Making Sense Out of Uncertainty in Geospatial Data

Andrew Scott Foy

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

Doctor of Philosophy
In
Geospatial and Environmental Analysis

Laurence W. Carstensen (Chair)
James B. Campbell
Randel L. Dymond
Stephen P. Prisley

July 29, 2011
Blacksburg, VA

Keywords: GIS, uncertainty, spatial data fusion, error-band framework, GIS tool, overlays

Copyright 2011, Andrew Scott Foy
Making Sense Out of Uncertainty in Geospatial Data

Andrew Scott Foy

ABSTRACT

Uncertainty in geospatial data fusion is a major concern for scientists because society is increasing its use of geospatial technology and generalization is inherent to geographic representations. Limited research exists on the quality of results that come from the fusion of geographic data, yet there is extensive literature on uncertainty in cartography, GIS, and geospatial data. The uncertainties exist and are difficult to understand because data are overlaid which have different scopes, times, classes, accuracies, and precisions. There is a need for a set of tools that can manage uncertainty and incorporate it into the overlay process. This research explores uncertainty in spatial data, GIS and GIScience via three papers. The first paper introduces a framework for classifying and modeling error-bands in a GIS. Paper two tests GIS users’ ability to estimate spatial confidence intervals and the third paper looks at the practical application of a set of tools for incorporating uncertainty into overlays. The results from this research indicate that it is hard for people to agree on an error-band classification based on their interpretation of metadata. However, people are good estimators of data quality and uncertainty if they follow a systematic approach and use their average estimate to define spatial confidence intervals. The framework and the toolset presented in this dissertation have the potential to alter how people interpret and use geospatial data. The hope is that the results from this paper prompt inquiry and question the reliability of all simple overlays. Many situations exist in which this research has relevance, making the framework, the tools, and the methods important to a wide variety of disciplines that use spatial analysis and GIS.
Acknowledgments

Graduate school was an interesting chapter in my life, and one that I am glad to be finishing. Try to imagine four-plus years studying uncertainty. The subject matter leads one to question many things. In the end, the journey is what stands out and there are so many people that have helped along the way, some with encouraging words and others brute force. Without the help of my colleagues, friends and family, I would have not been able make it this far. My education and graduate career have been a joy.

First, I would like to thank my graduate committee for their insight and patience. Their combined knowledge was critical to the progression of this work. It was very difficult teaching fulltime while working to finish a dissertation. I am thankful for a committee that understood the balance between research, teaching, and life. My committee chair, Dr. Carstensen, brought me on board as a programmer with Dr. Prisley on a National Geospatial Intelligence Agency grant to enhance geospatial data fusion. It sounds fancy, but the reality is that I was a programmer that became a geographer and found a niche. My work with them was like feeding a hungry dog. It was a great environment for creativity and problem solving. The other committee members also need mention because it was a team effort and truth be told we needed outside opinion on the certainty of our uncertainty ideas. Dr. Campbell’s vast knowledge in remote sensing opened many doors that extend beyond this research, specifically in balloon-based remote sensing. Dr. Dymond’s expertise in civil and environmental engineering helped mesh real world problems with academic research. They helped mold me into what I am today.

It would be improper to mention my committee without reference to the Geography Department at Virginia Tech. The lower floor of Major Bill was like a second home for a long time. I thank all of the faculty and students because they are the reason I followed this path.

The Geospatial Science Department at Radford University was also a tremendous resource during this process. I received my undergraduate degree in Geography from RU and had the wonderful opportunity to return there and teach. The faculty members at RU have gone above the call of duty as mentors. Dr. Kuennecke and Dr. Woodward have been so helpful that I cannot express my gratitude. It was with their guidance that I embarked down the rewarding path of academia, found Virginia Tech, and became a professor of Geospatial Science. It is rare that your undergraduate advisors continue to help you in your graduate schooling and professorship.

Lastly, I need to thank the Family (parents, girlfriend, friends, and the animals). I want to thank my girlfriend Cynthia who endlessly took care of me and most of the time managed to be kind, even when I was frustrated and stressed. She has been great, convincing me I would be done soon and supporting my decision to continue my graduate education. I also want to thank my parents, Mark and Sue, who instilled confidence and motivation. Finally, I want to thank the rest of my family and friends who have all supported my work and graduate career. Without all of this support, I would not be writing this.
# Table of Contents

ABSTRACT .................................................................................................................. ii  
Acknowledgments ....................................................................................................... iii  
List of Tables ............................................................................................................... vi  
List of Figures ............................................................................................................ vii  

Chapter 1: Introduction ............................................................................................... 1  
  1. Research Context and Justification ................................................................. 1  
  2. Dissertation Organization and Research Questions ...................................... 3  
  3. References ......................................................................................................... 6  

Chapter 2: An evaluation of a GIS framework for generating spatial confidence  
intervals to enhance geospatial data fusion ................................................................. 7  
  Abstract .................................................................................................................... 7  
  1. Introduction ........................................................................................................ 8  
  2. Review ................................................................................................................ 12  
    2.1 Modeling Error and Uncertainty ................................................................. 12  
    2.2 Data Quality Standards and Metadata ..................................................... 17  
    2.3 Relevant Statistics ...................................................................................... 20  
  3. Methods ............................................................................................................. 21  
    3.1 The Framework ........................................................................................... 21  
    3.2 Methodology for the Experiment ............................................................... 25  
  4. Results ............................................................................................................... 28  
  5. Discussion and Conclusions ............................................................................. 32  
  6. References ......................................................................................................... 35  

Chapter 3: GIS and statistical analysis of digitized data’s spatial confidence intervals  
for a forest clearing ..................................................................................................... 39  
  Abstract .................................................................................................................... 39  
  1. Introduction ........................................................................................................ 40  
  2. Uncertainty Basics ........................................................................................... 43  
    2.1 GIS vs. Cartography ................................................................................... 46  
    2.2 GIScience .................................................................................................... 50  
    2.3 Spatial Confidence Intervals, Errors, and Epsilon Bands ....................... 51  
  3. Methods ............................................................................................................. 53  
    3.1 Study Area .................................................................................................. 53  
    3.2 Experimental Design ............................................................................... 55  
    3.3 Data Analysis ............................................................................................. 57  
      3.3.1 Questions 1 and 2 ................................................................................. 59  
      3.3.2 Question 3 ........................................................................................... 61  
      3.3.3 Question 4 ........................................................................................... 62  
  4. Discussion and Conclusions ............................................................................. 64  
  5. References ......................................................................................................... 69
Chapter 4: Applying GIS uncertainty tools to floodplain mapping and decision making

