Maximum Value	3.2	2.9

Values that are computed for positional errors in X and Y coordinates are converted to ".dat" file format to calculate standardised semivariograms and cross semivariogram. Like in the previous analysis GSlib software tool is utilized to construct standardised semivariograms.

As in the previous sample region of 95 polygons, optimum lag distance is selected according to interval between sample points. As stated, 2 meters is applied as lag distance while forming semivariograms and optimum number of lags for autocorrelation is employed as 60 for X and Y axis and for the cross – correlation structure semivariograms. All three semivariograms are displayed on Figure 4.18.

In model fitting, three different alternatives are assessed, then Spherical and Gaussian semivariogram models following models are selected as the best fitting model for X and Y coordinate errors (Figure 4.23). Spherical model is observed as the best fitting model for cross - semivariogram of positional errors. Indicative Goodness of Fit is used as quality indicator for assessing fitting performance of semivariogram models (Table 4.13). Parameters that are obtained from modelling process are entered as uncertainty model parameters within DUE to generate simulations.

Table 4.13 Model Parameters and IGF Results for 95 Polygon Sample Region

Model Parameters	**X Models**		**Y Models**		**XY Cross Model**
	Gaussian	**Spherical**	**Gaussian**	**Spherical**	**Spherical**
Nugget	0.1	0.07	0.11	0.05	0
Sill	1	1	1	1	**0.3**
Range	38	**51**	61	**75**	**41**
I.G.F	0.038575	**0.018980**	0.136200	**0.107400**	**0.2074**

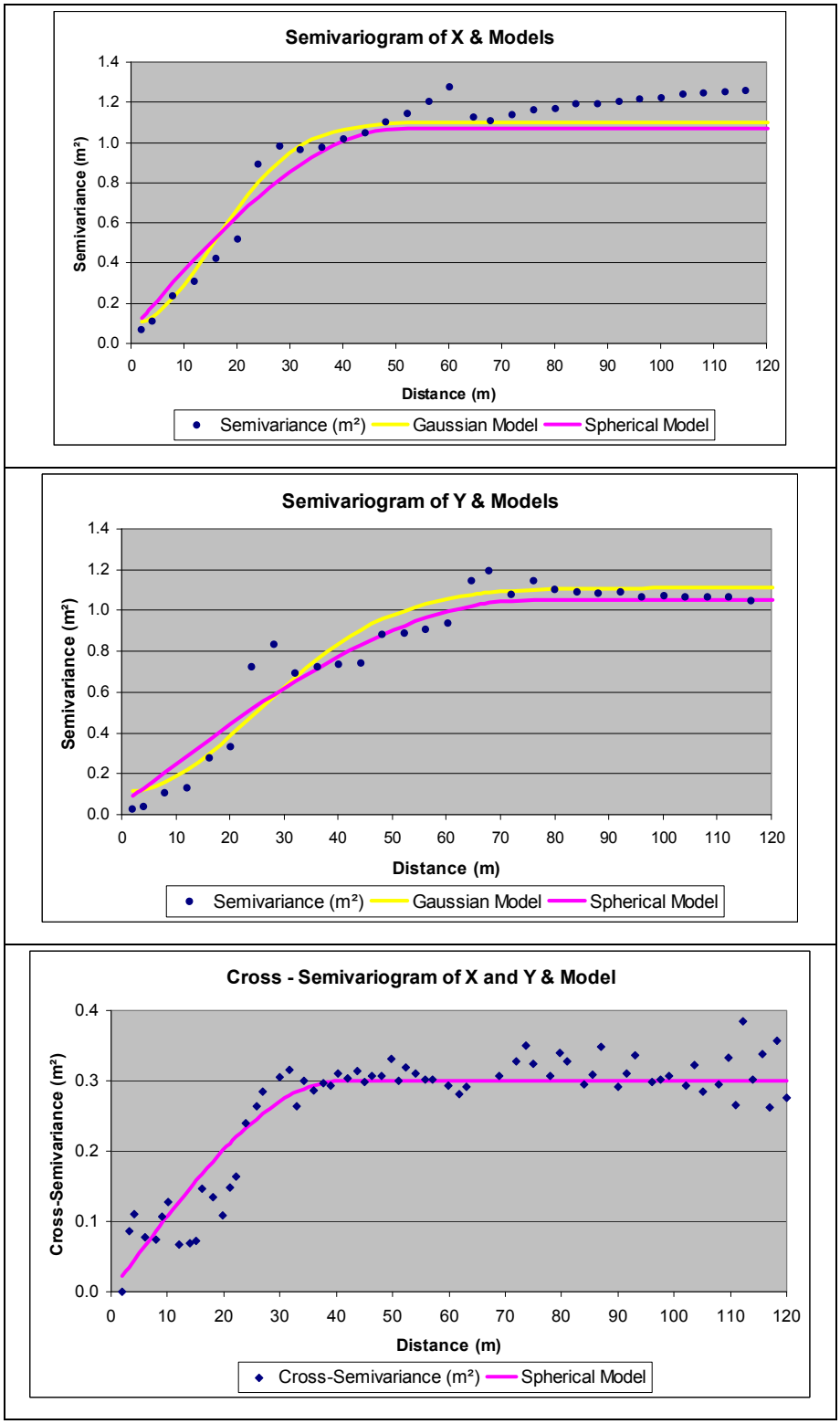

Figure 4.23 Semivariograms for Positional Errors and Best Fitting Models

4.4.2.1. Deformable Object Model with Cross Correlation

Using uncertainty model parameters, which are calculated from semivariogram models, 250 realisations are generated for each polygon within DUE. Parcel boundaries are defined to be deformable; where concept of deformability is explained previously.

Model parameters are supplied as “expert judgement” since DUE does not allow utilization of training data. Standard deviations or spreads of normally distributed errors, σ_X and σ_Y, are both considered as 1 metre (Table 4.12). Normal distribution curves, regarding the coordinate errors, are centred on the object coordinates in case of $\mu_X = \mu_Y = 0$. Auto-correlation and cross-correlation structures are defined by using semivariogram parameters in the modelling window of DUE 3.1. Range parameters for X and Y are entered as 51 and 75 meters respectively. For auto correlation structure, sill is considered as constant on DUE and thus taken as 1 meter. Based on this restriction, semivariograms are standardised so that 1 meter of sill can be utilized. For this case, since cross – correlation displays a positive structure, it can be modelled within DUE. Parameters of range and sill are employed as 41 meters and 0.3 respectively.

Regarding the uncertainty model described above, simulation results performed in DUE can be seen in Figure 4.24.

