Modeling Error in Geographic Information Systems

Kimberly R. Love

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Keying Ye, Chair
Eric P. Smith, Co-Chair
Stephen P. Prisley
George R. Terrell

December 3, 2007
Blacksburg, Virginia

Keywords: GIS, vector data, Bayesian statistics, positional error, Douglas-Peucker

Modeling Error in Geographic Information Systems

Kimberly R. Love

(ABSTRACT)

Geographic information systems (GISs) are a highly influential tool in today's society, and are used in a growing number of applications, including planning, engineering, land management, and environmental study. As the field of GISs continues to expand, it is very important to observe and account for the error that is unavoidable in computerized maps. Currently, both statistical and non-statistical models are available to do so, although there is very little implementation of these methods.

In this dissertation, I have focused on improving the methods available for analyzing error in GIS vector data. In particular, I am incorporating Bayesian methodology into the currently popular G-band error model through the inclusion of a prior distribution on point locations. This has the advantage of working well with a small number of points, and being able to synthesize information from multiple sources. I have also calculated the boundary of the confidence region explicitly, which has not been done before, and this will aid in the eventual inclusion of these methods in GIS software. Finally, I have included a statistical point deletion algorithm, designed for use in situations where map precision has surpassed map accuracy. It is very similar to the Douglas-Peucker algorithm, and can be used in a general line simplification situation, but has the advantage that it works with the error information that is already known about a map rather than adding unknown error. These contributions will make it more realistic for GIS users to implement techniques for error analysis.

This work received support from the National Geospatial-Intelligence Agency.

Acknowledgements

I would like to thank my advisor, Dr. Keying Ye, for all the time he devoted to helping me complete this work. Despite relocation to San Antonio, Texas, he was always the first person to respond when I needed help. I am very grateful for all of his assistance. I would also like to thank Dr. Eric Smith, my co-advisor, for finding the time to work with me, even while becoming department head. I appreciate Dr. Stephen Prisley's expertise in the areas of forestry and GISs, without which I never would have been able to write this dissertation. Finally, I thank Dr. George Terrell, for his input and outside viewpoint as my only committee member who was not also working under our research grant.

I am grateful to the National Geospatial-Intelligence Agency for the financial support it provided during this project. Because of this support I was able to dedicate several years of graduate study to this project, and to share the results at several nationwide and statewide conferences.

I would like to thank my fiance, Jon, for his understanding and encouragement during this time. I would also like to thank my mother, Celeste, for always having time for me when I was feeling overwhelmed.

This dissertation is dedicated to my father, W. Dwight Love. Thank you for all your support, both in this work and throughout the years.

Contents

1 Introduction **1**

2 Literature Review **4**

2.1 GIS Data 5

2.1.1 How GIS software stores data 5

2.1.2 Error in GIS data 7

2.2 Some Statistical Tools 10

2.2.1 Bayesian statistical theory 10

2.2.2 Multivariate normal distribution 12

2.3 Error Models for Vector Data 14

2.3.1 Points 15

2.3.2 Line segments 17

2.3.3 Polygons 19

2.3.4 Current problems in vector data 19

2.4 Error models for raster data 28

2.4.1 Problems in raster data 30

2.4.2 Current techniques for communicating error 32

2.4.3 Fuzzy models 34

2.4.4 Statistical models 36

2.5 Discussion 40

3 Bayesian Methods for Vector Data 43

3.1 Introduction 43

3.2 Bayesian Methods in Vector Data 44

3.2.1 Bayesian error models for points 44

3.2.2 Bayesian error model for line segments and polygons 45

3.2.3 Form of the posterior mean and variance for a point on a line segment 48

3.2.4 Choosing an appropriate prior distribution 58

3.3 Examples 60

3.3.1 Point data 60

3.3.2 Line segment data 67

3.4 Discussion 72

4 Calculation for the Boundary of the Confidence Region of a Line Segment 74

4.1 Introduction 74

4.2 Confidence Region Boundary Calculation 75

4.3 Examples 84

4.3.1 Example 1 84

4.3.2 Example 2 85

4.4 General Case 87

4.5 Discussion 93

5 A Probabilistic Point Deletion Algorithm 95

5.1 Introduction 95

5.2 The Douglas-Peucker Algorithm 96

5.3 Statistical Point Deletion Algorithm 99

5.3.1 Error-based Point Deletion 100

5.3.2 General Point Deletion and Comparison to the Douglas-Peucker Algorithm 103

5.4 Discussion 104

6 Conclusion 107

6.1 Discussion 107

6.2 Future Work 108

6.2.1 Application of the Statistical Error Model 109

6.2.2 Further Theoretical Development of the Statistical Error Model . . . 110

6.2.3 The Missing Data Problem 111

6.2.4 Raster Data Error Models 113

6.2.5 Integration with GIS Software 114

Bibliography 115

List of Figures

2.1 Line representation in vector (a) and raster (b) format. 7

2.2 Blakemore's vision of uncertainty in a polygon boundary ([6]). Figure reproduced with permission. 14

2.3 Shi's point distribution ([54]). Figure reproduced with permission. 15

2.4 G-band error model from Shi et al. for line segments ([54]). Figures reproduced with permission. 19

2.5 Polygon confidence region. 20

2.6 Example of point-in-polygon problem. 21

2.7 $R((x, y))$ with (x, y) in A^* (a) and (x, y) outside A^* (b) ([41]). From Springer and Geoinformatica, volume 1, 1997, page 99, Point-in-Polygon Analysis Under Certainty and Uncertainty, Y. Leung and J. Yan, Figure 3; with kind permission from Springer Science and Business Media. 23

2.8 Random polygon with a simple realization (a) and a non-simple realization (b). 25

2.9 Figure describing Leung et al. method for polygon triangulation ([39]). From Springer and the Journal of Geographic Systems, volume 6, 2004, page 370, A general framework for error analysis in measurement-based GIS Part 2: The algebra-based probability model for point-in-polygon analysis, Y. Leung, J.-H. Ma, and M. F. Goodchild, Figure 6; with kind permission from Springer Science and Business Media. 25

2.10 Polygon overlay problem as presented by Chrisman ([15]). Figure reproduced with permission. 26

2.11 Another version of the polygon overlay problem. 27

2.12 Problem of aggregated data. 29

2.13 Example of a contingency table. 33

2.14 Generic contingency table. 34

2.15 Visualization of fuzzy boundaries from Jiang ([35]). Figure reproduced with permission. 35

3.1 Comparison of traditional error ellipse (grey) and Bayesian error ellipse (black). 63

3.2 Comparison of traditional error ellipse (grey) and Bayesian error ellipse (black). 65

3.3 Comparison of traditional confidence region (grey) and Bayesian credible region (black). 72

4.1 Example of confidence region drawn with ellipses. 77

4.2 Demonstration of each ellipse's contribution to the boundary as more and more ellipses are included in the confidence region representation. 78

4.3 Visual demonstration of the determination of the y-coordinate matching the x-coordinate according to line segment slope. 82

4.4 Demonstration of the explicit formula for the boundary of the G-band line segment confidence region. 85

4.5 Demonstration of the explicit formula for the boundary of the Bayesian line segment confidence region. 85

4.6 Bayesian line segment boundary compared to the traditional G-band boundary. 86

4.7 Figure (a) shows the error in the field boundary using error ellipses; figure (b) shows the error in the field boundary using the direct boundary calculation. 86

5.1 Demonstration of the Douglas-Peucker line simplification algorithm. 97

5.2 Map of Virginia, with arc nodes indicated. 98

5.3 Demonstration of the statistical adjustment to the Douglas-Peucker algorithm. 100

5.4 Demonstration of the statistical point deletion algorithm. 102

5.5 A problem with the Douglas-Peucker algorithm; note that an important bend in the line is removed with the DP algorithm (a), but remains with the probabilistic alternative (b). 104

5.6 The statistical alternative to the DP algorithm retains points toward the centre of the line segment (b), whereas the DP algorithm does not (a). 105

Chapter 1

Introduction

The field of geographic information systems, or GISs, is a highly influential force in today's geographic and environmental disciplines. As a computerized system of maps and geographic data analysis, it has found many applications in today's world, including research, government, military, and private uses. Although there are multiple methods of storing and analyzing data, the most general way to describe a GIS is that it uses geographic coordinates to store and display location data, and employs an associated database to identify and describe regional properties.

Despite its relatively long-term use of more than 50 years, until recently there was not much thought given to the types of errors that might occur in such a system. Any information about error included with a GIS was generally brief and over-simplified, leading to a general ignorance of the effect it might have on a final product. Within the last ten to twenty years, however, because of the growing applications and ability of computers to process data, many in the field have become more concerned about the inclusion and appropriate use of detailed error information.

I begin with a review of current work, which has reached varying stages of development depending on the type of data and the application. Some applications are lacking in statistical

modeling and have embraced alternative methods of error communication and display. My discussion includes the merits and downfalls of some of the current models.

I then discuss my work in the area of geographic error modeling. One of my contributions is the introduction of Bayesian methodology into the field, which has multiple advantages. A Bayesian model for map error allows one to introduce expert knowledge, historical data, or information from other maps in the form of a prior distribution. In addition, a Bayesian model can perform well with only a small number of observations. Both of these qualities are essential to the field of GISs, where multiple observations are rare, and outside knowledge can be very informative. I explore this addition, including a discussion on choosing a prior distribution, and provide examples based on my calculations.

Another contribution I have made is the direct calculation of error regions. A popular statistical model for error in line features involves the depiction of confidence regions around line segments. Currently, however, the regions are composed entirely of individual ellipses drawn around many points along a line segment. Here, I have completed a calculation for the direct computation of the boundary for the confidence region. This allows for faster calculation and requires less computational time to draw confidence regions, and should make it easier to include statistical error models in future releases of GIS software.

Finally, I have developed a statistical point deletion algorithm (to be used when simplifying line features in maps), based on error models and the currently popular Douglas-Peucker algorithm. This alternative algorithm has the advantage that there is statistical justification for each point that is removed. Rather than inducing further error, a common complaint about current point deletion algorithms, this algorithm takes advantage of the error structure already present in the data. This algorithm also retains points in situations that the Douglas-Peucker algorithm does not, in particular, points that result in sharp angles and that are roughly centered between their two nearest neighbors. These retained points could provide significant information for the user of a map product, and so the statistical algorithm is a valuable practical alternative.

This work is very important in fields that utilize GIS data. Many of its applications determine such things as allocation of resources, use of funding, and even issues of safety. Without a clear understanding of the error involved in geographic measurement processes and how to interpret it, GIS users may suffer loss and hardship. This work will provide the GIS community with practical tools for the analysis and understanding of error in a GIS.

Chapter 2

Literature Review

The field of geographic information systems, or GISs, is constantly growing in both functionality and application. Although there are many ways of defining a GIS (D. J. Maguire [42] gives some examples, as does D. F. Marble [44]), most would agree that it is a database system for storing, analyzing and manipulating spatially-referenced geographic data. Although some might expand the term to include paper maps and information, users generally understand it to refer to a computerized database structure. A GIS project usually consists of one or more of the following components: maps, information databases, and spatial analysis ([42]).

Computer-based map and data storage go back as far as the late 1950s, but it was not until the 1980s that people commonly used GIS to analyze spatial data ([17]). Now GISs are in widespread use by research institutions, divisions of government (for example, the U.S. Census Bureau and U.S. Geological Survey), private industries, and the U.S. military. Its still-growing number of applications includes planning, engineering, land management, and environmental study ([16]). As GISs continues to gain popularity, some users are paying more attention to the issues of data precision and accuracy.

2.1 GIS Data

2.1.1 How GIS software stores data

One of the most popular commercially available GIS software packages today is ArcGIS, produced by ESRI. It is the most recent incarnation of the Arc/Info software that ESRI began publishing in the early 1980s ([16]). As with any GIS, there are essentially two basic formats for storing map data: *vector* and *raster*.

Vector data is a system of *points*, *lines*, and *polygons* that represent geographic objects. Each *point* that appears in a vector map represents a specific coordinate pair. *Lines* connecting points across a map represent larger and more complicated features, such as rivers or roads. A *polygon* is defined as a closed set of lines. Polygons can represent features with sharp boundaries, like buildings and countries. Another common use for polygon features is to denote areas that are classified differently from other nearby areas, which in reality may have very rough boundaries. Examples of this include maps of soil types or land cover. Each polygon, line, and point feature is linked to a database through a unique identification number, which enables the user to find information about that feature.

This system of storing data is good in the sense that it can store coordinates with as much precision as desired. It has the flexibility to model straight lines with a small number of vertices, or more vertices can be used to plot a more complicated line. It is a very good format to describe objects with well-defined boundaries. It also does not require the use of a lot of computing storage space, since a relatively small number of points need to be stored ([16]).

Raster data is a grid system where each grid cell represents a range of geographic x-,y-coordinates in regular increments. Each raster cell is called a *map unit* and is comparable to the concept of a pixel on a computer screen ([16]). Each map unit has one or more values, generally based on real-world properties over that cell's range of coordinates. Each value

can represent a continuous variable, such as altitude or temperature, or it can represent a categorical variable, like land type.

We interpret the value through the use of a database connected to the raster map. To aid visual display and interpretation, we often associate values with colors or a color scale. One example of a continuous-variable raster map that most people are familiar with is a temperature map, common on the evening news, in which a range of colors on the map represent current temperatures. A good example of a categorical raster map is a land cover map, where each map unit's value represents a summary of the environment within its boundaries.

The raster data format has many advantages. Raster maps are often easy to interpret, as evidenced by the above familiar examples. They also work well in computer applications because of the relationship of map units to pixels on a screen ([16]). When compared to vector data, it is clear that raster data is much better at representing continuous fields (like elevation and temperature) since these fields do not have regular boundaries.

There are also disadvantages to using raster data. One such disadvantage is that each cell has a predefined size that determines the resolution of a map–for example, each grid map unit might represent a square of land that is 10 km by 10 km. Once we establish this, it is not possible to obtain more detail because each cell can only have one value over its entire area. In other words, a map that someone made to reflect the population distribution of the United States in blocks of 10 square kilometers would be useless when trying to determine the distribution of the population in a particular county. Vector data obviously has the advantage here, since it has potential for much higher precision ([16], [24]).

Raster data also falls short of vector data in terms of storage space. In a vector application, software only has to store a relatively small number of coordinates to represent an object, whereas each cell must hold a value in a raster map. This has become less of an issue in recent years as memory storage continues to become cheaper and more easily available. Nonetheless, it is still a consideration since raster maps have the potential to contain a very

large amount of data ([16]), and this can cause certain data operations to be computationally expensive and time-consuming.

Finally, boundaries and well-defined objects are much easier to illustrate using vector data. Although we can draw a line with a raster representation, it is often a clumsy impression and can only be as good as the resolution of the map ([16], [43]). In Figure 2.1, the figure on the left is a vector representation of a line; the figure on the right is a possible raster interpretation of that same line.

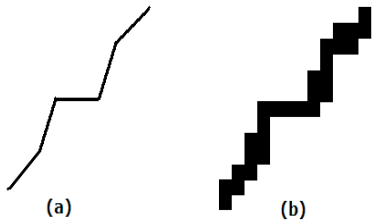

Figure 2.1: Line representation in vector (a) and raster (b) format.

2.1.2 Error in GIS data

As the field of GISs continues to expand, users are paying more attention to error that may be present in a data set. As early as 1984, Chrisman ([11]) and others began to recognize that computers were capable of storing data much more precise than the source of the data could provide. To quote Chrisman, "Storage of detail finer than the error inherent in the source document is a means of fooling yourself." Echoing this early protest, in 1999 he made the observation that data quality has still not increased as quickly as computer storage capacity ([14]). It has perhaps taken us so long to recognize this fault because many users may not wish to acknowledge that there is some variability or uncertainty in their data, since it may affect how their clients or the public view their product ([46], [47]) Many users confused the term "error" with the term "mistake", and think that error indicates sloppiness on the part

of the map producer. I use the term error in this dissertation synonymously with the term "uncertainty".

Recently, however, many users of GIS software are realizing that it is necessary to have a system in place to account for and describe the uncertainty in GIS data. It can be misleading and sometimes dangerous to disregard geographic error. For example, in 2005 when Hurricane Charley struck, many Punta Gorda residents were taken by surprise when the hurricane struck Port Charlotte instead of Tampa. This is because a GIS map released by the National Hurricane Center used a single black line to depict the hurricane's path, and this line went through Tampa which was 100 miles away from Port Charlotte. Although the NHC already acknowledged there was some error in that projection, the average map user probably did not understand how to incorporate that into what they saw on the map. Because of this misunderstanding the NHC committed to incorporating a buffer zone into future hurricane path projections ([32]).

There are many sources of error in GIS applications. Positional inaccuracies, for example, can arise from photogrammetry, map digitization, and survey work, as well as other sources. Photogrammetry refers to the process of creating a map from an aerial photograph or similar picture. This process is somewhat objective, even when performed by trained experts. In order to create a map based on a photograph, it is necessary to determine the true dimensions of objects based on their dimensions in the picture. It is often more complicated than simply tracing objects. The angle of the photograph, distance from the ground, method used to create the image, and other details can distort the interpretation ([7]).

Map digitization is another basis for error in a GIS. The original sources of many maps are paper maps that map producers scan and digitize with software programs. An operator will use his or her personal judgment to outline and classify features such as rivers, roads, buildings, land areas, etc. by denoting points and edges. This procedure is certainly open to operator error, in addition to already existing map or photo quality error ([7]).

Finally, survey work can be a source of error ([7]). In the case of vector data it is often

a matter of judgment where a boundary exists. For example, it may not be clear where something classified as a "forest" or a "swamp" actually begins and ends, and different surveyors may choose differently. Different surveyors may also choose different places to take coordinate readings. GPS equipment is also commonly problematic. Depending on the time of day, number of satellites available, and quality of the equipment, coordinate measurements may be very nearly correct or they may differ quite a bit from the actual coordinates of the spot where the measurement was taken. In the case of *thematic* (classification) data, this can result in misclassification errors (like classifying a map unit as "forest" when it is actually "swamp"). Readings on altitude, temperature, and similar continuous classification variables are also likely to be misread due to quality of equipment. Additionally, when no measurement is available for a particular cell or range of cells, users commonly must interpolate values from the surrounding cells, and there is no method that is consistent across users to do this.

In addition to these initial errors, errors can become compounded through different techniques of analysis and map transformation. For example, maps are commonly overlaid (visually combined) in a GIS project, and individual errors can result in larger errors or in disagreements between the layers of a map. This is also a problem when converting a map to a larger scale, especially in the case of raster data. When users combine small cells to make larger cells, they must make decisions as to how to combine those small cells, and there is no uniform method across users.

Most of the error sources above can influence both vector and raster data. Because of the structure of each type of data, however, vector data is especially prone to positional error, and thematic error is associated most directly with raster data. For the purpose of my research, I will consider vector error to be primarily positional, and raster error to be primarily thematic, although this is not strictly true.

2.2 Some Statistical Tools

2.2.1 Bayesian statistical theory

Bayesian methodology has recently become a larger force in applied statistics, assumedly because of the increased ability of computers to handle complicated distributions (especially those with no closed form). Gelman et al. ([23]) wrote an introductory-level textbook on Bayesian analysis. I provide here a brief summary of the Bayesian premise, based on their text.

Traditional frequentist methodology assumes that the parameter of interest in a problem (often denoted by θ) is a fixed value. From the Bayesian point of view, however, this is not true, and we assume that the parameter has some kind of prior probability distribution. There are several advantages to this approach. At a fundamental level, methods of estimating the parameter differ. Frequentists must use an estimate founded on repeated sampling (for example, the sample mean as an estimate of the population mean), whereas Bayesians may directly estimate the parameter from its distribution. This may not be critical in the case of the population mean, but can have very different results when considering other parameters (for example, the coefficient of variation, or CV).

From a less theoretical point of view, Bayesian methods allow one to consider the "state of knowledge" of the parameter. In the case of geographic data in particular, this makes it easy to incorporate historical or expert knowledge on the location of certain landmarks, or interactions of environmental factors, into the error model. Another general advantage of the Bayesian approach is the common sense interpretation of confidence intervals. Rather than using frequentist logic to quote a *level of confidence* regarding the location of an unknown parameter, one can justifiably quantify the *probability* that the parameter is in a particular interval. This is easier for statisticians and users alike to understand and explain.

Assuming that the parameter of interest does have a distribution, Bayesian analysis is based

on the fundamental *Bayes Rule*. Suppose that the parameter θ and the data y have a joint probability distribution, $p(\theta, y)$. It is a fact that $p(\theta, y) = p(\theta)p(y|\theta) = p(y)p(\theta|y)$. Manipulating these equations results in Bayes Rule,

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}.$$

We refer to the term $p(\theta)$ as the *prior* distribution, $p(y|\theta)$ as the *sampling* or *data* distribution, and finally $p(\theta|y)$ as the *posterior* distribution.

In the above formula, $p(y)$ is a constant since we assume that we have already observed the data y, and there is no dependence on the unknown quantity θ. Therefore the most common form of the posterior distribution is

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{2.1}$$

The right side of equation (2.1) is the *unnormalized posterior density*. We can recover the constant if necessary, since the integral over the range of the parameter must be equal to 1.

Several other issues present themselves when dealing with Bayesian analysis. Briefly, there are two basic types of prior distributions: *informative* and *noninformative* priors. Generally speaking, informative priors are used when there is some population or other basis one can use to justify a particular prior distribution. Noninformative priors are more useful when there is no such basis for inference, and one would prefer that the prior distribution has as little impact on the posterior distribution as possible.

To simplify posterior distributions, many statisticians will use a *conjugate* prior whenever possible. We define a conjugate prior as follows:

$$p(\theta|y) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

where $\mathcal{P}$ is a class of prior distributions and $\mathcal{F}$ is a class of sampling distributions. To put it simply, a conjugate prior is a distribution that, when used with a particular type of sampling distribution, results in a posterior distribution from the same class as the prior. This is very convenient, especially when taking multiple samples, since this guarantees a

posterior that will be easy to work with. Using a conjugate prior when it is not clearly the "right" distribution is similar to a statistician assuming a normal distribution for data that may not be "exactly" normal because it is easy to work with, and is often a good approximation. Of course in cases where one cannot justify this, a *nonconjugate* prior may also be used.

Finally, there is the issue of a *proper* versus *improper* prior distribution. A proper prior does not rely on the data, integrates to 1 (as is required of a probability distribution), and has a proper joint probability distribution with the sampling distribution. An improper prior does not meet these requirements, but is sometimes useful as a noninformative prior when it results in a proper posterior distribution.

This is a very basic overview of Bayesian techniques. We will provide further information on Bayesian analysis throughout, as needed.

2.2.2 Multivariate normal distribution

One distribution in particular that is useful to understand for the traditional treatment of the vector data error model is the multivariate normal distribution. Rencher ([48]) presents a standard treatment in his textbook, which we will follow. Rencher begins with a review of the univariate normal density function, which is

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

where y is the random variable, and μ and σ^2 are the mean and variance of that random variable, respectively.

Applying some well-known results in linear models, we get the density function of the multivariate normal model for p variables, which is

$$f(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^p |\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})/2}$$

where $\mathbf{y}$ is a $p \times 1$ vector of the p random variables, $\boldsymbol{\mu}$ is a $p \times 1$ vector of means for those random variables, and $\boldsymbol{\Sigma}$ is the $p \times p$ variance/covariance matrix of those random variables. $\boldsymbol{\Sigma}$ has the form

$$\begin{pmatrix} \sigma^2_{y_1} & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_p} \\ \sigma_{y_2 y_1} & \sigma^2_{y_2} & \cdots & \sigma_{y_2 y_p} \\ \vdots & \vdots & & \vdots \\ \sigma_{y_p y_1} & \sigma_{y_p y_2} & \cdots & \sigma^2_{y_p} \end{pmatrix}$$

where $\sigma^2_{y_i}$ is the variance of variable y_i, and $\sigma_{y_i y_j}$ is the covariance between the variables y_i and y_j. When a set of p variables has the multivariate normal distribution, we write it as

$$\boldsymbol{y} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).$$

One particularly important case of the multivariate normal distribution is the *bivariate normal* distribution, where $p = 2$. When two variables have the bivariate normal distribution, we sometimes write

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \right). \tag{2.2}$$

One result related to the bivariate normal distribution that is outstanding in the vector literature is the formula for the bivariate 100(1-α)% *confidence ellipse*, which is defined by the set of points $\boldsymbol{\mu}$ satisfying

$$(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \leq \chi^2_{1-\alpha}. \tag{2.3}$$

In a traditional frequentist sense, we interpret this to mean that over a large number of random repetitions of an experiment, this region will contain the true mean $\boldsymbol{\mu}$ 100(1-α)% of the time, on average. We interpret the interval more loosely in Bayesian analysis as a 100(1-α)% *credible region*, which means there is a 100(1-α)% probability that the true $\boldsymbol{\mu}$ is contained in that region.

2.3 Error Models for Vector Data

There have been many ideas on how to account for and display positional error in vector data. As far back as 1966, Perkal attempted to "generalize" a line using the concept of an *epsilon band*, which was simply a band of width epsilon on either side of the line ([6]). There was no statistical calculation involved, simply a buffer zone that he meant to capture all points that could possibly be on the true line. The GIS community mainly used this to determine whether a point from one map could possibly be inside a polygon on another map after an overlay operation had combined the two. Blakemore ([6]) added to this description by classifying a point as *definitely in* the polygon if it is inside the polygon as depicted and not in the buffer zone, *possibly in* if it is inside the polygon and in the buffer zone, *possibly out* if it is outside the polygon but in the buffer zone, and *definitely out* if it is outside the polygon and the buffer zone (Figure 2.2, ([6])).

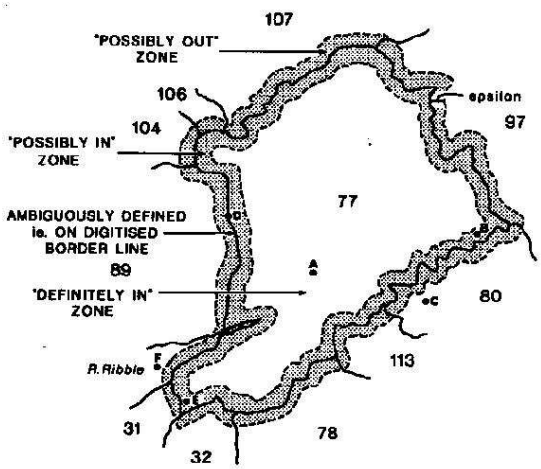

Figure 2.2: Blakemore's vision of uncertainty in a polygon boundary ([6]). Figure reproduced with permission.

Recently, GIS users are attempting to use more statistical analyses to describe error in vector features. Although the epsilon band model is still prominent ([1]), there is new interest in using probabilistic models to describe the distribution of vector features.

2.3.1 Points

Because of the way that vector data is structured, we can trace essentially any type of positional error back to error in vertices. If a line or a polygon is depicted at the wrong coordinates, it could mean that (1) the coordinates of the vertices in line segments or polygons were recorded erroneously, and/or (2) the line segments between the recorded points are not truly straight because they are curved or some vertices have been omitted.

Wenzhong Shi and his colleagues ([50], [54], [52], [10]) have made a lot of progress in this area. In one of his early papers ([50]) he develops a method for modeling points based on a two-dimensional normal distribution. Figure 2.3 from Shi et al. ([54]) visually presents the distribution of a point.

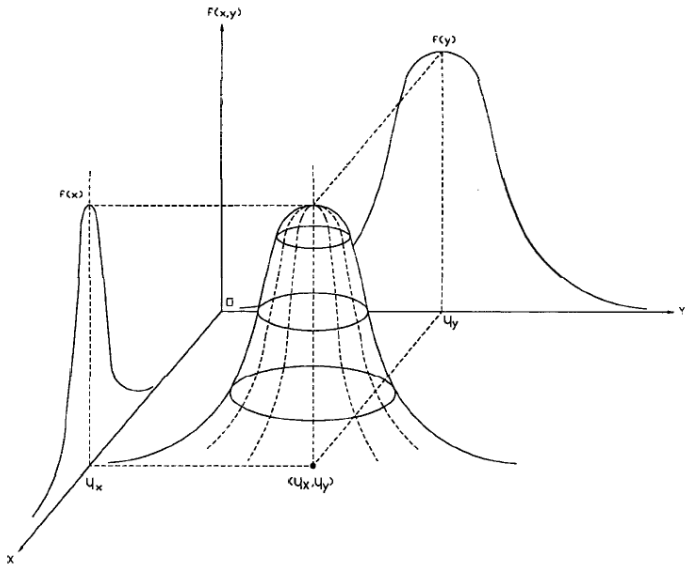

Figure 2.3: Shi's point distribution ([54]). Figure reproduced with permission.

He describes a point with a bivariate normal distribution as follows:

$$\boldsymbol{Q}_{20} = \begin{bmatrix} X_{10} \\ X_{20} \end{bmatrix} \sim N_2 \left[\begin{bmatrix} \mu_{10} \\ \mu_{20} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \right]$$

For geographic clarity, Shi et al. later chose to simply denote X_{10} and X_{20} as X_0 and Y_0, respectively. Note that this equation allows for variation in the X and Y directions as well as correlation between the two, which may be quite likely in some GIS applications. They went on to use this model as the basis for error analysis of lines and polygons in this and future papers. Shi and others often use the (1-α) error ellipse in equation (2.3) or some variation as a means of visually portraying the distribution of the point. This point model and its extension to lines was quickly accepted as a probabilistic alternative to the original epsilon band and other earlier models by Alesheikh ([1]) and others in the academic community.

Although I will be dealing with variations of the probabilistic model above, the academic GIS community has developed other models for uncertain points. As an example, Leung and Yan ([41]) work with an *imprecise* or *fuzzy* model. Under a fuzzy model, a *membership* or *characteristic* function describes each point. Characteristic functions do not follow the rules of probability, and they were originally designed to accommodate some amount of vagueness. Instead of a variable having a probability associated with certain values, it has a *degree of membership*.

Leung and Yan offer an example. Suppose we have a point (x, y), and its membership function is

$$\mu_p(x, y) = exp\left(-\frac{(x-a)^2 + (y-b)^2}{\lambda}\right),$$

with $\lambda > 0$. We interpret this as "(x, y) is approximately (a, b) to the degree $\mu_p(x, y)$". To put this idea in simple terms, we do not consider a point to be one specific pair of coordinates, but a range of coordinate pairs, each to varying degrees. While this approach may have some interpretive advantages, it doesn't offer the ability to make probabilistic statements about GIS features, and therefore I will not focus on the fuzzy model.

Fuzzy methods are becoming a common approach to raster data in GISs. For more examples of fuzzy methods, please refer to the section on raster data error models (2.4).

2.3.2 Line segments

The most common probabilistic error model for line segments in a GIS is a direct extension of the Shi model as I have described it. Shi and Liu ([54]) begin by describing a line segment Z_0Z_1 as a line connecting two endpoints Z_0 and Z_1. We can geometrically represent a point on the line, $Z_t = (X_t, Y_t)$, with the equations

$$\begin{cases} X(t) = (1-t)X_0 + tX_1 \\ Y(t) = (1-t)Y_0 + tY_1 \end{cases}, \tag{2.4}$$

where $0 \leq t \leq 1$.