Abstract .....................................................................................................................75
1. Introduction ........................................................................................................76
2. Flood Mapping History .......................................................................................79
3. Uncertainty in Flood Maps ..................................................................................87
4. Methods ...............................................................................................................91
   4.1 Study Area ......................................................................................................91
   4.2 Data Analysis .................................................................................................92
      4.2.1 Base Data ..............................................................................................93
      4.2.2 The Processing Toolset ..........................................................................94
      4.2.3 Results ....................................................................................................98
5. Discussion and Conclusions ..............................................................................102
6. References .........................................................................................................106

Chapter 5: Conclusion ............................................................................................111
1. Conclusions .......................................................................................................111
2. Future Research .................................................................................................115
3. References .........................................................................................................117

Appendix A: Test questions from the experiment in Chapter 1 ..............................118
### List of Tables

**Table 2.1.** Likelihood ratio chi-squares test results using the Pareto plateform (* significant at 0.05 alpha)........................................................................................................................................ 30

**Table 3.1.** Vegetation boundary error-band categories, r = the radius that contained ~95% of the interpretations from Bley and Haller’s (2006) findings........................................................................ 53

**Table 3.2.** Summary statistics for the transect and whole boundary................................................................. 58

**Table 3.3.** Wilcoxon Signed Rank test results identifying transect F as the only estimate that was not significantly different from the sample transect F 90% confidence interval....... 62

**Table 4.1.** The table displays the number of buildings that intersect the 100-year floodplain per municipality by treatment ........................................................................................................ 99

**Table 4.2.** The percent increase and decrease in building counts between a crisp overlay and ones that involved in the inner and outer buffer extents, and the percent difference ...... 99
List of Figures

Figure 2.1. Traditional sample mean estimate with a confidence interval ........................................10

Figure 2.2. A comparison of a traditional point in polygon problem (left), with a spatial confidence interval overlaid (right); note that the error-band can vary in width, it can be wider or narrower in certain places.................................................................11

Figure 2.3. Depiction of the Epsilon Generalization Model .................................................................14

Figure 2.4. Example of fuzzy set membership for “near” decreasing with distance .......................15

Figure 2.5. Generalized error-band modeled from point errors of a non-generalized line ........16

Figure 2.6. Simplified illustrations depicting the different classes of error-band geometry in the framework used for geospatial-data uncertainty classification..................................................22

Figure 2.7. Depicts GPS points recorded at approximately the same time and with the same devices having non-uniform or variable error-band widths, due to tree cover and buildings affecting satellite reception (2002 VGIN image [online] http://geog-alexandria.radford.edu/Website/va_base_6_1/viewer.htm [accessed July 15, 2011]) .....23

Figure 2.8. Conceptual model of the classification system showing the potential responses in the test..................................................................................................................................................25

Figure 2.9. Mosaic bar chart of answers from test subjects grouped by the question ..............29

Figure 2.10. Agreement plot with the highest percentage agreement and effectiveness for each question displayed. Significant differences in the response frequency proportions and geometry are symbolized ............................................................................................................29

Figure 2.11. Agreement and effectiveness within raters.................................................................32

Figure 3.1. A typical point-in-polygon operation; displays the original and the generalized representation of a park boundary ........................................................................................................48

Figure 3.2. Depiction of Perkal’s Epsilon Generalization Model.........................................................52

Figure 3.3. Map of the study area with the orthophotograph and transects layer used in the experiment (orthophotograph [online] http://geoserve.asp.radford.edu/doqq/B/blacksburg_1_01.htm, Photo 8, [accessed April 2 2009]) .........................................................................................................................54

Figure 3.4. Pictures of different boundary conditions in the field from the study area; A is near transect E and B near transect G (used with permission from Stephen Prisley, 2011) ....55

Figure 3.5. An example of using the measure tool to estimate the width of a confidence interval
at transect and clearing boundary intersection...............................................................56

**Figure 3.6.** Individual realizations of the forest clearing and the point of intersection between each person’s line and transects (orthophotograph [online]
http://geoserve.asp.radford.edu/doqq/B/blacksburg_1_01.htm, Photo 8, [accessed April 2 2009])........................................................................................................................................58

**Figure 3.7.** Graph of ANOVA results on Deviation and EstBand and Transects as the grouping variable. The top and bottom of each diamond represent the 95% confidence interval for each group. The mean line across the middle of each diamond represents the group mean. Overlapping marks indicate that the group means are not significantly different.................................................................60

**Figure 3.8.** ANOM results indicating which group means are statistically different from the overall mean........................................................................................................................................61

**Figure 3.9.** Final map showing the different spatial confidence interval widths and a uniform error-band (orthophotograph [online]
http://geoserve.asp.radford.edu/doqq/B/blacksburg_1_01.htm, Photo 8, [accessed April 2 2009])........................................................................................................................................63

**Figure 3.10.** Bivariate analysis on EstBand and MeasBand with a regression line and normal correlation ellipse; the horizontally stretched band, points’ close proximity to the regression line, and upward trend show a strong positive association between EstBand and MeasBand........................................................................................................................................64

**Figure 4.1.** Depiction of FEMA’s Special Flood Hazard Area and normal stream bank .........83

**Figure 4.2.** The NRVPD, the region used for the floodplain uncertainty analysis .................92

**Figure 4.3.** The toolbar for the uncertainty tools ....................................................................96

**Figure 4.4.** The dialog box for Module 1, used to select feature class error-band geometry .....96

**Figure 4.5.** The dialog box for Module 2 used to create fields and for storing and calculating error-band widths.................................................................................................96

**Figure 4.6.** The *DataFusion* module (Module 3), defines the type of overlay, layer involved and displays the parameter that should have been already set or available in the attribute table. ........................................................................................................97

**Figure 4.7.** Buffer overlay example and visualization of buffer IDs .....................................98

**Figure 4.8.** The map identifies building always in the 100-year floodplain and ones that may or may not be in for the Town of Pulaski.........................................................................................101
**Figure 4.9.** Kernel density plot showing the areas with high-density clusters of uncertain buildings.
Chapter 1: Introduction

1. Research Context and Justification

Uncertainty in geospatial data has become a central issue in the GIScience community (Goodchild 2010). Zhang and Goodchild (2002) describe uncertainty as ambiguity, inexactness, and vagueness, as opposed to error which is inaccuracy or imprecision. The two terms are related and often incorrectly used as synonyms, but are central to debates over data quality and metadata standards. Geospatial data uncertainty is a major concern because society is increasing its use of geospatial technology without truly understanding the limitations of the data. Geographic Information Systems (GIS), for example, combine data from different origins, in different scales, from different dates, and with varying levels of quality, which can greatly complicate analysis and confuse the analyst.

The diversity of geospatial data and spatial analysis methods has caused much trepidation in scientists who employ spatial analysis. As a result, GIS uncertainty research proliferated, but the need remains for a unifying conceptual data model for representing geographic data and its uncertainty (Couclelis 1992, Goodchild et al. 2007, Goodchild, 2010, Voudouris 2010), especially at the conceptual and logical levels of abstraction (Voudouris 2010).