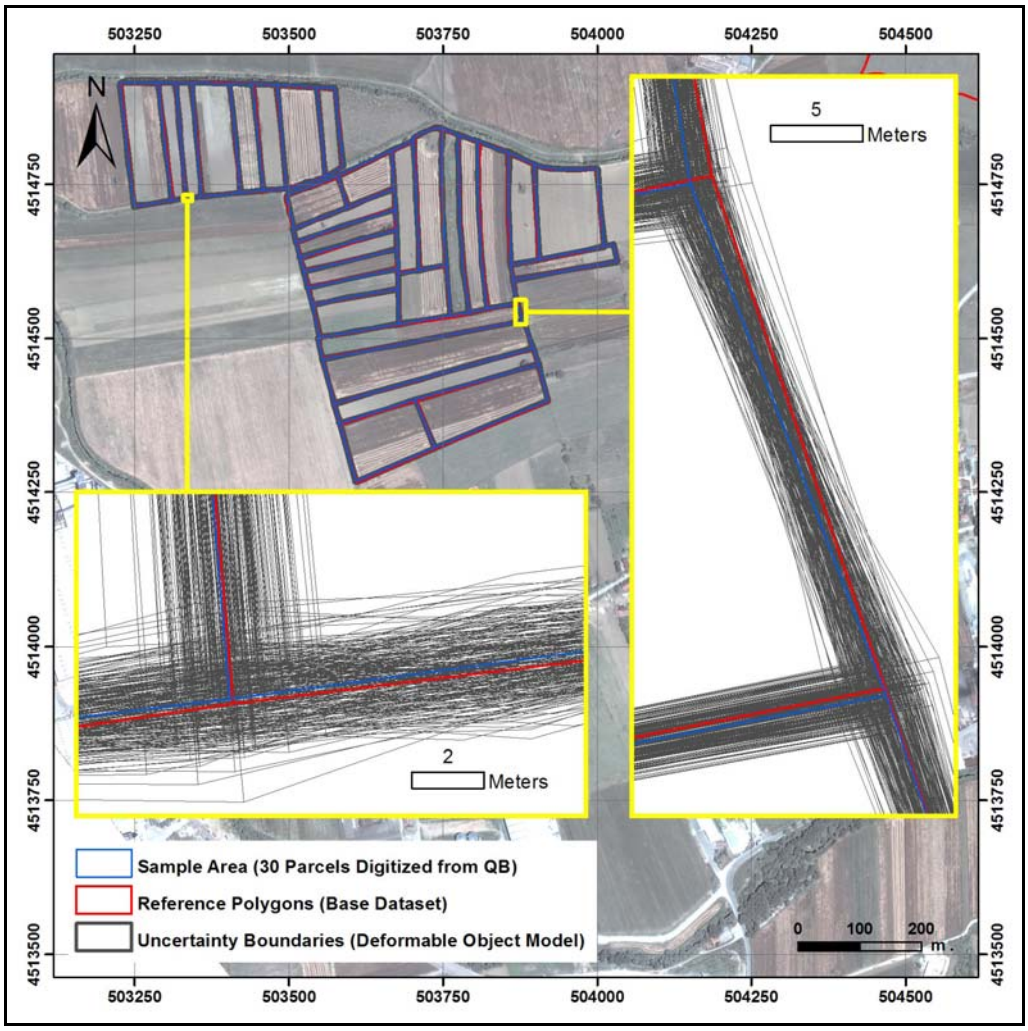

Figure 4.24 Simulations Derived from Uncertainty Model by Deformable Object

Totally, 7500 realisations are evaluated in order to figure the spatial distribution around reference data. These uncertain boundaries are displayed in Figure 4.25.