Suppose now that each endpoint has the bivariate normal point distribution, that is,

$$\boldsymbol{z}_i \sim N_2(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i z_i})$$

where i = 0,1. As they point out, we can further generalize this concept by allowing for the two endpoints of the line segment to be correlated. Shi and Liu characterize the joint distribution of the two endpoints as

$$\boldsymbol{z}_{01} \sim N_4(\boldsymbol{\mu}_{z_{01}}, \boldsymbol{\Sigma}_{z_{01} z_{01}})$$

where

$$\boldsymbol{z}_{01} = \begin{pmatrix} x_0 \\ y_0 \\ x_1 \\ y_1 \end{pmatrix}, \boldsymbol{\mu}_{z_{01}} = \begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_{z_{01} z_{01}} = \begin{bmatrix} \sigma^2_{x_0} & \sigma_{x_0 y_0} & \sigma_{x_0 x_1} & \sigma_{x_0 y_1} \\ \sigma_{y_0 x_0} & \sigma^2_{y_0} & \sigma_{y_0 x_1} & \sigma_{y_0 y_1} \\ \sigma_{x_1 x_0} & \sigma_{x_1 y_0} & \sigma^2_{x_1} & \sigma_{x_1 y_1} \\ \sigma_{y_1 x_0} & \sigma_{y_1 y_0} & \sigma_{y_1 x_1} & \sigma^2_{y_1} \end{bmatrix}.$$

Using some basic results from linear models and equation (2.4), Shi and Liu derive the distribution of a point on the line segment to be

$$\boldsymbol{Z}(t) = (X(t), Y(t))' \sim N_2(\boldsymbol{\mu}_z(t), \boldsymbol{\Sigma}_{zz}(t)), \tag{2.5}$$

where $0 \leq t \leq 1$,

$$\boldsymbol{\mu}_z(t) = \begin{bmatrix} \mu_x(t) \\ \mu_y(t) \end{bmatrix} = \begin{bmatrix} (1-t)\mu_{x_0} + t\mu_{x_1} \\ (1-t)\mu_{y_0} + t\mu_{y_1} \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_{zz}(t) = \begin{bmatrix} \sigma_x^2(t) & \sigma_{xy}(t) \\ \sigma_{yx}(t) & \sigma_y^2(t) \end{bmatrix},$$

where

$$\sigma_x^2(t) = (1-t)^2\sigma_{x_0}^2 + 2t(1-t)\sigma_{x_0x_1} + t^2\sigma_{x_1}^2,$$

$$\sigma_{xy}(t) = (1-t)^2\sigma_{x_0y_0} + t(1-t)(\sigma_{x_1y_0} + \sigma_{x_0y_1}) + t^2\sigma_{x_1y_1},$$

$$\sigma_{yx}(t) = (1-t)^2\sigma_{y_0x_0} + t(1-t)(\sigma_{y_1x_0} + \sigma_{y_0x_1}) + t^2\sigma_{y_1x_1}, \text{ and}$$

$$\sigma_y^2(t) = (1-t)^2\sigma_{y_0}^2 + 2t(1-t)\sigma_{y_0y_1} + t^2\sigma_{y_1}^2.$$

They use this distribution to develop what is referred to as the *generic error band* or *G-band* model. Based on the general error ellipse for the bivariate normal model in equation (2.3), they place an error ellipse at each point along the line segment, resulting in an infinite number of ellipses along the segment. Figure 2.4 (a) from Shi and Liu's paper demonstrates the distribution of the G-band, and Figure 2.4 (b) gives examples of the G-band over various values of $\boldsymbol{\mu}_{z_{01}}$ and $\boldsymbol{\Sigma}_{z_{01}z_{01}}$ ([54]). The collective "outer bound" of these ellipses is what they refer to as the G-band (sometimes called a *confidence region*). The interpretation of this region is that for any particular point along the segment, we are 95% confident that it is contained within this region. The region could be adjusted to account for more than one point if necessary.

Note that we can change the error rate α in equation (2.3) to change the confidence level of the G-band. Also, as far as I can ascertain, no one has yet determined an explicit formula for the confidence region separate from the individual error ellipses.

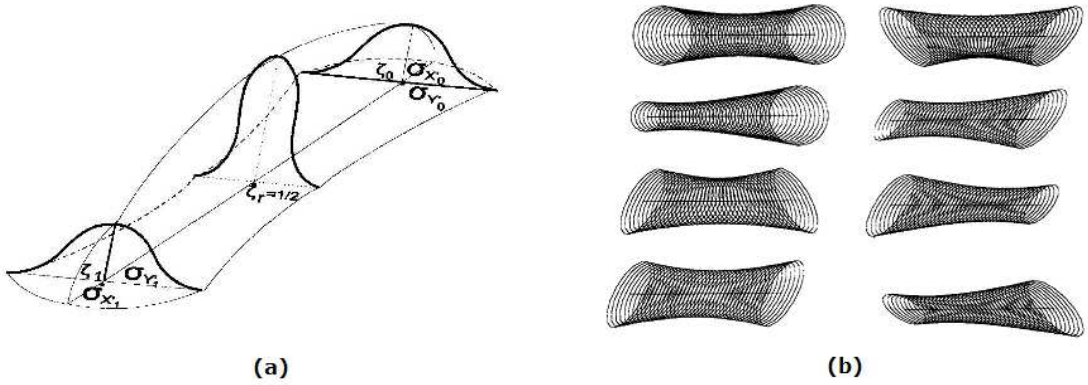

Figure 2.4: G-band error model from Shi et al. for line segments ([54]). Figures reproduced with permission.

2.3.3 Polygons

The extension of the error model for lines to polygons is not a difficult transition. Because a polygon in a GIS is, by definition, a closed set of lines, we can simply model the error in each line segment on the border of a polygon (Figure 2.5, from Leung et al. [39]). Shi used this concept in his work immediately ([50], [54]) and others quickly followed suit ([41], [39], [10], [1]).

2.3.4 Current problems in vector data

Although there are many potential error-related problems for GIS users working with vector data, several particularly prevalent issues have risen to the surface. One problem that exists in current GIS applications is the problem of *missing* or *omitted vertices*, or *line simplification*. Note that in the current approach to describing the confidence region around a line, as developed by Shi, we assume that the line segment between two endpoints is truly straight. That is, we assume the only error is in the endpoints that we measured. This is not necessarily true for all cases. In reality, many geographic features are curved or have

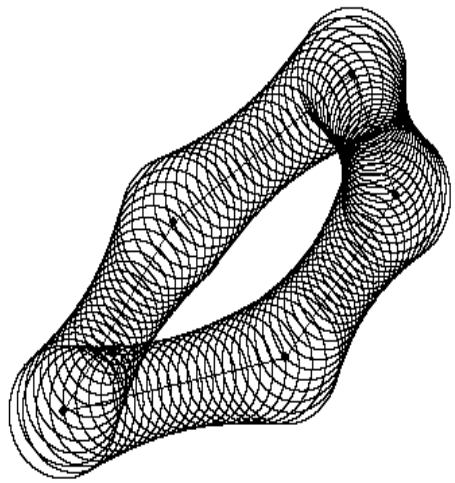

Figure 2.5: Polygon confidence region.

more vertices than the small number that a user is able to record (for example, a river, a road, or a forest boundary). Additionally, users frequently convert small scale maps to small scale maps, which is not always conducive to retaining all the original vertices. An example of this is using detailed local maps of tributaries to make a small scale map of a river. The small scale map simply isn't able to incorporate as much local detail as the original maps, and so we might simplify the tributaries through a point deletion algorithm, such as the Douglas-Peucker algorithm. For more information on this algorithm and alternatives, see Jackson and Woodsford ([34]), Veregin ([60], [61]), or the original article by Douglas and Peucker ([19]). I will also discuss the Douglas-Peucker algorithm further in Chapter 5.

The current model results in a concave error band, as Shi demonstrated (review Figure 2.4), with the most room for error at the endpoints. Some GIS users feel this is not appropriate, however, because intuitively the most information is lacking in the middle of a line segment where data was not taken or was removed with some algorithm. This would suggest that an error band around a line segment should be convex rather than concave, which would reflect a lack of certainty far from the endpoints. This problem has long been recognized ([1]), and some authors have begun to work on the problem ([61], [57], [9]), although largely from a

non-statistical point of view.

Another problem commonly encountered in vector GISs is the *point-in-polygon* problem. Usually this problem occurs when we combine a polygon and a point layer in an overlay operation. The relationship between the points and polygons is affected by the error present in both layers. One important example in the field of defense is attempting to locate an enemy camp, relative to the border of a state or other territory. The information used to assign coordinates to the center of the camp will presumably come from a different source than the information used to determine the border. If the two appear "close" on the resulting layered map, there is a chance the camp may actually exist on the other side of the border from where it appears. See Figure 2.6 for an illustration; the solitary point represents an enemy camp from the point layer, and the partial polygon boundary represents the border of a territory. The curves around each represent their error regions.

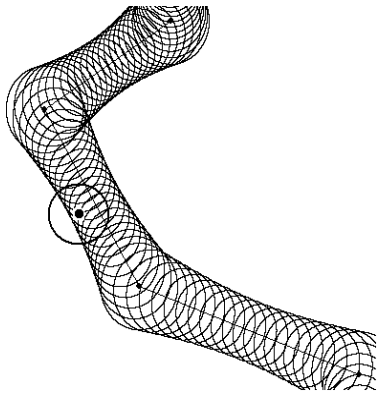

Figure 2.6: Example of point-in-polygon problem.

There have been various attempts to deal with this problem, including Blakemore's early approach discussed at the beginning of this section (review Figure 2.2). Until recently, however, none of these approaches were probability based. Several papers have since been published using probability models to solve this problem.

Leung and Yan ([41]) attempt to combine various notions of uncertainty to quantify the chances that a fixed, fuzzy or random point is inside a fixed, fuzzy, or random polygon. Their

definition of a random point is identical to Shi's definition (bivariate normal distribution) but their definition of a random polygon is not based on the distribution of its vertices. They instead choose the model

$$F_{A^*}(r) = 1 - exp\left(-\frac{r^2}{2\sigma^2}\right)$$

to denote the probability that the boundary of polygon A^* as depicted is located within a distance r of the actual boundary of the polygon. (It is not clear in what capacity they are referring to the boundary, whether this function applies to single points or to the entire boundary in some sense.) This is different from the Shi model, in which we determine the distribution of the boundary solely by the distance from the endpoints. This seems equivalent to the stronger assumption that entire line segments (or even the entire polygon, depending on interpretation) is observed rather than the vertices, and is perhaps less justifiable.

In the particular case of determining whether a random point is inside a random polygon, Leung and Yan calculate what they call the *upper-bound expectation*:

$$\int\int_{R^2} f_{p*}(x,y) \cdot \text{Prob}((x,y) \in A^*)dxdy.$$

They describe this as "the expected probability that $(x,y) \in A^*$, under the [point] distribution density $F_{p*}(x,y)$." Unfortunately they claim they are unable to calculate $\text{Prob}((x,y) \in A^*)$ and therefore use the formula $\int\int_{R^2} f_{p*}(x,y) \cdot [1 - \text{Prob}(A^* \in R((x,y)))]dxdy$. In this equation, they define $R((x,y))$ differently depending on whether (x,y) is located inside or outside the boundary of A^*. If (x,y) is within the boundary of A^*, $R((x,y))$ is the region *inside* the polygon A^* with edges located at the same distance as (x,y) from the edge of A^*. If (x,y) is located outside the boundary of A^*, $R((x,y))$ is the region extending *outside* of the polygon A^* with edges located at the same distance as (x,y) from the edge of A^*. For clarification see Figure 2.7 from Leung and Yan ([41]); in their notation, L_A represents the polygon A^*.

According to Leung and Yan, $1 - \text{Prob}(A^* \in R((x,y)))$ is the "upper limit" of $\text{Prob}((a,b) \subset A^*)$. They must then phrase their result by saying that the upper bound expectation that a

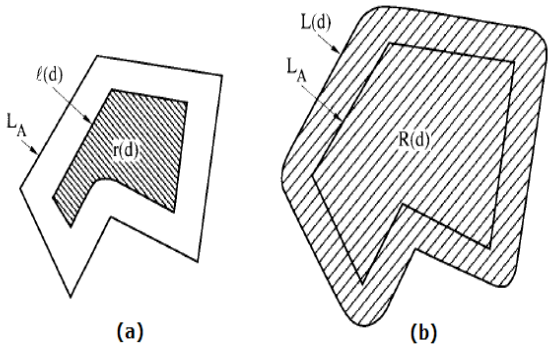

Figure 2.7: $R((x, y))$ with (x, y) in A^* (a) and (x, y) outside A^* (b) ([41]). From Springer and Geoinformatica, volume 1, 1997, page 99, Point-in-Polygon Analysis Under Certainty and Uncertainty, Y. Leung and J. Yan, Figure 3; with kind permission from Springer Science and Business Media.

random point P^* is inside a random polygon A^* is at most

$$\int\int_{R^2} f_{p*}(x, y) \cdot [1 - \text{Prob}(A^* \in R((x, y)))]dxdy.$$

This is not exactly the strict probabilistic result we are looking for, and it does not adhere to Shi's model for the polygon boundary.

Cheung, Shi, and Zhou published an article on the point-in-polygon problem in 2004 ([10]) that answered the problem more directly. Cheung et al. propose that to find the probability of a random point being inside a random polygon, one could first find the probability of a random point P being within a fixed polygon A with the double integral

$$Pr(P \in A) = \int\int_{(x,y)\in A} f_P(x, y)dxdy,$$

where $f_P(x, y)$ is the pdf of P. This is a basic use of the rules of probability.

Assuming now that the polygon A is random, we can use the above equation to calculate the probability that P is in A, given a particular realization of the vertices of A. That is, we can say this is $Pr(P \in A|A)$. In order to find the general probability that an uncertain point P is

in an uncertain polygon A, we can integrate the conditional probability above over the range of possible vertices for A. More basic probability tells us that $Pr(P \in A) = E(Pr(P \in A))$

$$= \int \dots \int h_A(x_1, y_1, ..., x_{NP}, y_{NP}) \times Pr(P \in A|A) dx_1 dy_1 \dots dx_{NP} dy_{NP}.$$

This is essentially a complete solution to the problem, but the authors pointed out that as the number of vertices in the polygon A increases, the computation time involved in the integral becomes overwhelming. Consequently, the authors suggest that it is possible to modify the problem in most situations. Since the error band around the polygon and the point will generally be relatively small, we can consider only the edges of the polygon that are "close" to the point P. The authors therefore recommend considering only the edges of the polygon that intersect the error ellipse of the point be considered.

The authors proceed to break the integral down into cases based on the number of polygon edges the error ellipse of the point intersects, as well as the way those edges are related in the polygon (joined, separate, angle at intersection, etc). This results in dozens of cases, each requiring a different integration. Most certainly this paper is extremely theoretically sound, but the end result is fairly complicated and not necessarily practical or easily programmable for future implementation in GIS software.

One more paper that attempts to solve the point-in-polygon problem was written by Leung, Ma, and Goodchild in 2004 ([39]). This paper focuses on the difficulty introduced by the concept of a random polygon. In particular, a random polygon that is simple, i.e. has a well-defined interior, may not retain this property in certain random realizations. See Figure 2.8 for an example of this phenomenon.

In order to counter this problem, the authors point out that triangulating a polygon by dividing the polygon into multiple triangles circumvents this dilemma, since the interior of a triangle must always be well-defined, even in the random case. They recommend completing the analysis by calculating the probability that the point is in any triangle for which their error regions intersect. See Figure 2.9, originally from Leung et al. ([39]), for a visual

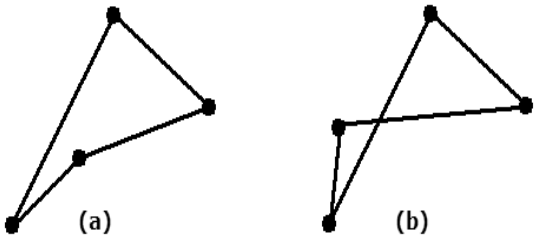

Figure 2.8: Random polygon with a simple realization (a) and a non-simple realization (b).

explanation.

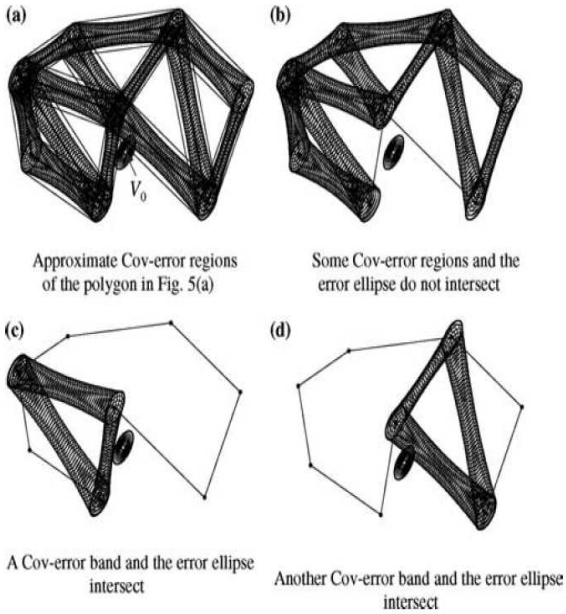

Figure 2.9: Figure describing Leung et al. method for polygon triangulation ([39]). From Springer and the Journal of Geographic Systems, volume 6, 2004, page 370, A general framework for error analysis in measurement-based GIS Part 2: The algebra-based probability model for point-in-polygon analysis, Y. Leung, J.-H. Ma, and M. F. Goodchild, Figure 6; with kind permission from Springer Science and Business Media.

This method does have several failings, however. For one, any polygon does not necessarily have one possible triangulation, but it can have many. Another issue here is that trian-

gulation may not be necessary at all. In most cases, manipulating the covariance between the points in the polygon boundary should result in almost zero probability of randomly generating a non-simple polygon from a known simple polygon.

To conclude what is known about the point-in-polygon problem at present, there is no one specific satisfactory method of solving the problem. Some authors have made attempts to solve the problem, but these solutions are either inaccurate or somewhat complicated in terms of implementation.

One final issue that frequently arises in GIS vector data is the *polygon sliver* or *fractionated polygon* problem. There are several reasons one might want to overlay two polygon maps. One reason, a case discussed by Chrisman et al. ([12],[13],[15]), is to compare or combine two different interpretations of the same area. See Figure 2.10 for an illustration. Map 1 and Map 2 are two different users' interpretations of the land cover of the same area, and the final map is made up of black polygons on a white background to demonstrate where the two disagree ([15]).

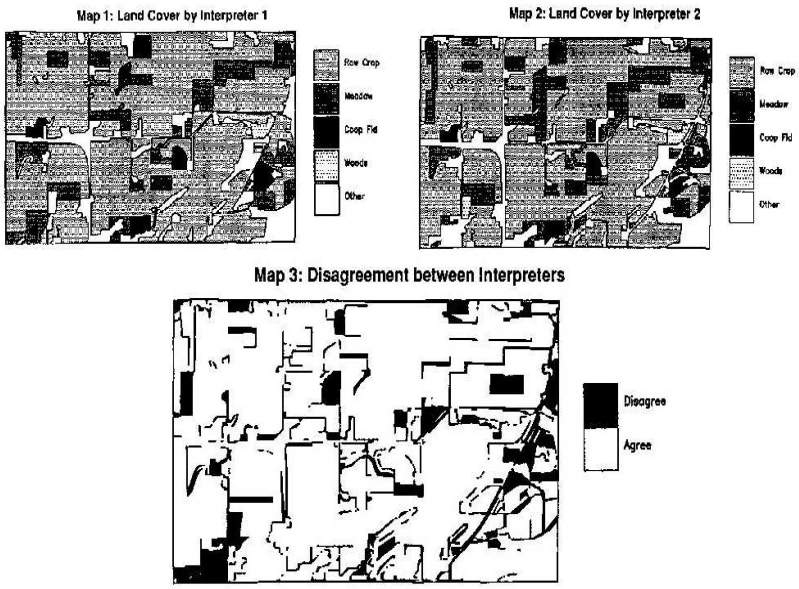

Figure 2.10: Polygon overlay problem as presented by Chrisman ([15]). Figure reproduced with permission.

These black polygons can be considered *sliver polygons*, or small polygons that in all likelihood do not exist. For example, there are areas that one interpreter thought were row crops while the other thought they were meadow. This area is obviously not both, therefore anyone attempting to combine the two maps must make a decision as to how to deal with the error and classify this area. While in some instances this may be a result of misclassification, there are some areas where this is clearly a result of measurement error–the two interpreters did not agree on the boundary of an area of land cover. Chrisman proposed a test to determine the amount of error involved, but did not suggest a method based on probability.

Another example of a sliver polygon problem would be when two maps of different but related applications are overlayed. Veregin ([58]) discusses this type of problem, and I provide an example. Figure 2.11 shows a map of land cover and a map of land use that have been overlaid. The resulting map is a composite polygon map where polygons are classified with a land cover and a land use.

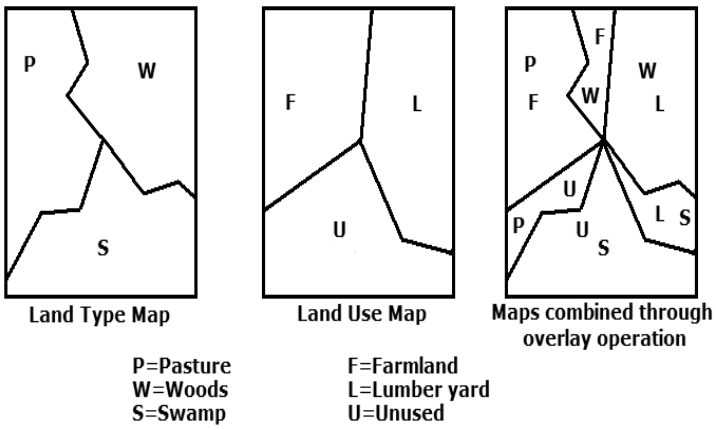

Figure 2.11: Another version of the polygon overlay problem.

The question in this case is whether all the resulting polygons truly exist, or if some of them are created by location error in the two original maps. In Figure 2.11, there are small polygons that in all likelihood do not truly exist; for example, there is a polygon where the swamp appears to be used for lumber. Veregin discusses an algorithmic method to deal

with this situation based on deleting polygons that are "small" relative to others, but again there is no probabilistic approach mentioned. As far as I am aware, there is no current probability-based method available to help in the decision-making process when performing polygon overlays. There has, however, been some probabilistic work on error propagation models for intersections in vector overlay operations that should prove helpful in designing such a method ([40],[51]).

2.4 Error models for raster data

Raster data, unlike vector data, necessitates the use of classification information since each cell involved must be associated with a value in a database. Although positional error may exist, the end result for the user is misclassification error. The reason for this is easy to explain. Suppose a surveyor doing field work stands at the edge of a forested area next to a clearing, and uses GPS equipment to record the location of the boundary. If the GPS reading is incorrect, and the surveyor records coordinates that are actually located in the clearing, the resulting map will show a forested area where there is actually none. This can have many unwanted effects on a subsequent analysis, including incorrect distance calculations to the boundary of the forest and incorrect estimates of the area of the forest. The literature on GIS data contains further discussion and viewpoints on this topic, for example, Sunila et al. ([56]).

There are many origins for error in raster data. According to Bolstad and Smith ([7]), most error in categorical data comes from such things as remote sensing (cameras, scanners and video) or from field inventory. Each of the remote sensing methods involves error from the quality of the original product, whether it be a picture or a paper map, as well as error on the part of the interpreter, who must decide how to classify the information from the source. Field inventory relies on the quality of the instruments and training of the surveyors, as well as the sampling scheme they use–for example, surveyors often evaluate and classify small

areas, then apply these measurements to larger areas for the final product.

Additional error comes from operations that others may have performed before the current user inherited the product. For example, a producer sometimes *aggregates* large scale data into fewer cells, and information on the details of the smaller-area cells is lost ([56]). Additionally, Shi ([53]) makes the point that large cells may be combinations of small cells from different categories, and even if there are nearly equal proportions of each category present, a user must choose one category as the value for the large cell. Major features from the large scale map may also be lost, depending on how the cells are aggregated ([29]). For an example, see Figure 2.12, in which a large scale land cover map is aggregated to a small scale map in which nine of the original cells make up one large cell. The shaded cells represent a stream and the light cells represent sand. In Figure 2.12 (a), the six stream cells happen to fall in one large-area cell, and because they make up the majority of that cell the new cell is labeled stream. In Figure 2.12 (b), however, the stream cells are split between two large-area cells and do not make up the majority land cover in either of them, and so the stream no longer appears on the small scale map. If a user only possessed the aggregated map, they would not be aware that a stream existed.

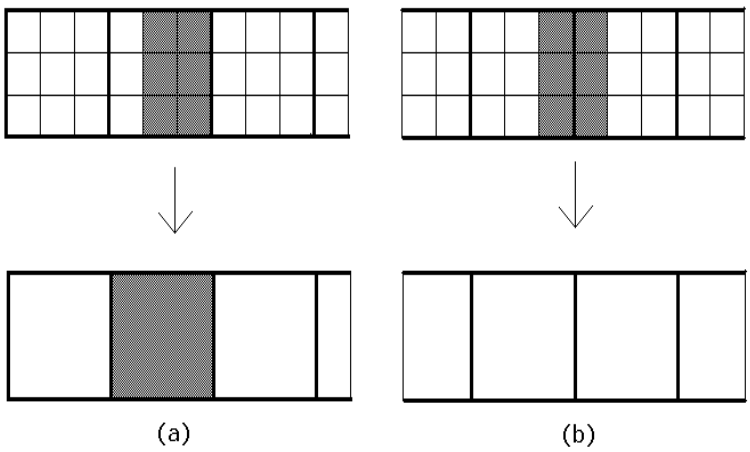

Figure 2.12: Problem of aggregated data.

2.4.1 Problems in raster data

There are many difficulties that arise from classification error in raster data. For one thing, the two types of classification data in GISs, continuous and categorical, generally require different statistical error models. For another thing, when modeling classification data, we must recognize that cells are usually spatially correlated (for example, elevation in a cell is likely to be related to the elevation in nearby cells). The amount and type of error in a cell is most likely affected by location, both in general and relative to other cells. This sometimes causes probabilistic methods to become complicated, mathematically and conceptually. This has lead to the development of alternative models.

Many authors choose to model errors in thematic data not with a probabilistic or statistical approach at all, but with fuzzy sets and related logic. This approach has recently gained popularity in the GIS community because it lends itself well to pictorial representations of error, which are easily interpreted by the average map user. It does, however, have certain downfalls. In particular, it does not allow statements regarding the probability of certain events. I will later discuss fuzzy methods in further detail.

No matter what the approach, it is common for raster error modeling to be tailored to specific problems. Often authors put forth models in relation to certain GIS operations. Some popular problems center around map overlay, DEM conversion operations, and boundary detection/representation. It is also important to remember that many GIS users utilize these final products in some way to calculate loss or gain in time, resources, and money–calculations that will include error from the source maps and operations.

I have already described map overlay briefly as a combination of two maps, but there are multiple procedures in the raster case that we can use to do so, and numerous setbacks can arise from a map overlay. Veregin ([58]) discusses several of the operations that we can use in a raster map overlay. When we combine two or more cells from different maps, there are various mathematical functions to choose from: addition, subtraction, or taking a ratio, to name a few. There are also set theory possibilities for the map overlay, including the AND

and OR operators. There are additional complications when the maps involved in the overlay are not at the same scale or do not have the same size cells. In the event that the maps are not at the same scale, one or both of them must be adjusted. If they do not have the same cell sizes, a user must decide how the overlapping cells from each map will be averaged or otherwise combined to facilitate the overlay. Each of these operations will induce error in the overlay process, which adds to the error already present in the source maps. Determining how error is propagated and the best way to communicate this to a client is the focus of most map overlay error studies. For more discussion of these problems, see Veregin ([58]) or Griffith et al. ([27],[2]).

A *DEM*, or *digital elevation model*, is a map which assigns a terrain elevation to each cell. (Some authors choose to refer to such maps as *digital terrain models*, or *DTMs*, as a matter of personal preference.) Usually a DEM is created through field work with a GPS unit or from digitizing a paper elevation map. For some basic information on DEMs, see Clarke's introductory text in GISs ([16]). DEMs are particularly useful in a GIS because of the many ways users can manipulate the information. By using the elevations in each cell and the relationships between these elevations, a user can obtain a map reflecting the land's convexity, drainage bins or water sheds, viewsheds (models indicating what points on the land are in the direct line of sight from a particular location), hillshading, and perspective, among many other applications. Each of these operations will reflect the original error in the DEM, as well as inflict further error on the final product through the methods employed. For more information on these operations and the type of error they can incur, see Weibel and Heller ([63]).

Boundary detection is a common problem in GISs as well. This is a problem for both vector and raster data. As I have already discussed, polygon data is an excellent way to display areas with sharp boundaries. In the case of classification data, however, there are many situations in which objects do not have sharp boundaries–examples include maps of land cover and soil type. Many objects involve *transition zones*, or areas between objects that are not easily classified as either object. In polygon data, this means the sharp boundaries as

depicted in the map product are contrived and unrepresentative of the objects as they appear in nature. In raster data, this usually means that cells in transition zones are artificially "forced" into a particular category and do not reflect reality. Besides being difficult to display and communicate to users, uncertain boundary data leads to error in common operations, like measuring distance from other locations to that boundary, and determining the total area on a map composed of a certain class of data. For more information on boundary error, see Greve and Greve ([26]), Sunila et al. ([56]), or Jiang ([35]).

Now, I will discuss some of the various error models in the literature.

2.4.2 Current techniques for communicating error

Since the GIS community has long recognized the existence of error in raster data, there are already some rudimentary methods in place for communicating this error to users. One measure of error for continuous data that GIS maps usually include is the *RMSE*, or *Root Mean Squared Error*. The formula for the RMSE is $\sqrt{\sum_{j=1}^{m} (p_j - s_j)^2}$, where $j = 1 \ldots m$ is the number of cells used in the calculation (generally a small number that the map producer has re-sampled with a high degree of accuracy), p_j is the value recorded in the map for cell j, and s_j is the "true" value of cell j recorded in a re-sampling procedure. It is a valuable and well-understood method of communicating error, but it is not very useful for modeling error in cases where there is systematic bias or trend in the map error ([22]).

One very common method for expression of error for categorical data is the *contingency table*, *confusion matrix*, or *classification error matrix*, which is based on a small re-sampling of the cells in a map. It is a table accompanying the data that tells the user how many times the map's producer included a cell in one class (row) while it is actually in the category specified by the re-sampling procedure (column) ([28], [55]). See Figure 2.13 for an example of a contingency table. (GIS users sometimes alternately write contingency tables in terms of proportions, where each numeric cell is divided by the total number of re-sampled cells.)

Recorded category from original map	"True" category from re-sampled data: Forest	Swamp	Pasture	Recorded category totals
Forest	28	2	1	31
Swamp	6	12	3	21
Pasture	2	0	18	20
"True" category totals	36	14	22	

Figure 2.13: Example of a contingency table.

Several measures of accuracy based on contingency tables are common. Two of the most widespread are the overall Proportion Correctly Classified (PCC) and the Kappa coefficient of agreement. Referencing the generic contingency table in Figure 2.14 , define p_{ij} to be the proportion of cells initially in class i and in re-sampled reference class j, $p_{i+} = \sum_{j=1}^{q} p_{ij}$ is the proportion of cells the original map included in class i, and $p_{+j} = \sum_{i=1}^{q} p_{ij}$ is the proportion of cells belonging to class j in the re-sampled data. The overall proportion of area correctly classified is then $P_c = \sum_{i=1}^{q} p_{ii}$, and the Kappa coefficient of agreement is $\kappa = \frac{P_c - \sum_{i=1}^{q} p_{i+}p_{+i}}{1 - \sum_{i=1}^{q} p_{i+}p_{+i}}$. Two additional measures are the *user's accuracy* for class i, $P_{Ui} = \frac{p_{ii}}{p_{i+}}$, which is the proportion of cells a user interprets as belonging to class i and truly belong to that class according to the reference data, and the *producer's accuracy* for class j, $P_{Aj} = \frac{p_{jj}}{p_{+j}}$, which is the proportion of cells the map's producer classified as belonging to class j and truly belong to that class according to the reference data. For these and further accuracy measures based on contingency tables, see Stehman ([55]).