The expansions of GIS, remote sensing, location-based services, and other geo-enabled technologies have transformed earth sciences, social sciences, business, and even daily life, and created reliance upon geospatial data. Rapid innovation in geospatial technology has begun an information technology revolution, driven by the ability to encode data geographically. Geo-enabled mobile devices, such as cell phones, and internet software, such as GoogleEarth©, have spread the use of geospatial technology. Historically, only the government, military, academia,
and wealthy private sectors had access to and the economic means with which to use GIS, but now geospatial technologies in various forms are easily used and widespread.

Geospatial technologies are powerful and useful because they provide the ability to fuse maps with other geographic and tabular data. A geographic or map overlay that involves a topological operation (such as, “intersects” or “is within”) is geospatial data fusion, a concept fundamental to GIS. Such an overlay is essentially a database join that uses spatial coordinates as a key or unique identification number. This connection links data used to create maps and graphical displays. For example, location-based services on a GPS-enabled device can provide a person with the closest restaurant and a satisfaction rating; and an interactive map can display a real-estate parcel with tax information, a zoning map overlay, and orthophotographs. Countless real-estate agents, insurance companies, police, emergency responders, engineers, medical doctors, earth scientists, and ordinary citizens rely on geospatial technology.

Despite its fast rate of adoption, major limitations in geospatial data fusion exist and are often overlooked by the average user of the technology. The limitations exist because geographic data overlaid are of different scopes, times, classes, accuracies, and precisions. Monmonier (1991) calls the artifacts or errors in geographic representations “white lies,” but there are varying levels of seriousness concerning error and uncertainty. People getting lost while navigating with GPS in a car has become a common occurrence and usually is inconvenient, but not terribly important. Whereas, a doctor erroneously interpreting a Magnetic Resonance Image (MRI), or a construction crew digging into an underground fiber optic cable, have dire consequences. These examples come from different fields that use the same technologies and theories, demonstrating that an understanding of uncertainty in geospatial data fusion is a concern well beyond academic geography and has universal application. As geospatial technologies continue to develop, the
methods for fusion and overlay analysis must improve to ensure geospatial analysis is a rigorous and valid scientific process.

Although a vast literature on uncertainty in maps, GIS, and geospatial data exists (Goodchild 2010), no tools utilize current metadata standards and make it easy for a geospatial analyst or the average GPS-enabled cell phone user to understand the confidence levels and statistical uncertainty in results. Brown and Huevelink’s (2007) Data Uncertainty Engine (DUE) is the best working system for addressing uncertainty in geospatial data, but the DUE lacks the ability to create and visualize spatial confidence intervals in a GIS. Not only is statistical significance questionable when using geospatial data, so is its practical importance of conducting a geospatial analysis and using the results. Most geographic representations are single realizations (i.e. a sample size of one), a factor which severely limits the methods available for statistical analysis. This manuscript is an intense exploration into uncertainty in spatial data, spatial statistics, GIS uncertainty, and GIScience that seeks to expand the current knowledge base and create GIS tools to help understand uncertainty in geospatial data.

2. Dissertation Organization and Research Questions

This dissertation comprises three stand-alone manuscripts written for submission to peer-reviewed academic journals. The papers build on one another to present a progressive and cohesive study that defines a simplified error-band framework for uncertainty classification and develops GIS tools to incorporate geospatial data uncertainty into overlay operations.

Paper 1 (Chapter 2) is a literature-based synthesis of geospatial statistics and uncertainty that concludes with a new framework for classifying uncertainty in geospatial data. Two research questions designed to test the application of a GIS error-band framework for modeling spatial
confidence intervals are the foundation of the paper: 1) do GIS users feel comfortable making an assumption about the error and uncertainty of geospatial data from metadata, and 2) can GIS users determine geospatial data’s error-band geometry from metadata?

Building from the literature and the experimental results presented in Paper 1, the second paper (Chapter 3) analyzes tests on the ability of people to estimate spatial confidence intervals for data they create. It explores a person’s ability to use an estimate of confidence for use in the framework in the absence of metadata or with inadequate metadata. The objectives of the second paper are: 1) to determine if people digitizing a forest-clearing boundary can estimate 90% spatial confidence intervals for their geographic representation, and 2) to explore the geometry of a forest-clearing boundary’s error-band.

Finally, Paper 3 (Chapter 4) incorporates the framework and implements the findings of the previous two papers in a toolset for ArcGIS10©. The tools utilize current metadata standards such as National Standard for Spatial Data Accuracy (NSSDA) to simulate error, propagate error-bands, and incorporate spatial confidence intervals into the overlay process. The toolset was tested on a real-life dataset to determine if the number of buildings in a 100-year floodplain was significantly different on a crisp overlay compared to overlays that involved various spatial confidence intervals.

Together these papers provide a unique perspective, methods, and a set of tools that have the potential to improve geospatial analysis, data fusion, and map use. Each part of this dissertation sets forth its own recommendations for improving uncertainty analysis in GIS, but the final contribution of the dissertation is a toolset that has the potential to change how people interpret GIS overlay results, and how people make decisions based on them. The hope is that these papers prompt inquiry and question the reliability of all simple overlays. Many situations exist
in which this research has relevance, making the framework, the tool, and the methods important to a wide variety of disciplines that use spatial analysis and GIS. Chapter 5 reviews the findings of all three papers and concludes with recommendations on the use of the framework and toolset, as well as areas for future research.
3. References


Chapter 2:

An evaluation of a GIS framework for generating spatial confidence intervals to enhance geospatial data fusion

Andrew S. Foy, Laurence W. Carstensen, Stephen P. Prisley, James B. Campbell, and Randel L. Dymond

Department of Geography, College of Natural Resources and Environment, Virginia Tech, Blacksburg, VA, USA; Department of Forestry Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, USA; Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, USA

Abstract: Advances in computer technologies and web GIS have improved the quality of maps, making map comparison and analysis easier. Yet, uncertainty and error still exist because of inherent problems, such as ones in accuracy and precision of measuring devices, and introduced problems, such as user inputs, algorithms, definitions, generalizations, etc. These uncertainties are highly prevalent in GIS, where the additions of tabular data, frequent scale changes, and data purpose discrepancies compound this problem. The fusion of geographic data with multiple or unknown confidence levels creates statistical problems and concerns. Therefore, the methods and interpretation of GIS estimates must evolve to capture and communicate uncertainty. This paper reviews error-band literature in the context of cartography and GIS for the purpose of creating GIS tools that generate spatial confidence intervals for overlays. GIS users classified data into groups from a framework using metadata and maps to determine if they can agree on applications of frequentist or Bayesian methods to model the data’s error-band geometry, in defining the uncertainty of the data. There was great diversity in the test subjects’ responses to the data, but there was correlation between user agreement and the geometry of data. The evidence suggests point-based data are better understood in terms of uncertainty. In conclusion, common metadata and methods for conveying data quality are hard to incorporate into this model for spatial confidence intervals, and GIS experience is important for assessing uncertainty. For a tool to be practical and useful users need more guidance and training, whether a wizard, series of questions, or measurement devices in a GIS tool that helps them assess data-quality.