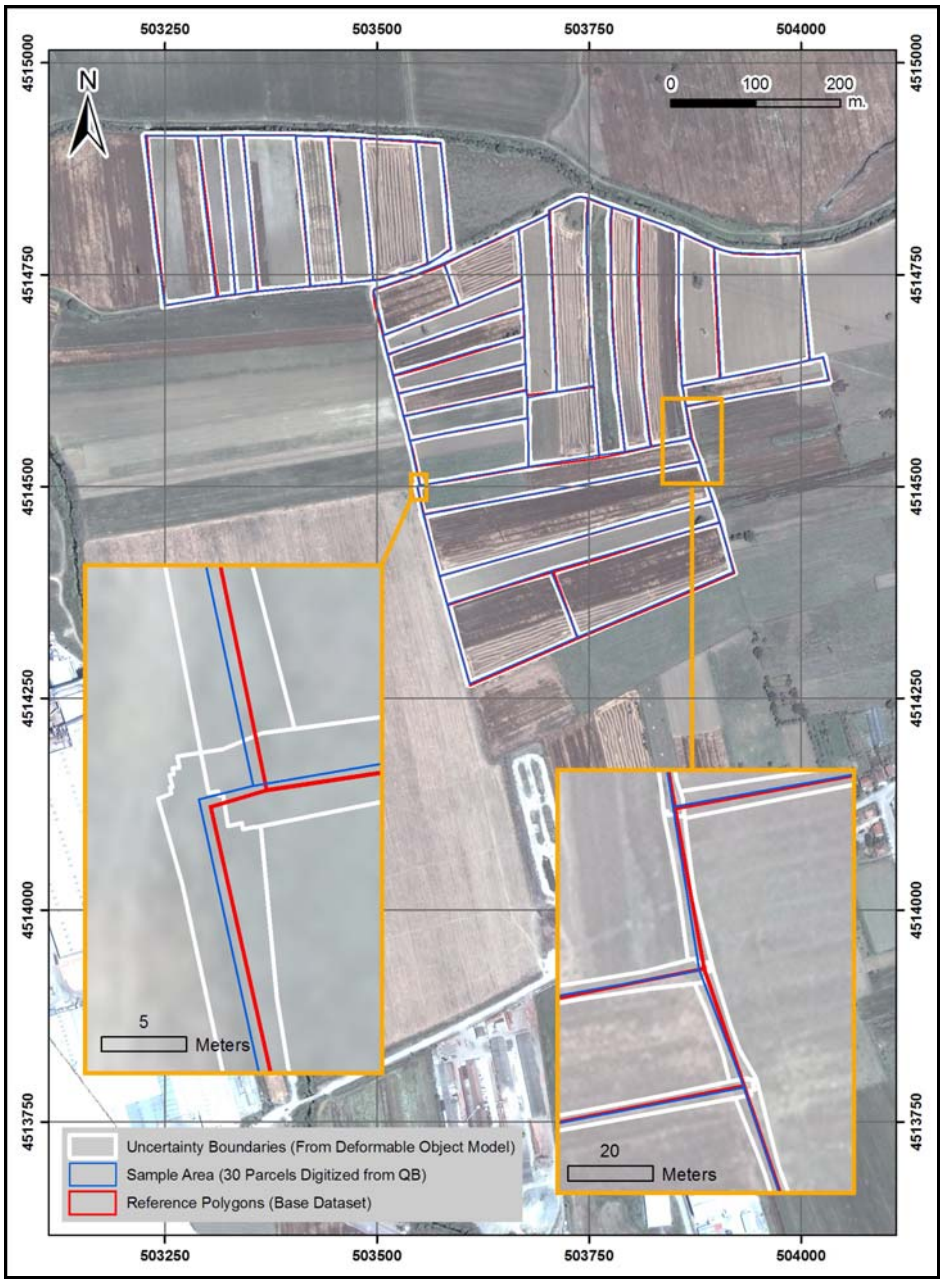

Figure 4.25 Uncertainty Boundaries Derived from Simulations

Number of simulations that resides in multiple buffer zones, each separated with 0.5 meters, is counted by performing a simple spatial query in ArcGIS Desktop v9.3. Spatial dispersion characteristics of these simulations, a histogram and a table, are given in Figure 4.26 and Table 4.14.

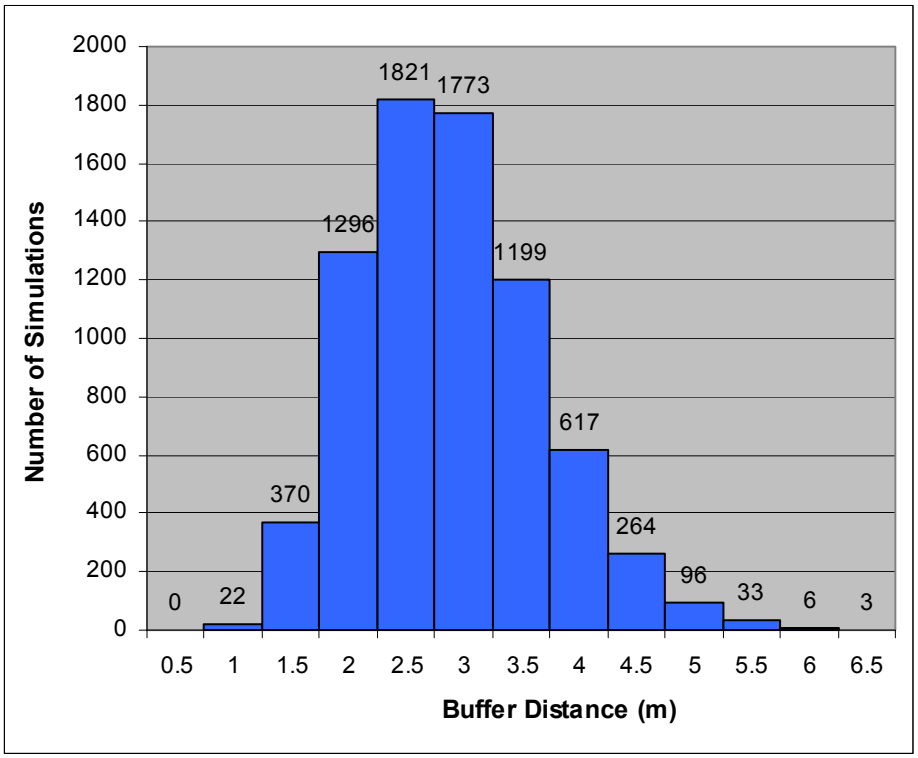

Figure 4.26 Deformable Object Model's Distribution Graphic of Simulations

Table 4.14 Deformable Object Model's Percentage and Cumulative Percentage of Distributions

Buffer Distance (m)	Number of Simulations	Percentage	Cumulative Percentage
0.5	0	0.00%	0.00%
1	22	0.29%	0.29%
1.5	370	4.93%	5.23%
2	1296	17.28%	22.51%
2.5	1821	24.28%	46.79%
3	1773	23.64%	70.43%
3.5	**1199**	**15.99%**	**86.41%**
4	**617**	**8.23%**	**94.64%**
4.5	264	3.52%	98.16%
5	96	1.28%	99.44%
5.5	33	0.44%	99.88%
6	6	0.08%	99.96%
6.5	3	0.04%	100.00%

4.4.2.2. Rigid Object Model with Cross Correlation

Similar procedures and model parameters utilized while exploring the spatial distribution for deformable object model are also carried for the rigid object model case. Rigid object model with translation is selected when defining uncertainty model. Distribution of simulations around reference data is given in Figure 4.27 and Table 4.15.

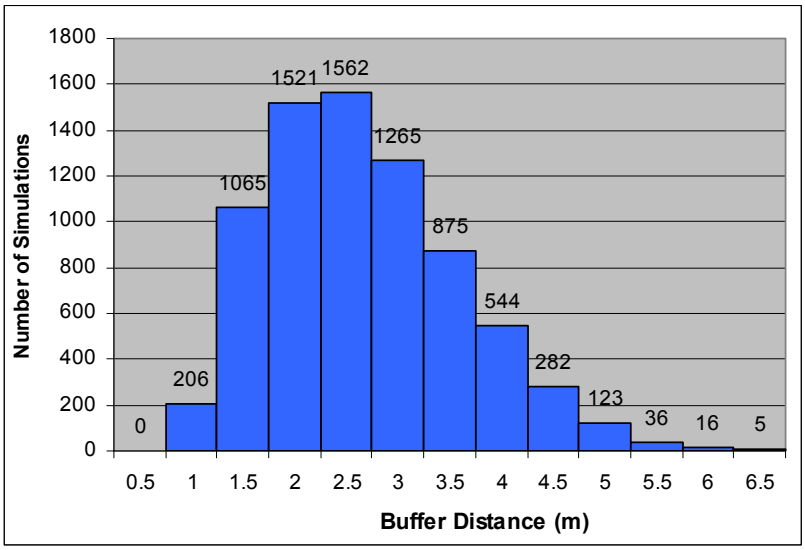

Figure 4.27 Rigid Object Model's Distribution Graphic of Simulations

Table 4.15 Rigid Object Model's Percentage and Cumulative Percentage of Distributions

Buffer Distance (m)	Number of Simulations	Percentage	Cumulative Percentage
0.5	0	0.00%	0.00%
1	206	2.75%	2.75%
1.5	1065	14.20%	16.95%
2	1521	20.28%	37.23%
2.5	1562	20.83%	58.05%
3	1265	16.87%	74.92%
3.5	**875**	**11.67%**	**86.59%**
4	**544**	**7.25%**	**93.84%**
4.5	282	3.76%	97.60%
5	123	1.64%	99.24%
5.5	36	0.48%	99.72%
6	16	0.21%	99.93%
6.5	5	0.07%	100.00%

4.4.2.3. Comparison of Different Object Model Results

In order to compare the results of deformable and rigid object model scenarios, uncertainty boundaries of simulations obtained from each model are overlaid in Figure 4.28. Besides, distribution of simulations around reference dataset are tried to be compared, however since location of vertices are scattered, this leads implication. Similar to G-Band Model, for deformable objects, it is assumed that uncertainty band width gets narrower towards the midpoints, by other means in un-sampled locations.