While these traditional measures are generally useful, they are not comprehensive and do not offer error modeling for individual cells. Because of the GIS community's recent attention to these issues, however, more detailed error models are evolving. Below, I discuss some examples.

		Reference				
		1	2	...	q	
Map	1	p_{11}	p_{12}		p_{1q}	p_{1+}
	2	p_{21}	p_{22}		p_{2q}	p_{2+}
	...					
	q	p_{q1}	p_{q2}		p_{qq}	p_{q+}
		p_{+1}	p_{+2}		p_{+q}	

Figure 2.14: Generic contingency table.

2.4.3 Fuzzy models

As I have already mentioned, many researchers are turning to fuzzy models as a method for modeling and visualizing error in data. Fuzzy models have found many uses in GIS error modeling (for one of the more unusual uses, refer back to Section 2.3 for Leung and Yan's [41] fuzzy model for vector data points). Perhaps the most popular application of fuzzy methods, however, is boundary representation. There are many examples in the literature that deal with this concept ([36], [56], [22], [3], [35], [26]).

One particularly popular example is soil classification maps. Soil boundaries are particularly unclear. Not only are sampling points often scarce due to the expense of testing (leading to a large amount of interpolation between points), but soil types do not have sharp, immediate transitions. Often there are large transition zones between soil types where the soil may not fit well in either category. Because of this, not only do different soil experts designate the boundary differently, but the map user may not ever become aware of the size or even the presence of the transition zone between soil types.

The fuzzy solution to this problem is to assign a *degree of belonging* to each soil class to each map unit in the transition zone. Jiang ([35]) explains this by discussing the *membership function* of a fuzzy set, which defines the grade of membership x in a set A. The value x must be between 0 and 1 for any set. In the case of soil maps, this means a cell may simultaneously belong to two or more soil types. The largest difference between fuzzy set theory and probability is that a map unit may have varying degrees of belonging to multiple

sets. This is in contrast to the Boolean probability approach, in which each map unit can belong to exactly one set (a value of 1) and all other sets are assigned a value of 0.

The fuzzy set approach does have a natural appeal, especially for the soil boundary situation where transition zones clearly will have some characteristics of each surrounding soil class. In fact, a large factor in the conception of fuzzy sets was its ability to mimic human perception ([38]), perhaps explaining why so many in the GIS community are drawn to it. In addition, it is easy to represent fuzzy boundaries in an intuitive fashion with the use of color and shading ([35]). For example, if a unit's membership is close to 1 it may be a dark or solid color, and if it is closer to 0 it may be lightly shaded, or even be a different color representing a set in which it has a higher degree of membership. For an example of this, see Figure 2.15 from Jiang ([35]).

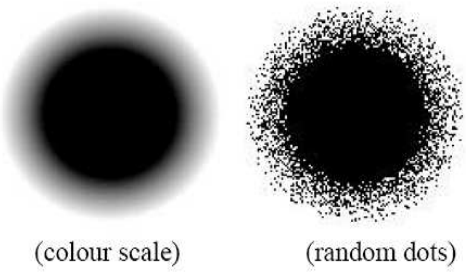

Figure 2.15: Visualization of fuzzy boundaries from Jiang ([35]). Figure reproduced with permission.

On the other hand, there are significant limitations to fuzzy logic models. In particular, they have no probability interpretation. Although they may be successful in communicating the existence of uncertainty, and even provide some quantification, there is no way to calculate the probability of an event taking place or putting limits on related estimates. For example, in the soil boundary case, there is no way to estimate the probability that a point is located in one type of soil or another, and there is no calculation for an upper and lower confidence bound on the total map area made up of a particular soil class. As Laviolette et al. ([38]) point out, fuzzy methods *describe* error, but do not *prescribe* a way to handle that error.

They note that while this type of vagueness may resemble human boundary perception, it is not acceptable between a contractor and a client. Additionally, the authors give several examples of fuzzy models that are not difficult to closely approximate by probability functions, in an attempt to demonstrate the flexibility in probability applications. It is primarily for these reasons that probability models are the best alternative to model uncertainty in cases where one hopes to do more than communicate and describe error.

2.4.4 Statistical models

Although fuzzy models are recently gaining popularity, some authors have developed statistical models to deal with error in thematic data. There are no all-encompassing statistical models for error in raster data. Oftentimes, the statistical alternative in GIS operations for continuous data is Monte Carlo simulation, in which error is randomly simulated from some appropriate distribution over multiple trials, incorporated into the recorded map, and the operation of interest is performed on the map variation that has been generated. Summary statistics are then computed from the resulting sample of product maps.

As an example, suppose we are converting a DEM to a viewshed map; we would choose an error distribution for the elevation data and randomly generate error for each cell. After incorporating the elevation error into the map, we then convert the new DEM to a viewshed. We repeat this process M times, and in the end we can compute, for example, the proportion of times for each cell that it was in the line of sight. Fisher ([21]) has successfully implemented such a simulation. As another example, Krivoruchko and Gotway-Crawford ([37]) suggest using Monte Carlo methods to perform a sensitivity analysis relating to various map operations such as buffering, overlay, union/intersection, and interpolation. A sensitivity analysis is, essentially, a way to measure how reliable the data resulting from a geoprocessing operation will be when faced with a specific size and type of error.

Many in the GIS community advocate Monte Carlo methods for the reasons that finding a continuous differentiable function to describe the error in certain data operations would be

very difficult, the results of a Monte Carlo analysis are easy to interpret (summary statistics and tests of significance), and it is a general and flexible method requiring few assumptions ([47]). Monte Carlo simulation is in fact a valuable tool, although problems occur when dealing with spatially correlated data. It can be very hard to accurately simulate this kind of data, and when the model becomes complicated and a map has many cells, simulations become very time- and resource-consuming and may not even be possible. (Fisher [22] in fact commented afterward that his viewshed simulation was too computationally intensive for widespread use.) Monte Carlo simulation is also difficult to apply to categorical data, and so a different method of statistically evaluating raster error should be found for general use.

Veregin has done some general work exploring the propagation of error through map overlay operations ([58]). In the case of continuous data, he has studied the propagation of error through various methods of overlay for two or more maps. For example, he calculates the covariance of two maps using the addition operation as $S_{ij} = \frac{1}{M}\sum_{m=1}^{M}(Z_{mi}-\hat{Z}_{mi})(Z_{mj}-\hat{Z}_{mj})$, where M = number of cells in a layer, $Z_{mi}-\hat{Z}_{mi}$= difference in actual versus estimated value for a cell (which is only known in theory), and i and j are the two map layers. (This operation can, of course, be extended to the situation where more than two maps are involved.) Veregin notes that, perhaps contrary to first thought, negative covariance actually implies the final map product may be more accurate than the individual maps. Veregin also examines the propagation of error in the categorical case through the use of a standard contingency table. As an example, for the AND operator, suppose $P(\bar{E}_i)$ is the proportion of cells in layer i that are correctly classified. The error for two maps combined with the AND operator is then $P(\bar{E}_c) = P(\bar{E}_1 \cap \bar{E}_2) = P(\bar{E}_1)P(\bar{E}_2|\bar{E}_1)$. Expanding this result to multiple maps, Veregin demonstrates that error rises exponentially with the number of maps.

Many authors have similarly worked with the classification error matrices already in existence to study map error, and have attempted to improve the contingency table concept. For example, Veregin ([59]) has done some work to increase the accuracy of the Proportion Correctly Classified (PCC) and other statistics from the error matrix for overlay operations.

Shi et al. ([53]) give an alternate table that is not based on the idea that a cell is either correctly or incorrectly classified, as in a traditional error classification matrix, but assumes instead that cells are not homogeneous. They take the approach that while a cell must be classified as a particular category for mapping purposes, its actual ground location may be made up of several categories. Their table gives a frequency distribution for each category on the map based on a detailed re-sampling of a small number of cells. The table may indicate, for example, the number of cells that are less than 10% composed of the indicated type, 10-20% composed of the indicated type, 20-30%, etc. The goal of the table is to communicate whether the cells on the map are largely made up of the category they are recorded as, or if the cells are fairly divided between classes. This is different from the common contingency table, which simply records whether or not the cells are erroneous.

Carmel and Dean ([8]) have created the Combined Location and Classification error matrix, or CLC. They account for locational and classification error separately, and then combine it into the CLC matrix. They optionally incorporate temporal maps as separate layers of a project. The basic assumptions of the CLC are that the error in location and classification are uncorrelated, and that error is uncorrelated over time. In testing the robustness of this model, they found it was essentially insensitive to spatial correlation in any error and was still an accurate summary. It did, however, react to moderate correlation in error between locational and classification error, as well as moderate correlation between time steps.

Czaplewski ([18]) discusses the issue of the accuracy of the contingency table itself. He notes that in reality it is very difficult or even impossible to take a nearly error-free sample to use in the contingency table. He focuses on creating contingency tables based on sampling methods that are not completely random–for example, stratified sampling, which is less expensive than randomized sampling. He also works with methods that take into account the error in the re-sampled map data used in the contingency table.

Several error models and discussions focus on the proportion of a cell made up of a particular class or probability of belonging to a class. This is different from a contingency table, which

generally assumes that cells are classified correctly or incorrectly, and does not attempt to assign any distributions to error. Hughes et al. ([31]) experimented with this idea for Landsat data (a satellite image) of a suburban neighborhood. They used polygon vector data and co-registered it with the satellite data so that he was able to attach the proportion values of each class to every cell through the associated database. Next, they used Monte Carlo simulation to perturb the pixels based on locational error in the map, and recalculated the proportions of each class in every pixel. They then used this information to calculate various statistics about each cell and provide multiple visualizations of the area based on those statistics.

Goodchild et al. ([25]) give a stochastic process error model for categorical data. The main focus of their paper is accurately generating spatially autocorrelated data for use in simulation studies. That is, they assume that the probability that a cell is from a particular category is related to the probability of the cells around it. Their simulation model therefore takes correlation with neighboring cells into account. They also make the distinction between a map in which all cells are autocorrelated and a map in which boundaries are the most variable pixels. The idea was to develop a model in which the marginal distributions of the variables in each cell is known, and spatial dependence is controlled. The model they use is $\boldsymbol{X} = \rho \boldsymbol{W} \boldsymbol{X} + \boldsymbol{\epsilon}$ where $\boldsymbol{X}$ is a vector of length N (total number of cells in the map) and entries are in $[0, 1]$; $\rho \in [0, .25]$ is a parameter determining the amount of spatial autocorrelation between cells; $\boldsymbol{W}_{uv} = 1$ if cells u and v share common edge, else 0; and $\boldsymbol{\epsilon} \sim N(0, 1)$. A quick calculation to solve for $\boldsymbol{X}$ gives $\boldsymbol{X} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{\epsilon}$. They allow for the possibility that ρ might vary regionally, to create a boundary. The method is advantageous in the sense that it is very general, and can be used in many applications. For example, this type of simulation was successfully used by Horttanainen and Virrantaus ([30]) for an analysis of the uncertainty in soil and military terrain.

Molsen et al. ([45]) use a Generalized Linear Mixed Model (GLMM) to examine the relationship between error in blackbrush cover-type and some topographical and heterogeneity components of a satellite vegetation map. They use the model $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = 0$ or 1,

with a logit link function $g(\mu) = log(\frac{\mu}{1-\mu}) = \boldsymbol{X\beta} + \boldsymbol{Z\nu}$ where $\boldsymbol{X}$ is a matrix of fixed effects and $\boldsymbol{Z}$ is a matrix of random effects; $E(\boldsymbol{\nu}) = 0$ and $cov(\boldsymbol{\nu}) = \boldsymbol{G}$ (unknown); and $E(\boldsymbol{\epsilon}|\boldsymbol{\mu}) = \boldsymbol{R}_\mu^{1/2}\boldsymbol{R}\boldsymbol{R}_\mu^{1/2}$ where $\boldsymbol{R}_\mu$ is a diagonal matrix containing evaluations at μ of the variance function $V(\mu) = \mu(1-\mu)$. The authors try several covariance structures for $\boldsymbol{G}$ and used Akaike's Information Criterion (AIC) and Schwartz's Baysian criterion to choose the most appropriate model. The end result is a probability of error in each cell based on variables in the model, such as the ground slope at the cell and whether it appears near a road.

Finally, Bayesian belief networks are another general tool gaining some popularity in GIS error models. This is actually a term that encompasses a range of different, but related, techniques for dealing with uncertainty ([37]). Three basic stages are involved in the creation of a Bayesian belief network. The first stage involves creating a graphical model with expert knowledge indicating potential dependencies among the rasterized values. Next, an uncertainty model specifying probabilities of various situations needs to be addressed. Finally, the relationships between the potential dependencies and prior probabilities in the first two steps need to be established to find a joint distribution over all variables. The data can then be introduced to find the posterior probability of events of interest. This is not necessarily an easy task, nor is it objective, but it is useful for computationally fast error assessments and situations in which data are scarce (an advantage to many Bayesian applications).

2.5 Discussion

I have discussed the current work and error models available for both vector and raster data. In each case, both probabilistic and non-probabilistic models have gained some popularity in the field. One particularly popular alternative is the fuzzy model, which describes points or cells as belonging to multiple locations or classes simultaneously, to some varying degree. Fuzzy models are not preferable, however, because they do no truly lay down methods

for dealing with error or giving probabilities that an event will occur, and therefore this is unacceptable in a field largely based on making decisions regarding use of resources, monetary losses and gains, and even human safety in some applications.

As far as vector data is concerned, there is already a solid, unified, probabilistic model for location error. This model is the bivariate normal model for error in points, as developed by Wenzhong Shi and his colleagues. There are some issues that still need to be addressed, however, particularly the direct calculation of probabilities and a way to model the case of omitted vertices.

In contrast to this situation, there is no universal and outstanding statistical model for error in raster data. The work here is more disjoint; for example, models for error propagation and for boundary error tend to be separate in the literature. Current models are therefore very application-dependent, as well as reliant on data simulation. There is a lot of work still to be done in this field, in terms of unifying error models and even dealing appropriately with such things as spatial autocorrelation in data.

I should point out, before leaving the literature review, that my research is not entirely comprehensive. There are two ways of looking at error in data. The way I have chosen to consider GIS data is to separate vector data from raster data. Another way to handle this, however, is to look at data in terms of location data and thematic data. The two concepts are certainly related–vector data is often synonymous with location data and raster associated with thematic data. There are exceptions, however. For example, many soil classification and land cover maps are displayed in a vector format, where polygons are used to represent the boundaries between classes. There can also be location error in raster data. Cells can be misclassified not only through improper interpretation or measurement, but also by misplaced measurements. Additionally, there are some applications that involve vector and raster data fusion which I have chosen not to consider.

Given the many forms of GIS information and the error that can occur, though, I feel I have adequately covered the wide range of literature on GIS error. In the remaining chapters, I

will introduce my own additions and ideas for modeling error in GIS data.

Chapter 3

Bayesian Methods for Vector Data

3.1 Introduction

I propose here to introduce an empirically-based Bayesian method to analyze the error in vector data. This line segment model is very similar to the G-band error model, with the addition of a prior distribution for individual points. This is certainly a feasible idea, since users can draw information for a prior distribution from many sources, including previous vector maps, other types of maps, and expert or historical knowledge about the location of particular points. The Bayesian methodology even has an advantage in the sense that we can still make inferences when there is only a small number of observations, which is a common situation when dealing with GIS maps.

3.2 Bayesian Methods in Vector Data

3.2.1 Bayesian error models for points

We begin with a point $\boldsymbol{z}_0 = (x_0, y_0)'$. We assume the true location of this point is $\boldsymbol{\mu}_0 = (\mu_{x_0}, \mu_{y_0})'$, and that we have n random observations of this value, $\boldsymbol{z}_{i0} = (x_{i0}, y_{i0})'$, $\overline{\boldsymbol{z}}_0 = (\overline{x}_0, \overline{y}_0)'$, $i = 1, \ldots, n$. We will assume that $\boldsymbol{\mu}_0$ has a bivariate normal prior distribution with parameters $(\boldsymbol{\mu}_{0_0}, \boldsymbol{\Lambda}_0)$, where $\boldsymbol{\Lambda}_0$ is the variance-covariance matrix of the point in our prior distribution, and that the data points $\boldsymbol{z}_{i0}$ have a normal distribution with parameters $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$,

$$\boldsymbol{\mu}_{0_0} = \begin{pmatrix} \mu_{x_{00}} \\ \mu_{y_{00}} \end{pmatrix}, \boldsymbol{\Lambda}_0 = \begin{pmatrix} \tau^2_{\mu_{x_0}} & \tau_{\mu_{x_0}\mu_{y_0}} \\ \tau_{\mu_{y_0}\mu_{x_0}} & \tau^2_{\mu_{y_0}} \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma^2_{x_0} & \sigma_{x_0 y_0} \\ \sigma_{y_0 x_0} & \sigma^2_{y_0} \end{pmatrix}.$$

Note that I have used the Greek letter τ to represent the variance terms in the prior distribution, rather than the usual symbol σ, since I have used this to represent the data variance. Also I denote the data through the matrix $\boldsymbol{Z}_0 = (\boldsymbol{z}_{10}, \boldsymbol{z}_{20}, \cdots, \boldsymbol{z}_{n0})$. We have the following results.

Theorem 3.1. Under the above assumption, the posterior distribution of a point $\boldsymbol{\mu}_0$ is

$$\boldsymbol{\mu}_0 | \boldsymbol{Z}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{g}_0, \boldsymbol{H}_0),$$

where

$$\boldsymbol{g}_0 = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_{0_0} + n\boldsymbol{\Sigma}_0^{-1}\overline{\boldsymbol{z}}_0), \text{ and } \boldsymbol{H}_0 = (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1}.$$

Proof. This is an immediate result of equation (2.1), and appears in Gelman et al. ([23]). □

Corollary 3.1. The equation for the $100(1-\alpha)\%$ confidence ellipse for a point $\boldsymbol{\mu}_0$ is

$$(\boldsymbol{\mu}_0 - \boldsymbol{g}_0)'\boldsymbol{H}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{g}_0) \leq \chi^2_{2,1-\alpha},$$

where $\boldsymbol{\mu}_0$ is the set of points in the ellipse, $\boldsymbol{g}_0$ is the Bayesian posterior mean for $\boldsymbol{\mu}_0$, and $\boldsymbol{H}_0$ is the Bayesian posterior variance matrix.

Proof. Using the mean and variance of the posterior distribution for $\boldsymbol{\mu}_0$, and the fact that this distribution is bivariate normal, the corollary follows immediately. $\square$

3.2.2 Bayesian error model for line segments and polygons

The basic concept of the Bayesian error model for line segments is also very similar to the frequentist approach. Recall that any point on a line segment can be described as a function of its endpoints. Suppose then that the endpoints of a particular line segment are $\boldsymbol{z}_0$ and $\boldsymbol{z}_1$, with coordinates $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. We can then describe any point on the line with the equation

$$\boldsymbol{\mu}_\gamma = (1-\gamma)\boldsymbol{\mu}_0 + \gamma\boldsymbol{\mu}_1,$$

where $0 \leq \gamma \leq 1$.

Suppose each of these coordinates have a bivariate normal prior distribution with parameters $(\boldsymbol{\mu}_{0_0}, \boldsymbol{\Lambda}_0)$ and $(\boldsymbol{\mu}_{1_0}, \boldsymbol{\Lambda}_1)$ respectively. Assume further that these two points may be correlated, so that the joint prior distribution of the endpoints is

$$\boldsymbol{\mu}_{01} \sim N\left(\boldsymbol{\mu}_{01_0}, \boldsymbol{\Lambda}_{01}\right),$$

where

$$\boldsymbol{\mu}_{01} = \begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix}, \boldsymbol{\mu}_{01_0} = \begin{pmatrix} \mu_{x_{00}} \\ \mu_{y_{00}} \\ \mu_{x_{10}} \\ \mu_{y_{10}} \end{pmatrix}, \text{ and } \boldsymbol{\Lambda}_{01} = \begin{pmatrix} \tau^2_{\mu_{x_0}} & \tau_{\mu_{x_0}\mu_{y_0}} & \tau_{\mu_{x_0}\mu_{x_1}} & \tau_{\mu_{x_0}\mu_{y_1}} \\ \tau_{\mu_{y_0}\mu_{x_0}} & \tau^2_{\mu_{y_0}} & \tau_{\mu_{y_0}\mu_{x_1}} & \tau_{\mu_{y_0}\mu_{y_1}} \\ \tau_{\mu_{x_1}\mu_{x_0}} & \tau_{\mu_{x_1}\mu_{y_0}} & \tau^2_{\mu_{x_1}} & \tau_{\mu_{x_1}\mu_{y_1}} \\ \tau_{\mu_{y_1}\mu_{x_0}} & \tau_{\mu_{y_1}\mu_{y_0}} & \tau_{\mu_{y_1}\mu_{x_1}} & \tau^2_{\mu_{y_1}} \end{pmatrix}.$$

Assume we also have n independent observations on each of these endpoints, $\boldsymbol{z}_{i01} \sim N(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01})$ where

$$\boldsymbol{z}_{i01} = \begin{pmatrix} x_{i0} \\ y_{i0} \\ x_{i1} \\ y_{i1} \end{pmatrix}, \overline{\boldsymbol{z}}_{01} = \begin{pmatrix} \overline{x}_0 \\ \overline{y}_0 \\ \overline{x}_1 \\ \overline{y}_1 \end{pmatrix}, \boldsymbol{\mu}_{01} = \begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_{01} = \begin{pmatrix} \sigma^2_{x_0} & \sigma_{x_0y_0} & \sigma_{x_0x_1} & \sigma_{x_0y_1} \\ \sigma_{y_0x_0} & \sigma^2_{y_0} & \sigma_{y_0x_1} & \sigma_{y_0y_1} \\ \sigma_{x_1x_0} & \sigma_{x_1y_0} & \sigma^2_{x_1} & \sigma_{x_1y_1} \\ \sigma_{y_1x_0} & \sigma_{y_1y_0} & \sigma_{y_1x_1} & \sigma^2_{y_1} \end{pmatrix}.$$

According to equation (2.1), the joint conditional posterior distribution for the endpoints is $N(\boldsymbol{g}_{01}, \boldsymbol{H}_{01})$, where

$$\boldsymbol{g}_{01} = (\boldsymbol{\Lambda}_{01}^{-1} + n\boldsymbol{\Sigma}_{01}^{-1})^{-1}(\boldsymbol{\Lambda}_{01}^{-1}\boldsymbol{\mu}_{01_0} + n\boldsymbol{\Sigma}_{01}^{-1}\overline{\boldsymbol{z}}_{01}), \boldsymbol{H}_{01} = (\boldsymbol{\Lambda}_{01}^{-1} + n\boldsymbol{\Sigma}_{01}^{-1})^{-1}.$$

For clarity of notation, I will indicate the individual elements of the posterior mean and variance as follows:

$$\boldsymbol{g}_{01} = \begin{pmatrix} \boldsymbol{g}_0 \\ \boldsymbol{g}_1 \end{pmatrix} = \begin{pmatrix} g_{x_0} \\ g_{y_0} \\ g_{x_1} \\ g_{y_1} \end{pmatrix},$$

$$\boldsymbol{H}_{01} = \begin{pmatrix} h_{x_0}^2 & h_{x_0y_0} & h_{x_0x_1} & h_{x_0y_1} \\ h_{y_0x_0} & h_{y_0}^2 & h_{y_0x_1} & h_{y_0y_1} \\ h_{x_1x_0} & h_{x_1y_0} & h_{x_1}^2 & h_{x_1y_1} \\ h_{y_1x_0} & h_{y_1y_0} & h_{y_1x_1} & h_{y_1}^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{H}_0 & \begin{matrix} h_{x_0x_1} & h_{x_0y_1} \\ h_{y_0x_1} & h_{y_0y_1} \end{matrix} \\ \begin{matrix} h_{x_1x_0} & h_{x_1y_0} \\ h_{y_1x_0} & h_{y_1y_0} \end{matrix} & \boldsymbol{H}_1 \end{pmatrix}.$$

Theorem 3.2. Using the above notation, the posterior conditional distribution of a point on a line segment, $\boldsymbol{\mu}_\gamma = \gamma\boldsymbol{\mu}_1 + (1-\gamma)\boldsymbol{\mu_0}$, is

$$\boldsymbol{\mu}_\gamma | \boldsymbol{\Sigma}_{01}, \boldsymbol{Z}_{01} \sim N(\boldsymbol{g}_\gamma, \boldsymbol{H}_\gamma)$$

where

$$\begin{aligned} &\boldsymbol{g}_\gamma = (1-\gamma)\boldsymbol{g}_0 + \gamma\boldsymbol{g}_1, \ \boldsymbol{H}_\gamma = \begin{pmatrix} h_{x_\gamma}^2 & h_{x_\gamma y_\gamma} \\ h_{y_\gamma x_\gamma} & h_{y_\gamma} \end{pmatrix}, \\ &h_{x_\gamma}^2 = (1-\gamma)^2 h_{x_0}^2 + 2\gamma(1-\gamma)h_{x_0x_1} + \gamma^2 h_{x_1}^2, \\ &h_{y_\gamma}^2 = (1-\gamma)^2 h_{y_0}^2 + 2\gamma(1-\gamma)h_{y_0y_1} + \gamma^2 h_{y_1}^2, \\ \text{and} \quad &h_{x_\gamma y_\gamma} = h_{y_\gamma x_\gamma} = (1-\gamma)^2 h_{x_0y_0} + \gamma(1-\gamma)(h_{x_0y_1} + h_{y_0x_1}) + \gamma^2 h_{x_1y_1}. \end{aligned}$$

Proof. Following some basic facts from linear models, the function $\gamma\boldsymbol{\mu}_1 + (1-\gamma)\boldsymbol{\mu}_0$ can be

written as

$$\begin{pmatrix} (1-\gamma) & 0 & \gamma & 0 \\ 0 & (1-\gamma) & 0 & \gamma \end{pmatrix} \begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix}.$$

Since $\boldsymbol{\mu}_{01}$ is normally distributed, we know that this linear function of $\boldsymbol{\mu}_{01}$ is normally distributed. The mean of this distribution is

$$\begin{pmatrix} (1-\gamma) & 0 & \gamma & 0 \\ 0 & (1-\gamma) & 0 & \gamma \end{pmatrix} \begin{pmatrix} g_{x_0} \\ g_{y_0} \\ g_{x_1} \\ g_{y_1} \end{pmatrix} = (1-\gamma)\boldsymbol{g}_0 + \gamma \boldsymbol{g}_1.$$

The variance of this distribution is

$$\begin{pmatrix} (1-\gamma) & 0 & \gamma & 0 \\ 0 & (1-\gamma) & 0 & \gamma \end{pmatrix} \begin{pmatrix} h^2_{x_0} & h_{x_0y_0} & h_{x_0x_1} & h_{x_0y_1} \\ h_{y_0x_0} & h^2_{y_0} & h_{y_0x_1} & h_{y_0y_1} \\ h_{x_1x_0} & h_{x_1y_0} & h^2_{x_1} & h_{x_1y_1} \\ h_{y_1x_0} & h_{y_1y_0} & h_{y_1x_1} & h^2_{y_1} \end{pmatrix} \begin{pmatrix} (1-\gamma) & 0 \\ 0 & (1-\gamma) \\ \gamma & 0 \\ 0 & \gamma \end{pmatrix}$$

$$= \begin{pmatrix} h^2_{x_\gamma} & h_{x_\gamma y_\gamma} \\ h_{y_\gamma x_\gamma} & h_{y_\gamma} \end{pmatrix}$$

where the individual terms are as written in the theorem. □

Note that the result in Theorem 3.2 is very similar to the result in the frequentist case ([54]).

Corollary 3.2. The equation for the $100(1-\alpha)\%$ confidence ellipse for a point $\boldsymbol{\mu}_\gamma$ is

$$(\boldsymbol{\mu}_\gamma - \boldsymbol{g}_\gamma)' \boldsymbol{H}_\gamma^{-1} (\boldsymbol{\mu}_\gamma - \boldsymbol{g}_\gamma) \leq \chi^2_{1-\alpha}$$

where $\boldsymbol{\mu}_\gamma$ is the set of points in the ellipse, $\boldsymbol{g}_\gamma$ is the Bayesian Posterior mean for $\boldsymbol{\mu}_\gamma$, and $\boldsymbol{H}_\gamma$ is the Bayesian posterior variance matrix.

Proof. Using the mean and variance of the posterior distribution for $\boldsymbol{\mu}_\gamma$, and the fact that this distribution is bivariate normal, I reference equation (2.3). The corollary follows immediately. □

3.2.3 Form of the posterior mean and variance for a point on a line segment

Theorem 3.2 states the general form of the posterior mean and variance of a point on a line segment. In some cases, however, the form of the posterior distribution can be greatly simplified.

Suppose we want to find an explicit form for the posterior mean and variance of a point on a line, $\boldsymbol{\mu}_\gamma$. In the most general case, that is for correlated x- and y- data with

$$\boldsymbol{\Lambda}_{01} = \begin{pmatrix} \tau^2_{\mu_{x_0}} & \tau_{\mu_{x_0}\mu_{y_0}} & \tau_{\mu_{x_0}\mu_{x_1}} & \tau_{\mu_{x_0}\mu_{y_1}} \\ \tau_{\mu_{y_0}\mu_{x_0}} & \tau^2_{\mu_{y_0}} & \tau_{\mu_{y_0}\mu_{x_1}} & \tau_{\mu_{y_0}\mu_{y_1}} \\ \tau_{\mu_{x_1}\mu_{x_0}} & \tau_{\mu_{x_1}\mu_{y_0}} & \tau^2_{\mu_{x_1}} & \tau_{\mu_{x_1}\mu_{y_1}} \\ \tau_{\mu_{y_1}\mu_{x_0}} & \tau_{\mu_{y_1}\mu_{y_0}} & \tau_{\mu_{y_1}\mu_{x_1}} & \tau^2_{\mu_{y_1}} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{01} = \begin{pmatrix} \sigma^2_{x_0} & \sigma_{x_0y_0} & \sigma_{x_0x_1} & \sigma_{x_0y_1} \\ \sigma_{y_0x_0} & \sigma^2_{y_0} & \sigma_{y_0x_1} & \sigma_{y_0y_1} \\ \sigma_{x_1x_0} & \sigma_{x_1y_0} & \sigma^2_{x_1} & \sigma_{x_1y_1} \\ \sigma_{y_1x_0} & \sigma_{y_1y_0} & \sigma_{y_1x_1} & \sigma^2_{y_1} \end{pmatrix},$$

we can only substitute terms directly into the equations from Theorem 3.2 to calculate the posterior mean and variance. In many cases, however, it is possible to simplify the results, and it is advantageous to do so. I present these cases as corollaries to the result in Theorem 3.2.

Corollary 3.3. Suppose the endpoints of a line segment, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, are independent from one another, and the variance/covariance matrices of the prior and sampling distributions are

$$\boldsymbol{\Lambda}_{01} = \begin{pmatrix} \tau^2_{\mu_{x_0}} & \tau_{\mu_{x_0}\mu_{y_0}} & 0 & 0 \\ \tau_{\mu_{y_0}\mu_{x_0}} & \tau^2_{\mu_{y_0}} & 0 & 0 \\ 0 & 0 & \tau^2_{\mu_{x_1}} & \tau_{\mu_{x_1}\mu_{y_1}} \\ 0 & 0 & \tau_{\mu_{y_1}\mu_{x_1}} & \tau^2_{\mu_{y_1}} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{01} = \begin{pmatrix} \sigma^2_{x_0} & \sigma_{x_0y_0} & 0 & 0 \\ \sigma_{y_0x_0} & \sigma^2_{y_0} & 0 & 0 \\ 0 & 0 & \sigma^2_{x_1} & \sigma_{x_1y_1} \\ 0 & 0 & \sigma_{y_1x_1} & \sigma^2_{y_1} \end{pmatrix}.$$

The mean of the posterior distribution of $\boldsymbol{\mu}_\gamma$ is then

$$\boldsymbol{g}_\gamma = \gamma(\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1}(\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\mu}_{1_0} + n\boldsymbol{\Sigma}_1^{-1}\overline{\boldsymbol{z}}_1) + (1-\gamma)(\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_{0_0} + n\boldsymbol{\Sigma}_0^{-1}\overline{\boldsymbol{z}}_0),$$

and the posterior variance is $\boldsymbol{H}_\gamma = \gamma^2\boldsymbol{H}_1 + (1-\gamma)^2\boldsymbol{H}_0$.