Keywords: GIS, geospatial data fusion, spatial confidence intervals, data quality, uncertainty, meta-data
1. Introduction

Geographic Information Systems (GIS) technology has revolutionized how scientists study the earth, but its expansion has brought to light inherent and operational factors that lead to uncertainty in geospatial data and GIS analyses. GIS is essentially a tool for scientists to use for overlaying geospatial data, for inference, and for estimation about geographic phenomena. Uncertainty is a major concern within the GIS and GIScience community (Goodchild 2010). For example, procedures such as estimating confidence intervals for the coordinates of a geographic feature are difficult because the data are spatial, do not always have frequency distributions for coordinates, can be mobile, or are areal. Other complications with GIS analyses exist because data are overlaid having different scopes, times, classes, accuracies and precisions, which leads to vast uncertainties in geospatial data fusion (Burrough 1986, Goodchild 2000, Plewe 2002). Therefore, the methods and interpretation of GIS estimates must evolve to capture and communicate uncertainty.

The purpose of this paper is to review and synthesize GIS uncertainty literature relevant to models for generating spatial confidence intervals for geospatial data. The synthesis of the literature contained in this paper led to the development of a new framework for classifying the size and shape of spatial confidence intervals (error-band geometry) in geospatial data and for classifying data quality standards and/or relevant statistics for geospatial uncertainty (uncertainty knowledge). The utility of the framework and concept was tested via experimental evaluation of GIS users’ ability to apply the framework using summarized metadata and maps. Two questions guide the theoretical progression of this paper:

1. Do GIS users feel comfortable making an assumption about the error and uncertainty of geospatial data from metadata?
2. Can GIS users determine geospatial data’s *error-band geometry* from metadata?

The limited GIS tools available for computing and visualizing spatial confidence intervals confirm the need for a framework easily understood at the conceptual level and readily incorporated into a tool for analyzing uncertainty in overlays. The concerns over uncertainty and other trepidations over the proper usage of GIS stem from models often relying on overlays (i.e. geospatial data fusion) because the data come from numerous geographically referenced sources. Limited research exists on the quality of results that come from the fusion of geographic data, which raises question about the reliability of even simple overlays. There is an obvious gap in the literature because uncertainty in overlays is a real problem, but hard to quantify.

For a simple example, consider the uncertainty of GIS data using probabilities. If you have four layers that are 95% accurate and follow the basic rules of probability, the confidence in the output is only 82% \((0.95 \times 0.95 \times 0.95 \times 0.95 = 0.82)\); for eight layers confidence is down to 66%. This logic and the previously referenced concerns over the reliability of GIS analyses resulted in various data quality standards and policies that attempted to institutionalize the management of geospatial data uncertainty (USBB 1947, ASPRS 1990, FGDC 1998). However, a major obstacle remains. Most maps or depictions of geographic information are single realizations. That is samples of one, and this reality severely limits the options for statistical analysis. For instance, in data gathering for a GIS, how often does one re-digitize a stream to form a frequency distribution from which to assess positional errors?

Without a distribution, it is not possible to compute confidence intervals, power, and significance of estimates, all of which are critical metrics to validate scientific conclusions. A scientist would not estimate the mean weight of 50-year-old men by weighing only one man. It is necessary to weigh many men of that age and express the mean weight with a confidence
interval and level of confidence (Figure 2.1), most often at 90%, 95%, or 99%. Confidence intervals are closely associated with margin of error, where the range of a confidence interval is the sample statistic ± margin of error. For example, we could be 95% certain that the mean weight of men, age 50, is 84kg ± 35. The same logic suggests that a single geographic object needs its positional estimate expressed with a confidence interval (for example, we could be 95% certain that an estimated location is X = 89.6° W, Y = 31.2° N ± 0.11°). The problem is that to compute the margin of error the location must be recorded multiple times and preferably with multiple devices to ensure independence.

Figure 2.1. Traditional sample mean estimate with a confidence interval.

In science, the expectation is that estimates will have confidence intervals. However, in GIS there are limited data, small sample sizes, and inadequate tools to provide such. This makes it difficult to answer questions such as; what is the confidence interval for the mapped location of this parcel boundary? How far away should the utility company dig from the mapped location of the underground gas-line? What is the margin of error in the estimate of a forest stand’s acreage? These questions should be easy to answer in a GIS, but they are not, or at least, the GIS process is not intuitive. This is perplexing because estimating a geographic location with a confidence interval is no different from calculating a mean weight with 95% confidence. The sample data is what makes it possible.
Geospatially-based sciences need to develop a framework, methods, and GIS tools to create, overlay, and visualize spatial confidence intervals. Figure 2.2 illustrates a geographic boundary with and without a hypothetical spatial confidence interval, often referred to as an error-band in GIS uncertainty literature. The realization that boundaries are not crisp and possibly indeterminate can lead to many questions in a GIS analysis (Burrough and Frank 1996).

**Figure 2.2.** A comparison of a traditional point in polygon problem (left), with a spatial confidence interval overlaid (right); note that the error-band can vary in width, it can be wider or narrower in certain places.

The most widely adopted method to model positional uncertainty is through an error-band, a confidence-like region around the “true” position of a point, line, or polygon (Perkal 1956, Chrisman 1982, Blakemore 1984, Dutton 1992, Shi and Liu 2000). In most situations the band size is subjective because the typical practice is to treat X and Y positional variance as an arbitrary distance parameter, though some have used the law of error propagation (Caspary and Scheuring 1992, Shi 1994). Other methods for estimating variance include using Monte Carlo simulation for generating multiple realizations of data (Lei et al. 2006) and parametric spatial bootstrapping to simulate error (Cressie 1993).

Despite the advancements in modeling and simulating error distributions, the human dimension still plays a strong role in introducing and interpreting spatial uncertainty in data and
GIS analysis (Jenks 1981). Studies of spatial reality cognition describe great variations in how human beings perceive geographical reality (Portugali 1996). Thus, it is likely that people interpret accuracy, error, and uncertainty information differently.

Rapid technological innovations are changing how society uses geographic data, bringing new knowledge requirements to the modeling, analysis, and visualization of spatial objects. These changes stipulate a new comprehension of geographical reality in terms of the geometry and properties of a spatial object. No basic GIS framework exists for conceptualizing error-bands because of the diversity of geospatial data and its unique properties. Understanding how GIS users apply metadata and other geospatial data-quality information to error-band generation takes a critical step forward in GIS research by tying basic research on GIS uncertainty theory to an applied research field.