Main objective here is achieving the smallest uncertainty bandwidth containing reference data. Thus, uncertainty boundaries representing error models should exhibit the possible optimum bandwidth. As wider bandwidths create loose regions, false inclusion and exclusion regions; narrower bandwidths probably end up with exclusion of reference data, so that uncertainty band fails to represent uncertainties in manual digitizing.

In this study, deformable object model gives the best possible uncertainty boundaries for parcel boundaries. As can be followed from Figure 4.28, rigid object model generates wider bands but also includes superfluous regions.

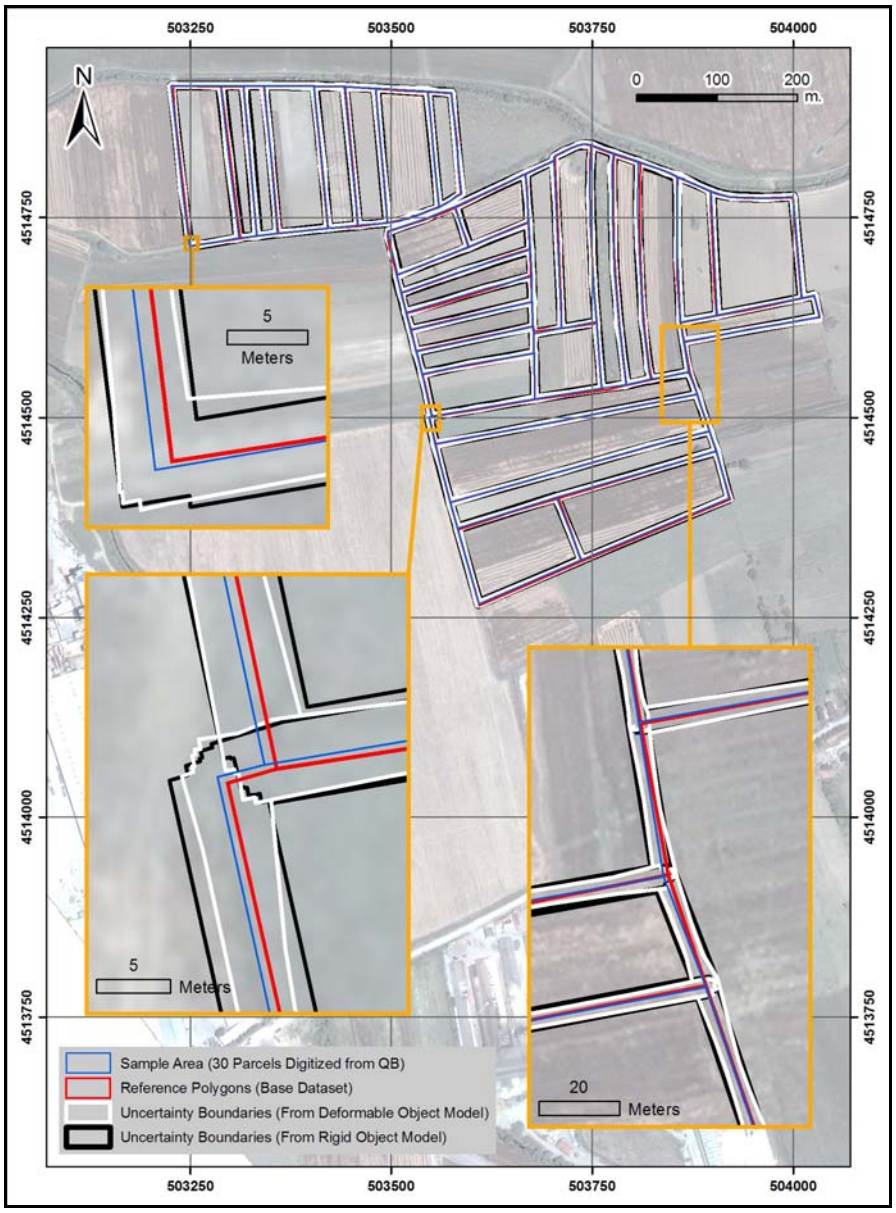

Figure 4.28 Overlaid Uncertainty Boundaries of Rigid and Deformable Object Model

4.5. Area Uncertainty Analysis

Since areas of agricultural parcels are sensitive to the changes in polygon positions, effect of positional uncertainty on area attribute is also investigated in this study.

Inner and outer simulation boundaries are converted into polylines, merged into new shapefile and dissolved as multipart polyline objects representing inner and outer simulation boundaries which resemble a donut like shape (Figure 4.29). This donut like shapes is overlaid with the reference data and possible inclusion, exclusion areas are calculated for each of 30 polygons (Table 4.16).