Proof. First, note that we can write $\boldsymbol{\Lambda}_{01} = \begin{pmatrix} \boldsymbol{\Lambda}_0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_1 \end{pmatrix}$ and $\boldsymbol{\Sigma}_{01} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{1}} \end{pmatrix}$. Linear models results tell us

$$\boldsymbol{\Lambda}_{01}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_1^{-1} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{01}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{1}}^{-1} \end{pmatrix}.$$

Next, the formula for the posterior mean from equation (2.1) tells us

$$\boldsymbol{g}_{01} = (\boldsymbol{\Lambda}_{01}^{-1} + n\boldsymbol{\Sigma}_{01}^{-1})^{-1}(\boldsymbol{\Lambda}_{01}^{-1}\boldsymbol{\mu}_{01_0} + n\boldsymbol{\Sigma}_{01}^{-1}\overline{\boldsymbol{z}}_{01}).$$

By writing the equation explicitly in terms of our assumed values for $\boldsymbol{\Lambda}_{01}$ and $\boldsymbol{\Sigma}_{01}$ and applying some linear algebra, we find

$$\begin{aligned}
\boldsymbol{g}_{01} &= \left[\begin{pmatrix} \boldsymbol{\Lambda}_0 & 0 \\ 0 & \boldsymbol{\Lambda}_1 \end{pmatrix}^{-1} + n \begin{pmatrix} \boldsymbol{\Sigma}_0 & 0 \\ 0 & \boldsymbol{\Sigma}_1 \end{pmatrix}^{-1} \right]^{-1} \\
&\left[\begin{pmatrix} \boldsymbol{\Lambda}_0 & 0 \\ 0 & \boldsymbol{\Lambda}_1 \end{pmatrix}^{-1} \boldsymbol{\mu}_{01_0} + n \begin{pmatrix} \boldsymbol{\Sigma}_0 & 0 \\ 0 & \boldsymbol{\Sigma}_1 \end{pmatrix}^{-1} \overline{\boldsymbol{z}}_{01} \right] \\
&= \left[\begin{pmatrix} \boldsymbol{\Lambda}_0^{-1} & 0 \\ 0 & \boldsymbol{\Lambda}_1^{-1} \end{pmatrix} + n \begin{pmatrix} \boldsymbol{\Sigma}_0^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_1^{-1} \end{pmatrix} \right]^{-1} \\
&\left[\begin{pmatrix} \boldsymbol{\Lambda}_0^{-1} & 0 \\ 0 & \boldsymbol{\Lambda}_1^{-1} \end{pmatrix} \boldsymbol{\mu}_{01_0} + n \begin{pmatrix} \boldsymbol{\Sigma}_0^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_1^{-1} \end{pmatrix} \overline{\boldsymbol{z}}_{01} \right] \\
&= \begin{pmatrix} \boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1} & 0 \\ 0 & \boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1} \end{pmatrix}^{-1} \left[\begin{pmatrix} \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_{0_0} \\ \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\mu}_{1_0} \end{pmatrix} + n \begin{pmatrix} \boldsymbol{\Sigma}_0^{-1}\overline{\boldsymbol{z}}_0 \\ \boldsymbol{\Sigma}_1^{-1}\overline{\boldsymbol{z}}_1 \end{pmatrix} \right] \\
&= \begin{pmatrix} (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1} & 0 \\ 0 & (\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_{0_0} + n\boldsymbol{\Sigma}_0^{-1}\overline{\boldsymbol{z}}_0 \\ \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\mu}_{1_0} + n\boldsymbol{\Sigma}_1^{-1}\overline{\boldsymbol{z}}_1 \end{pmatrix} \\
&= \begin{pmatrix} (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_{0_0} + n\boldsymbol{\Sigma}_0^{-1}\overline{\boldsymbol{z}}_0) \\ (\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1}(\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\mu}_{1_0} + n\boldsymbol{\Sigma}_1^{-1}\overline{\boldsymbol{z}}_1) \end{pmatrix}.
\end{aligned}$$

By applying the results of Theorem 3.2, $\boldsymbol{g}_\gamma = (1-\gamma)\boldsymbol{g}_0 + \gamma\boldsymbol{g}_1$, we arrive at the stated conclusion.

Next, to get the result for the posterior variance, we again use equation (2.1) to find $\boldsymbol{H}_{01} = (\boldsymbol{\Lambda}_{01}^{-1} + n\boldsymbol{\Sigma}_{01}^{-1})^{-1}$. We can then write

$$\boldsymbol{H}_{01} = \begin{pmatrix} \boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1} \end{pmatrix}$$
$$= \begin{pmatrix} \boldsymbol{H}_0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_1 \end{pmatrix}$$

where $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ are defined as in Theorem 3.1.

Our posterior variance is then

$$\begin{pmatrix} (1-\gamma) & 0 & \gamma & 0 \\ 0 & (1-\gamma) & 0 & \gamma \end{pmatrix} \begin{pmatrix} h_{x_0}^2 & h_{x_0y_0} & 0 & 0 \\ h_{y_0x_0} & h_{y_0}^2 & 0 & 0 \\ 0 & 0 & h_{x_1}^2 & h_{x_1y_1} \\ 0 & 0 & h_{y_1x_1} & h_{y_1}^2 \end{pmatrix} \begin{pmatrix} (1-\gamma) & 0 \\ 0 & (1-\gamma) \\ \gamma & 0 \\ 0 & \gamma \end{pmatrix},$$

from the proof of Theorem 3.2.

Multiplying through this equation gives us

$$\begin{pmatrix} (1-\gamma)^2 h_{x_0}^2 + \gamma^2 h^2 x_1 & (1-\gamma)^2 h_{x_0y_0} + \gamma^2 h_{x_1y_1} \\ (1-\gamma)^2 h_{y_0x_0} + \gamma^2 h_{y_1x_1} & (1-\gamma)^2 h_{y_0}^2 + \gamma^2 h^2 y_1 \end{pmatrix} = \gamma^2 \boldsymbol{H}_1 + (1-\gamma)^2 \boldsymbol{H}_0,$$

which is the result stated in the corollary. □

Corollary 3.3 provides the posterior mean and variance for a general situation that allows the x- and y- coordinates within each endpoint to be correlated, without correlation between endpoints. This may or may not be a valid assumption. In the case of GPS instrument error, for example, an instrument may be more likely to always read a high x- coordinate when the satellites calibrating the instrument are in a certain location, meaning endpoints will have correlated error if they are taken one after another at the same time. If points were taken at different times, however, or the satellites were well placed, there may be no such correlation. Additionally, in the case of manual map digitization, a well-trained technician

may not demonstrate any trend in error between endpoints. The following corollaries provide details for the subset of situations in which there is no correlation between or within the endpoints.

Corollary 3.4. Suppose now that there is no correlation between the endpoints $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, and additionally, there is no correlation between the x- and y- coordinates at each endpoint. That is,

$$\boldsymbol{\Lambda}_{01} = \begin{pmatrix} \tau^2_{\mu_{x_0}} & 0 & 0 & 0 \\ 0 & \tau^2_{\mu_{y_0}} & 0 & 0 \\ 0 & 0 & \tau^2_{\mu_{x_1}} & 0 \\ 0 & 0 & 0 & \tau^2_{\mu_{y_1}} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{01} = \begin{pmatrix} \sigma^2_{x_0} & 0 & 0 & 0 \\ 0 & \sigma^2_{y_0} & 0 & 0 \\ 0 & 0 & \sigma^2_{x_1} & 0 \\ 0 & 0 & 0 & \sigma^2_{y_1} \end{pmatrix}.$$

The mean of the posterior distribution of $\boldsymbol{\mu}_\gamma$ is

$$\gamma \left[\begin{pmatrix} \frac{\sigma^2_{x_1}}{\sigma^2_{x_1}+n\tau^2_{\mu_{x_1}}} & 0 \\ 0 & \frac{\sigma^2_{y_1}}{\sigma^2_{y_1}+n\tau^2_{\mu_{y_1}}} \end{pmatrix} \boldsymbol{g}_1 + \begin{pmatrix} \frac{\tau^2_{x_1 0}}{\sigma^2_{x_1}+n\tau^2_{\mu_{x_1}}} & 0 \\ 0 & \frac{\sigma^2_{y_1}}{\sigma^2_{y_1}+n\tau^2_{\mu_{y_1}}} \end{pmatrix} \overline{\boldsymbol{z}}_1 \right]$$
$$+(1-\gamma) \left[\begin{pmatrix} \frac{\sigma^2_{x_0}}{\sigma^2_{x_0}+n\tau^2_{\mu_{x_0}}} & 0 \\ 0 & \frac{\sigma^2_{y_0}}{\sigma^2_{y_0}+n\tau^2_{\mu_{y_0}}} \end{pmatrix} \boldsymbol{g}_0 + \begin{pmatrix} \frac{\tau^2_{\mu_{x_0}}}{\sigma^2_{x_0}+n\tau^2_{\mu_{x_0}}} & 0 \\ 0 & \frac{\sigma^2_{y_0}}{\sigma^2_{y_0}+n\tau^2_{\mu_{y_0}}} \end{pmatrix} \overline{\boldsymbol{z}}_0 \right]$$

and the variance of the posterior distribution is

$$\gamma^2 \begin{pmatrix} \frac{\tau^2_{\mu_{x_1}}\sigma^2_{x_1}}{\sigma^2_{x_1}+n\tau^2_{\mu_{x_1}}} & 0 \\ 0 & \frac{\tau^2_{\mu_{y_1}}\sigma^2_{y_1}}{\sigma^2_{y_1}+n\tau^2_{\mu_{y_1}}} \end{pmatrix} + (1-\gamma)^2 \begin{pmatrix} \frac{\tau^2_{\mu_{x_0}}\sigma^2_{x_0}}{\sigma^2_{x_0}+n\tau^2_{\mu_{x_0}}} & 0 \\ 0 & \frac{\tau^2_{\mu_{y_0}}\sigma^2_{y_0}}{\sigma^2_{y_0}+n\tau^2_{\mu_{y_0}}} \end{pmatrix}.$$

Proof. Because this is a special case of Corollary 3.3, we know that the mean of the posterior distribution is

$$\boldsymbol{g}_\gamma = \gamma(\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1}(\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\mu}_{1_0} + n\boldsymbol{\Sigma}_1^{-1}\overline{\boldsymbol{z}}_1) + (1-\gamma)(\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_{0_0} + n\boldsymbol{\Sigma}_0^{-1}\overline{\boldsymbol{z}}_0),$$

and the posterior variance is $\boldsymbol{H}_\gamma = \gamma^2\boldsymbol{H}_1 + (1-\gamma)^2\boldsymbol{H}_0$. Imposing the additional condition that the x- and y- coordinates at each endpoint are uncorrelated, we know

$$\boldsymbol{\Lambda}_0^{-1} = \begin{pmatrix} \frac{1}{\tau^2_{\mu_{x_0}}} & 0 \\ 0 & \frac{1}{\tau^2_{\mu_{y_0}}} \end{pmatrix}, \ \boldsymbol{\Lambda}_1^{-1} = \begin{pmatrix} \frac{1}{\tau^2_{\mu_{x_1}}} & 0 \\ 0 & \frac{1}{\tau^2_{\mu_{y_1}}} \end{pmatrix}, \ \boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} \frac{1}{\sigma^2_{x_0}} & 0 \\ 0 & \frac{1}{\sigma^2_{y_0}} \end{pmatrix},$$

and

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} \frac{1}{\sigma_{x_1}^2} & 0 \\ 0 & \frac{1}{\sigma_{y_1}^2} \end{pmatrix}.$$

Inserting these results into our previous equations, we get

$$\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} \frac{1}{\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{1}{\tau_{\mu_{y_0}}^2} \end{pmatrix} + \begin{pmatrix} \frac{n}{\sigma_{x_0}^2} & 0 \\ 0 & \frac{n}{\sigma_{y_0}^2} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_{x_0}^2 + n\tau_{\mu_{x_0}}^2}{\tau_{\mu_{x_0}}^2 \sigma_{x_0}^2} & 0 \\ 0 & \frac{\sigma_{y_0}^2 + n\tau_{\mu_{y_0}}^2}{\tau_{\mu_{y_0}}^2 \sigma_{y_0}^2} \end{pmatrix}$$

$$\Rightarrow \left(\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1}\right)^{-1} = \begin{pmatrix} \frac{\tau_{\mu_{x_0}}^2 \sigma_{x_0}^2}{\sigma_{x_0}^2 + n\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{\tau_{\mu_{y_0}}^2 \sigma_{y_0}^2}{\sigma_{y_0}^2 + n\tau_{\mu_{y_0}}^2} \end{pmatrix}.$$

Calculation of $\left(\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1}\right)^{-1}$ is similar.

We can now calculate the posterior mean and variance for the general no-covariance case.

$$\begin{aligned}
\boldsymbol{g}_\gamma &= \gamma \begin{pmatrix} \frac{\tau_{\mu_{x_1}}^2 \sigma_{x_1}^2}{\sigma_{x_1}^2 + n\tau_{\mu_{x_1}}^2} & 0 \\ 0 & \frac{\tau_{\mu_{y_1}}^2 \sigma_{y_1}^2}{\sigma_{y_1}^2 + n\tau_{\mu_{y_1}}^2} \end{pmatrix} \left[\begin{pmatrix} \frac{1}{\tau_{\mu_{x_1}}^2} & 0 \\ 0 & \frac{1}{\tau_{\mu_{y_1}}^2} \end{pmatrix} \boldsymbol{\mu}_{1_0} + \begin{pmatrix} \frac{1}{\sigma_{x_1}^2} & 0 \\ 0 & \frac{1}{\sigma_{y_1}^2} \end{pmatrix} \overline{\boldsymbol{z}}_1 \right] \\
&\quad + (1-\gamma) \begin{pmatrix} \frac{\tau_{\mu_{x_0}}^2 \sigma_{x_0}^2}{\sigma_{x_0}^2 + n\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{\tau_{\mu_{y_0}}^2 \sigma_{y_0}^2}{\sigma_{y_0}^2 + n\tau_{\mu_{y_0}}^2} \end{pmatrix} \left[\begin{pmatrix} \frac{1}{\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{1}{\tau_{\mu_{y_0}}^2} \end{pmatrix} \boldsymbol{\mu}_{0_0} + \begin{pmatrix} \frac{1}{\sigma_{x_0}^2} & 0 \\ 0 & \frac{1}{\sigma_{y_0}^2} \end{pmatrix} \overline{\boldsymbol{z}}_0 \right] \\
&= \gamma \left[\begin{pmatrix} \frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + n\tau_{\mu_{x_1}}^2} & 0 \\ 0 & \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + n\tau_{\mu_{y_1}}^2} \end{pmatrix} \boldsymbol{\mu}_{1_0} + \begin{pmatrix} \frac{\tau_{\mu_{x_1}}^2}{\sigma_{x_1}^2 + n\tau_{\mu_{x_1}}^2} & 0 \\ 0 & \frac{\sigma_{y_1}^2}{\sigma_{y_1}^2 + n\tau_{\mu_{y_1}}^2} \end{pmatrix} \overline{\boldsymbol{z}}_1 \right] \\
&\quad + (1-\gamma) \left[\begin{pmatrix} \frac{\sigma_{x_0}^2}{\sigma_{x_0}^2 + n\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{\sigma_{y_0}^2}{\sigma_{y_0}^2 + n\tau_{\mu_{y_0}}^2} \end{pmatrix} \boldsymbol{\mu}_{0_0} + \begin{pmatrix} \frac{\tau_{\mu_{x_0}}^2}{\sigma_{x_0}^2 + n\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{\sigma_{y_0}^2}{\sigma_{y_0}^2 + n\tau_{\mu_{y_0}}^2} \end{pmatrix} \overline{\boldsymbol{z}}_0 \right], \\
\boldsymbol{H}_\gamma &= \gamma^2 \begin{pmatrix} \frac{\tau_{\mu_{x_1}}^2 \sigma_{x_1}^2}{\sigma_{x_1}^2 + n\tau_{\mu_{x_1}}^2} & 0 \\ 0 & \frac{\tau_{\mu_{y_1}}^2 \sigma_{y_1}^2}{\sigma_{y_1}^2 + n\tau_{\mu_{y_1}}^2} \end{pmatrix} + (1-\gamma)^2 \begin{pmatrix} \frac{\tau_{\mu_{x_0}}^2 \sigma_{x_0}^2}{\sigma_{x_0}^2 + n\tau_{\mu_{x_0}}^2} & 0 \\ 0 & \frac{\tau_{\mu_{y_0}}^2 \sigma_{y_0}^2}{\sigma_{y_0}^2 + n\tau_{\mu_{y_0}}^2} \end{pmatrix}.
\end{aligned}$$

□

Endpoints certainly may have correlation between their x- and y- coordinates. If, for example, similar instruments were used at the same endpoint to collect data at different times,

this could cause some correlation between the errors in its coordinates. Depending on the time of day, for instance, GPS instruments relying on satellite information may have similar types of error in each coordinate, based on the changing positions of the satellites. In some cases, though, for instance manual digitization, it is probably a valid assumption that there is no correlation in error between coordinates.

Corollary 3.5. Suppose there is no correlation between the endpoints $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, and there is no correlation between the coordinates at each endpoint. Additionally, suppose each endpoint has equal variance at its x- and y- coordinates (although it may be different at each endpoint); that is, we know that $\boldsymbol{\Sigma}_1 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}$, $\boldsymbol{\Sigma}_0 = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2 \end{pmatrix}$, $\boldsymbol{\Lambda}_1 = \begin{pmatrix} \tau_1^2 & 0 \\ 0 & \tau_1^2 \end{pmatrix}$, and $\boldsymbol{\Lambda}_0 = \begin{pmatrix} \tau_0^2 & 0 \\ 0 & \tau_0^2 \end{pmatrix}$.

The posterior mean is then

$$\boldsymbol{g}_\gamma = \gamma \left(\frac{\sigma_1^2}{\sigma_1 + n\tau_1^2} \boldsymbol{\mu}_{1_0} + \frac{n\tau_1^2}{\sigma_1 + n\tau_1^2} \overline{\boldsymbol{z}}_1 \right) + (1-\gamma) \left(\frac{\sigma_0^2}{\sigma_0 + n\tau_0^2} \boldsymbol{\mu}_{0_0} + \frac{n\tau_0^2}{\sigma_0 + n\tau_0^2} \overline{\boldsymbol{z}}_0 \right),$$

and the posterior variance is

$$\boldsymbol{H}_\gamma = \left[\gamma^2 \left(\frac{\tau_1^2 \sigma_1^2}{\sigma_1^2 + n\tau_1^2} \right) + (1-\gamma)^2 \left(\frac{\tau_0^2 \sigma_0^2}{\sigma_0^2 + n\tau_0^2} \right) \right] \boldsymbol{I}.$$

Proof. It is easy to show first that $(\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1}$ and $(\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1}$.

$$\boldsymbol{\Lambda}_1^{-1} = \begin{pmatrix} \frac{1}{\tau_1^2} & 0 \\ 0 & \frac{1}{\tau_1^2} \end{pmatrix}, \ \boldsymbol{\Lambda}_0^{-1} = \begin{pmatrix} \frac{1}{\tau_0^2} & 0 \\ 0 & \frac{1}{\tau_0^2} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_1^2} \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} \frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{\sigma_0^2} \end{pmatrix}.$$

We can then calculate

$$\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} \frac{1}{\tau_1^2} + n\frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\tau_1^2} + n\frac{1}{\sigma_1^2} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_1^2 + n\tau_1^2}{\tau_1^2 \sigma_1^2} & 0 \\ 0 & \frac{\sigma_1^2 + n\tau_1^2}{\tau_1^2 \sigma_1^2} \end{pmatrix}$$

and

$$\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} \frac{1}{\tau_0^2} + n\frac{1}{\sigma_0^2} & 0 \\ 0 & \frac{1}{\tau_0^2} + n\frac{1}{\sigma_0^2} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_0^2 + n\tau_0^2}{\tau_0^2\sigma_0^2} & 0 \\ 0 & \frac{\sigma_0^2 + n\tau_0^2}{\tau_0^2\sigma_0^2} \end{pmatrix}.$$

This gives us

$$(\boldsymbol{\Lambda}_1^{-1} + n\boldsymbol{\Sigma}_1^{-1})^{-1} = \begin{pmatrix} \frac{\tau_1^2\sigma_1^2}{\sigma_1^2 + n\tau_1^2} & 0 \\ 0 & \frac{\tau_1^2\sigma_1^2}{\sigma_1 + n\tau_1^2} \end{pmatrix} = \left(\frac{\tau_1^2\sigma_1^2}{\sigma_1^2 + n\tau_1^2}\right)\boldsymbol{I}$$

and

$$(\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}_0^{-1})^{-1} = \begin{pmatrix} \frac{\tau_0^2\sigma_0^2}{\sigma_0^2 + n\tau_0^2} & 0 \\ 0 & \frac{\tau_0^2\sigma_0^2}{\sigma_0 + n\tau_0^2} \end{pmatrix} = \left(\frac{\tau_0^2\sigma_0^2}{\sigma_0^2 + n\tau_0^2}\right)\boldsymbol{I}.$$

We can now calculate the posterior mean and variance.

$$\begin{aligned}
\boldsymbol{g}_\gamma &= \gamma\left[\left(\frac{\tau_1^2\sigma_1^2}{\sigma_1^2 + n\tau_1^2}\right)\boldsymbol{I}\right]\left[\frac{1}{\tau_1^2}\boldsymbol{\mu}_{1_0} + \frac{n}{\sigma_1^2}\overline{\boldsymbol{z}}_1\right] + (1-\gamma)\left[\left(\frac{\tau_0^2\sigma_0^2}{\sigma_0^2 + n\tau_0^2}\right)\boldsymbol{I}\right]\left[\frac{1}{\tau_0^2}\boldsymbol{\mu}_{0_0} + \frac{n}{\sigma_0^2}\overline{\boldsymbol{z}}_0\right] \\
&= \gamma\left(\frac{\sigma_1^2}{\sigma_1 + n\tau_1^2}\boldsymbol{\mu}_{1_0} + \frac{n\tau_1^2}{\sigma_1 + n\tau_1^2}\overline{\boldsymbol{z}}_1\right) + (1-\gamma)\left(\frac{\sigma_0^2}{\sigma_0 + n\tau_0^2}\boldsymbol{\mu}_{0_0} + \frac{n\tau_0^2}{\sigma_0 + n\tau_0^2}\overline{\boldsymbol{z}}_0\right), \\
\boldsymbol{H}_\gamma &= \gamma^2\left(\frac{\tau_1^2\sigma_1^2}{\sigma_1^2 + n\tau_1^2}\right)\boldsymbol{I} + (1-\gamma)^2\left(\frac{\tau_0^2\sigma_0^2}{\sigma_0^2 + n\tau_0^2}\right)\boldsymbol{I} \\
&= \left[\gamma^2\left(\frac{\tau_1^2\sigma_1^2}{\sigma_1^2 + n\tau_1^2}\right) + (1-\gamma)^2\left(\frac{\tau_0^2\sigma_0^2}{\sigma_0^2 + n\tau_0^2}\right)\right]\boldsymbol{I}.
\end{aligned}$$

□

This may be a valid situation, especially in the case of human error. Suppose, for example, that two maps of adjoined areas are combined into a larger map. Each map may have been created by a different agency–possibly with different methods–and each agency may have emphasized a different standard of accuracy in terms of identifying exact point location. Therefore, some points on the large map may have small variance in error while others have larger error variance. It would, however, be likely within an agency that error in the x- and y- directions at each point would be quite similar.

Corollary 3.6. Suppose again that there is no correlation between the endpoints $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, and there is no correlation between the x- and y- coordinates at each endpoint. Suppose also

that the variance in the x coordinates is similar between the endpoints, as is the variance in the y coordinates; that is, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 = \mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$, and $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_0 = \mathbf{\Lambda} = \begin{pmatrix} \tau_x^2 & 0 \\ 0 & \tau_y^2 \end{pmatrix}$.

In this case, the posterior mean for a point $\boldsymbol{\mu}_\gamma$ is

$$\boldsymbol{g}_\gamma = \begin{pmatrix} \frac{\sigma_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} [\gamma\boldsymbol{\mu}_{1_0} + (1-\gamma)\boldsymbol{\mu}_{0_0}] + \begin{pmatrix} \frac{\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} [\gamma\overline{\boldsymbol{z}}_1 + (1-\gamma)\overline{\boldsymbol{z}}_0],$$

and the posterior variance is

$$\boldsymbol{H}_\gamma = (2\gamma^2 - 2\gamma + 1) \begin{pmatrix} \frac{\sigma_x^2\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix}.$$

Proof. Again calculating the individual terms involved in $(\mathbf{\Lambda}^{-1} + n\mathbf{\Sigma}^{-1})^{-1}$, we have

$$\mathbf{\Lambda}^{-1} = \begin{pmatrix} \frac{1}{\tau_x^2} & 0 \\ 0 & \frac{1}{\tau_y^2} \end{pmatrix}, \text{ and } \mathbf{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix}.$$

We can then calculate

$$\mathbf{\Lambda}^{-1} + n\mathbf{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\tau_x^2} + n\frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\tau_y^2} + n\frac{1}{\sigma_y^2} \end{pmatrix} = \begin{pmatrix} \frac{\sigma_x^2+n\tau_x^2}{\tau_x^2\sigma_x^2} & 0 \\ 0 & \frac{\sigma_y^2+n\tau_y^2}{\tau_y^2\sigma_y^2} \end{pmatrix}.$$

This gives us

$$(\mathbf{\Lambda}^{-1} + n\mathbf{\Sigma}^{-1})^{-1} = \begin{pmatrix} \frac{\tau_x^2\sigma_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\tau_y^2\sigma_y^2}{\sigma_y+n\tau_x^2} \end{pmatrix}.$$

We can now calculate the posterior mean and variance.

$$
\begin{aligned}
\boldsymbol{g}_\gamma &= \gamma \begin{pmatrix} \frac{\tau_x^2\sigma_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\tau_y^2\sigma_y^2}{\sigma_y+n\tau_x^2} \end{pmatrix} \left[\begin{pmatrix} \frac{1}{\tau_x^2} & 0 \\ 0 & \frac{1}{\tau_y^2} \end{pmatrix} \boldsymbol{\mu}_{1_0} + \begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix} \overline{\boldsymbol{z}}_1 \right] \\
&\quad + (1-\gamma) \begin{pmatrix} \frac{\tau_x^2\sigma_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\tau_y^2\sigma_y^2}{\sigma_y+n\tau_x^2} \end{pmatrix} \left[\begin{pmatrix} \frac{1}{\tau_x^2} & 0 \\ 0 & \frac{1}{\tau_y^2} \end{pmatrix} \boldsymbol{\mu}_{0_0} + \begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix} \overline{\boldsymbol{z}}_0 \right] \\
&= \begin{pmatrix} \frac{\sigma_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} [\gamma\boldsymbol{\mu}_{1_0} + (1-\gamma)\boldsymbol{\mu}_{0_0}] + \begin{pmatrix} \frac{\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} [\gamma\overline{\boldsymbol{z}}_1 + (1-\gamma)\overline{\boldsymbol{z}}_0], \\
&\quad + (1-\gamma) \left[\begin{pmatrix} \frac{\sigma_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} \boldsymbol{\mu}_{0_0} + \begin{pmatrix} \frac{\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} \overline{\boldsymbol{z}}_0 \right] \\
\boldsymbol{H}_\gamma &= \gamma^2 \begin{pmatrix} \frac{\sigma_x^2\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} + (1-\gamma)^2 \begin{pmatrix} \frac{\sigma_x^2\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix} \\
&= (2\gamma^2 - 2\gamma + 1) \begin{pmatrix} \frac{\sigma_x^2\tau_x^2}{\sigma_x^2+n\tau_x^2} & 0 \\ 0 & \frac{\sigma_y^2\tau_y^2}{\sigma_y^2+n\tau_y^2} \end{pmatrix}.
\end{aligned}
$$

□

This is a possibility in a case where a map has been digitized in a situation where x- and y-distances are not displayed at the same scale. This can happen locally, for example, when using certain coordinate systems. A single technician digitizing a map is likely to make the same size error at all points based on the visual display available, in both the x- and y-directions. This means that the true size of error as measured on the ground will be different for x- and y- coordinates at a single point, but similar over the scope of the map.

Corollary 3.7. Suppose again that there is no correlation between the endpoints $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, and there is no correlation between the x- and y- coordinates at each endpoint. Suppose additionally that all coordinate variances are similar; that is, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, and $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_0 = \boldsymbol{\Lambda} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix}$.

In this case, the posterior mean for a point $\boldsymbol{\mu}_\gamma$ is

$$\boldsymbol{g}_\gamma = \left(\frac{\tau^2\sigma^2}{\tau^2+n\sigma^2}\right)\left[(1-\gamma)\left(\frac{1}{\tau^2}\boldsymbol{\mu}_{0_0}+\frac{n}{\sigma^2}\overline{\boldsymbol{z}}_0\right)+\gamma\left(\frac{1}{\tau^2}\boldsymbol{\mu}_{1_0}+\frac{n}{\sigma^2}\overline{\boldsymbol{z}}_1\right)\right],$$

and the posterior variance is

$$\boldsymbol{H}_\gamma = (2\gamma^2-2\gamma+1)\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I}.$$

Proof. Again calculating the individual terms involved in the computations, we have

$$\boldsymbol{\Lambda}^{-1} = \begin{pmatrix} \frac{1}{\tau^2} & 0 \\ 0 & \frac{1}{\tau^2} \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}.$$

We can then calculate

$$\boldsymbol{\Lambda}^{-1}+n\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{1}{\tau^2}+n\frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\tau^2}+n\frac{1}{\sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2+n\tau^2}{\tau^2\sigma^2} & 0 \\ 0 & \frac{\sigma^2+n\tau^2}{\tau^2\sigma^2} \end{pmatrix}.$$

This gives us

$$(\boldsymbol{\Lambda}^{-1}+n\boldsymbol{\Sigma}^{-1})^{-1} = \begin{pmatrix} \frac{\tau^2\sigma^2}{\sigma^2+n\tau^2} & 0 \\ 0 & \frac{\tau^2\sigma^2}{\sigma^2+n\tau^2} \end{pmatrix} = \frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I}. \tag{3.1}$$

We can now easily calculate the posterior mean and variance.

$$\begin{aligned}
\boldsymbol{g}_\gamma &= \gamma\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I}\left[\begin{pmatrix} \frac{1}{\tau^2} & 0 \\ 0 & \frac{1}{\tau^2} \end{pmatrix}\boldsymbol{\mu}_{1_0}+\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}\overline{\boldsymbol{z}}_1\right] \\
&\quad + (1-\gamma)\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I}\left[\begin{pmatrix} \frac{1}{\tau^2} & 0 \\ 0 & \frac{1}{\tau^2} \end{pmatrix}\boldsymbol{\mu}_{0_0}+\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}\overline{\boldsymbol{z}}_0\right] \\
&= \left(\frac{\tau^2\sigma^2}{\tau^2+n\sigma^2}\right)\left[\gamma\left(\frac{1}{\tau^2}\boldsymbol{\mu}_{1_0}+\frac{n}{\sigma^2}\overline{\boldsymbol{z}}_1\right)+(1-\gamma)\left(\frac{1}{\tau^2}\boldsymbol{\mu}_{0_0}+\frac{n}{\sigma^2}\overline{\boldsymbol{z}}_0\right)\right], \\
\boldsymbol{H}_\gamma &= \gamma^2\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I}+(1-\gamma^2)\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I} \\
&= (2\gamma^2-2\gamma+1)\frac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I}.
\end{aligned}$$

□

This is something of a best-case scenario, and would certainly simplify subsequent calculations using error variance. It may even be a realistic situation, for example, when all maps involved have been digitized by a single technician. It seems reasonable that human error would tend to be randomly and evenly distributed at all points. There are many cases in which this assumption is not valid, however, which I have already mentioned. One example is multiple technicians digitizing separate parts of a map. Error in instrument readings is another example, since those types of idiosyncracies in an instrument are likely to be correlated across a map. Therefore, anyone hoping to accurately discuss positional error in a GIS product should seriously consider the types of correlation and variance that may occur, as it may change the model considerably.