The hope beyond this paper is to design software to improve geospatial data fusion within a GIS. Before designing software, it is necessary to know if quality information about data provided to a GIS user can be incorporated into this GIS framework for spatial confidence intervals. This research is critical for the development of tools and protocols to quantify and manage geospatial data-fusion uncertainty.

2. Review

2.1. Modeling Error and Uncertainty

Uncertainty is unavoidable in GIS because the real world is diverse, complex, and dynamic. It can arise in any of the following six aspects: lineage, positional uncertainty, attribute uncertainty, logical consistencies, completeness, and temporal uncertainty; and errors can be random or systematic. Some uncertainties are inherent (inherent error) to spatial data while others are
introduced (operational error) through user input, generalization, or other subjective decisions (Leung et al. 2004a)

Early research in what is today GIS uncertainty focused on hard computing, for example probability theory (Arthurs 1965), adjustment of observed data (Mikhail and Ackermann 1976), evidence theory (Shafer 1976), and spatial statistics (Cressie 1993). More specific to spatial confidence intervals was research on error/epsilon bands (Perkal 1956, Chrisman 1982, Blakemore 1984), s-bands (Shi and Ehler 1993, Shi 1994) and the g-donut (Dajun et al. 2003). Others discussed using systems that store the measurements themselves and not the derived data, i.e. coordinates. These ideas led to measurement-based geographic information systems (Buyong and Kuhn 1992, Goodchild 1999, Leung et al. 2004a). Unfortunately, there is no such working system. Soft computer theories were also explored, for example, fuzzy logic (Zadeh 1965), genetic algorithms (Zhang and Goodchild 2002), and sensitivity analysis (Crosetto and Tarantola 2001). Others combined the hard and soft computing approaches, for example, the cloud model combined randomness and fuzziness (Wang et al. 2003). Even with the expansive literature and GIScientists’ devotion to understanding accuracy and error in GIS data, there are no tools in general use for geospatial confidence interval estimation. The most notable attempt is Brown and Huevelink’s (2007) paper in which they released their Data Uncertainty Engine (DUE). The engine is software for assessing and simulating uncertainty, but it is lacking in GIS compatibility, is not intuitive, and cannot make spatial confidence intervals in GIS.

In an effort to better define and represent the boundaries of the vector model Perkal published the idea of utilizing bands or buffers to model uncertainty in 1956. The Epsilon Generalization Model (EGM) uses circular bands to represent the level of confidence in a boundary (Figure 2.3).
Figure 2.3. Depiction of the Epsilon Generalization Model

In that model the margin of error is equal across the length of a line, meaning the uncertainty is the same at the end and middle of that line. The EGM is conceptually easy to understand, but it is limited in its scope by a lack of suitable methods for calculating the width of the band, referred to as epsilon. Chrisman (1982) and Blakemore (1984) used the EGM to study cartographic error and more specifically categorized error generically; for example, the error of a digitized line is 1.5m (refer to Jenks [1981] for a review of digitizing error).

Research that is more recent suggests using a normal or Gaussian distribution to simulate error (Dutton 1992, Shi and Liu 2000). However, a major problem remains; the variance of positional error for most spatial data is unknown. Monte Carlo simulation is one method to generate probability density functions from which one can generate error-bands, but this method still requires making assumptions (i.e. arbitrarily defining the mean and variance). Leung et al. (2004a) explored another error-band modeling approach, using the approximate law of error propagation to generate covariance-based error-bands, but their model does not work with coordinate-based GIS or derived products (Leung et al 2004a, pp. 327). One last criticism of most error-band models is they do not provide a gradient of confidence levels: 90%, 95%, and 99%. Probabilistic epsilon generation as explored by Honeycutt (1986) is a good example of the inability to infer error from a sample of one, but none of these methods is available to the average GIS user or analyst.

An alternative to the EGM is the fuzzy error-band model. It represents a probability-like
surface for locations in a continuous manner (Laviolette et al. 1995). The probability-like values are created using “fuzzy set” membership that mimics human thinking and terms such as “near” and “far” or “more” and “less.” In error-band applications, the values range from 0-1, much like a probability. Based on the type of topologic operation, the fuzzy values either increase or decrease with distance extending outward from a boundary, point, or line. For example, if you are trying to determine if a point is inside a polygon, by overlaying the point with a fuzzy error-band the value of the point’s location can be identified (Wenbao et al. 2001). If the point value were a one, this would indicate that location is inside the polygon, but if it were a 0.2, you would be “less” certain. Figure 2.4 is an example of a linear decreasing fuzzy error-band, so the values decrease with distance from the line. Although the fuzzy error-band model has some advantages, people have been hesitant to incorporate it in GIS tools because fuzzy logic is often based on subjective decisions. Fuzzy set membership works by using a series of distance and standardizing equations, based on user input to provide fuzzy values.

![Figure 2.4. Example of fuzzy set membership for “near” decreasing with distance](image)

Further development on error-bands included deriving the variance in the X and Y directions from an arbitrary point on a line segment based on the law of error propagation (Zhang and Tulip 1990, Caspary and Scheuring 1992). Dutton (1992) used Monte Carlo simulation to model the error distribution of geometric objects, but the above methods were not useful in a GIS. Shi and
Liu (2000) and Dajun et al. (2003), using strictly geometric methods, developed the g-band and g-donut, respectively, which estimate error at points or vertices along a line to compute the uncertainty of that line, as opposed to using the line itself. Shi (1994), used a similar approach (s-band), but modeled thematic and positional uncertainty. These models assume that errors occur at points and that the center of a line is actually more accurate than the ends. This is problematic because it suggests more confidence in a point’s location along a line where points do not exist (i.e. the mid-point, Figure 2.5). In effect, it infers that places we do not measure are more accurate than those we do measure. Geometry alone however does not add processes such as line generalization to the equation, so such point-based models have serious limitations.

![Figure 2.5. Generalized error-band modeled from point errors of a non-generalized line.](image)

Many methods to quantify the uncertainty of boundaries in vector GIS are available, but the variety of approaches and failure to establish standards suggests that generating error-bands for individual data is difficult. A better approach may be setting error-band widths based on categories of data or based on the needs of a GIS user. Couclelis (1996) developed a typology for geographic boundaries in which 120 different types of boundaries were identified. Boundary situations exist in three dimensions: empirical nature of the data, mode of observation and user purpose. These dimensions and the examples of boundary categories help in categorizing geospatial error, but do not mesh with the mathematics of calculating error-band width and make programming labor intensive. Thus, further exploration into the basic geometry of error-bands in a GIS follows.
2.2. *Data Quality Standards and Metadata*

Over the past few decades, there have been many efforts to define geospatial-data quality standards and their core elements, but the efforts have struggled to keep pace with the technological innovation in the geospatial sciences. The most notable and earliest published standard for spatial data was the National Map Accuracy Standard (USBB 1947), which applied to all U.S. federal agencies that produce maps. It predates digital GIS data, as it was intended for paper maps, but due to its continued reference in GIS, there are many questions on how to apply the standard to digitally converted maps that adhere to NMAS.