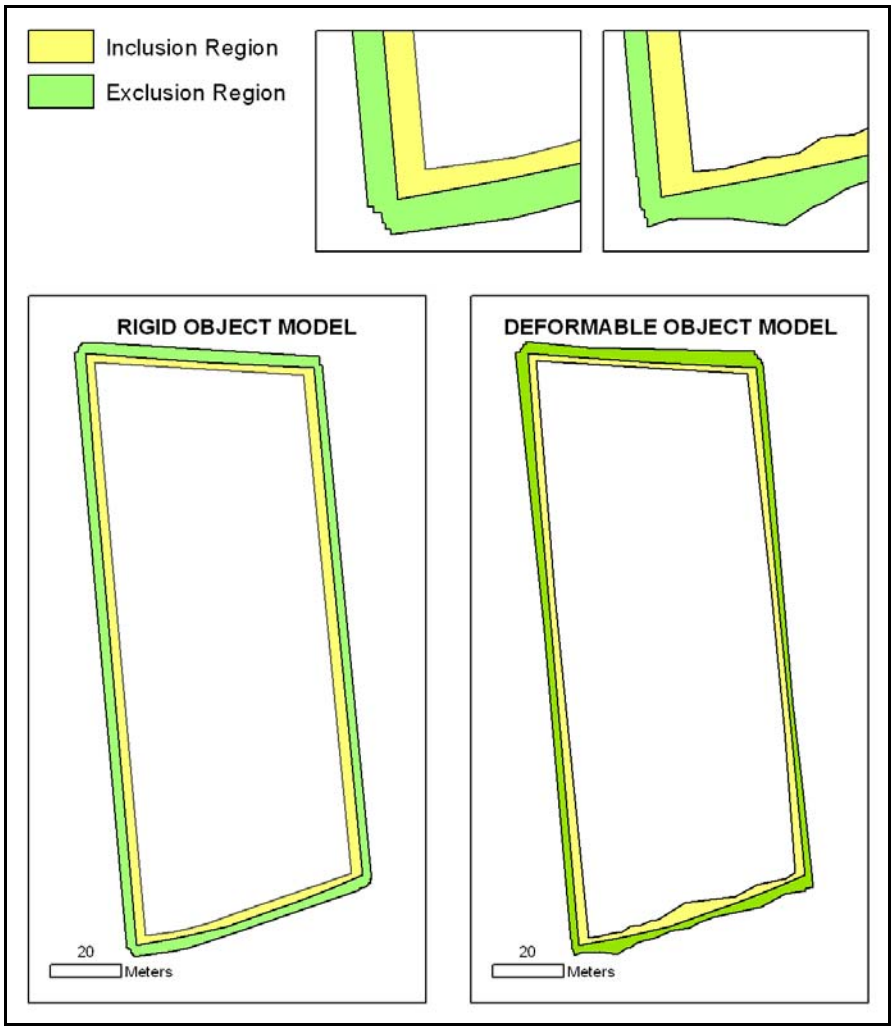

Figure 4.29 Inclusion and Exclusion Regions for Parcel Boundaries

Accuracy of area is influenced by vague locations of point positions. Resulting areas calculated with respect to these inaccurate input values generate unreliable results. However, these results create awareness about the accuracy of data by supplying preliminary information about the true nature of the data.

With reference to Table 4.16, it can be figured out that deformable object model creates more accurate and reliable results compared to the rigid object model. It should be noticed that mean values of rigid method, both for excluded and included regions, are higher than the mean values observed for deformable one.

Table 4.16 Inclusion and Exclusion Areas for Deformable and Rigid Object Models

	Deformable Object Model		**Rigid Object Model**	
ID	**Area Exclusion**	**Area Inclusion**	**Area Exclusion**	**Area Inclusion**
1	776.13	1205.25	1053.45	2493.88
2	1114.05	811.02	1317.23	2284.05
3	1061.48	955.80	1256.42	2433.24
4	1280.05	954.44	1274.28	2263.12
5	937.93	993.73	1073.90	2293.52
6	1305.03	669.91	2005.57	3314.13
7	1121.93	1251.78	1354.11	2715.53
8	823.13	672.04	1023.96	1839.04
9	667.04	672.57	788.07	1593.64
10	736.25	537.28	1057.90	1795.75
11	1020.54	837.80	1308.52	2465.44
12	670.14	1077.09	971.51	2318.05
13	1167.92	511.53	1260.54	1928.18
14	872.14	680.85	1193.71	2150.67
15	844.90	964.35	1159.57	2280.50
16	686.60	890.96	846.93	2019.91
17	988.01	1252.21	1117.19	2604.36
18	538.47	929.72	679.92	1728.15
19	1491.69	975.17	1817.64	3033.95
20	1224.32	1777.55	1633.03	3810.13
21	1304.21	1592.90	1522.46	3355.23
22	1910.96	883.99	2173.50	3380.62
23	2021.77	1131.85	2334.34	3905.54
24	1605.35	1795.78	1861.34	3685.96
25	979.90	1086.94	1350.17	2842.16
26	1468.78	1848.64	1716.06	4053.26
27	1236.54	1229.21	1621.60	3222.04
28	1022.14	770.57	1237.37	2221.20
29	1098.41	1140.65	1408.55	2993.29
30	1039.11	938.22	1248.23	2278.88
Mean	**1100.50**	**1034.66**	**1355.57**	**2643.45**
Std. Deviation	**350.75**	**351.30**	**396.42**	**686.73**

4.6. Discussion of Results

In this chapter a general framework of error analysis for vector polygonal fields are introduced using Data Uncertainty Engine. Previous studies mainly concentrated on points, lines or line segments, because of computational difficulty in defining positional uncertainty of polygon data model.

Generating simulations with Monte Carlo Simulation (MCS) method of possible feature locations from a defined probability distribution function (pdf) is the main principle running behind DUE.

When analyzing positional uncertainty, geostatistical estimation of spatial correlation structure which is developed by Zhang and Kirby (2000) is employed. According to Tobler's first law of geography "Everything is related to everything else, but near things are more related than distant things", spatial dependence in geographical objects requires more than just statistical evaluation. Semivariogram analysis – which is a major part of geostatistics - accounts for distances between spatial objects when representing correlation structures. Because of reasons that are explained, use of geostatistics in uncertainty analysis improved error modeling in GIS.

Three different case studies using the same area are conducted in the study. First case concentrates on modeling positional uncertainty within large dataset. 221 and 479 training and test parcels, respectively are managed using corner points of polygons. In large dataset, it is computationally hard to model cross correlation structure when compared to relatively small ones. It is observed that in the absence of cross – correlation structure uncertainty boundaries showed an epsilon error band model structure, which gives a buffer region around data with a distance of ε.

Departure of 55250 simulations from true value is examined. From all the realisations derived, with 85.90% probability, digitized vectors contain an error of 4 meters in training data. In test data number of simulations per each polygon is

selected as 1000, so that resulting 479,000 simulations are evaluated. Compared to training set, it is observed that distribution of simulations display a more dispersed range of 12.5 meters. In training data, 4 meters distance contain 85.902% of all simulations, but in test data 4 meters of distance cover up 89.87% of all realisations.

Also efficiency of different semivariogram models (namely exponential, gaussian and spherical) is compared on this dataset. Semivariogram models' parameter values and goodness of fit results presented similar results, so that in the uncertainty analysis simulations derived from these parameters shows slight differences.

In second and third case studies a different sampling method is employed, instead of corner points, 5% of population selected as training data and divided into line segments. End points of these line segments are used to calculate positional errors. Small magnitude positional errors are provided with this method because line segments produce consequent sample points (2 meters and 5 meters) within the same polygons which can not divert far away from each other.