3.2.4 Choosing an appropriate prior distribution

There are many ways to find prior information. Some points on a map, for example, may have well-known coordinates that have been measured many times. Consulting with an expert can then result in an appropriate prior distribution. Often, previous maps already exist, and using the information from such a map as a prior distribution can result in a more accurate point estimate than current information alone. Additionally, one can use information from related non-vector maps to create a prior distribution. For example, a digital elevation map (DEM), which uses a system of equal-area grid cells to represent the land, can be used to determine the flow path of a river through its recorded elevation values and cell relationships. You could use this flow path, along with the error of the DEM, to determine a prior distribution for points on the river. There are many other ways to determine a prior distribution.

In this dissertation, I have only included examples and calculations that assume the prior distribution of a point is normal. While I feel that this is generally acceptable, given its similarity to the commonly accepted normal data distribution, there certainly may be situations in which it is not appropriate. The normal distribution does have the advantage that it is

a conjugate prior for the mean when combined with the normal data distribution; that is, the combination of normal prior and normal data result in a normal posterior distribution. The Bayesian method is widely adaptable, however, and Bayes' Rule can be used with any combination of prior and data distributions. In the case that the posterior distribution does not have a closed form, the wide availability of computing resources, for instance, the MCMC Gibbs sampling algorithm, see ([49]), makes it possible to use simulation to approximate the posterior distribution.

Furthermore, it is still possible to use Bayesian methods when no truly informative prior information is available; one can use a non-informative prior. The advantage of using a non-informative prior when you have no useful information, rather than using a non-Bayesian approach, is primarily conceptual. Traditional frequentist methodology requires focusing on an estimator of the parameter, which can only be found through repeated sampling, while Bayesian methodology treats the parameter of interest as a variable and can estimate it directly from its distribution. This means that we should still be able to use fewer observations through a Bayesian method than a non-Bayesian one. Mostly, the ability to use a non-informative prior serves to illustrate the flexibility of the Bayesian method. It can be adapted to use full or only partial prior information, depending on what is available for use.

I would like to make a cautionary note here regarding the usage of these formulas. I have assumed that we know the true variance-covariance matrix of the data, $\mathbf{\Sigma}$. In the event that this information is unknown, and must be estimated from the data, the chi-square error region formulas are only approximate. When only a small number of observations are available, the approximation is not a very good one, and other more appropriate error ellipse approximations may be useful. For more discussion on this issue, see Section 6.2. The following examples, however, will assume the data have a known standard deviation.

3.3 Examples

I now provide some theoretical examples of the incorporation of Bayesian methodology into GIS vector data error analysis.

3.3.1 Point data

A professor in the geography department at a university wants to find an accurate measurement of a particular landmark. A graduate student, using a GPS instrument, measures the coordinates at $(551466.47, 4119762.54)$ (in meters, on the UTM Nad 1927 coordinate system). Studies on comparable GPS instruments have shown that the standard deviation of the error is 4.5 m. in both the x and y directions, and the errors are uncorrelated.

They also have access to a previously digitized map that includes the landmark feature. The map depicts the point at coordinates $(551465.9, 4119758.6)$. The map was digitized at a scale of 1:20,000. The map claims to meet the National Map Accuracy Standards set by the USGS, which states that no more than than 10% of the points will be off by more than 1/50 of an inch at map scale. At the 1:20,000 scale, 1/50 of an inch is representative of 400 inches, or 33.33 feet. This is equivalent to 10.16 meters. Therefore, on this map, we assume at least 90% of the points on the map will be within 10.16 meters of their true location. Assuming the points are normally distributed in the x and y directions, the number of standard deviations equivalent to a 90% confidence interval is 1.645. Therefore, the standard deviation for this map in both directions is no more than $10.16/1.645 \approx 6.18$ meters. We assume there is no covariance between the x and y directions.

Frequentist Method

Traditionally, only the current information on the coordinates would be used to estimate the true location of the landmark. The estimate of the landmark location is $(551466.47, 4119762.54)$.

The variance of these estimates is $\Sigma = \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}$. Finally, we can represent the error pictorially by drawing the error ellipse at the 95% confidence level,

$$\begin{aligned}(\boldsymbol{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) &= \begin{pmatrix} 551466.47-\mu_{x_0} \\ 4119762.54-\mu_{y_0} \end{pmatrix}' \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}^{-1} \begin{pmatrix} 551466.47-\mu_{x_0} \\ 4119762.54-\mu_{y_0} \end{pmatrix} \\ &\leq \chi^2_{2,.95},\end{aligned}$$

shown in grey in figure 3.1.

Bayesian Method: Prior 1

The Bayesian methodology I have proposed here allows us to use the additional information we have about the landmark—the previously digitized map. Suppose we will use this as our prior distribution on $\boldsymbol{\mu}_0$:

$$\begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \end{pmatrix} \sim \left(\begin{pmatrix} 551465.9 \\ 4119758.6 \end{pmatrix}, \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix} \right).$$

Combining this prior distribution with the information from the data, we can give the posterior distribution of the point as $\boldsymbol{\mu_0}|\boldsymbol{Z}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{g}_0, \boldsymbol{H}_0)$, where

$$\begin{aligned}\boldsymbol{g}_0 &= \left(\begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} + \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}^{-1} \right)^{-1} \\ &\left(\begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \begin{pmatrix} 551465.9 \\ 4119758.6 \end{pmatrix} + \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}^{-1} \begin{pmatrix} 551466.47 \\ 4119762.54 \end{pmatrix} \right) \\ &= \begin{pmatrix} 551466.3 \\ 4119761.2 \end{pmatrix},\end{aligned}$$

and

$$\boldsymbol{H_0} = \begin{bmatrix} 13.23 & 0 \\ 0 & 13.23 \end{bmatrix}.$$

We can represent the 95% probability error ellipse in this situation with the equation

$$(\boldsymbol{\mu}_0 - \boldsymbol{g}_0)'\boldsymbol{H}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{g}_0) = \begin{pmatrix} \mu_{x_0} - 551466.3 \\ \mu_{y_0} - 4119761.2 \end{pmatrix}' \begin{bmatrix} 13.23 & 0 \\ 0 & 13.23 \end{bmatrix}^{-1} \begin{pmatrix} \mu_{x_0} - 551466.3 \\ \mu_{y_0} - 4119761.2 \end{pmatrix}$$
$$\leq \chi^2_{2,.95},$$

shown in black in figure 3.1.

Method Comparison

When we compare the results from the two methods, we can see that the Bayesian result has smaller variance, due to the inclusion of the information from the Bayesian prior. In situations where this type of information is available, it is clear how the Bayesian prior can improve the inference process. Figure 3.1 compares the error ellipses around the landmark for the frequentist (grey) and Bayesian (black) methods. Not only is the Bayesian ellipse smaller, but it has an easier interpretation—rather than being a confidence region it is a probability region. Instead of being 95% "confident" that the coordinates are in the ellipse, we can say there is a 95% "probability" that the coordinates of the landmark are in the Bayesian ellipse. It is also important to note that the Bayesian result is compatible and in agreement with the traditional result, since the probability ellipse here falls well within the boundary of the traditional confidence region.

Bayesian Method: prior 2

Suppose instead that the digitized map had been created at a much smaller scale, say 1:5,000. The map still meets National Map Accuracy Standards. Assuming a normal distribution and using the same calculations as in the previous case, this is equivalent to an error standard deviation of approximately 1.54 meters. Suppose we will use this as our prior distribution

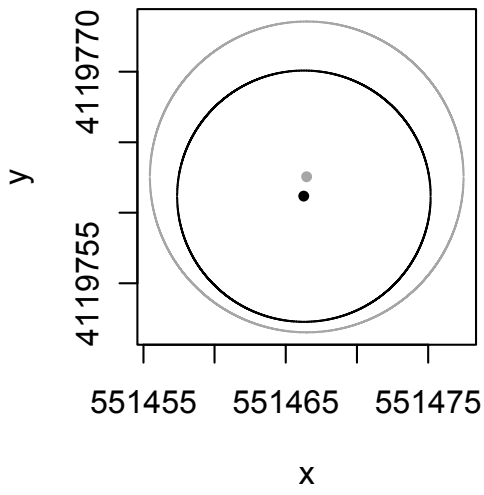

Figure 3.1: Comparison of traditional error ellipse (grey) and Bayesian error ellipse (black).

on $\boldsymbol{\mu}_0$:

$$\begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \end{pmatrix} \sim \left(\begin{pmatrix} 551465.9 \\ 4119758.6 \end{pmatrix}, \begin{bmatrix} 2.37 & 0 \\ 0 & 2.37 \end{bmatrix} \right).$$

Combining this prior distribution with the information from the data, we can give the posterior distribution of the point as $\boldsymbol{\mu_0}|\boldsymbol{Z}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{g}_0, \boldsymbol{H}_0)$, where

$$\begin{aligned} \boldsymbol{g}_0 &= \left(\begin{bmatrix} 2.37 & 0 \\ 0 & 2.37 \end{bmatrix}^{-1} + \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}^{-1} \right)^{-1} \\ &\left(\begin{bmatrix} 2.37 & 0 \\ 0 & 2.37 \end{bmatrix}^{-1} \begin{pmatrix} 551465.9 \\ 4119758.6 \end{pmatrix} + \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}^{-1} \begin{pmatrix} 551466.47 \\ 4119762.54 \end{pmatrix} \right) \\ &= \begin{pmatrix} 551466.0 \\ 4119759.0 \end{pmatrix}, \end{aligned}$$

and

$$\boldsymbol{H_0} = \begin{bmatrix} 2.12 & 0 \\ 0 & 2.12 \end{bmatrix}.$$

We can represent the 95% probability error ellipse in this situation with the equation

$$\begin{aligned} (\boldsymbol{\mu}_0 - \boldsymbol{g}_0)' \boldsymbol{H}_0^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{g}_0) &= \begin{pmatrix} \mu_{x_0} - 551466.0 \\ \mu_{y_0} - 4119759.0 \end{pmatrix}' \begin{bmatrix} 2.12 & 0 \\ 0 & 2.12 \end{bmatrix}^{-1} \begin{pmatrix} \mu_{x_0} - 551466.0 \\ \mu_{y_0} - 4119759.0 \end{pmatrix} \\ &\leq \chi^2_{2,.95}, \end{aligned}$$

shown in black in figure 3.2.

Method Comparison

Again, when we compare the results from the two methods, we can see that the Bayesian result has smaller variance, due to the inclusion of the information from the Bayesian prior. Figure 3.2 compares the error ellipses around the landmark for the frequentist (grey) and

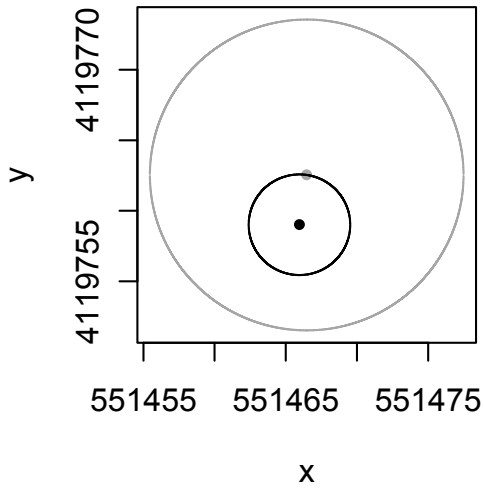

Figure 3.2: Comparison of traditional error ellipse (grey) and Bayesian error ellipse (black).

Bayesian (black) methods. Additionally, when comparing Figure 3.2 to Figure 3.1, we see that using the Bayesian prior with smaller variance gives the expected result, and the posterior ellipse and variance is smaller.

Using a non-informative prior

Up to this point, we assumed we had good information to create a prior distribution on our data. In fact, my argument for using Bayesian methodology has been that good prior information is often available in GIS applications, and large numbers of observations are not. In the case where there is no good prior information available, however, we can still use Bayesian methodology. We do this by using a *noninformative* prior distribution, designed to have as little impact on the inference as possible, and therefore allow the data to "speak for themselves" ([23]). The reasons for doing so are primarily theoretical, although there are

many real examples across the field of statistics in which noninformative prior distributions have been shown to work quite well. I offer an example here to demonstrate how the Bayesian method can still be applied, even when there is no good prior information.

There are several ways to determine the noninformative prior to use for a particular problem. There are some noninformative priors that have been created to meet certain requirements; for example, Jeffreys's prior is founded on the principle that any rule for determining the prior density should yield an equivalent posterior distribution for any one-to-one transformation of the parameter ([23]). One simple and very popular prior distribution, which I will use in the example, is the uniform prior. The uniform distribution assigns equal probability to any value of the parameter in its admissible range.

Suppose we have the same situation as in the previous example, in which we are examining the position of a landmark. Our GPS instrument, as before, gives the apparent location of the landmark as $(551466.47, 4119762.54)$. The variance of these estimates is $\Sigma = \begin{bmatrix} 20.25 & 0 \\ 0 & 20.25 \end{bmatrix}$. Suppose now, however, we do not have access to a pre-existing map. We can still implement a Bayesian method, if desired, by the inclusion of a uniform prior distribution.

In order to find the posterior distribution using uniform distributions on the mean variables, we turn to Bayes' Rule (2.1), which tells us that the posterior distribution of the point is

$$p(\boldsymbol{\mu}|y) = \frac{p(\boldsymbol{\mu})p(y|\boldsymbol{\mu})}{p(y)} = \frac{\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}\right)}{p(y)},$$

where $p(y)$ is the marginal distribution of y, and is a constant, given that y has been observed. We recognize this distribution as the normal distribution. Therefore the result of using a uniform prior on the data is that the posterior distribution is equal to the normal data distribution. An error ellipse drawn from this posterior distribution would therefore be identical to the error ellipse drawn from Shi's data distribution model ([50]). The difference would be only in the interpretation: the ellipse would represent a probability region rather

than a confidence region.

We can see from this example that it is possible to use a noninformative prior that will minimally impact the information gathered from the data. Bayesian methods are therefore certainly compatible with the frequentist methods currently in use, even when there is no good prior information available.

3.3.2 Line segment data

Suppose you are interested in learning the coordinates of two adjacent corners of a particular building from a computerized GIS map. The map scale is 1:20,000, and so according to National Map Accuracy Standards and assuming normally distributed errors, this is equivalent to a standard deviation of 6.18 meters. On the map, the coordinates have been placed at $(376106.36, 2798635.52)$ and $(376230.06, 2798628.07)$, in the state plane coordinate system. Having been digitized by hand, we assume no correlation between any coordinates.

You also have access to the original plans for the building, which instructed that the two corners of the building should have been placed at $(376100, 2798635)$ and $(376220, 2798635)$. Based on construction difficulties at the actual building cite (for example, condition of the ground), the construction workers often find it necessary to vary slightly from the original plans. From their own records, the company estimates the corners of the building independently vary from the original plans under a normal distribution with a standard deviation of 5 meters.

Frequentist Method

From the frequentist standpoint, the estimates of the vertices of the side of this building are simply the parameters on the digitized map itself. The standard deviation of the estimates is therefore the standard deviation of the digitizer's accuracy. We have

$$\boldsymbol{z}_{01} = \begin{pmatrix} 376106.36 \\ 2798635.52 \\ 376230.06 \\ 2798628.07 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_{01} = \begin{bmatrix} 38.19 & 0 & 0 & 0 \\ 0 & 38.19 & 0 & 0 \\ 0 & 0 & 38.19 & 0 \\ 0 & 0 & 0 & 38.19 \end{bmatrix}.$$

We can follow this up with a 95% confidence ellipse around each endpoint. The ellipse around the first endpoint is

$$\begin{aligned}(\boldsymbol{\mu}_0 - \boldsymbol{z}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{z}_0) &= \begin{pmatrix} \mu_{x_0} - 376106.36 \\ \mu_{y_0} - 2798635.52 \end{pmatrix}' \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \begin{pmatrix} \mu_{x_0} - 376106.36 \\ \mu_{y_0} - 2798635.52 \end{pmatrix} \\ &\leq \chi^2_{2,.95},\end{aligned}$$

and the ellipse around the second endpoint is

$$\begin{aligned}(\boldsymbol{\mu}_1 - \boldsymbol{z}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{z}_1) &= \begin{pmatrix} \mu_{x_1} - 376230.06 \\ \mu_{y_1} - 2798628.07 \end{pmatrix}' \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \begin{pmatrix} \mu_{x_1} - 376230.06 \\ \mu_{y_1} - 2798628.07 \end{pmatrix} \\ &\leq \chi^2_{2,.95}.\end{aligned}$$

To create a confidence region around the entire line segment, we note that we can create a confidence region around any point on the line segment with mean estimate

$$\begin{pmatrix} (1-t)x_0 + (t)x_1 \\ (1-t)y_0 + (t)y_1 \end{pmatrix} = \begin{pmatrix} (1-t)376106.36 + (t)376230.06 \\ (1-t)2798635.52 + (t)2798628.07 \end{pmatrix},$$

and variance

$$\boldsymbol{\Sigma}_z(t) = \begin{bmatrix} 38.19\left((1-t)^2 + (t)^2\right) & 0 \\ 0 & 38.19\left((1-t)^2 + (t)^2\right) \end{bmatrix}.$$

The ellipse-based picture of the confidence region around the line segment is shown in grey in Figure 3.3.

Bayesian Method

The Bayesian method allows us to include information from both the original design plans and the current digitized map. According to our formulas, our Bayesian posterior estimate of the value of the coordinates for the first point is

$$\begin{aligned}
\boldsymbol{g}_0 &= \left(\begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}^{-1} + \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \right)^{-1} \\
&\left(\begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}^{-1} \begin{pmatrix} 376100 \\ 2798635 \end{pmatrix} + \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \begin{pmatrix} 376106.36 \\ 2798635.52 \end{pmatrix} \right) \\
&= \begin{pmatrix} 376102.5 \\ 2798635.2 \end{pmatrix},
\end{aligned}$$

and the posterior variance of this estimate is $\boldsymbol{H_0} = \begin{bmatrix} 15.11 & 0 \\ 0 & 15.11 \end{bmatrix}$.

Our Bayesian posterior estimate of the value of the coordinates for the second point is

$$\begin{aligned}
\boldsymbol{g}_1 &= \left(\begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}^{-1} + \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \right)^{-1} \\
&\left(\begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}^{-1} \begin{pmatrix} 376220 \\ 2798635 \end{pmatrix} + \begin{bmatrix} 38.19 & 0 \\ 0 & 38.19 \end{bmatrix}^{-1} \begin{pmatrix} 376230.06 \\ 2798628.07 \end{pmatrix} \right) \\
&= \begin{pmatrix} 376224.0 \\ 2798632.3 \end{pmatrix},
\end{aligned}$$

and the posterior variance of this estimate is $\boldsymbol{H_0} = \begin{bmatrix} 15.11 & 0 \\ 0 & 15.11 \end{bmatrix}$.

We can now describe a 95% probability ellipse around each endpoint. The ellipse around the first endpoint is

$$
\begin{aligned}
(\boldsymbol{\mu}_0 - \boldsymbol{g}_0)'\boldsymbol{H}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{g}_0) &= \begin{pmatrix} \mu_{x_0} - 376102.5 \\ \mu_{y_0} - 2798635.2 \end{pmatrix}' \begin{bmatrix} 15.11 & 0 \\ 0 & 15.11 \end{bmatrix}^{-1} \begin{pmatrix} \mu_{x_0} - 376102.5 \\ \mu_{y_0} - 2798635.2 \end{pmatrix} \\
&\leq \chi^2_{2,.95},
\end{aligned}
$$

and the ellipse around the second endpoint is

$$
\begin{aligned}
(\boldsymbol{\mu}_1 - \boldsymbol{g}_1)'\boldsymbol{H}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{g}_1) &= \begin{pmatrix} \mu_{x_1} - 376224.0 \\ \mu_{y_1} - 2798632.3 \end{pmatrix}' \begin{bmatrix} 15.11 & 0 \\ 0 & 15.11 \end{bmatrix}^{-1} \begin{pmatrix} \mu_{x_1} - 376224.0 \\ \mu_{y_1} - 2798632.3 \end{pmatrix} \\
&\leq \chi^2_{2,.95}.
\end{aligned}
$$

To create a probability region around the entire line segment, we note that we can create a probability region around any point on the line segment with mean estimate

$$
\begin{pmatrix} (1-\gamma)g_{x_0} + (\gamma)g_{x_1} \\ (1-\gamma)g_{y_0} + (\gamma)g_{y_1} \end{pmatrix} = \begin{pmatrix} (1-\gamma)376102.5 + (\gamma)376224.0 \\ (1-\gamma)2798635.2 + (\gamma)2798632.3 \end{pmatrix},
$$

and variance

$$
\begin{aligned}
\boldsymbol{H}_\gamma &= \begin{bmatrix} h^2_{x_\gamma} & h_{x_\gamma y_\gamma}(t) \\ h_{y_\gamma x_\gamma} & h^2_{y_\gamma} \end{bmatrix} = \begin{bmatrix} 15.11\left((1-\gamma)^2 + (\gamma)^2\right) & 0 \\ 0 & 15.11\left((1-\gamma)^2 + (\gamma)^2\right) \end{bmatrix} \\
&= 15.11(2\gamma^2 - 2\gamma + 1)\boldsymbol{I}.
\end{aligned}
$$

Alternatively, we can find the formula for the posterior mean and variance of a point on the line segment by consulting corollary 3.7, and the results agree.

$$
\begin{aligned}
\boldsymbol{g}_\gamma &= \left(\frac{\tau^2\sigma^2}{\tau^2+n\sigma^2}\right)\left[(1-\gamma)\left(\frac{1}{\tau^2}\boldsymbol{\mu}_{0_0}+\frac{n}{\sigma^2}\overline{\boldsymbol{z}}_0\right)+\gamma\left(\frac{1}{\tau^2}\boldsymbol{\mu}_{1_0}+\frac{n}{\sigma^2}\overline{\boldsymbol{z}}_1\right)\right] \\
&= \left(\frac{25\cdot 38.19}{25+38.19}\right)\left[(1-\gamma)\left(\frac{1}{25}\begin{pmatrix}376100\\2798635\end{pmatrix}+\frac{1}{38.19}\begin{pmatrix}376106.36\\2798635.52\end{pmatrix}\right)\right. \\
&\left.+\gamma\left(\frac{1}{25}\begin{pmatrix}376220\\2798635\end{pmatrix}+\frac{1}{38.19}\begin{pmatrix}376230.06\\2798628.07\end{pmatrix}\right)\right] \\
&= 15.11\left[(1-\gamma)\begin{pmatrix}24892.29\\185227.29\end{pmatrix}+\gamma\begin{pmatrix}24900.33\\185227.10\end{pmatrix}\right] \\
&= \begin{pmatrix}(1-\gamma)376102.5+(\gamma)376224.0\\(1-\gamma)2798635.2+(\gamma)2798632.3\end{pmatrix},
\end{aligned}
$$

and $\boldsymbol{H}_\gamma = (2\gamma^2-2\gamma+1)\dfrac{\tau^2\sigma^2}{\sigma^2+n\tau^2}\boldsymbol{I} = (2\gamma^2-2\gamma+1)\dfrac{25\cdot 38.19}{38.19+25}\boldsymbol{I} = 15.11(2\gamma^2-2\gamma+1)\boldsymbol{I}$.

The ellipse-based picture of the credible region around the line segment is shown in black in Figure 3.3.

Method Comparison

As in our first example, we see here that the variance of our estimate is considerably smaller when using the prior information about the building in combination with the data currently available on the map. We can also see that our final posterior estimate of the mean is much closer to that specified by the original business plans (thanks to the inclusion of this information), and therefore probably more accurate. We can also examine Figure 3.3, shown here, which demonstrates both the difference in placement of the line segment (where the two endpoint vertices appear) and the variance of our estimate (the region surrounding the posterior distribution's line segment is smaller than the region surrounding the data distribution's line segment). Notice that like with the ellipses in example 1, the Bayesian region is completely compatible with the traditional region since it falls almost entirely within the traditional boundary. Furthermore, we again consider the difference in the meaning of

the two regions—the traditional region is a confidence region, and we are 95% "confident" that the line segment is contained in that region, and the Bayesian region is a "credible" region, and there is a 95% probability that the true line segment is contained in that region. Once more, the benefits of Bayesian methods are clear.

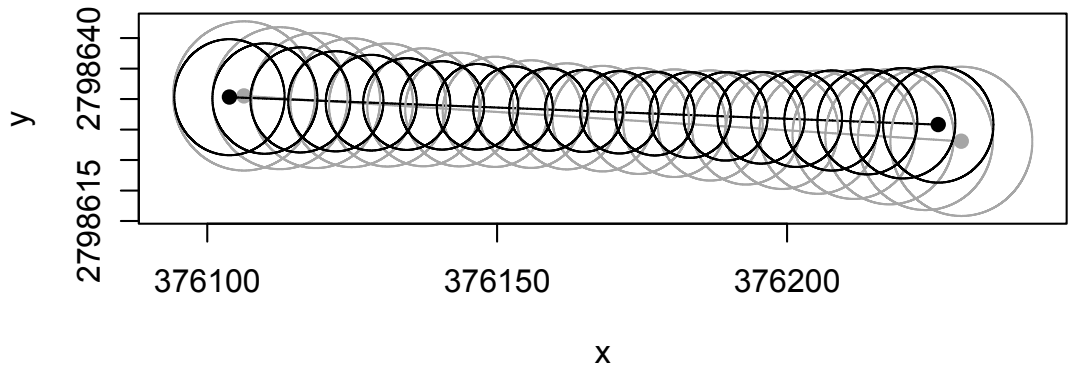

Figure 3.3: Comparison of traditional confidence region (grey) and Bayesian credible region (black).

3.4 Discussion

In this chapter, I have explored the addition of Bayesian methodology to current methods for analyzing error in vector GIS data. There are multiple advantages to this addition. For one thing, Bayesian analysis does not rely strongly on the present data to develop an error distribution or to estimate the location of a point. This is because the distribution of a point can be made to rely strongly on a prior distribution that can be based on expert or historical knowledge. This is not unreasonable in the geographic disciplines, where a lot of knowledge may already exist to indicate coordinate locations (for example, the possible path of a stream

or the locations of certain well-studied landmarks). This is a big advantage because often only one sample observation is available for each point on a map. Additionally, Bayesian methods can increase the accuracy of an analysis when prior information is more reliable than the data distribution, which could often be the case in GIS applications. Bayesian methodology is also completely compatible with traditional methods, as seen in our examples.

Bayesian analysis is also good from the standpoint of understandability. Often, a traditional confidence interval is interpreted as a probability interval. The interpretation of a Bayesian credible region, on the other hand, is direct and accessible to many users.

There are many possible applications of Bayesian error analysis. For example, it could be applied to the point-in-polygon and the sliver polygon problems mentioned in Section 2.3.4. The flexibility and interpretation of Bayesian methods should prove ideal for such scenarios.

Chapter 4

Calculation for the Boundary of the Confidence Region of a Line Segment

4.1 Introduction

In the previous chapter, I focused on a method for analyzing error in vector data known as the G-band model. It was originated by Shi et al. ([50, 53, 52, 10]), and is a statistical method based on the error distribution of vector data points. I expanded on this error model by introducing a Bayesian prior distribution, which creates additional flexibility in the model.

This method creates elliptical error regions around individual geographic points. Confidence regions around line segments are the composite of an infinite number of elliptical regions surrounding the points that make up the line segment. Currently, however, there is no explicit calculation for this confidence region; any subsequent analysis must be performed by continually drawing a large number of ellipses. This is computationally intense, which makes it very difficult to include in GIS software. The large number of calculations that must be performed to draw the confidence region also makes it difficult to perform further

error analyses.

I have calculated an explicit formula for the boundary of the G-band model's line segment confidence region. This will cut back on the amount of computation involved in drawing the confidence region. This will allow the G-band error region to be more easily included in software applications. It will also create a clearer picture that will be more appealing to the average GIS user. Additionally, it will help pave the way for more in-depth error analysis, including such things as point-in-polygon analysis and fractionated polygon analysis.

I first present an explicit confidence region calculation for the most simple error distribution, where there is no covariance between any pair of coordinates and all variances are the same. I then provide several examples of this calculation. Finally, I present a method for finding the x-coordinate of the boundary for the general case.

4.2 Confidence Region Boundary Calculation

Theorem 4.1. In the most simple case of the line segment ellipse confidence region, that is, when

$$\begin{pmatrix} x_0 \\ y_0 \\ x_1 \\ y_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix}, \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}\right),$$

and

$$\begin{pmatrix} x_\gamma \\ y_\gamma \end{pmatrix} = \gamma \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + (1-\gamma) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

the confidence bound can be written explicitly as a function of γ, $0 \leq \gamma \leq 1$, as follows:

$$x_{b1\gamma} = \frac{x_\gamma - \frac{1}{m}c + \sqrt{(x_\gamma - \frac{1}{m}c)^2 - (1 + \frac{1}{m^2})(c^2 - \chi^2_{2,1-\alpha}\sigma^2_\gamma + x^2_\gamma)}}{1 + \frac{1}{m^2}}$$
$$y_{b1\gamma} = y_\gamma - \sqrt{\chi^2_{2,1-\alpha}\sigma^2_\gamma - (x_{b1\gamma} - x_\gamma)^2}$$

if

· $m \geq 0$, $x_1 < x_0$, and x_{b1i} is decreasing with i at $i = \gamma$,

· $m \geq 0$, $x_1 \geq x_0$, and x_{b1i} is not decreasing with i at $i = \gamma$,

· $m < 0$, $x_1 < x_0$, and x_{b1i} is not decreasing with i at $i = \gamma$,

· $m < 0$, $x_1 \geq x_0$, and x_{b1i} is decreasing with i at $i = \gamma$; or

$$x_{b1\gamma} = \frac{x_\gamma - \frac{1}{m}c + \sqrt{(x_\gamma - \frac{1}{m}c)^2 - (1 + \frac{1}{m^2})(c^2 - \chi^2_{2,1-\alpha}\sigma^2_\gamma + x^2_\gamma)}}{1 + \frac{1}{m^2}}$$
$$y_{b1\gamma} = y_\gamma + \sqrt{\chi^2_{2,1-\alpha}\sigma^2_\gamma - (x_{b1\gamma} - x_\gamma)^2}$$

if none of the above conditions hold;

and

$$x_{b2\gamma} = \frac{x_\gamma - \frac{1}{m}c - \sqrt{(x_\gamma - \frac{1}{m}c)^2 - (1 + \frac{1}{m^2})(c^2 - \chi^2_{2,1-\alpha}\sigma^2_\gamma + x^2_\gamma)}}{1 + \frac{1}{m^2}}$$
$$y_{b2\gamma} = y_\gamma + \sqrt{\chi^2_{2,1-\alpha}\sigma^2_\gamma - (x_{b2\gamma} - x_\gamma)^2}$$

· $m \geq 0$, $x_1 < x_0$, and x_{b2i} is decreasing with i at $i = \gamma$,

· $m \geq 0$, $x_1 \geq x_0$, and x_{b2i} is not decreasing with i at $i = \gamma$,

· $m < 0$, $x_1 < x_0$, and x_{b2i} is not decreasing with i at $i = \gamma$,

· $m < 0$, $x_1 \geq x_0$, and x_{b2i} is decreasing with i at $i = \gamma$; or

$$x_{b2\gamma} = \frac{x_\gamma - \frac{1}{m}c - \sqrt{(x_\gamma - \frac{1}{m}c)^2 - (1 + \frac{1}{m^2})(c^2 - \chi^2_{2,1-\alpha}\sigma^2_\gamma + x^2_\gamma)}}{1 + \frac{1}{m^2}}$$
$$y_{b2\gamma} = y_\gamma - \sqrt{\chi^2_{2,1-\alpha}\sigma^2_\gamma - (x_{b2\gamma} - x_\gamma)^2}$$

if none of the above conditions hold. Note these conditions may hold for part of the boundary and not for the other part, so the calculation of the y_b term may not be the same for every point γ.