The NMAS states that for maps on publication scales larger than 1:20,000, not more than 10% of the points tested shall be in error by more than 1/30 of an inch, and that for maps on publication scales of 1:20,000 or smaller, 1/50 of an inch. This translates to an error of approximately 12.19m for a map at a scale of 1:24,000. If this standard were used to create error-bands for all data that adhere to NMAS, many GIS operations would be far more uncertain. Even more so, when one considers the standard only applies to well-defined points. Therein lies an interesting question; how accurate is a road on a map that adheres to NMAS? As computers advanced and geospatial data transitioned to digital forms, many governmental agencies developed their own standards for their data, most of which exceeded the NMAS.

To meet the growing needs of the geospatial community, in 1990 the Office of Management and Budget (OMB) established the Federal Geographic Data Committee (FGDC) for the coordination of geographic information and related spatial data activities for state and local government, industry, and professional organizations. Through the FGDC, the National Spatial Data Infrastructure (NSDI) was developed; a physical, organizational, and virtual network designed to enable the development and sharing of the United States digital geographic...
information resources.

In 1994, the United States Spatial Data Transfer Standard was published, defining five key components to spatial data quality: positional accuracy, attribute accuracy, lineage, logical consistency, and currency. These components were expanded along with the NMAS for the publication of the National Standard for Spatial Data Accuracy (NSSDA), a statistical and testing methodology for estimating the positional accuracy of points on maps and in digital geospatial data, with respect to georeferenced ground positions of higher accuracy (FGDC 1998). The NSSDA uses root-mean-square error (RMSE) to designate error. This uncertainty standard emphasizes spatial confidence interval estimation. The reported accuracy value reflects all uncertainties; including those introduced by geodetic-control coordinate compilation and final computation of ground coordinate values in the product, but expressed at the dataset level (FGDC 1998). While this is a robust standard, the design is more of a template for storing geospatial-data quality information rather than a mandated procedure.

Other than the FGDC standards, which are guidelines at best, the issues with data quality are left for the producers and consumers. For instance, how accurate could we make the data? How accurate do the data need to be? Many federal and state organizations publish data following FGDC standards, but smaller governmental organizations, academia, industry and many new opensource geospatial-data providers may not follow any standard for spatial data quality. Large amounts of GIS data are produced and used at local scales, in which the mindset is simply that digital is better, or the data are collected to comply with a code or regulation. Inevitably, these data are acquired by other agencies, used in research or for other purposes. Clearly, the GIS community requires better methods to manage geospatial data quality.

Other issues arise because metadata are published in a variety of formats or are often absent,
as is the case with GoogleEarth\(^1\), WikiMapia\(^2\), OpenStreetMap\(^3\), many web mapping services, geocoded information, mobile phones, and spatially intelligent software. These new geospatial-data are used in a variety of new applications from underground utility location, to land grading, and traffic control, and even Facebook\(^\text{©}\); yet there is limited liability placed on that information. For example, a line locator service is legally responsible if it spray paints an underground utility line location incorrectly, but if the GPS data they collect, produces an incorrect map there can be no legal recourse. Von der Dunk (2008) and Awang \textit{et al.} (2009) provide a thorough review of the legal implications of geospatial information.

Not only is liability debatable, to add to the crisis little is known about how individuals use quality information, especially given there are no commercially available tools that create error-bands from metadata in a GIS. The reality of maps, GIS and all of the new geospatial media is that all the data are geographic representations and simply estimates, driven by the needs of the user and their ability to measure phenomena. Therefore, a basic framework on error-band geometry needs to incorporate user perception of data quality and type of knowledge.

Increased GIS exposure and the accumulation of geospatial data are beneficial to geography, but as a portion of the discipline has evolved into a new discipline, GIScience, there is a great need to understand and address a fundamental statistical limitation of our data. Maps and most digital geospatial data are single representations of an object or phenomenon. Such small sample sizes (n=1) are undesirable because they greatly reduce the power of an analysis. Frequency distributions and probability density functions are fundamental to statistical inference, but require a larger sample size. To infer is to use evidence to form conclusions. With less

\(^1\)http://earth.google.com/
\(^2\)http://wikimapia.org/
\(^3\)http://www.openstreetmap.org/
evidence, power is reduced. Greater power is important because it reduces the chance of a Type II Error, thereby increasing the probability of correctly identifying a trend or phenomenon.

Difficulties determining geospatial power, small sample sizes, missing metadata on accuracy and precision, geo-referencing errors, overlay techniques, and many other uncertainties in data highlight the need to explore other statistical techniques and utilize both frequentist and Bayesian approaches to quantifying uncertainty in geospatial data.

2.3. Relevant Statistics

Both Neyman–Pearson (frequentist) and Bayesian statistics can offer insight into uncertainty estimation in geospatial data fusion, as both disciplines incorporate probability and significance testing, but their approaches and interpretation of the data are quite different. The frequentists see probability as the long-run expected frequency of occurrence (Little 2006). The Bayesian view of probability is related to degree of belief, measuring the plausibility of an event given incomplete knowledge.

The frequentist believes that a population mean is real, but unknown, and unknowable, it can only be estimated from a sample of the data. For example, the frequentists’ view is that you could never really know the average weight of 50-year-old men, but you could estimate it by sampling the weight of a representative group of males and computing the sample's mean. By knowing the distribution for the sample mean, a confidence interval can be constructed, centered at that sample mean, and used to describe the population mean. Nevertheless, the population mean is a single fixed value without a distribution, making statements referencing the probability of the population mean illogical. As such, the frequentist cannot say there is a 95% probability the population mean is in a confidence interval. Rather the probability refers to the weight of the
i-th randomly sampled 50-year-old male or the mean of a random sample of 50-year-old men.

Unlike frequentist logic that suggests the population mean is real, Bayesian logic implies only the data are real, and that the population mean is merely an abstraction. As such, certain estimates are more believable than others based on the data and prior beliefs. The Bayesian approach is to construct a credible interval, centered around the sample mean, but adjusted using Bayes theorem and beliefs or expert knowledge about the mean (Howson and Urbach 2005).

Bayesian logic allows one to say that there is a given probability the interval contains the mean, while frequentist cannot (although this mistake is often made using frequentists’ techniques). The Bayesian view of probability is that it is the degree of believability. For example, Bayesians can refer to tomorrow's weather as having a 30% chance of snow, whereas this would not make sense to a frequentist because tomorrow’s weather is just one unique event, and should not be referred to as a relative frequency. It does not snow 30% of the days out of the year, rather Bayes theorem incorporates prior belief and evidence to come up with the probability. Granted that is a simple example and not all meteorologists are Bayesians. Neither of the two schools of statistical thoughts is best for geospatial data. A “Calibrated Bayes” approach, in which one uses Bayesian methods for inference in a particular model, but frequentist methods for model assessment might prove useful in a GIS (Little 2006).