In second example cross correlation structure in positional uncertainty modelling is evaluated on 95 polygons. When compared with the previous model, which ignores cross – correlation, nearly 85 % of simulations dispersed within 4 metres. However if cross – correlation structure is considered in the model, about 86% of simulations can be covered within a range of 3.5 metres. Modelling cross correlation improves resulting uncertainty bands.

Last case area focuses on deformable and rigid object model results. First two examples use rigid object model due to model constraints. In this case a region of 30 polygons are modelled both rigid and deformable model with the same error model, and results are compared.

Distribution of simulations are scattered around vertices in deformable object model, this is because of separately calculated pdf for each vertices in the object model

however rigid object model assumes one pdf for all errors. Similar to G-Band Model, for deformable objects, it is assumed that uncertainty band width gets narrower towards the midpoints, by other means in un-sampled locations.

In this study it is aimed to attain the smallest uncertainty bandwidth containing true position of data. Wide bandwidths create loose regions and narrow bandwidths probably end up with exclusion of reference data, so that uncertainty band fails to represent uncertainties in manual digitizing. Deformable object model gives the best possible uncertainty boundaries for parcel boundaries.

Weighted averages of realisations' departure from true locations for 3 case studies are computed. Number of simulations lying in each buffer distance is multiplied with buffers' distance to reference polygons. Afterwards, summation of these multiplications is divided to total number of simulations in each case. A summary of all cases with results are given in Table 4.17. Weighted average positional error for this study is between 2.66 to 2.91 meters for all cases.

To assess inclusion and exclusion areas of uncertainty bands, area analysis is performed comparable with rigid and deformable model results. Rigid object model's falsely included and excluded regions are higher than deformable model. This also proves efficiency of deformable model for representing uncertain polygonal vector data.

Another issue observed in this study that when generating simulation, rigid object model gives a shift to polygon according to a distance and direction; however deformable model concentrates on errors in vertices. This implies the assumption of de Bruin et al. (2008) that rigid object model can be used to assess geometric correction, rectification of aerial photographs or satellite imagery; however deformable model represents human error better.

Table 4.17 Summary of All Three Case

	Number of Test Data	Number of Training Data	Sampling & Number of Sample Polygons and Vertices	Correlation Structure & Object Model	Number of Simulations Derived	Percentage of Distribution Covered in 4 Meter Buffer Distance	Weighted Average Positional Error
Case 1	479 Polygons	221 Polygons	Corner points of Polygons & 479 Vertices	Auto-correlations in X and Y Axis Errors & Rigid Object Model	1000 for each polygon, total number of 479000 simulations	89.87%	2.91 meters
Case 2	95 Polygons	5 Polygons	5% of Population & 5 meter interval 387 Vertices	Auto- and Cross Correlations in X and Y Axis Errors & Rigid Object Model	250 for each polygon, total number of 23750 simulations	92.59%	2.75 meters
Case 3A	30 Polygons	2 Polygons	5% of Population & 2 meter interval 409 Vertices	Auto- and Cross Correlations in X and Y Axis Errors & Rigid Object Model	250 for each polygon, total number of 7500 simulations	93.84%	2.66 meters
Case 3B	30 Polygons	2 Polygons	5% of Population & 2 meter interval 409 Vertices	Auto- and Cross Correlations in X and Y Axis Errors & Deformable Object Model	250 for each polygon, total number of 7500 simulations	94.64%	2.88 meters

CHAPTER 5

CONCLUSIONS AND RECCOMENDATIONS

In this chapter, the conclusions of the conducted study are described together with the recommendations for further studies related with GIS and uncertainty analysis.

5.1. Conclusions

In conventional analyses, where agricultural field boundaries are obtained from field surveys, GPS measurements and digitizing from high resolution imagery are assumed to have exact spatial boundaries, however in reality there is an amount of uncertainty about the border lines.

A general positional uncertainty analysis method is demonstrated in this study that can be used to verify positional accuracy of manual digitizing operations. In this implementation, Data Uncertainty Engine (DUE) is employed which is a free software that aids user in defining probability distributions for uncertain spatial objects and draws random samples from these distributions.

On the previous studies of uncertainty analysis, mainly point and line feature spatial entities are handled within different applications. In this study, uncertainty analyses are performed for polygonal vector data that are manually digitized from satellite imagery.

If user does not have any previous lineage information about digitized data and using the cadastral information of studied region, positional accuracy must be studied to raise the awareness about data. But GIS applications are not eroor-free due to the error may occur from natural variation. Error-aware datasets allows user to describe more reliable outputs from GIS operations.

Use of geostatistical techniques as semivariograms for defining autocorrelation and cross - correlation structures increased the consistency of model outputs.

Two different sampling methods - first is using the corner vertices from the selected training set (221 polygons) and second using selecting 5% of population with spatial random sampling and then using the 2 and 5 metered interval vertices - are used to identify uncertainty model within DUE and all worked well on the data.

DUE's functionality works well within relatively small regions. In the example region of 30 polygons, both deformable and rigid object model are applied easily. However as the number of data increased application can't create variance – covariance matrix to derive simulations. Also the cross – correlation structure gets computationally hard to define in large number of samples. In this study, the smallest sample of 30 polygons can be easily modelled with cross – correlation structure and both in deformable and rigid object model, but in 95 sampled example uncertainty model can only be applied to rigid object model with cross – correlation.

Cross correlation structure increased the reliability and precision of uncertainty boundaries around reference data. Resulting uncertain bands are narrower compared to uncertainty models without cross – correlation structure between X and Y positional errors. In the model applied without cross correlation, the bandwidths expand at most 12.5 meters but in the examples modelled with cross correlation, first region maximum bandwidth is 10 meters and in the second sample area maximum bandwidth is 6 meters.

Best results are found for deformable object model because in deformable model it is assumed that every vertex has its own error distribution and points are considered independent from each other. However in rigid object model, point movements in simulations are modelled with the same joint probability distribution function. Deformable object model gives the best possible uncertainty boundaries for the parcel boundaries in the study. Because in deformable object model, similar to G-Band model output, it is assumed that, uncertainty bandwidth gets narrower towards

the midpoints, by other means un-sampled locations. Rigid object model generate wider bands but also includes superfluous regions.

It can be concluded that human digitising errors, which is the main focus of this study, are represented better with deformable object model, however rigid object model is more convenient for ortho-rectification alignment errors.