Here, $c = \chi^2_{2,1-\alpha}\sigma^2 \frac{1-2\gamma}{y_0 - y_1} - \frac{1}{m}x_\gamma$, and m is the slope of the line segment.

Proof. The traditional line segment confidence region is defined by a series of ellipses with the formula

$$(\boldsymbol{z}_\gamma - \boldsymbol{\mu}_\gamma)'\boldsymbol{\Sigma}_\gamma^{-1}(\boldsymbol{z}_\gamma - \boldsymbol{\mu}_\gamma) = \chi^2_{2,1-\alpha}.$$

In order to depict this, a limited number of γ values are chosen and ellipses are drawn. Figure 4.1 demonstrates this concept with 21 values of γ evenly spaced across the line segment.

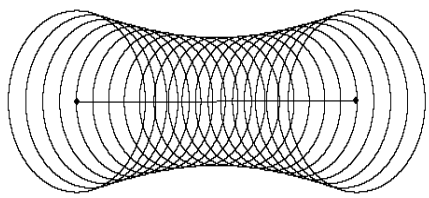

Figure 4.1: Example of confidence region drawn with ellipses.

From this picture, we can see that the point at which one ellipse ceases to be on the boundary and the ellipse next to it becomes part of the boundary is the point at which the two ellipses

intersect. As we include more ellipses in the picture, the portion of any ellipse that is included in the boundary becomes smaller and smaller. As we approach the true situation, with an infinite number of ellipses along the line segment, it becomes apparent that each ellipse will contribute exactly one point to the boundary. See Figure 4.2.

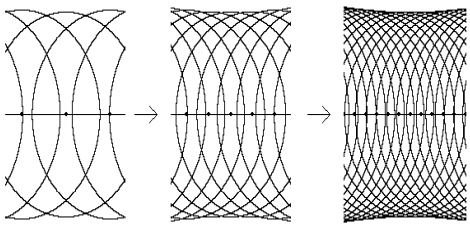

Figure 4.2: Demonstration of each ellipse's contribution to the boundary as more and more ellipses are included in the confidence region representation.

Because an infinite number of ellipses are involved, we cannot compute the value of this point as an intersection, but rather as the limit of the intersection of two ellipses as the distance between them approaches 0. We calculate the limit of the intersection here for the simple case only, as in the statement of the theorem.

We will expand the formula for a confidence ellipse at point γ in two ways.

$$\mu_{y_\gamma} = \frac{(\mu_{x_\gamma} - x_\gamma)\sigma_{x_\gamma y_\gamma} \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma} - (\mu_{x_\gamma} - x_\gamma)^2)(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma_{x_\gamma}{y_\gamma}^2)}}{\sigma^2_{x_\gamma}} + y_\gamma$$

$$(\mu_{y_\gamma} - y_\gamma)^2\sigma^2_{x_\gamma} - 2(\mu_{x_\gamma} - x_\gamma)(\mu_{y_\gamma} - y_\gamma)\sigma_{x_\gamma y_\gamma} + (\mu_{x_\gamma} - x_\gamma)^2\sigma_{y_\gamma} = \chi^2_{2,1-\alpha}(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma_{x_\gamma y^2_\gamma})$$

In the case discussed in the statement of the theorem, these equations simplify to

$$\mu_{y_\gamma} = \pm\sqrt{(\chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2)} + y_\gamma \tag{4.1}$$

$$(\mu_{y_\gamma} - y_\gamma)^2 + (\mu_{x_\gamma} - x_\gamma)^2 = \chi^2_{2,1-\alpha}\sigma^2_\gamma \tag{4.2}$$

In order to find the intersection between two ellipses, one at point γ and one at point $\gamma + \delta$ with $0 \leq \gamma, \gamma + \delta \leq 1$, we set $\boldsymbol{\mu}_\gamma = \boldsymbol{\mu}_{\gamma+\delta}$ and solve equations 4.1 and 4.2 simultaneously.

$$\mu_{y_\gamma} = \pm\sqrt{(\chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2)} + y_\gamma$$

$$(\mu_{y_\gamma} - y_{\gamma+\delta})^2 + (\mu_{x_\gamma} - x_{\gamma+\delta})^2 = \chi^2_{2,1-\alpha}\sigma^2_{\gamma+\delta}$$

$$\Rightarrow \left(y_\gamma \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2)} - y_{\gamma+\delta}\right)^2 + (\mu_{x_\gamma} - x_{\gamma+\delta})^2 = \chi^2_{2,1-\alpha}\sigma^2_{\gamma+\delta}$$

$$\Rightarrow \left(y_\gamma - y_{\gamma+\delta} \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2)}\right)^2 + (\mu_{x_\gamma} - x_{\gamma+\delta})^2 = \chi^2_{2,1-\alpha}\sigma^2_{\gamma+\delta}$$

$$\Rightarrow (y_\gamma - y_{\gamma+\delta})^2 \pm (y_\gamma - y_{\gamma+\delta})\sqrt{(\chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2)} + \chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2$$

$$+(\mu_{x_\gamma} - x_{\gamma+\delta})^2 = \chi^2_{2,1-\alpha}\sigma^2_{\gamma+\delta}$$

$$\Rightarrow \chi^2_{2,1-\alpha}(\sigma^2_\gamma - \sigma^2_{\gamma+\delta}) + (y_\gamma - y_{\gamma+\delta})^2 - (\mu_{x_\gamma} - x_\gamma)^2 + (\mu_{x_\gamma} - x_{\gamma+\delta})^2$$

$$= \mp 2(y_\gamma - y_{\gamma+\delta})\sqrt{\chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2}$$

Note that

$$(\mu_{x_\gamma} - x_{\gamma+\delta})^2 - (\mu_{x_\gamma} - x_\gamma)^2 = 2\mu_{x_\gamma}(x_\gamma - x_{\gamma+\delta}) - (x_\gamma - x_{\gamma+\delta})(x_\gamma + x_{\gamma+\delta})$$

$$\Rightarrow \left(\frac{\chi^2_{2,1-\alpha}(\sigma^2_\gamma - \sigma^2_{\gamma+\delta}) + (y_\gamma - y_{\gamma+\delta})^2 + 2\mu_{x_\gamma}(x_\gamma - x_{\gamma+\delta}) - (x_\gamma - x_{\gamma+\delta})(x_\gamma + x_{\gamma+\delta})}{2(y_\gamma - y_{\gamma+\delta})}\right)^2$$

$$= \chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2$$

$$\Rightarrow \left(\frac{\chi^2_{2,1-\alpha}(\sigma^2_\gamma - \sigma^2_{\gamma+\delta}) + (y_\gamma - y_{\gamma+\delta})^2 - (x_\gamma - x_{\gamma+\delta})(x_\gamma + x_{\gamma+\delta})}{2(y_\gamma - y_{\gamma+\delta})}\right)^2$$

$$= \chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2$$

$$\Rightarrow \left(\frac{2\mu_{x_\gamma}(x_\gamma - x_{\gamma+\delta})}{2(y_\gamma - y_{\gamma+\delta})} + c'\right)^2 = \chi^2_{2,1-\alpha}\sigma^2_\gamma - (\mu_{x_\gamma} - x_\gamma)^2,$$

where

$$c' = \frac{\chi^2_{2,1-\alpha}(\sigma^2_\gamma - \sigma^2_{\gamma+\delta}) + (y_\gamma - y_{\gamma+\delta})^2 - (x_\gamma - x_{\gamma+\delta})(x_\gamma + x_{\gamma+\delta})}{2(y_\gamma - y_{\gamma+\delta})},$$

$$\Rightarrow \mu_{x_\gamma}^2 \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)^2 + 2\mu_{x_\gamma}\left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c' + c'^2 = \chi^2_{2,1-\alpha}\sigma_\gamma^2 - (\mu_{x_\gamma} - x_\gamma)^2$$

$$\Rightarrow \mu_{x_\gamma}^2 \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)^2 + 2\mu_{x_\gamma}\left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c' + c'^2 = \chi^2_{2,1-\alpha}\sigma_\gamma^2 - \mu_{x_\gamma}^2 + 2\mu_{x_\gamma}x_\gamma - x_\gamma^2$$

$$\Rightarrow \mu_{x_\gamma}^2 \left[1 + \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)\right]^2 + 2\mu_{x_\gamma}\left[\left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c' - x_\gamma\right] + c'^2 = \chi^2_{2,1-\alpha}\sigma_\gamma^2 - x_\gamma^2$$

$$\Rightarrow \mu_{x_\gamma}^2 \left[1 + \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)\right]^2 + 2\mu_{x_\gamma}\left[\left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c' - x_\gamma\right] + \left[c'^2 - \chi^2_{2,1-\alpha}\sigma_\gamma^2 - x_\gamma^2\right] = 0$$

We can now solve for μ_{x_γ} by solving the quadratic equation.

$$\frac{x_\gamma - \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c' \pm \sqrt{\left(x_\gamma - \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c'\right)^2 - \left(1 + \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)^2\right)\left(c'^2 - \chi^2_{2,1-\alpha}\sigma_\gamma^2 + x_\gamma^2\right)}}{1 + \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)^2}$$

This gives us the x-coordinate of the point where the two ellipses at points γ and $\gamma + \delta$ intersect. In order to find the point that ellipse γ contributes to the boundary region, we now find the limit of this intersection as δ approaches 0.

$$\lim_{\delta \to 0} \left(\frac{x_\gamma - \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c' \pm \sqrt{\left(x_\gamma - \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)c'\right)^2 - \left(1 + \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)^2\right)\left(c'^2 - \chi^2_{2,1-\alpha}\sigma_\gamma^2 + x_\gamma^2\right)}}{1 + \left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)^2}\right)$$

We start by observing that the quantity $\left(\frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}\right)$ is a constant function equal to $\frac{1}{m}$, where m is the slope of the line segment, except where $\gamma = \gamma + \delta$. By definition, then, the limit of this quantity as $\delta \to 0$ is equal to $\frac{1}{m}$.

Next, it remains to find the limit of c' as $\delta \to 0$.

$$\lim_{\delta\to 0} \left(\frac{\chi^2_{2,1-\alpha}(\sigma^2_\gamma - \sigma^2_{\gamma+\delta}) + (y_\gamma - y_{\gamma+\delta})^2 - (x_\gamma - x_{\gamma+\delta})(x_\gamma + x_{\gamma+\delta})}{2(y_\gamma - y_{\gamma+\delta})} \right)$$

$$= \lim_{\delta\to 0} \frac{1}{2} \left(\frac{\chi^2_{2,1-\alpha}(\sigma^2_\gamma - \sigma^2_{\gamma+\delta})}{y_\gamma - y_{\gamma+\delta}} + \frac{(y_\gamma - y_{\gamma+\delta})^2}{y_\gamma - y_{\gamma+\delta}} - \frac{x_\gamma - x_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}}(x_\gamma + x_{\gamma+\delta}) \right)$$

$$= \frac{1}{2} \left(\chi^2_{2,1-\alpha} \lim_{\delta\to 0} \left(\frac{(\sigma^2_\gamma - \sigma^2_{\gamma+\delta})}{(y_\gamma - y_{\gamma+\delta})} \right) + 0 - \frac{2x_\gamma}{m} \right)$$

$$\lim_{\delta\to 0} \left(\frac{\sigma^2_\gamma - \sigma^2_{\gamma+\delta}}{y_\gamma - y_{\gamma+\delta}} \right) = \lim_{\delta\to 0} \frac{\left(2\gamma^2 - 2\gamma + 1 - 2(\gamma+\delta)^2 + 2(\gamma+\delta) - 1\right)\sigma^2}{\gamma y_1 + (1-\gamma)y_0 - (\gamma+\delta)y_1 - (1-(\gamma+\delta))y_0}$$

$$= 2 \lim_{\delta\to 0} \frac{(\gamma^2 - (\gamma+\delta)^2 - \gamma + (\gamma+\delta))\sigma^2}{(\gamma - (\gamma+\delta))y_1 + ((\gamma+\delta) - \gamma)y_0}$$

$$= 2 \lim_{\delta\to 0} \frac{(\gamma^2 - (\gamma+\delta)^2 - \gamma + (\gamma+\delta))\sigma^2}{(\gamma - (\gamma+\delta))y_1 + ((1-\gamma) - (1-(\gamma+\delta)))y_0}$$

$$= 2 \lim_{\delta\to 0} \frac{\left(\gamma^2 - (\gamma^2 + 2\gamma\delta + \delta^2) + \delta\right)\sigma^2}{-\delta y_1 + \delta y_0} = 2 \lim_{\delta\to 0} \left(\frac{\delta(1-2\gamma)\sigma^2}{\delta(y_0 - y_1)} - \frac{\delta^2\sigma^2}{\delta(y_0 - y_1)} \right)$$

$$= \frac{(2\gamma - 1)\sigma^2}{y_1 - y_0}$$

So,

$$\lim_{\delta\to 0} (c') = \frac{1}{2} \left(2\chi^2_{2,1-\alpha}\sigma^2 \frac{(1-2\gamma)}{y_0 - y_1} - \frac{2x_\gamma}{m} \right) = \chi^2_{2,1-\alpha}\sigma^2 \frac{2\gamma - 1}{y_1 - y_0} - \frac{1}{m}x_\gamma = c$$

and therefore the x-coordinate of the boundary point is

$$x_{b\gamma} = \frac{x_\gamma - \frac{1}{m}c \pm \sqrt{(x_\gamma - \frac{1}{m}c)^2 - (1 + \frac{1}{m^2})(c^2 - \chi^2_{2,1-\alpha}\sigma^2_\gamma + x^2_\gamma)}}{1 + \frac{1}{m^2}}$$

as written in the theorem.

The y-coordinate is given by the formula

$$\mu_{y_\gamma} = \frac{(\mu_{x_\gamma} - x_\gamma)\sigma_{x_\gamma y_\gamma} \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma} - (\mu_{x_\gamma} - x_\gamma)^2)(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma_{x_\gamma y_\gamma}{}^2)}}{\sigma^2_{x_\gamma}} + y_\gamma,$$

and so there are two possible y-coordinates for each x-coordinate. The y-coordinate on the boundary can be determined by examining the geometry of the confidence region. When the line segment and the point γ meet one of the conditions given in the theorem for $x_{b1\gamma}$ or $x_{b2\gamma}$, the sign in the equation for the y-coordinate will be the opposite of the sign in the equation for the x-coordinate. If one of these conditions is not met, then the sign will be the same. Figure 4.3 demonstrates this conclusion for several different regions.

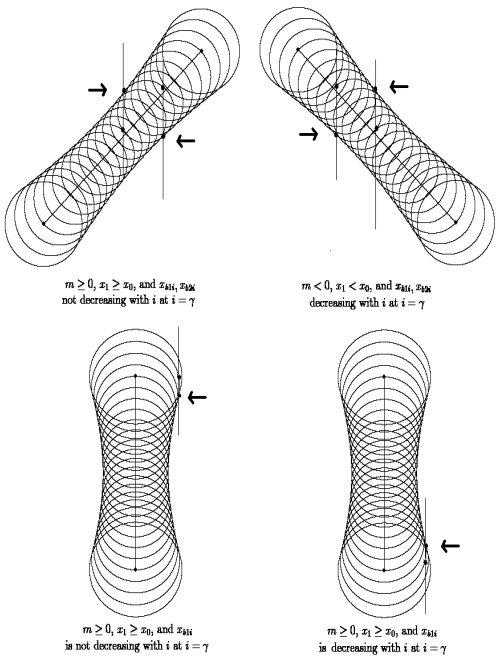

Figure 4.3: Visual demonstration of the determination of the y-coordinate matching the x-coordinate according to line segment slope.

□

Corollary 4.1. In the Bayesian case, with a prior distribution

$$\begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{x_{00}} \\ \mu_{y_{00}} \\ \mu_{x_{10}} \\ \mu_{y_{10}} \end{pmatrix}, \begin{bmatrix} \tau_0^2 & 0 & 0 & 0 \\ 0 & \tau_0^2 & 0 & 0 \\ 0 & 0 & \tau_0^2 & 0 \\ 0 & 0 & 0 & \tau_0^2 \end{bmatrix} \right)$$

and a data distribution as in theorem 4.1, leading to a posterior distribution

$$\begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix} \sim N\left(\begin{pmatrix} g_{x_0} \\ g_{y_0} \\ g_{x_1 0} \\ g_{y_1 0} \end{pmatrix}, \begin{bmatrix} h^2 & 0 & 0 & 0 \\ 0 & h^2 & 0 & 0 \\ 0 & 0 & h^2 & 0 \\ 0 & 0 & 0 & h^2 \end{bmatrix} \right),$$

the confidence bound can be written explicitly as a function of γ, $0 \leq \gamma \leq 1$, as follows:

$$\begin{aligned} x_{b\gamma} &= \frac{g_{x_\gamma} - \frac{1}{m}c_b \pm \sqrt{(g_{x_\gamma} - \frac{1}{m}c_b)^2 - (1 + \frac{1}{m^2})(c_b^2 - \chi^2_{2,1-\alpha}h_\gamma^2 + g_{x_\gamma}^2)}}{1 + \frac{1}{m^2}} \\ y_{b\gamma} &= g_{y_\gamma} \mp \sqrt{\chi^2_{2,1-\alpha}h_\gamma^2 - (x_{b\gamma} - g_{x_\gamma})^2}. \end{aligned}$$

Here, $c_b = \chi^2_{2,1-\alpha}h^2\frac{1-2\gamma}{g_{y_0}-g_{y_1}} - \frac{1}{m}g_{x_\gamma}$, m is again the slope of the line segment, g_{x_γ} and g_{y_γ} are the posterior means of μ_{x_γ} and μ_{y_γ}, respectively, and h_γ is the posterior variance of both μ_{x_γ} and μ_{y_γ}.

Proof. Substituting the terms from the posterior mean and variance of the point $\begin{pmatrix} \mu_{x_\gamma} \\ \mu_{y_\gamma} \end{pmatrix}$ into the result of theorem 4.1, in place of the data mean and variance of $\begin{pmatrix} x_\gamma \\ y_\gamma \end{pmatrix}$, provides a direct proof of this corollary. □

A Note on Calculating Confidence Regions

Note that because of the way in which the slope enters the formula, when the two y-coordinates are the same (that is, $y_0 = y_1$), the formula will fail to produce an x-coordinate. This problem can be fixed by switching the x- and y-coordinates in the formula because the calculations are symmetrical (making sure to also adjust the slope by using its reciprocal). Simply switch the x- and y-coordinates back after the calculations have been completed to draw the confidence bound correctly.

4.3 Examples

We now provide several examples of our calculation, including an example of the Bayesian confidence region.

4.3.1 Example 1

Suppose we have a line segment with x-coordinates at $x_0 = 0$ and $x_1 = 10$, and y-coordinates at $y_0 = 0$ and $y_1 = 1$. The error structure is as described in Theorem 4.1, and the common variance parameter is $\sigma = 1$. We calculate the confidence region using the formula in Theorem 4.1. Figure 4.4 demonstrates the calculation of the boundary region and compares it to the ellipse confidence region. The boundary is represented as a solid line, and the ellipses are dashed lines.

Bayesian variation

Suppose we have the same data as above, and we now have a prior distribution on the points. The prior distribution of the x-coordinates places them at $\mu_{x_{00}} = 1$ and $\mu_{x_{10}} = 12$, and the prior distribution of the y-coordinates places them both at $\mu_{y_{00}} = \mu_{y_{10}} = 0$.

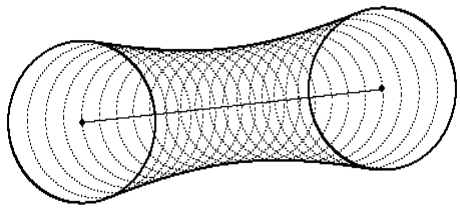

Figure 4.4: Demonstration of the explicit formula for the boundary of the G-band line segment confidence region.

Additionally, the prior value for the common variance parameter is $\tau = 1.5$. We calculate the posterior distribution from this information, using Corollary 4.1. Figure 4.5 demonstrates the Bayesian version of the confidence region. Figure 4.6 compares the boundary-only versions of the Bayesian and frequentist methods, demonstrating one advantage of using the prior information: including more information creates a smaller confidence region. The traditional boundary is drawn with larger dashed lines.

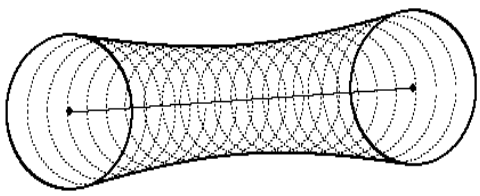

Figure 4.5: Demonstration of the explicit formula for the boundary of the Bayesian line segment confidence region.

4.3.2 Example 2

This example uses a map adapted from actual GPS data. The map is a polygon outline of a field located in Blacksburg, Virginia. Research has shown that GPS error often is reasonably normally distributed, with a standard deviation of approximately 3.5 meters in

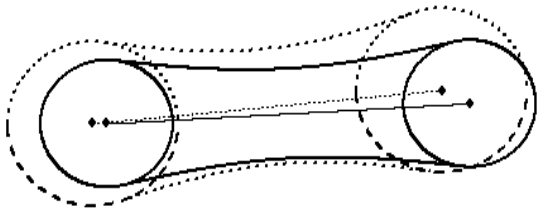

Figure 4.6: Bayesian line segment boundary compared to the traditional G-band boundary.

each direction. Figure 4.7 (a) demonstrates the error in the boundary of the field using the individual error ellipses. Figure 4.7 (b) demonstrates the boundary using the direct boundary calculation. It is clear from this picture that using the direct calculation is advantageous in several ways. First, it involves much less computational time and much less drawing time. This makes it easier to include in GIS software, and therefore more accessible to GIS users. Secondly, the picture is clearer and probably less confusing to the average GIS user, who is most likely familiar with error regions in the form of epsilon bands (refer back to Section 2.3 for more on epsilon bands).

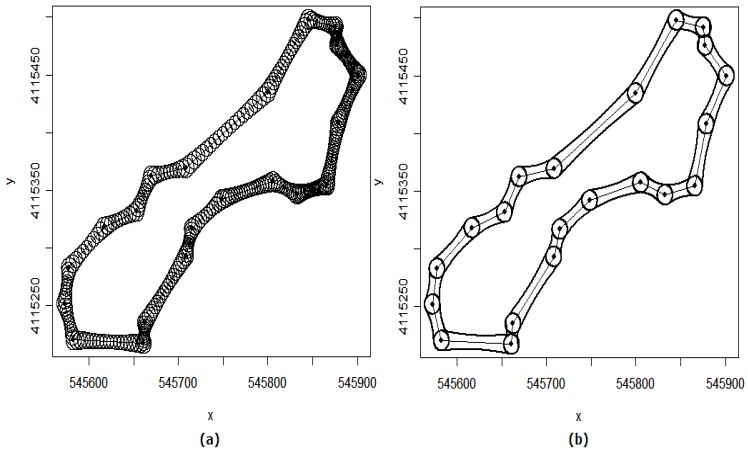

Figure 4.7: Figure (a) shows the error in the field boundary using error ellipses; figure (b) shows the error in the field boundary using the direct boundary calculation.

4.4 General Case

Theorem 4.2. In the general case, when the endpoints of a line segment have a 4-dimensional normal distribution,

$$\begin{pmatrix} x_0 \\ y_0 \\ x_1 \\ y_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{x_0} \\ \mu_{y_0} \\ \mu_{x_1} \\ \mu_{y_1} \end{pmatrix}, \begin{bmatrix} \sigma^2_{x_0} & \sigma_{x_0y_0} & \sigma_{x_0x_1} & \sigma_{x_0y_1} \\ \sigma_{x_0y_0} & \sigma^2_{y_0} & \sigma_{x_1y_0} & \sigma_{y_0y_1} \\ \sigma_{x_0x_1} & \sigma_{x_1y_0} & \sigma^2_{x_1} & \sigma_{x_1y_1} \\ \sigma_{x_0y_1} & \sigma_{y_0y_1} & \sigma_{x_1y_1} & \sigma^2_{y_1} \end{bmatrix}\right),$$

the x-coordinates of the two boundaries of the line segment confidence region calculated at point $(x_\gamma, y_\gamma) = (\gamma x_1 + (1-\gamma)x_0, \gamma y_1 + (1-\gamma)y_1)$, where $0 \le \gamma \le 1$, are equal to the real solutions x_b of the equation

$$A1(x_b - x_\gamma)^4 + A2(x_b - x_\gamma)^3 + A3(x_b - x_\gamma)^2 + A4(x_b - x_\gamma) + A5 = 0, \tag{4.3}$$

where

$$\begin{aligned} A1 = {} & 4(\sigma^2_{x_\gamma})^3\sigma^2_y c^2_{xy} - 4(\sigma^2_{x_\gamma})^2\sigma^2_{y_\gamma}\sigma_{x_\gamma y_\gamma}c_x c_{xy} + (\sigma^2_{x_\gamma})^2(\sigma^2_{y_\gamma})^2 c^2_x + (\sigma^2_{x_\gamma})^4 c^2_y \\ & +4(\sigma^2_{x_\gamma})^2(\sigma_{x_\gamma y_\gamma})^2 c_x c_y - 2(\sigma^2_{x_\gamma})^3\sigma^2_{y_\gamma}c_x c_y - 4(\sigma^2_{x_\gamma})^3\sigma_{x_\gamma y_\gamma}c_y c_{xy}, \end{aligned}$$

$$\begin{aligned} A2 = {} & 4(x_0 - x_1)(\sigma^2_{x_\gamma})^4\sigma^2_{y_\gamma}c_y - 4(x_0 - x_1)(\sigma^2_{x_\gamma})^3(\sigma^2_{y_\gamma})^2 c_x + 4(x_0 - x_1)(\sigma^2_{x_\gamma})^2\sigma^2_{y_\gamma}(\sigma_{x_\gamma y_\gamma})^2 c_x \\ & -4(x_0 - x_1)(\sigma^2_{x_\gamma})^3(\sigma_{x_\gamma y_\gamma})^2 c_y - 8(y_0 - y_1)(\sigma^2_{x_\gamma})^4\sigma^2_{y_\gamma}c_{xy} + 8(y_0 - y_1)(\sigma^2_{x_\gamma})^3(\sigma_{x_\gamma y_\gamma})^2 c_{xy} \\ & +8(y_0 - y_1)(\sigma^2_{x_\gamma})^3\sigma^2_{y_\gamma}\sigma_{x_\gamma y_\gamma}c_x - 8(y_0 - y_1)(\sigma^2_{x_\gamma})^2(\sigma_{x_\gamma y_\gamma})^3 c_x, \end{aligned}$$

$$
\begin{aligned}
A3 = {} & 4(x_0-x_1)^2(\sigma_{x_\gamma}^2)^4(\sigma_{y_\gamma}^2)^2 - 4(x_0-x_1)^2(\sigma_{x_\gamma}^2)^3\sigma_{y_\gamma}^2(\sigma_{x_\gamma y_\gamma})^2 + 8(x_0-x_1)(y_0-y_1)(\sigma_{x_\gamma}^2)^3(\sigma_{x_\gamma y_\gamma})^3 \\
& -8(x_0-x_1)(y_0-y_1)(\sigma_{x_\gamma}^2)^4\sigma_{y_\gamma}^2\sigma_{x_\gamma y_\gamma} + 4(y_0-y_1)^2(\sigma_{x_\gamma}^2)^5\sigma_{y_\gamma}^2 - 4(y_0-y_1)^2(\sigma_{x_\gamma}^2)^4(\sigma_{x_\gamma y_\gamma})^2 \\
& -6(\sigma_{x_\gamma}^2)^3(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 c_x c_y + 2(\sigma_{x_\gamma}^2)^4\sigma_{y_\gamma}^2\chi_{2,1-\alpha}^2 c_x c_y - 4(\sigma_{x_\gamma}^2)^3(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 c_{xy}^2 \\
& +4(\sigma_{x_\gamma}^2)^2(\sigma_{x_\gamma y_\gamma})^3\chi_{2,1-\alpha}^2 c_x c_{xy} + 8(\sigma_{x_\gamma}^2)^4\sigma_{x_\gamma y_\gamma}\chi_{2,1-\alpha}^2 c_y c_{xy} - 2(\sigma_{x_\gamma}^2)^5\chi_{2,1-\alpha}^2 c_y^2 \\
& -4(\sigma_{x_\gamma}^2)^4\sigma_{y_\gamma}^2\chi_{2,1-\alpha}^2 c_{xy}^2 - 2(\sigma_{x_\gamma}^2)^2\sigma_{y_\gamma}^2(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 c_x^2 + 4(\sigma_{x_\gamma}^2)^3\sigma_{y_\gamma}^2\sigma_{x_\gamma y_\gamma}\chi_{2,1-\alpha}^2 c_x c_{xy},
\end{aligned}
$$

$$
\begin{aligned}
A4 = {} & 4(x_0-x_1)(\sigma_{x_\gamma}^2)^4(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 c_y - 4(x_0-x_1)(\sigma_{x_\gamma}^2)^5\sigma_{y_\gamma}^2\chi_{2,1-\alpha}^2 c_y \\
& +4(x_0-x_1)(\sigma_{x_\gamma}^2)^3\sigma_{y_\gamma}^2(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 c_x - 4(x_0-x_1)(\sigma_{x_\gamma}^2)^2(\sigma_{x_\gamma y_\gamma})^4\chi_{2,1-\alpha}^2 c_x \\
& +8(y_0-y_1)(\sigma_{x_\gamma}^2)^5\sigma_{y_\gamma}^2\chi_{2,1-\alpha}^2 c_{xy} - 8(y_0-y_1)(\sigma_{x_\gamma}^2)^4(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 c_{xy} \\
& -8(y_0-y_1)(\sigma_{x_\gamma}^2)^4\sigma_{y_\gamma}^2\sigma_{x_\gamma y_\gamma}\chi_{2,1-\alpha}^2 c_x + 8(y_0-y_1)(\sigma_{x_\gamma}^2)^3(\sigma_{x_\gamma y_\gamma})^3\chi_{2,1-\alpha}^2 c_x,
\end{aligned}
$$

$$
\begin{aligned}
A5 = {} & 4(\sigma_{x_\gamma}^2)^4(\sigma_{x_\gamma y_\gamma})^2(\chi_{2,1-\alpha}^2)^2 c_{xy}^2 + (\sigma_{x_\gamma}^2)^2(\sigma_{x_\gamma y_\gamma})^4(\chi_{2,1-\alpha}^2)^2 c_x^2 + (\sigma_{x_\gamma}^2)^6(\chi_{2,1-\alpha}^2)^2 c_y^2 \\
& -4(\sigma_{x_\gamma}^2)^3(\sigma_{x_\gamma y_\gamma})^3(\chi_{2,1-\alpha}^2)^2 c_x c_{xy} - 4(\sigma_{x_\gamma}^2)^5\sigma_{x_\gamma y_\gamma}(\chi_{2,1-\alpha}^2)^2 c_y c_{xy} + 2(\sigma_{x_\gamma}^2)^4(\sigma_{x_\gamma y_\gamma})^2(\chi_{2,1-\alpha}^2)^2 c_x c_y \\
& -4(x_0-x_1)^2(\sigma_{x_\gamma}^2)^4\sigma_{y_\gamma}^2(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2 + 4(x_0-x_1)^2(\sigma_{x_\gamma}^2)^3(\sigma_{x_\gamma y_\gamma})^4\chi_{2,1-\alpha}^2 \\
& +8(x_0-x_1)(y_0-y_1)(\sigma_{x_\gamma}^2)^5\sigma_{y_\gamma}^2\sigma_{x_\gamma y_\gamma}\chi_{2,1-\alpha}^2 - 8(x_0-x_1)(y_0-y_1)(\sigma_{x_\gamma}^2)^4(\sigma_{x_\gamma y_\gamma})^3\chi_{2,1-\alpha}^2 \\
& -4(y_0-y_1)^2(\sigma_{x_\gamma}^2)^6\sigma_{y_\gamma}^2\chi_{2,1-\alpha}^2 + 4(y_0-y_1)^2(\sigma_{x_\gamma}^2)^5(\sigma_{x_\gamma y_\gamma})^2\chi_{2,1-\alpha}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
c_x &= 2\gamma/n(\sigma_{x_1}^2 - 2\sigma_{x_0x_1} + \sigma_{x_0}^2) + 2/n(\sigma_{x_0x_1} - \sigma_{x_0}^2), \\
c_y &= 2\gamma/n(\sigma_{y_1}^2 - 2\sigma_{y_0y_1} + \sigma_{y_0}^2) + 2/n(\sigma_{y_0y_1} - \sigma_{y_0}^2), \\
c_{xy} &= 2\gamma/n(\sigma_{x_0y_0} - \sigma_{x_0y_1} - \sigma_{x_1y_0} + \sigma_{x_1y_1}) + 1/n(\sigma_{x_0y_1} + \sigma_{x_1y_0} - 2\sigma_{x_0y_0}).
\end{aligned}
$$

Note that in the Bayesian case, we can simply substitute the mean and variance terms from the posterior distribution for those from the data distribution to draw the appropriate confidence region.