3. Methods

3.1. The Framework

A synthesis of error-band literature, expert knowledge in GIS and a need for tools to model spatial confidence intervals in a GIS, led to the development of the Error-Band Geometry Framework. It consists of eight simple categories (Figure 2.6), based on geometry (point, line,
and polygon) and relationship with other objects and their components within a GIS data layer (uniform, non-uniform and segmented). The classes cover all two-dimensional GIS data, but also could have application to three-dimensional data. Butler (1999) had interesting findings, concluding that confidence intervals had different shapes dependent upon the vertical or horizontal planes of stereotactic brain imagery. GIScientists must have a fundamental grasp on confidence intervals for geospatial data and continue to expand on this GIS framework for spatial confidence intervals.

![Diagram](image)

Figure 2.6. Simplified illustrations depicting the different classes of error-band geometry in the framework used for geospatial-data uncertainty classification.

Procedurally, a bivariate normal or Gamma distribution can be used to model the uncertainty of coordinates in a GIS. Most standards deal with error at points and are good examples of point-uniform data for this model (USBB 1947, FGDC 1998) (Figure 2.6a). Other error-band applications expand on this concept by constructing bands for individual points and connecting
them to create line and polygon buffers (Dutton 1992, Shi and Liu 2000, Dejuan et al. 2003, Love 2009); see Figures 2.6d and 2.6g.

Specifically when dealing with point data, quality information is sometimes expressed in metadata at the dataset level, for example 95% of the points are within 3m horizontal positional accuracy. When this is the case, following the framework, the point data have error-bands with uniform geometry, i.e. the same width and shape for all the points. However, not all collections are that simple. Frequent GPS users know accuracy of each GPS point varies. The map in Figure 2.7 illustrates the large variance in the margin of error for GPS recordings taken at the approximately the same time. This illustrates a situation in which point data are non-uniform in their error-band geometry (Figure 2.6b). It is not possible for a topological connection to other points in a point dataset; as a result, points never have segmented geometry (Figure 2.6c).

![Figure 2.7](image)

Figure 2.7. Depicts GPS points recorded at approximately the same time and with the same devices having non-uniform or variable error-band widths, due to tree cover and buildings affecting satellite reception (2002 VGIN image [online] http://geog-alexandria.radford.edu/Website/va_base_6_1/viewer.htm [accessed July 15, 2011]).

Lines and polygons are more complicated because they can be segmented and are frequently generalized. The concern is that not all point locations (vertices) in geospatial lines are known -
the vector model dictates that we sample lines and generalize them - and therefore should model error for the smallest object with data quality information. For a generalized line, it would not be its points, rather it is mostly likely to be the line feature or line dataset. A useful method for generalized data is to model uncertainty at the dataset level (*uniform*, Figure 2.6d). For example, you could construct error-bands for all the roads in a roads centerline shapefile.

Expanding on this concept the model can be refined to include non-uniform line error-bands. For instance, if the roads shapefile was a composite from different counties and one county’s data was higher quality than the other counties then the roads from each county would have error-bands of different widths (Figure 2.6e). If the metadata allow, the line error-band class can be further refined to individual segments of individual lines. Imagine you are digitizing roads for a layer in a GIS using orthophotography. At certain areas on the map, the road is not visible due to tree canopy cover. Therefore, it is logical that you would be more certain of the correct position of the line where it were clearly visible and less certain along the segment of the line that was not visible (Figure 2.6f).

Leung *et al.* (2004a) illustrates how under different cases error-bands can take on different shapes. The different shapes are critical to the segmented error-band geometry class, for example some segments could be modeled with a uniform shape (*line type*, Figure 6f and 6i), while others might have a wedge shape and be modeled using the error-band width of the endpoints (*wedge type*, Figure 6f and 6i). While polygons are topologically different from lines, they are still collections of points and lines. The various shapes of a polygon error-bands vary in the same ways as lines (Leung *et al.* 2004b, Dajun 2003, Love *et al.* 2009); see Figures 2.6h, Figure 2.6i, and line versus wedge types of segmented error-band geometry.
3.2 Methodology for the Experiment

It is critical to understand if GIS users can differentiate among the eight categories of error-band geometry and if they can determine it best to create a confidence interval or credible interval based on available data quality information. A prototype tool for ArcGIS 9©, using VBA has been developed to implement such procedures.

To evaluate the ways in which GIS users interpret metadata and apply that information to the GIS error-band framework (Figure 2.8), 26 students from an intermediate GIS class were administered an internet-based survey, asking them to classify 12 examples of GIS data and applications. The classification system has two categories (2x3), therefore six potential answers (Figure 2.8).

Figure 2.8. Conceptual model of the classification system showing the potential responses in the test.

The two categories capture data for the purpose of identifying uncertainty knowledge (Figure2.8a), which measures if a person “knows” something about the uncertainty in an example dataset (objective and useful in frequentist models); or does not “know” uncertainty, but has a belief (subjective and useful in Bayesian models). The second category is error-band geometry (Figure 2.8b), which classifies data into three categories (uniform, non-uniform, or segmented). For example, an NMAS based dataset of well-known points has a reported RMSE
horizontal positional error, therefore the user should know the error at the points, and because RMSE is at the dataset level the error-band geometry should be uniform (e.g. all points with the same error-band width). In another case, the precise locations for a boundary line between a forest stand and a defunct logging operation is difficult to determine as trees gradually return to clear-cut areas over time. This might require an analyst use “belief” that the boundary is in a particular location rather than having a measured value from which to estimate error.

The test examples were constructed from typical GIS operations or were published from various government and academic institutions. There were three sets of tests with random question orders, distributed at random. The survey involved viewing an image representing common GIS data and applications, answering multiple-choice questions, and providing feedback. Each example (Appendix A) included metadata from which the subject classified the example based on two categories: uncertainty knowledge and error-band geometry.

Uncertainty knowledge (known or unknown) describes whether you are aware of the data quality.

- **Known**- information on accuracy, precision, standard deviation, error, resolution or other appropriate measure is provided with the data, or you feel comfortable estimating error.
- **Unknown**- no information on accuracy, precision, etc. is provided, or you are not comfortable estimating error and need assistance, but may have some beliefs.

Error-band geometry (uniform, non-uniform or segmented) describes the shape of an error-band around geographic objects. It can only be uniform or non-uniform for point layers, but all three types for lines and polygons (as illustrated in Figure 2.6).