5.2. Recommendations

Correlations are assumed to depend on the distance between points. But in the study conducted by De Bruin et.al.(2008) a field boundaries are measured with different GPS equipments – both real time kinematic, hand held differential GPS – and data collected from these measurements are modelled as time series data. Another study can be executed using GPS measurements of four corner vertices of field boundaries and the discrepancy between the selected points can be modelled for uncertainty analysis

In this study, main focus is on the human digitising errors so that the area is selected with respect to the agricultural field density where there is low altitude variation, which is also corrected by ortho-rectification. In remotely sensed images height of geographical objects may cause positional bias in terms of X and Y coordinates of feature. Therefore potential errors that may occur depending on the height of buildings due to perspective distortion can also be modelled within DUE.

Data quality studies for multiple user systems can also be studied for private sector initiatives to improve metadata characteristics. Average positional errors acquired from different users can be included in lineage information.

Also the error propagation of DEM, used in ortho-rectification, can be modelled in future studies. Combined uncertainty of error generated via orthorectication, DEM error and operator's error of digitisation is should be studied within such a study.

However the data used in the study has only the areas as attributes, attributes which are affected by the positions of data can also be evaluated. Amount of fertilizer needed for harvesting, mineral deposit amount of the land or soil salinity can be listed among the examples of attributes that are affected by the position and area of the objects.

Since crop heights in agricultural fields affect the visibility of the borderlines, uncertainty of the digitisation process that is arising with respect to the agricultural crop pattern can be investigated. For instance, long crops tend to create vagueness and the uncertainty increases parallel to the height of the crop. Satellite imagery utilized in this study include 4 bands, if such an NDVI study is willing to be executed especially near infrared and red bands should be used to classify crop pattern on the region. For instance wheat as a long crop type, wheat and harvested wheat exhibits nearly the same reflectance value either in visible or near infrared band. From this perspective, in addition to spectral operations (e.g. NDVI), a field survey study should be undertaken in order to properly analyze the effects of crop pattern on uncertainty studies.

REFERENCES

1. Bogaert , P., Delincé, J. and Kay, S. , (2005), Assessing the Error of Polygonal Area Measurements : A General Formulation with Applications to Agriculture, Measurement Science and Technology, v.16 pp.1170-1178

2. Bolstad, P.V., Gessler P., and Lillesand T.M., (1990), Positional uncertainty manually digitised map data, International Journal of Geographical Information Systems, v. 4, no.4, pp. 399-412.

3. Brown, J.D. and Heuvelink, G.B.M., (2006), Data Uncertainty Engine (DUE) User's Manual.

4. Brown J.D. and Heuvelink, G.B.M., (2007), The Data Uncertainty Engine (DUE): A Software Tool for Assessing and Simulating Uncertain Environmental Variables, Computers and Geoscience, v. 33, pp.172–190.

5. Caspary, W. and Scheuring, R., (1992), Error-Bands as Measures of Geometrical Accuracy, EGIS *92*, pp. 226–233.

6. Cheng T. C., (2003), Assessing Positional and Modelling Uncertainties in Vector-Based Spatial Processes and Analyses in Geographical Information Systems, P.H.D. Thesis, The Hong Kong Polytechnic University Department of Land Surveying and Geo-Informatics

7. Chrisman N., (1982). A Theory of Cartographic Error and It's measurement in Digital Database, In Proceedings of Auto Cartography, pp. 159–168.

8. de Bruin S., Heuvelink G. B. M., and Brown J.D., (2008), Propagation of Positional Measurement Errors to Agricultural Field Boundaries and Associated Costs, Computers and Electronics in Agriculture, v.63, pp. 245 – 256

9. Dutton G., (1992), Handling Positional Uncertainty in Spatial Databases, Proceedings Spatial Data Handling Symposium 5, v.2, pp. 460 – 469, Charleston, South Carolina

10. Fisher P.F., (2005), Models of Uncertainty in Spatial Data, Geographical Information Systems: Principles, Techniques, Management and Applications (second abridged edition), Eds Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. 2nd edn. Wiley, New York pp. 69 - 83

11. Goodchild M., (2007), Encyclopaedia of GIS, Imprecision and Uncertainty, pp. 480-481 Springer

12. Goodchild, M.F., and Hunter G. J., (1997), A simple positional accuracy measure for linear features, International Journal of Geographical Information Science, v.11, pp. 299-306

13. Goovaerts, P., (1997), Geostatistics for Natural Resources Evaluation, Oxford University Press, New York.

14. Heuvelink G. B. M. , Burrough P. , and Stein A., (1989), Propagation of Errors in Spatial Modelling with GIS, International Journal of Geographical Information Systems, v. 3(4), pp. 303–322.

15. Heuvelink, G.B.M., (1998), Error Propagation in Environmental Modeling with GIS, Taylor and Francis,London.

16. Heuvelink, G.B.M.,(2007), Error-Aware GIS at Work: Real – World Applications of the Data Uncertainty Engine, Proceedings of ISSDQ

17. Heuvelink G.B.M., Brown J.D., Van Loon E.E., (2007), A Probabilistic Framework for Representing and Simulating Uncertain Environmental Variables, International Journal of Geographic Information Science, v.21, pp.497–513.

18. Isaaks E.H., and Srivastava R.M., (1989) Applied Geostatistics, Elsevier Science, New York.

19. Kanevski M., and Maignan M., (2004) Analysis and Modelling of Spatial Environmental Data, EPFL Press, Lausanne, Switzerland

20. Kiiveri H.T., (1997), Assessing, Representing and Transmitting Positional Uncertainty in Maps, International Journal of Geographic Information Science, v. 11(1), pp. 33–52.

21. King, J.P., (2002), Modelling Boundaries of Influence among Positional Uncertainty Fields, M.S. Thesis, The University of Maine Graduate School in Spatial Information Science and Engineering

22. Klir G.J., and B. Yuan., (1995), Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall, Englewood Cliffs

23. Kurtar A. K., (2006), Uncertainty Models for Vector Based Functional Curves and Assessing the Reliability of G-band, Master's thesis, Middle East Technical University Geodetic and Geographic Information Technologies Department.

24. Leung, Y., and Yan, J.P., (1998), A locational error model for spatial features, International Journal o fGeographical Information Science, v. 12, pp. 607-620

25. Mathworld (2009) Probability Density Function in Wolfram Mathworld retrieved from: http://mathworld.wolfram.com/ProbabilityDensityFunction.html Date of access: 29.12.2009

26. Perkal, J., (1956), On epsilon length, Bulletin de l'Academie Polonaise des Sciences v. 4, pp. 399–403

27. Perkal, J., (1966), On the length of empirical curves, Michigan Inter-University Community of Mathematical Geographers

28. Shi, W.Z., (1994), Modelling Positional and Thematic Uncertainty in Integration of GIS and Remote Sensing, ITC Publication 22

29. Shi W. and Liu W., (2000), A Stochastic Process-based Model for the Positional Error of Line Segments in GIS, International Journal of Geographic Information Science, v. 14(1), pp. 51–66.