Proof. We apply the same principle as in the simple case, which is that the boundary point is the limit of the intersection of two ellipses as the distance between them approaches 0. Review Figure 4.2. Also as in that case, we being by writing the equation for the confidence ellipse at point γ in two different ways.

$$\mu_{y_\gamma} = \frac{(\mu_{x_\gamma} - x_\gamma)\sigma_{x_\gamma y_\gamma} \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma} - (\mu_{x_\gamma} - x_\gamma)^2)(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma})}}{\sigma^2_{x_\gamma}} + y_\gamma$$

and

$$(\mu_{y_\gamma} - y_\gamma)^2\sigma^2_{x_\gamma} - 2(\mu_{x_\gamma} - x_\gamma)(\mu_{y_\gamma} - y_\gamma)\sigma_{x_\gamma y_\gamma} + (\mu_{x_\gamma} - x_\gamma)^2\sigma^2_{y_\gamma} = \chi^2_{2,1-\alpha}(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma})$$

Next, in order to find the point μ_{x_γ} where two ellipses at points $\boldsymbol{z}_\delta$ and $\boldsymbol{z}_{\delta+\gamma}$ intersect, we combine the two equations as follows:

$$\begin{aligned}
&\left(\frac{(\mu_{x_\gamma} - x_\gamma)\sigma_{x_\gamma y_\gamma} \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma} - (\mu_{x_\gamma} - x_\gamma)^2)(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma})}}{\sigma^2_{x_\gamma}} + y_\gamma - y_{\gamma+\delta}\right)^2 \sigma^2_{x_{\gamma+\delta}} \\
&- 2(\mu_{x_\gamma} - x_{\gamma+\delta}) \\
&\left(\frac{(\mu_{x_\gamma} - x_\gamma)\sigma_{x_\gamma y_\gamma} \pm \sqrt{(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma} - (\mu_{x_\gamma} - x_\gamma)^2)(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma})}}{\sigma^2_{x_\gamma}} + y_\gamma - y_{\gamma+\delta}\right) \sigma_{x_{\gamma+\delta} y_{\gamma+\delta}} \\
&+ (\mu_{x_\gamma} - x_{\gamma+\delta})^2\sigma^2_{y_{\gamma+\delta}} = \chi^2_{2,1-\alpha}(\sigma^2_{x_{\gamma+\delta}}\sigma^2_{y_{\gamma+\delta}} - \sigma^2_{x_{\gamma+\delta} y_{\gamma+\delta}}).
\end{aligned}$$

In order to begin solving this equation for μ_{x_γ}, we first isolate the portion of the equation under the square root to one side of the equation. Several manipulations result in

$$
\begin{aligned}
&\frac{(\mu_{x_\gamma}-x_\gamma)^2\sigma^2_{x_\gamma y_\gamma}}{(\sigma^2_{x_\gamma})^2}\sigma^2_{x_{\gamma+\delta}}+(y_\gamma-y_{\gamma+\delta})^2\sigma^2_{x_{\gamma+\delta}}+2\frac{(\mu_{x_\gamma}-x_\gamma)\sigma_{x_\gamma y_\gamma}}{\sigma^2_{x_\gamma}}(y_\gamma-y_{\gamma+\delta})\sigma^2_{x_{\gamma+\delta}}\\
&+\frac{\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})}{(\sigma^2_{x_\gamma})^2}\sigma^2_{x_{\gamma+\delta}}-\frac{(\mu_{x_\gamma}-x_\gamma)^2(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})}{(\sigma^2_{x_\gamma})^2}\sigma^2_{x_{\gamma_\delta}}\\
&-2(\mu_{x_\gamma}-x_{\gamma+\delta})\frac{(\mu_{x_\gamma}-x_\gamma)\sigma_{x_\gamma y_\gamma}}{\sigma^2_{x_\gamma}}\sigma_{x_{\gamma+\delta}y_{\gamma+\delta}}-2(\mu_{x_\gamma}-x_{\gamma+\delta})(y_\gamma-y_{\gamma+\delta})\sigma_{x_{\gamma+\delta}y_{\gamma+\delta}}\\
&+(\mu_{x_\gamma}-x_{\gamma+\delta})^2\sigma^2_{y_{\gamma+\delta}}-\chi^2_{2,1-\alpha}(\sigma^2_{x_{\gamma+\delta}}\sigma^2_{y_{\gamma+\delta}}-\sigma^2_{x_{\gamma+\delta}y_{\gamma+\delta}})\\
&=\pm\frac{\sqrt{(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}-(\mu_{x_\gamma}-x_\gamma)^2)(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})}}{\sigma^2_{x_\gamma}}\\
&\left(2(\mu_{x_\gamma}-x_{\gamma+\delta})\sigma_{x_{\gamma+\delta}y_{\gamma+\delta}}-2\frac{(\mu_{x_\gamma}-x_\gamma)\sigma_{x_\gamma y_\gamma}}{\sigma^2_{x_\gamma}}\sigma^2_{x_{\gamma+\delta}}-2(y_\gamma-y_{\gamma+\delta})\sigma^2_{x_{\gamma+\delta}}\right)
\end{aligned}
$$

We now multiply each term on either side by 1 in the form of 1, $\sigma^2_{x_\gamma}/\sigma^2_{x_\gamma}$, or $(\sigma^2_{x_\gamma})^2/(\sigma^2_{x_\gamma})^2$ so that we have a common denominator of $(\sigma^2_{x_\gamma})^2$ on each side of the equation and it can be removed completely from the equation. Additionally, we now square the equation on both sides to remove the square root from the equation. We now have

$$
\begin{aligned}
&\Big((\mu_{x_\gamma}-x_\gamma)^2\sigma^2_{x_\gamma y_\gamma}\sigma^2_{x_{\gamma+\delta}}+(y_\gamma-y_{\gamma+\delta})^2\sigma^2_{x_{\gamma+\delta}}(\sigma^2_{x_\gamma})^2+2(\mu_{x_\gamma}-x_\gamma)\sigma_{x_\gamma y_\gamma}\sigma^2_{x_\gamma}(y_\gamma-y_{\gamma+\delta})\sigma^2_{x_{\gamma+\delta}}\\
&+\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})\sigma^2_{x_{\gamma+\delta}}-(\mu_{x_\gamma}-x_\gamma)^2(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})\sigma^2_{x_{\gamma_\delta}}\\
&-2(\mu_{x_\gamma}-x_{\gamma+\delta})(\mu_{x_\gamma}-x_\gamma)\sigma_{x_\gamma y_\gamma}\sigma^2_{x_\gamma}\sigma_{x_{\gamma+\delta}y_{\gamma+\delta}}-2(\mu_{x_\gamma}-x_{\gamma+\delta})(y_\gamma-y_{\gamma+\delta})\sigma_{x_{\gamma+\delta}y_{\gamma+\delta}}(\sigma^2_{x_\gamma})^2\\
&+(\mu_{x_\gamma}-x_{\gamma+\delta})^2\sigma^2_{y_{\gamma+\delta}}(\sigma^2_{x_\gamma})^2-\chi^2_{2,1-\alpha}(\sigma^2_{x_{\gamma+\delta}}\sigma^2_{y_{\gamma+\delta}}-\sigma^2_{x_{\gamma+\delta}y_{\gamma+\delta}})(\sigma^2_{x_\gamma})^2\Big)^2\\
&=\Big(\chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})-(\mu_{x_\gamma}-x_\gamma)^2(\sigma^2_{x_\gamma}\sigma^2_{y_\gamma}-\sigma^2_{x_\gamma y_\gamma})\Big)\\
&\Big(2(\mu_{x_\gamma}-x_{\gamma+\delta})\sigma_{x_{\gamma+\delta}y_{\gamma+\delta}}\sigma^2_{x_\gamma}-2(\mu_{x_\gamma}-x_\gamma)\sigma_{x_\gamma y_\gamma}\sigma^2_{x_{\gamma+\delta}}-2(y_\gamma-y_{\gamma+\delta})\sigma^2_{x_{\gamma+\delta}}\sigma^2_{x_\gamma}\Big)^2
\end{aligned}
$$

In order to simplify calculations, rather than continue further and find the limit of μ_{x_γ} at the end of the process, we take the limit now as δ approaches 0. Finding the limit directly of each side of the equation results in the form $0=0$, which is not helpful. So, imposing the form $L^2=PR^2$ on the above equation, we take limits as follows, using l'Hospital's rule:

$$L^2/R^2 = P \Rightarrow \lim_{\delta \to 0} L^2/R^2 = \lim_{\delta \to 0} P = \lim_{\delta \to 0} L^2 / \lim_{\delta \to 0} R^2$$

$$\Rightarrow \left(\lim_{\delta \to 0} L / \lim_{\delta \to 0} R \right)^2 = (0/0)^2 \Rightarrow \left(\lim_{\delta \to 0} \frac{d}{d\delta} L / \lim_{\delta \to 0} \frac{d}{d\delta} R \right)^2 = \lim_{\delta \to 0} P$$

$$\Rightarrow \left(\lim_{\delta \to 0} \frac{d}{d\delta} L \right)^2 = \lim_{\delta \to 0} P \left(\lim_{\delta \to 0} \frac{d}{d\delta} R \right)^2 .$$

Finishing all the calculations to arrive at the final step in the above equation, we have

$$\begin{aligned}
&(x_b - x_\gamma)^4 \left[\sigma^2_{x_\gamma y_\gamma} c_x - \sigma^2_{x_\gamma} \sigma^2_{y_\gamma} c_x + \sigma^2_{x_\gamma y_\gamma} c_x - 2\sigma^2_{x_\gamma} \sigma_{x_\gamma y_\gamma} c_{xy} + (\sigma^2_{x_\gamma})^2 c_y \right]^2 \\
&+ (x_b - x_\gamma)^3 2 \left[2(x_0 - x_1)(\sigma^2_{x_\gamma})^2 \sigma^2_{y_\gamma} - 2\sigma^2_{x_\gamma} \sigma^2_{x_\gamma y_\gamma} (x_0 - x_1) \right] \\
&\left[\sigma^2_{x_\gamma y_\gamma} c_x - \sigma^2_{x_\gamma} \sigma^2_{y_\gamma} c_x + \sigma^2_{x_\gamma y_\gamma} c_x - 2\sigma^2_{x_\gamma} \sigma_{x_\gamma y_\gamma} c_{xy} + (\sigma^2_{x_\gamma})^2 c_y \right] \\
&+ (x_b - x_\gamma)^2 \Bigg(\left[2(x_0 - x_1)(\sigma^2_{x_\gamma})^2 \sigma^2_{y_\gamma} - 2\sigma^2_{x_\gamma} \sigma^2_{x_\gamma y_\gamma} (x_0 - x_1) \right]^2 \\
&+ 2 \left[\sigma^2_{x_\gamma y_\gamma} c_x - \sigma^2_{x-\gamma} \sigma^2_{y_\gamma} c_x + \sigma^2_{x_\gamma y_\gamma} c_x - 2\sigma^2_{x_\gamma} \sigma_{x_\gamma y_\gamma} c_{xy} + (\sigma^2_{x_\gamma})^2 c_y \right] \\
&\left[2\chi^2_{2,1-\alpha} (\sigma^2_{x_\gamma})^2 \sigma_{x_\gamma y_\gamma} c_{xy} - \chi^2_{2,1-\alpha} \sigma^2_{x_\gamma} \sigma^2_{x_\gamma y_\gamma} c_x - \chi^2_{2,1-\alpha} (\sigma_{x_\gamma})^3 c_y \right] \Bigg) \\
&+ (x_b - x_\gamma) 2 \left[2(x_0 - x_1)(\sigma^2_{x_\gamma})^2 \sigma^2_{y_\gamma} - 2\sigma^2_{x_\gamma} \sigma^2_{x_\gamma y_\gamma} (x_0 - x_1) \right] \\
&\left[2\chi^2_{2,1-\alpha} (\sigma^2_{x_\gamma})^2 \sigma_{x_\gamma y_\gamma} c_{xy} - \chi^2_{2,1-\alpha} \sigma^2_{x_\gamma} \sigma^2_{x_\gamma y_\gamma} c_x - \chi^2_{2,1-\alpha} (\sigma^2_{x_\gamma})^3 c_y \right] \\
&+ \left[2\chi^2_{2,1-\alpha} (\sigma^2_{x_\gamma})^2 \sigma_{x_\gamma y_\gamma} c_{xy} - \chi^2_{2,1-\alpha} \sigma^2_{x_\gamma} \sigma^2_{x_\gamma y_\gamma} c_x - \chi^2_{2,1-\alpha} (\sigma^2_{x_\gamma})^3 c_y \right]^2
\end{aligned}$$

$$
\begin{aligned}
&= -(x_b - x_\gamma)^4 \left[2\sigma^2_{x_\gamma} c_{xy} - 2\sigma_{x_\gamma y_\gamma} c_x\right]^2 \left[\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma}\right] \\
&- (x_b - x_\gamma)^3 2 \left[2\sigma^2_{x_\gamma} c_{xy} - 2\sigma_{x_\gamma y_\gamma} c_x\right] \left[2(x_0 - x_1)\sigma^2_{x_\gamma}\sigma_{x_\gamma y_\gamma} - 2(y_0 - y_1)(\sigma^2_{x_\gamma})^2\right] \\
&\left[\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma}\right] \\
&+ (x_b - x_\gamma)^2 \Bigg(\left[2\sigma^2_{x_\gamma} c_{xy} - 2\sigma_{x_\gamma y_\gamma} c_x\right]^2 \left[\chi^2_{2,1-\alpha}(\sigma^2_{x_\gamma})^2\sigma^2_{y_\gamma} - \chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}\sigma^2_{x_\gamma y_\gamma}\right] \\
&- \left[2(x_0 - x_1)\sigma^2_{x_\gamma}\sigma_{x_\gamma y_\gamma} - 2(y_0 - y_1)(\sigma^2_{x_\gamma})^2\right]^2 \left[\sigma^2_{x_\gamma}\sigma^2_{y_\gamma} - \sigma^2_{x_\gamma y_\gamma}\right] \Bigg) \\
&+ (x_b - x_\gamma) 2 \left[2\sigma^2_{x_\gamma} c_{xy} - 2\sigma_{x_\gamma y_\gamma} c_x\right] \left[2(x_0 - x_1)\sigma^2_{x_\gamma}\sigma_{x_\gamma y_\gamma} - 2(y_0 - y_1)(\sigma^2_{x_\gamma})^2\right] \\
&\left[\chi^2_{2,1-\alpha}(\sigma^2_{x_\gamma})^2\sigma^2_{y_\gamma} - \chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}\sigma^2_{x_\gamma y_\gamma}\right] \\
&+ \left[2(x_0 - x_1)\sigma^2_{x_\gamma}\sigma_{x_\gamma y_\gamma} - 2(y_0 - y_1)(\sigma^2_{x_\gamma})^2\right]^2 \left[\chi^2_{2,1-\alpha}(\sigma^2_{x_\gamma})^2\sigma^2_{y_\gamma} - \chi^2_{2,1-\alpha}\sigma^2_{x_\gamma}\sigma^2_{x_\gamma y_\gamma}\right]
\end{aligned}
$$

where x_b is the limit of μ_{x_γ}, the intersection of the two ellipses as $\delta \to 0$, and therefore the x-coordinate of a point on the boundary associated with $\boldsymbol{z}_\delta$. The terms c_x, c_y, and c_{xy} are as written in the theorem.

Now, we subtract the right side of this equation from the left side of this equation, and set the result equal to zero. This implies that the final solutions for the x-coordinates of the boundary point at $\boldsymbol{z}_\gamma$ are the real solutions of the equation

$$A1(x_b - x_\gamma)^4 + A2(x_b - x_\gamma)^3 + A3(x_b - x_\gamma)^2 + A4(x_b - x_\gamma) + A5 = 0,$$

where the terms are as written in the theorem.

□

4.5 Discussion

I have demonstrated a calculation that will give us directly the equation for the boundary of a line segment when using the G-band error model. The equation is valid for both the frequentist and Bayesian cases. It is entirely explicit for the simple case in which the variance at all coordinates is the same, and no covariance exists between points. This is likely a realistic distribution in most situations where a map was digitized by hand from a paper map. It is also realistic in some GPS data situations.

This is an important contribution for several reasons. It is currently very computationally expensive to calculate the boundaries and draw a series of confidence ellipses around a line segment. Consequently, it is difficult to include the G-band error model in GIS software. The explicit formula will help statistical error models become more accessible to the average GIS user. Additionally, the explicit formula makes it possible to draw the region very clearly, which may help to make the meaning of the region more obvious. At present, simple models such as the epsilon band model are still the most prominent ([1]), and by drawing the confidence region without the ellipses, the region is more likely to be understood.

The calculation of this boundary should also aid in the advancement of the G-band model. It is important to be able to use these statistical tools to perform analyses of confidence in high-level GIS operations, such as map overlay operations. By explicitly constructing the confidence region's boundary, we should be able to more easily approach such problems as the point-in-polygon problem or the fractionated polygon problem, as discussed in Section 2.3.4.

The calculation of the boundary region in the general case should be developed further to answer such questions as which roots will be real in various situations, and to determine the calculation of the proper y-coordinate to match the x-coordinate (it could be the same as for the simple case, based on the slope of the line segment, but this should be confirmed). We should also find out if the formula can be simplified further for situations in which some

covariances exist but not others, and situations in which the variances of each coordinate may be different. Additionally, it would be useful to prove that two of the roots of the quartic equation given in Theorem 4.2 will always be complex.

Chapter 5

A Probabilistic Point Deletion Algorithm

5.1 Introduction

When using vector-based data in a geographic information system (GIS) setting, users sometimes do not require the level of information provided by their data. In this case, users can perform line simplification, or line generalization, in order to reduce the computational time required for data operations. Line simplification methods usually involve the systematic deletion of points deemed unnecessary to a particular GIS user. Although manual simplification may provide the best results for an experienced GIS user, this cannot provide standardized results across users, and may be an impossibility in the case of a large data set ([62]). Therefore, computerized algorithms have been developed in order to standardize the process across users.

Currently, the most popular method for point deletion is the Douglas-Peucker, or DP, algorithm ([61], [20], [4]). It is also one of the earliest algorithms developed, having first been published in 1973 ([19]). It is one of only two algorithms currently used in the popular

Arc software (ESRI, Inc., ArcGIS v. 9.1). The ongoing popularity of the DP algorithm demonstrates that it has remarkable utility and accessibility.

However, some GIS users feel that the DP algorithm is not optimal in every situation. For example, the system documentation for the ArcGIS alternative to the DP algorithm, called "Bend Simplify", states that "[Bend Simplify] takes longer to process than [the DP algorithm], but the resulting line is more faithful to the original and shows better aesthetic quality" (ESRI, Inc., ArcGIS v. 9.1). Ebisch ([20]) and others have pointed out that the algorithm is ill-defined in certain situations where points along a line fall outside the reasonable scope of the line's endpoints in the parallel direction. The DP algorithm is also often criticized for inducing error, leaving out essential and recognizable features of lines, and distorting line geometry ([62], [60], [61], [4]).

In this chapter, I present a statistical adjustment to the DP algorithm, based on current methods of error modeling. This method provides statistical justification for point removal, based on the error distribution of the data and a user-chosen confidence level. This method can be generally applicable to any data set, if a pre-set level of confidence (such as 95%) is used, or can be used as a direct alternative to the DP algorithm with a user-specified level of confidence.

In the next section, I will discuss the details of the Douglas-Peucker algorithm. Afterward, I will discuss the details of this adjustment and the ways in which it may be applied.

5.2 The Douglas-Peucker Algorithm

The Douglas-Peucker algorithm was originally developed by D. H. Douglas and T. K. Peucker in 1973 ([19]). Although some adjustments and alternatives have since been developed, the DP algorithm is still the most widely-used point deletion algorithm, and is often used in its original form.

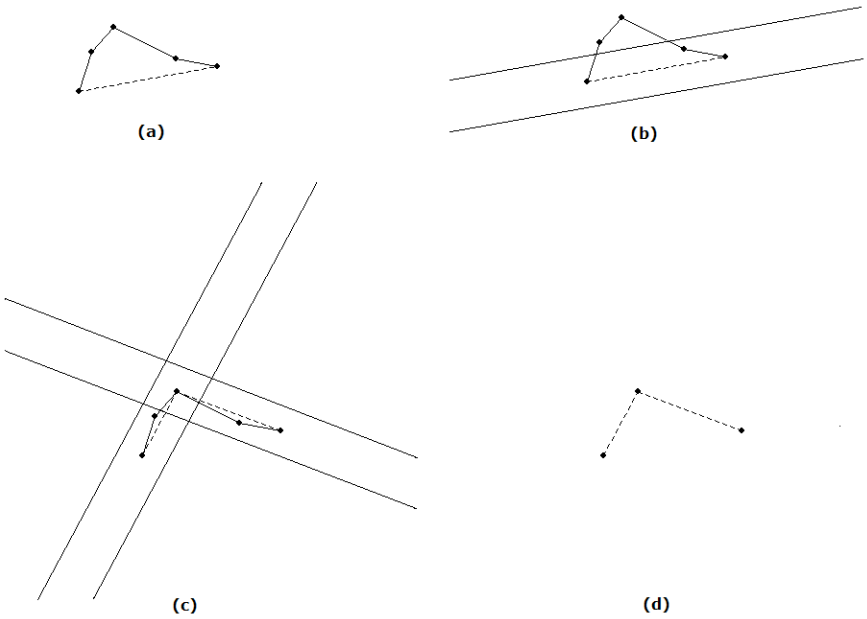

Figure 5.1: Demonstration of the Douglas-Peucker line simplification algorithm.

The DP algorithm is illustrated by Figure 5.1. The algorithm begins by identifying the endpoints of the line we are generalizing, and imposing a line segment connecting the two endpoints directly (a). If all point features in the line are within a perpendicular distance δ from the imposed line segment, all vertices are deleted and the imposed line segment becomes the new line.

If, however, there is at least one point feature in the line that does not fall within distance δ of our imposed line segment (b), we do not delete any points. Instead, we find the point on the original line that has the furthest perpendicular distance from the line segment we have imposed between the two endpoints. We then remove our imposed line segment and add two more, one between each endpoint and the furthest-distance point we have just identified (c). We begin the process again for each of these line segments. For each one, if all points on the line segment between its endpoints fall within distance δ of the line segment, we delete all points and keep the new segment (d). If not, we again identify the furthest point and repeat

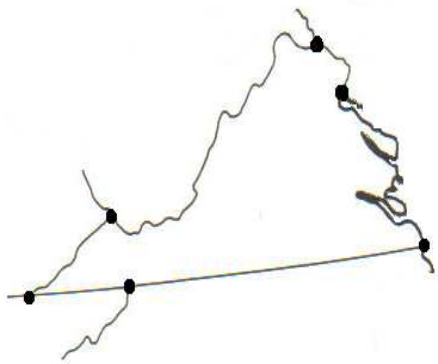

Figure 5.2: Map of Virginia, with arc nodes indicated.

the process. This continues until all points fall within distance δ of a line segment, and all extraneous endpoints have been deleted. The result is the generalized line.

When the DP algorithm is extended to a map containing polygon features, it is usually applied to individual arcs over the entire map. An arc is a portion of a line, and is generally multiple line segments connected by vertices. An arc usually arises as a line between two adjacent polygons. A polygon feature, then, is typically comprised of several arcs that form an enclosed figure. One example that demonstrates the formation of arcs is a map of the United States. Although the state of Virginia is a polygon formed by a closed polyline, the polyline is broken into several arcs. One arc is the border between Virginia and Kentucky. Another arc would be the border between Virginia and Maryland. The points where these arcs intersect are referred to as nodes. Figure 5.2 demonstrates how the polygon of Virginia is actually made up of six individual arcs. When a simplification algorithm is applied to this map, then, each arc would be simplified individually.

The Douglas-Peucker algorithm is generally very useful, and also flexible since users may select their own distance δ. However, some GIS users and authors feel that the DP algorithm may not be optimal in some respects. As I have already discussed, the DP algorithm receives criticism for being aesthetically unpleasing, removing important aspects of the data, and being ill-defined in some situations ([20], [62], [60], [61], [4]).

I next present my adjustment to the algorithm, and demonstrate my alternative. I will also discuss how the alternative copes with some of the criticisms of the DP algorithm.

5.3 Statistical Point Deletion Algorithm

I propose an adjustment to the Douglas-Peucker algorithm, based on the G-band error model for line segments. The DP algorithm determines whether or not points are deleted according to a perpendicular user-specified distance δ from a line segment. I propose instead to create a confidence region around the line segment based on the data's error variance matrix and a user-specified confidence level α. For a demonstration of this algorithm, see Figure 5.3.

The method begins in the same way as the DP algorithm, by imposing a line segment between the endpoints of the line we are simplifying (a). However, instead of considering whether the remaining points on the line fall within a distance δ of this imposed line segment, we determine whether they fall within the confidence region of this line segment, given a normal error distribution and a specified level of confidence. (If using Bayesian methodology, we would use the probability region from the posterior distribution.) If any point falls outside of this region (b), we continue as in the DP algorithm and impose new line segments between the endpoints and the point furthest from the line segment between the endpoints (c). If all points fall inside the confidence region, however, we delete all those points and the line segment becomes part of the generalized line (d). The algorithm completes when no more points are available for testing.

Users may select a level of confidence based on their personal requirements. In general, if a user chooses a generic significance level (such as .05) this method will remove points that many would agree are specified beyond the ability of the data source. On the other hand, a user may choose an α based on the distance at which points will be deleted, making it very similar to the DP algorithm. I will examine each of these interpretations and make a comparison of my algorithm to the DP algorithm.

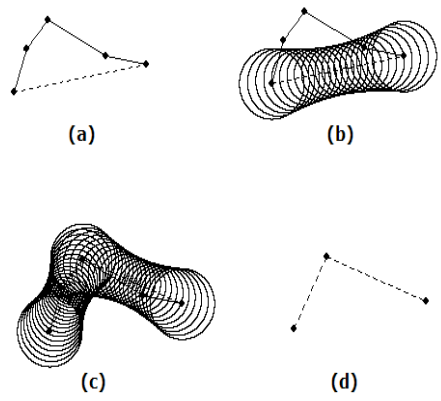

Figure 5.3: Demonstration of the statistical adjustment to the Douglas-Peucker algorithm.

5.3.1 Error-based Point Deletion

Many authors fault current line generalization techniques, such as the DP algorithm, for inducing error in the data. The probabilistic algorithm I have developed here, however, can be used to make data more compliant with the error that is already present. To quote Nicholas Chrisman, "Storage of detail finer than the error inherent in the source document is a means of fooling yourself" ([11]). Therefore, when we remove points that exceed the accuracy of the data, we are not necessarily contributing to the error of the map. Instead, we are simply reducing the data to a more appropriate level, given the error that already exists. I recommend using a standard confidence level to meet this aim, for example, a 95% confidence level. Note that different error structures can be used with different arcs on the same map, there is no restriction that dictates the entire map must be simplified with the same error structure.

The statistical interpretation of this method is clear, given the traditional interpretation of confidence intervals. When a point falls within the 95% confidence ellipse of the current line segment, the point is not statistically different from the line segment at the .05 level of significance, and therefore can be deleted. If, on the other hand, the point falls outside the

confidence region of the current line segment, the point is significantly different from the line segment, and should be included in further analyses.

I demonstrate this usage of the statistical point deletion algorithm in Figure 5.4. The data used in the figure is the outline of a Virginia Tech field in Blacksburg, Virginia, created by taking measurements with a GPS unit. The surveyor recorded points every few feet, which most GIS users immediately recognize results in a plurality of unnecessary points. By using the error distribution of the data and the statistical point deletion algorithm, we should be able to appropriately delete points that do not contribute to the map.

The field is largely a green clearing, but is partially composed of a bramble patch. I used this information to appropriately divide the field polygon into two arcs (a), which we then simplify using the statistical method. We start with the smaller arc, and create a line segment between the two endpoints, then draw the confidence region around that line segment (b). We base our confidence region on an error standard deviation of 3.5 at all endpoint coordinates, no correlation between coordinates, and a confidence level of .95 (α=.05). Parts (c)-(g) demonstrate the complete simplification of the first arc. This process was repeated on the second arc, and the final generalized polygon is presented in part (h).