30. Shi, W., and K. Tempfli, (1994), Modelling Positional Uncertainty of Line Features in GIS, In Proceedings of the ASPRS/ACMS Annual Convention, v. 1pp 696–705

31. Veregin, H.,(1999), Data Quality Parameters, Geographical Information Systems, Eds Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. 1st edn. Wiley, New York pp. 177 – 189.

32. Vullings W., De Vries M. and De Borman L., (2007) Dealing with Uncertainties in Spatial Planning, In Proceedings of 10th AGILE Conference on Geographic Information Science , Aalborg University, Denmark

33. Wikipedia (2009), Residual Sum of Squares (RSS) in en.wikipedia.org, retrieved from: http://en.wikipedia.org/wiki/Residual_sum_of_squares Date of access: 23.08.2009.

34. Wolf P.R., and Ghilani C.D., (1997), Adjustment Computations: Statistics and Least Squares in Surveying and GIS, Ed. Roy Minnick, Wiley, New York

35. Yarkınoğlu G. O., (2007), Modelling and Analyzing the Uncertainty Propagation in Vector-Based Network Structures in GIS , Master's thesis, Middle East Technical University Geodetic and Geographic Information Technologies Department

36. Zhang J. and Kirby R.P., (2000), A Geostatistical Approach to Modelling Positional Errors in Vector Data. Transactions in GIS, v.4 (2), pp.145–159.

APPENDIX A

RESIDUAL REPORT OF QUICKBIRD IMAGERY

Table A.1. Full residual report for image including GCPs

GCP/TP ID	Error	Error X	Error Y	Type	Image X	Image Y	Computed X	Computed Y
G0001	0,83	0,46	0,7	GCP	13274	9559	13274,5	9559,7
G0002	0,29	0,27	-0,11	GCP	16106	13201,5	16106,3	13201,4
G0005	0,92	-0,91	0,14	GCP	10002	9269	10001,1	9269,1
G0007	0,49	0,39	-0,3	GCP	14894	10443	14894,4	10442,7
G0008	0,7	-0,5	0,48	GCP	14649	7610	14648,5	7610,5
G0010	0,73	-0,17	0,71	GCP	9784,5	5200	9784,3	5200,7
G0011	1,02	-0,87	-0,53	GCP	6493	6770	6492,1	6769,5
G0013	1,27	-0,8	-0,99	GCP	3255	6714	3254,2	6713
G0014	0,4	0,31	-0,26	GCP	6502,5	1302	6502,8	1301,7
G0015	1,34	-1,14	0,7	GCP	4749	646	4747,9	646,7
G0018	0,9	0,53	0,73	GCP	11791	13109	11791,5	13109,7
G0024	0,76	0,73	-0,23	GCP	14069	10929	14069,7	10928,8
G0025	0,81	0,62	0,52	GCP	12693	10731	12693,6	10731,5
G0034	1,42	1,42	0,07	GCP	9052	6843	9053,4	6843,1
G0036	0,85	-0,79	-0,31	GCP	17503	13057	17502,2	13056,7
G0037	0,31	-0,24	-0,19	GCP	18078	15293	18077,8	15292,8
G0039	0,52	-0,1	-0,51	GCP	15783	14011	15782,9	14010,5
G0040	0,36	0,15	0,33	GCP	12733	12858	12733,2	12858,3
G0042	1,2	-1,17	-0,24	GCP	5342	5930	5340,8	5929,8
G0045	1,49	-0,61	-1,36	GCP	8908	12417	8907,4	12415,6
G0047	1,33	1,02	0,86	GCP	6962,5	4167	6963,5	4167,9
G0052	1,24	0,37	1,18	GCP	8323	9102,5	8323,4	9103,7
G0053	1,38	-1,37	-0,1	GCP	15141	6314,5	15139,6	6314,4
G0054	0,86	0,82	0,27	GCP	18010	5280	18010,8	5280,3
G0058	0,91	0,61	0,67	GCP	16491	6399	16491,6	6399,7
G0060	0,44	0,43	-0,06	GCP	15708	8343	15708,4	8342,9
G0062	0,9	0,22	-0,87	GCP	17791	5978	17791,2	5977,1
G0063	0,57	0,48	0,32	GCP	13805	6452	13805,5	6452,3
G0064	1,17	-0,58	1,02	GCP	13388	6947,5	13387,4	6948,5
G0067	1,19	-1,19	0,1	GCP	12999,5	7838,5	12998,3	7838,6
G0068	1,12	-0,54	-0,99	GCP	15812	4993	15811,5	4992
G0069	0,53	0,35	-0,39	GCP	14719	9359	14719,4	9358,6

G0070	0,62	-0,16	-0,6	GCP	14385	5548	14384,8	5547,4
G0071	0,64	-0,15	-0,62	GCP	14608	7108	14607,9	7107,4
G0072	1,19	0,28	1,16	GCP	5760	2713,5	5760,3	2714,7
G0074	1,18	0,11	1,17	GCP	3124,5	3750	3124,6	3751,2
G0075	1,21	0,72	-0,97	GCP	3364	5734	3364,7	5733
G0076	1,06	0,16	-1,05	GCP	3812	4838	3812,2	4837
G0079	1,41	0,64	-1,26	GCP	4345	1157	4345,6	1155,7
G0080	0,47	0,31	-0,35	GCP	16694	13144	16694,3	13143,6
G0081	0,65	0,63	-0,15	GCP	15673	13009	15673,6	13008,8
G0082	1,2	-0,95	0,73	GCP	13483	13109	13482,1	13109,7
G0083	0,94	-0,21	0,91	GCP	12315	12719,5	12314,8	12720,4
G0084	0,7	-0,55	-0,43	GCP	13319,5	9891,5	13318,9	9891,1
G0087	1,29	1,2	0,48	GCP	14026	11160,5	14027,2	11161
G0088	0,73	-0,71	0,17	GCP	13584,5	11512	13583,8	11512,2
G0090	1,29	1,27	-0,23	GCP	8412,5	11012,5	8413,8	11012,3
G0091	1,35	-1,27	-0,46	GCP	12644	9636	12642,7	9635,5
G0092	0,92	0,82	0,43	GCP	10959	6115	10959,8	6115,4
G0095	0,61	-0,53	-0,3	GCP	16268	5553	16267,5	5552,7

Residual Summary for image

GCPs: 50 XY RMS = 0,98 X RMS = 0.72 Y RMS = 0.67

** Residual units as image pixels*