Another option, which would result in a more conservative final map (more points removed), would be to take advantage of adjustments for multiple hypothesis tests. When testing n points for significant differences from the line segment, one could use the adjusted confidence level $\alpha/(n-2)$. The interpretation of all the points falling within this confidence region (for a .05 level of significance) is that we are confident at the 95% level that these points are simultaneously not significantly different from the line segment. When at least one point falls outside of the adjusted confidence region, the interpretation is that we are not confident at the 95% level that all of the points are not significantly different from the line segment. This will result in larger confidence regions and therefore the deletion of more points. The idea behind this type of test is that when many points have been taken, some of them will fall outside of the single-point confidence region due to random error. Simultaneous testing

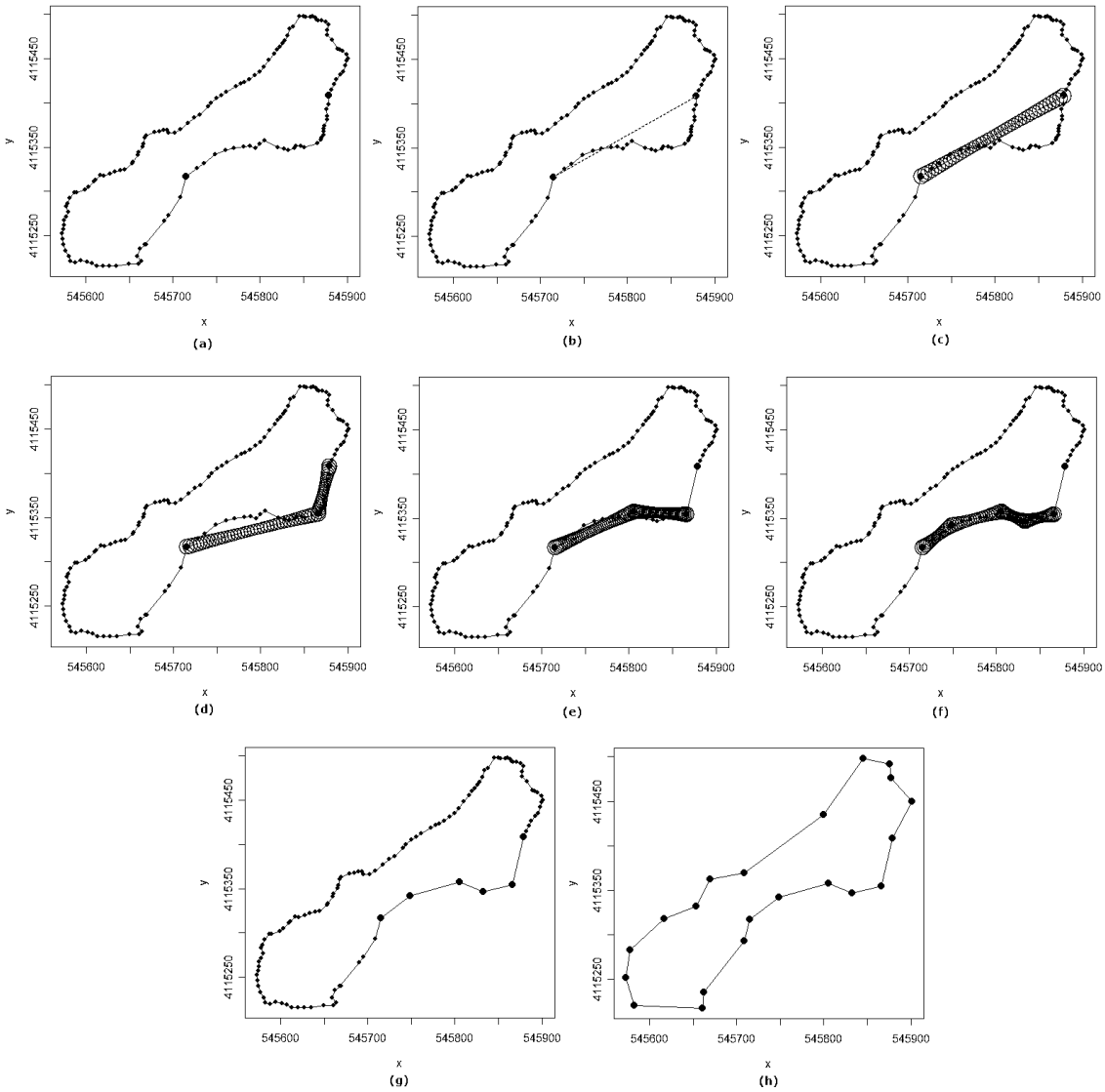

Figure 5.4: Demonstration of the statistical point deletion algorithm.

takes account of this fact.

It is clear from Figure 5.4 that the error-based point deletion algorithm results in a shape that many would agree represents the field well, but does not contain redundant points. The fact that we based this deletion on the error in the data implies that, rather than add to the error already present, we instead took account of this error and displayed the data more suitably.

5.3.2 General Point Deletion and Comparison to the Douglas-Peucker Algorithm

In general, the adjustment to the DP algorithm can be used at any confidence level with the data's error structure to produce any size confidence region. I compare this method to the DP algorithm when the confidence ellipses at the endpoints have a radius equal to the parameter δ. The alternative avoids several criticisms of the traditional DP algorithm. In addition to the issue addressed in the previous section regarding error in the data, the statistical algorithm avoids a common problem in the implementation of the DP algorithm, and also retains some key features of line segments that the DP algorithm would delete.

A common problem with the Douglas-Peucker algorithm is that of interpretation. The original paper claimed that a point should not be deleted when it does not fall within "the greatest perpendicular distance between it and the straight line defined by the anchor and the floater" ([19]). This definition leads to the deletion of very important features when interpreted literally. Figure 5.5 part (a) demonstrates this problem. For more discussion of this issue, see Ebisch ([20]). However, because the alternative presents a closed region in which points are to be deleted, this avoids the problem altogether. See Figure 5.5 part (b) for this resolution.

The alternative preserves other features that are important to the appearance of a line as well. For one thing, because the confidence region is smaller in the middle than at the

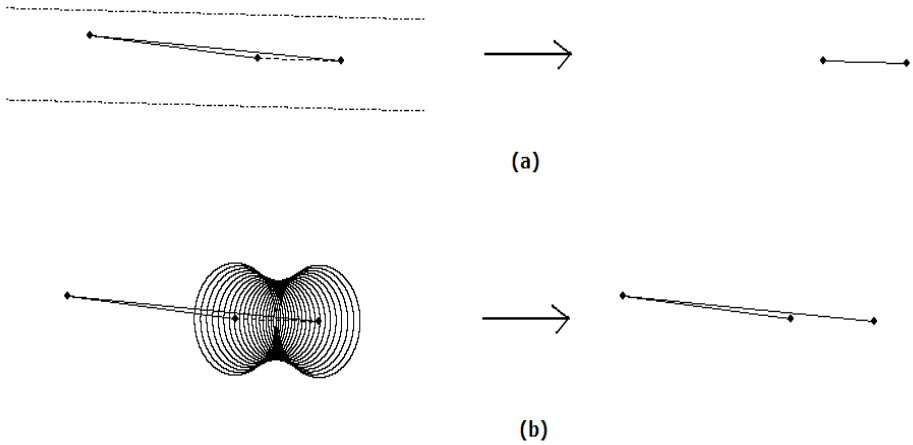

Figure 5.5: A problem with the Douglas-Peucker algorithm; note that an important bend in the line is removed with the DP algorithm (a), but remains with the probabilistic alternative (b).

endpoints, it will preserve points that are located toward the middle of the line segment that form sharp angles. Depending on the application this information may be important to retain. See Figure 5.6 for a demonstration.

5.4 Discussion

I have presented here an alternative to the Douglas-Peucker algorithm with statistical justification, based on the error structure of the data. The DP algorithm is somewhat arbitrary, due to the inclusion of the parameter δ which is entirely up to the map user. The alternative I have presented here, however, takes advantage of information from the data to perform line simplification. It is still flexible, however, in allowing the user to choose the confidence level at which they perform the point deletion algorithm. As I demonstrated in section 5.3, the algorithm can be used to reduce the map to an acceptable number of points, given the error structure in the data. It is flexible, since one can test using single-point confidence regions

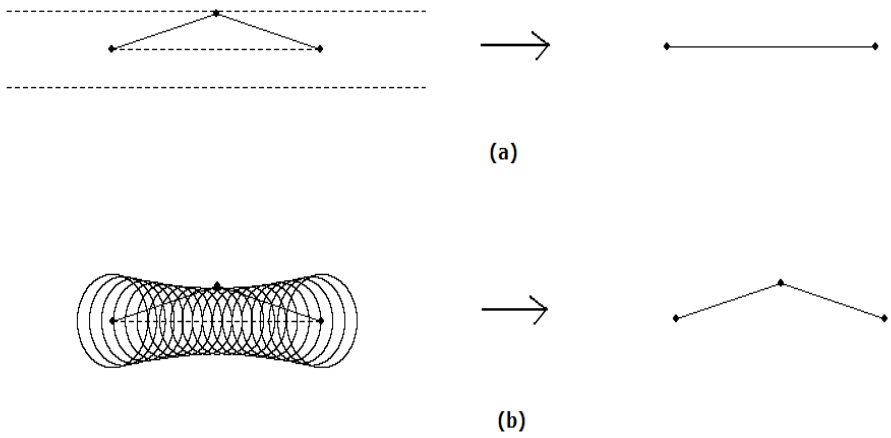

Figure 5.6: The statistical alternative to the DP algorithm retains points toward the centre of the line segment (b), whereas the DP algorithm does not (a).

or simultaneous multiple-point confidence regions. It can also be used more generally, as the DP algorithm, to remove points in accordance with an individual user's preferences.

This method also has the advantage of being able to incorporate error information from other sources besides the current map, through the use of a Bayesian prior in forming error regions around line segments. This method adds additional flexibility to the process, since it can replicate the results of using the G-band model only, or it can produce a more accurate error region. When a more accurate error region is used, this will obviously result in less deletion of important points, and more certainty about the usefulness of points that remains.

This algorithm also presents several advantages in addition to statistical justification. The algorithm retains certain important features of lines. It also solves a common problem in the implementation of the DP algorithm. Additionally, the algorithm can be used in certain situations to simplify the data according to the error involved, rather than exacerbating the error. This is a common complaint about the DP algorithm and other current point deletion algorithms.

There are several disadvantages to this method, however. For one, while the G-band error model is prominent in literature on vector map accuracy, it is probably not familiar to very many GIS users. This will certainly hamper the implementation of this statistical point deletion algorithm. The statistical adjustment I have presented here also does not answer every criticism about the DP algorithm. There are many situations in which they will delete similar points, and so this will not appeal to those in favor of radically different algorithms.

Chapter 6

Conclusion

6.1 Discussion

I began with a literature review in both vector and raster data, which revealed varying degrees of work in the field of GIS error modeling at present. Some researchers have already developed strong probabilistic error models for vector data, in particular Wenzhong Shi et al.'s bivariate normal model. On the other hand, models for raster data have turned largely toward fuzzy logic and fuzzy representations. While some statistical models do exist here, they are largely disjoint from one another and there is no method in which the field is united.

Next I presented my own work on error modeling. So far, Bayesian methodology is a tool that has been largely ignored in the literature on GIS error. The Bayesian concept is valuable to this work for several reasons. Primarily, Bayesian analysis is useful in situations where there is not a lot of data available; this is certainly the case in many GIS applications, where very few samples (or possibly only one) are available. Additionally, a realistic prior distribution should not be out of reach in a geographical discipline, where expert knowledge of location and many previous maps are often available.

My work in vector models builds on existing error models. I used the bivariate normal model

for points, and added the concept of a prior and posterior distribution. Many of the results, such as confidence ellipses around points and confidence regions around line segments, are parallel to the results already found in the frequentist case. I also calculated the posterior mean and variance of points on line segments for several specific instances that may be reflected in data, based on correlation between endpoints and variance of coordinates.

Additionally, I found an explicit formula for the boundary of the confidence region for a line segment for a simple error variance structure. I have also begun to develop a method for calculating the boundary directly for a general error structure. These will aid greatly in future probability calculations. These should also support the eventual implementation of statistical error models in to software programs. This is an essential step toward the regular use of statistical models, since it would make them available to the average GIS user.

My final vector data result is a probabilistic method for deleting points when simplifying a line. This algorithm is an alternative to the existing Douglas-Peucker algorithm, having the advantage that statistical theory justifies the deletion of points. I also demonstrated some other advantages that result from the properties of this algorithm.

Each of these accomplishments will make it easier for GIS users to communicate about map accuracy. As the field of GISs continues to expand, it is more important than ever that GIS users continue to gain tools for displaying and analyzing error. I now conclude with a discussion of possibilities for future research in both vector and raster data.

6.2 Future Work

The field of GIS is constantly expanding, as is the need for error analysis and interpretation. There are therefore many opportunities for important work in this field. I discuss several of those opportunities here.

6.2.1 Application of the Statistical Error Model

I have developed the statistical model to a point where it is now easy to use for displaying error found in a vector data set. This still ignores the problem, however, of using this information to learn about how the error is compounded through combining maps and performing map operations. As I discussed in section 2.3.4 as part of the literature review, there are many problems that extend well beyond the basic display of error.

Many problems, such as the point-in-polygon problem and the fractionated polygon problem, come from the overlay of two or more maps. Some solutions have been posed to these problems, but as I previously discussed, have presented complicated answers and/or not answered the questions directly.

In accordance with the rules of probability, the only way to correctly determine the exact probability of one object having a particular relationship to another is to integrate appropriately over the joint distribution of the two objects in the region in which this relationship occurs. This is what Cheung et al. ([10]) were referring to when they used the equation

$$Pr(P \in A) = \int\int_{(x,y)\in A} f_P(x,y)dxdy,$$

where $f_P(x,y)$ is the pdf of P, to solve the point-in-polygon problem. As their paper showed, this is not an easy equation to solve. There are other alternatives though, since the most likely result of these determinations is to choose a course of action based on whether or not the probability of some relationship is greater than a certain cutoff level.

As of now, the basis for modeling both line segments and polygons is the line segment confidence region. Therefore, the most basic problem of interest may be determining the probability that a point has a specific relationship to a line segment. Although there is still a lot of work to be done, one way to answer this question might be to examine the confidence regions of both the point and the line segment, and determine at what level(s) of confidence (or probability, in the Bayesian case) they overlap. For example, if the error in the point and the line segment are independent (a reasonable assumption when they come from different

data sources), we can create a 0.9745 confidence region around each. If they overlap, we can say the evidence that the point falls on a particular side of the line segment is not significant at the 95% level of confidence (0.9745 x 0.9745 $\approx$ 0.95). In the Bayesian case, we could claim that there is less than a 95% probability that the point is on a particular side of the line segment. Although there are additional complications (for example, we could use a 0.9595 and a 0.99 confidence region instead of two 0.9745 regions), this may be a good start to answering the problem in a realistic manner. The calculation of the confidence region boundary for the line segment I presented here should aid in this pursuit.

6.2.2 Further Theoretical Development of the Statistical Error Model

There are several ways in which the basic theory of the statistical error model (both the G-band and Bayesian versions) can be developed further. For example, in my work and previous work, I have assumed that the error variance of the data is known. This may be acceptable in some situations, but as in many other areas of statistics, is often a false assumption. When assuming a normal distribution with known variance, as I did here, the most appropriate confidence region around a point is created through the use of the chi-square distribution. If a t-distribution is more appropriate, however, the chi-square confidence region is a poor approximation when there is only a small amount of data available. In this case, a confidence region from the F-distribution may be much more realistic. Although, as I have stated many times, there is not often more than one data point from which to estimate a standard deviation, it would be well worth it to discover what the use of a t-distribution would change in the results.

Another development would be to expand the correlation structure between points to allow for correlation between any pair of points on a polygon. It is well known in spatial disciplines that data from two nearby sources are often related, and that relationship increases as the distance between points decreases. Additionally, it may be useful to incorporate longitudinal

correlation structures, to take into account when data points were collected consecutively.

The Bayesian method could also be developed in many respects. While I discussed the possibility and provided one example of using non-normal priors, I did not explore this option in detail. The method I presented here is also not an objective application of Bayesian methodology, but an empirical one. In order to make this more of an objective application, it is possible that one could put a hyper prior distribution on the mean and/or variance of the prior distribution, which could serve to add additional uncertainty. As the model currently stands, I treat the prior distribution as though it is a known fact, which is not entirely realistic since it will generally come from previously created GIS maps. I have also not discussed using a prior distribution for the variance of the data, which could help provide more realistic results.

6.2.3 The Missing Data Problem

One very important topic that I did not cover are situations in which points are not accurately represented. The G-band model is certainly appropriate if we can guarantee all the true points of the feature were included in the map; however, if points are missing due to omission or curvature in line and polygon features, this model is not correct. The G-band confidence region around a line segment is smaller in the middle than at the endpoints, implying greater confidence toward the center of the line segment where no data was taken. Therefore, many GIS users feel that the region should not be concave. This is a valid concern.

There are many ways one could possibly cope with this. One principle to keep in mind, however, is that we cannot create a realistic model without some further assumptions or knowledge of the data. Therefore, our ability to model missing points will only be as good as the information we are presented with.

One possibility is to assume we have a more detailed second map available that covers a sub-region of our large map. While this may not be very realistic in many cases, this could

possibly provide very good information. Simply by aligning the two maps and recording the average number of vertices excluded from the larger map, the average distance and average angles create by those vertices, and any other possibly pertinent information, I could develop a distribution of missing points for the whole map. This of course involves some amount of extrapolation, and should therefore be used with caution. The distribution of missing points on a river, for example, is probably not the same as the distribution for a road.

Another possibility is to act as though some basic point deletion algorithm was used on the map at some point prior to its creation. We could then learn about the parameters of this point deletion algorithm by removing points one at a time from the map and imposing a line segment between the neighbors of each point. Take measurements such as the distance from the removed point to the new line segment, the angle formed by the removed point, and its relative distance from each endpoint, and create a distribution for the missing points based on that information.

A final method involves the least amount of outside knowledge, and is therefore very flexible, but also the least subjective. In order to increase the size of the confidence region toward the middle of the line segment, we include an artificial "variance inflation factor" when calculating the confidence ellipses at each point along the line segment. We could base this factor on the position of the point relative to the endpoints of the line segment, so that the points closest the endpoints have the smallest inflation factor and the points closest to the middle have the largest. This could come from any function, including the inverse sum of the squares of the distances to the endpoints, or the minimum of the two distances. The following equations, where f_t is the variance inflation factor, show how this would increase the confidence region in the frequentist model (the Bayesian model is equivalent).

$$
\begin{aligned}
&(\overline{\boldsymbol{z}}_t - \boldsymbol{\mu}_t)' \left(f_t \boldsymbol{\Sigma}_t\right)^{-1} (\overline{\boldsymbol{z}}_t - \boldsymbol{\mu}_t) = \chi^2_{2,1-\alpha} \\
&\Rightarrow (\overline{\boldsymbol{z}}_t - \boldsymbol{\mu}_t)' \frac{1}{f_t} \boldsymbol{\Sigma}_t^{-1} (\overline{\boldsymbol{z}}_t - \boldsymbol{\mu}_t) = \chi^2_{2,1-\alpha} \\
&\Rightarrow (\overline{\boldsymbol{z}}_t - \boldsymbol{\mu}_t)' \boldsymbol{\Sigma}_t^{-1} (\overline{\boldsymbol{z}}_t - \boldsymbol{\mu}_t) = f_t \chi^2_{2,1-\alpha}
\end{aligned}
$$

6.2.4 Raster Data Error Models

I present here a general outline to pave the way for future work. Molsen et al. ([45]) presented a logistic mixed model regression for predicting the probability that a cell is classified correctly (review Section 2.4 for details). The model they used was $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is a vector of responses (agreement between map and ground cover = 1, disagreement = 0), and $\boldsymbol{\mu}$ is a vector of probabilities. These values are based on a set of fixed effects $\boldsymbol{X}$ and random effects $\boldsymbol{Z}$ at levels that a user has included in the model through a logit link function, $g(\mu) = log\{\mu/(1-\mu)\} = \boldsymbol{X\beta} + \boldsymbol{Z\nu}$.

I propose a similar model for prediction of error probabilities in individual cells. This model may or may not include a vector of random effects. The primary difference between the model presented here and the current one is that instead of a traditional frequentist model, I advocate the addition of a Bayesian prior distribution for the set of parameters $\boldsymbol{\beta}$. As with error models for vector data, I am in favor of Bayesian methodology because it is advantageous in situations with little available data, and in many cases we can obtain expert knowledge of the relationship between error and map variables to help determine a prior distribution.

There are many methods in the literature of obtaining a prior distribution for the parameters of a logistic regression model. Bedrick et al. ([5]) provide a general discussion of the issues at hand when dealing with a Bayesian binomial regression model. Gelman et al. ([23]) give an example of an independent, uniform distribution in each of the parameters, which they recommend for use as a noninformative distribution when not much is known about the parameters. Ibrahim and Laud ([33]) compare the use of several types of priors, including the uniform, normal, and Jeffreys's prior distribution. Beyond these examples, many more ideas are forthcoming in the literature, such as hierarchical models (in which more than one stage of sampling occurs) and empirical Bayes analysis (in which a prior is created from a partial sample of the data). I should point out that in many cases, simulation studies are needed to analyze properties of posterior distributions since many resulting from these types

of procedures do not have closed forms.

There are many issues to consider for raster data modeling. I have suggested a Bayesian logistic regression model to predict inappropriately-classified cells, but have yet to determine general guidelines such as a method for choosing a prior distribution. Even after we have done this and the model is explicit, it is certainly not the only consideration remaining. For example, for the model to be useful it is necessary that we determine its impact on error propagation in various situations, such as map overlay operations and map transformations.

Even beyond this, there are other ways to model raster data. For example, it may be useful to determine a model for predicting the proportion of each map class (i.e., land cover or soil type) within each cell. Another significant problem is modeling of continuous data, which cannot be modeled through a logistic regression because there is not a binary response variable.

6.2.5 Integration with GIS Software

Perhaps the most important step toward implementing the methods I have discussed here is to make them available to GIS software users. This is an area in which statisticians will have to work closely with the GIS community in order to make sure that solutions are presented appropriately, and in a manner that will make them attainable to the average GIS user. It is important to keep in mind that the end goal is to put these ideas in to practice, which is impossible without cooperation between the statistical and GIS communities.

Bibliography

[1] Alesheikh, A. A., Blais, J. A. R., Chapman, M. A., and Karimi, H. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 24: Rigorous Geospatial Data Uncertainty Models for GISs, pp. 195–202.

[2] Arbia, G., Griffith, D., and Hailing, R. Error propagation modelling in raster GIS: overlay operations. *International Journal of Geographical Information Science 12*, 2 (1998), 145–167.

[3] Bastin, L., Wood, J., and Fisher, P. F. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 18: Visualization of Fuzzy Spatial Information in Spatial Decision-Making, pp. 151–156.

[4] Beard, M. K. Theory of the cartographic line revisited / implications for automated generalization. *Cartographica 28*, 4 (1991), 32–58.

[5] Bedrick, E. J., C. R., and Johnson, W. Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician 51*, 3 (1997), 211–218.

[6] Blakemore, M. Part 4: Mathematical, algorithmic and data structure issues: Generalisation and error in spatial data bases. *Cartographica 21*, 2 (1984), 131–139.

[7] Bolstad, P. V., and Smith, J. L. Errors in GIS: Assessing spatial data accuracy. *Journal of Forestry 90*, 11 (1992), 21–29.

[8] Carmel, Y., and Dean, D. J. Performance of a spatio-temporal error model for raster datasets under complex error patterns. *International Journal of Remote Sensing - http://www.tandf.co.uk/journals* (2004).

[9] Cheung, C. K., and Shi, W. Positional error modeling for line simplification based on automatic shape similarity analysis in GIS. *Computers and Geosciences 32*, 4 (2006), 462–475.

[10] Cheung, C. K., Shi, W., and Zhou, X. A probability-based uncertainty model for point-in-polygon analysis in GIS. *GeoInformatica 8*, 1 (2004), 71–98.

[11] Chrisman, N. R. On storage of coordinates in geographic systems. *Geo-Processing 2* (1984), 259–170.

[12] Chrisman, N. R. *Accuracy of Spatial Databases.* Taylor & Francis, 1989, ch. 2: Modeling error in overlaid categorical maps, pp. 21–34.

[13] Chrisman, N. R. *Introductory Readings in Geographic Information Systems.* Taylor & Francis, 1990, ch. 22: The accuracy of map overlays: a reassessment, pp. 309–320.

[14] Chrisman, N. R. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 3: Speaking Truth to Power: An Agenda for Change, pp. 27–31.

[15] Chrisman, N. R., and Lester, M. A diagnostic test for error in categorical maps. In *Technical Papers - 1991 ACSM-ASPRS Annual Convention* (1991), vol. 6, ACSM-ASPRS, pp. 330–348.

[16] Clarke, K. C. *Getting Started with Geographic Information Systems*, fourth ed. Prentice Hall Pearson Education, Inc., 2003.

[17] Coppock, J. T., and Rhind, D. W. *Geographical Information Systems.* Longman Scientific and Technical, 1991, ch. 2: An Overview and Definition of GIS, pp. 9–20.

[18] Czaplewski, R. L. *Quantifying Spatial Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 7: Accuracy Assessments and Areal Estimates Using Two-Phase Stratified Random Sampling, Cluster Plots, and the Multivariate Composite Estimator, pp. 79–100.

[19] Douglas, D. H., and Peucker, T. K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer 10*, 2 (1973), 112–122.

[20] Ebisch, K. A correction to the Douglas-Peucker algorithm. *Computers and Geosciences 28*, 8 (2002), 995–997.

[21] Fisher, P. F. Simulation of the uncertainty of a viewshed. In *Technical Papers - 1991 ACSM-ASPRS Annual Convention* (1991), vol. 6, ACSM-ASPRS, pp. 205–217.

[22] Fisher, P. F. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 17: Set Theoretic Considerations in the Conceptualization of Uncertainty in Natural Resource Information, pp. 147–150.

[23] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, second ed. Chapman & Hall/CRC, 2003.

[24] Goodchild, M. F. *Accuracy of Spatial Databases.* Taylor & Francis, 1989, ch. 10: Modeling error in objects and fields, pp. 107–113.

[25] Goodchild, M. F., Guoqing, S., and Shiren, Y. Development and test of an error model for categorical data. *International Journal of Geographical Information Systems 6*, 2 (1992), 87 – 104.

[26] Greve, M. B., and Greve, M. H. Visualization of fuzzy boundaries of geographic objects.

[27] Griffith, D. A., Hailing, R. P., and Arbia, G. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 2:

Uncertainty and Error Propagation in Map Analyses Involving Arithmetic and Overlay Operations: Inventory and Prospects, pp. 11–25.

[28] Hess, G. R., and Bay, J. M. Generating confidence intervals for composition-based landscape indexes. *Landscape Ecology 12*, 5 (1997), 309–320.

[29] Hlavka, C. A. *Quantifying Spatial Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 13: Statistical Models of Landscape Pattern and the Effects of Coarse Spatial Resolution on Estimation of Area with Satellite Imagery, pp. 161–170.

[30] Horttanainen, P., and Virrantaus, K. Uncertainty evaluation of military terrain analysis results by simulation and visualization. In *Proceedings of the 12th International Conference on Geoinformatic - Geospatial Information Research: Bridging the Pacific and Atlantic* (2004), Geoinformatics.

[31] Hughes, M., Bygrave, J., Bastin, L., and Fisher, P. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 38: High Order Uncertainty in Spatial Information: Estimating the Proportion of Cover Types Within a Pixel, pp. 319–329.

[32] Hurricane forecasters introduce probability map. Associated Press. South Florida Sun Sentinel at http://ap.tbo.com/ap/florida/MGBIZ8TMG6E.html, Mar. 18, 2005.

[33] Ibrahim, J. G., and Laud, P. W. On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association* (1991).

[34] Jackson, M. J., and Woodsford, P. A. *Geographical Information Systems.* Longman Scientific and Technical, 1991, ch. 17: GIS Data Capture Hardware and Software, pp. 239–249.

[35] Jiang, B. Visualization of fuzzy boundaries of geographic objects. *Cartography 27*, 2 (1998), 41–46.

[36] Jordan, G. J., Fortin, M.-J., and Lertzman, K. P. Assessing spatial uncertainty associated with forest fire boundary delineation. *Landscape Ecology 20*, 6 (2005), 719–731.

[37] Krivoruchko, K., and Gotway-Crawford, C. A. *Spatial Analysis and Modeling.* ESRI Press, 2005, ch. 4: Assessing the Uncertainty Resulting from Geoprocessing Operations, pp. 67–92.

[38] Laviolette, M., Seaman, J. W., J., Barrett, J. D., and Woodall, W. A probabilistic and statistical view of fuzzy methods. *Technometrics 37*, 3 (1995), 249–292.

[39] Leung, Y., Ma, J.-H., and Goodchild, M. F. A general framework for error analysis in measurement-based GIS, part 2: The algebra-based probability model for point-in-polygon analysis. *Journal of Geographical Systems 6*, 4 (2004), 355–379.

[40] Leung, Y., Ma, J.-H., and Goodchild, M. F. A general framework for error analysis in measurement-based GIS part 3: Error analysis in intersections and overlays. *Journal of Geographical Systems 6*, 4 (2004), 381–402.

[41] Leung, Y., and Yan, J. Point-in-polygon analysis under certainty and uncertainty. *Geoinformatica 1*, 1 (1997), 93–114.

[42] Maguire, D. J. *Geographical Information Systems.* Longman Scientific and Technical, 1991, ch. 1: An Overview and Definition of GIS, pp. 9–20.

[43] Maguire, D. J., and Dangermond, J. *Geographical Information Systems.* Longman Scientific and Technical, 1991, ch. 21: The Functionality of GIS, pp. 319–335.

[44] Marble, D. F. *Introductory Readings in Geographic Information Systems.* Taylor & Francis, 1990, ch. 1: Geographic information systems: an overview, pp. 8–17.

[45] Molsen, G. G., Cutler, D. R., and Edwards, T. C. *Quantifying Spatial Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 3: Generalized Linear Mixed Models for Analyzing Error in a Satellite-Based Vegetation Map of Utah, pp. 37–43.

[46] Mowrer, H. T. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 1: Accuracy (Re)assurance: Selling Uncertainty Assessment to the Uncertain, pp. 3 – 10.

[47] Openshaw, S. *Accuracy of Spatial Databases.* Taylor and Francis, 1989, ch. 23: Learning to live with errors in spatial databases, pp. 263–276.

[48] Rencher, A. C. *Linear Models in Statistics.* Wiley Inter-Science, 2000.

[49] Robert, C. P., and Casella, G. *Monte Carlo Statistical Methods.* Springer-Verlag, 1999.

[50] Shi, W. A generic statistical approach for modelling error of geometric features in GIS. *International Journal of Geographic Information Science 12*, 2 (1998), 131–134.

[51] Shi, W., Cheung, C.-K., and Tong, X. Modelling error propagation in vector-based overlay analysis. *ISPRS Journal of Photogrammetry and Remote Sensing 59*, 1-2 (2004), 47–59.

[52] Shi, W., Cheung, C. K., and Zhu, C. Modelling error propagation in vector-based buffer analysis. *International Journal of Geographical Information Science 17*, 3 (2003), 251–271.

[53] Shi, W., Ehlers, M., and Tempfli, K. Analytical modelling of positional and thematic uncertainties in the integration of remote sensing and geographical information systems. *Transactions in GIS 3*, 2 (1999), 119–136.

[54] Shi, W., and Liu, W. A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science 14*, 1 (2000), 51–66.

[55] Stehman, S. V. *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources.* Ann Arbor Press, 1999, ch. 5: Alternative Measures for Comparing Thematic Map Accuracy, pp. 45–51.

[56] Sunila, R., Laine, E., and Kremenova, O. Fuzzy model and kriging for imprecise soil polygon boundaries. In *Proceedings of the 12th International Conference on Geoinformatic - Geospatial Information Research: Bridging the Pacific and Atlantic* (2004), Geoinformatics.

[57] Tong, X., Shi, W., and Liu, D. An error model of circular curve features in gis. In *Proceedings of the 11th ACM international symposium on Advances in geographic information systems* (2003), Association for Computing Machinery.

[58] Veregin, H. *Accuracy of Spatial Databases.* Taylor & Francis, 1989, ch. 1: Error modeling for the map overlay operation, pp. 3–18.

[59] Veregin, H. Developing and testing of an error propagation model for GIS overlay operations. *International Journal of Geographic Information Systems 9*, 6 (1995), 595–619.

[60] Veregin, H. Line simplification, geometric distortion, and positional error. *Cartographica 36*, 1 (1999), 25–39.

[61] Veregin, H. Quantifying positional error induced by line simplification. *International Journal of Geographical Information Science 14*, 2 (2000), 113–130.

[62] Visvalingam, M., and Whyatt, J. D. Cartographic algorithms: Problems of implementation and evaluation and the impact of digitising errors. *Computer Graphics Forum IO* (1991), 225–235.

[63] Weibel, R., and Heller, M. *Geographical Information Systems.* Longman Scientific and Technical, 1991, ch. 19: Digital Terrain Modelling, pp. 269–297.