Statistical analysis of the test subjects’ answers was conducted with JMP 9©, to look at
agreement on questions, and if any lack of agreement was symmetrical. The Categorical Response Analysis module for rater agreement was used to create a mosaic bar chart, calculate response frequencies, and plot agreement statistics. The Kappa for agreement and Bowker/McNemar tests for symmetry (marginal homogeneity), using this module, were inadequate because the classes in the survey were not equiprobable. All point data examples are limited in geometry to uniform and non-uniform, therefore only four potential answers. This makes the probability of random agreement 25% for point examples and 16.67% for line and polygon examples. It is important to account for these differences because such variances can make test statistics inflated and inconsistent.

When the above situation is the case and there are multiple responses, a more robust analysis method is to model each response category separately. The question is whether the response rates are the same across samples for groups. For each response category, we assume the frequency count has a random Poisson distribution. This rate, obtained using a Poisson regression of the frequency per unit modeled by the sample categorical variable, results in a likelihood ratio chi-square test of whether the response rates are different. This test used the Pareto platform, but can also be performed using the Categorical Response Analysis module or the Generalized Linear Modeling module. The statistical methods follow suggestions from Agresti and Coull (1998) and Agresti (2002).

Percent agreement between users for questions was calculated by comparing all pairs of raters by replicate combinations, for each question (1).

$$\%\text{Agreement for grouping variable}\, j = \frac{\sum_{i=1}^{\text{number of levels}} \left(\frac{\text{number of responses for level } i}{2}\right)}{\left(\frac{\text{number of users} \times \text{number of reps}}{2}\right)}$$

(1)

Effectiveness was also included in the analysis as an additional measure of user performance
because the examples from the survey were designed with a specific answer in mind and that function is part of the Rater Agreement Analysis options, it is defined as (# of correct decisions) / (Total # of opportunities for a decision). Agreement within users, or how often other users agreed with each other was calculated using (2):

\[
\text{%Agreement for rater } k = \frac{\sum_{i=1}^{n} \left( \sum_{j=1}^{r_i} \text{number of uncounted matching levels for this rater } k \text{ within part } i \text{ for rep } j \right)}{\sum_{i=1}^{n} \left( \sum_{j=1}^{m_i} r_i \times r_j - j \right)}
\]

(2)

Where,

- \( n \) = number of classes
- \( r_i \) = number of answers for question \( i \)
- \( m_i \) = number of raters for question \( i \)

### 4. Results

Visual interpretation of the test subject’s responses using Figure 2.9, a mosaic bar chart, shows great overall diversity in test subjects’ answers. Known; Non-Uniform was the most frequent response. However, the chart does not provide any statistical measures of user agreement and there is no unified test for the significance of percent agreement. Kappa is commonly used to compare categorical response data and more specifically inter-rater agreement, but the problem with kappa is that no value for significance can be regarded as universally acceptable (Bakeman et al. 1997). Therefore, these results contain various measures of agreement, interpreted in an attempt to understand the process by which raters responded. The difference in percent agreement and effectiveness of rater responses is most noticeable in Figure 2.10 because the highest percentage rate of agreement is plotted, along with effectiveness for each question. Effectiveness is a metric comparing user’s answers to a hypothesized correct answer. The
geometry of the questions can also be identified through that figure.

![Figure 2.9. Mosaic bar chart of answers from test subjects grouped by the question.](image)

Figure 2.9. Mosaic bar chart of answers from test subjects grouped by the question.

![Figure 2.10. Agreement plot with the highest percentage agreement and effectiveness for each question displayed. Significant differences in the response frequency proportions and geometry are symbolized.](image)

Figure 2.10. Agreement plot with the highest percentage agreement and effectiveness for each question displayed. Significant differences in the response frequency proportions and geometry are symbolized.

In this study, there was no evidence that GIS users agreed on a class consistently for all twelve examples, but a more detailed interpretation reveals an agreement bias by geometry of the data. Point data had the highest frequency (4) of statistically significant p-values for the likelihood ratio chi-square test, followed by line geometry, and polygons with two significant
values each (Table 2.1).

Table 2.1. Likelihood ratio chi-squares test results using the Pareto plateform (* significant at 0.05 alpha)

<table>
<thead>
<tr>
<th>Question ID</th>
<th>Chi-Square</th>
<th>DF</th>
<th>P-value</th>
<th>Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.9146</td>
<td>3</td>
<td>&lt;0.0001*</td>
<td>point</td>
</tr>
<tr>
<td>5</td>
<td>5.3962</td>
<td>3</td>
<td>0.1450</td>
<td>point</td>
</tr>
<tr>
<td>7</td>
<td>14.3658</td>
<td>3</td>
<td>0.0024*</td>
<td>point</td>
</tr>
<tr>
<td>11</td>
<td>11.5792</td>
<td>3</td>
<td>0.0090*</td>
<td>point</td>
</tr>
<tr>
<td>2</td>
<td>18.8181</td>
<td>5</td>
<td>0.0021*</td>
<td>polygon</td>
</tr>
<tr>
<td>3</td>
<td>7.8082</td>
<td>5</td>
<td>0.1671</td>
<td>polygon</td>
</tr>
<tr>
<td>4</td>
<td>4.7007</td>
<td>5</td>
<td>0.4535</td>
<td>polygon</td>
</tr>
<tr>
<td>8</td>
<td>12.1105</td>
<td>5</td>
<td>0.0333*</td>
<td>polygon</td>
</tr>
<tr>
<td>9</td>
<td>2.4693</td>
<td>5</td>
<td>0.7811</td>
<td>polygon</td>
</tr>
<tr>
<td>6</td>
<td>25.7187</td>
<td>5</td>
<td>&lt;0.0001*</td>
<td>line</td>
</tr>
<tr>
<td>10</td>
<td>8.3494</td>
<td>5</td>
<td>0.1380</td>
<td>line</td>
</tr>
<tr>
<td>12</td>
<td>14.2615</td>
<td>5</td>
<td>0.0140*</td>
<td>line</td>
</tr>
</tbody>
</table>

The grand mean for agreement between users was 26.67%, while the random chance for agreement on all twelve questions was only 19.42%, which would give an overall Kappa of 8.99%, which is not any better than chance, and not significant by any standard. However, because the questions acted as a grouping variable it was possible to look for statistical significance in responses for each question. Seven out of twelve questions had significant agreement between users greater than chance (Table 2.1). Nevertheless, only three questions were above 30% agreement (Figure 2.10). Each test subject commented on his or her responses. Frequently mentioned in the feedback was user experiences, and expert knowledge.

Even though the responses were highly variable, the evidence supports a bias towards agreement for point questions. Take for instance question one, seven, and eleven (Appendix A), these questions had the highest agreement and were all of point geometry. One might suggest it
is a result of the greater chance for random agreement, but the p-values for the chi square statistics in Table 2.1, confirm these point examples had statistically significant agreement. However, question five is a point example and had agreement above the grand mean agreement, yet it is not statistically significant because of point’s higher chance for random agreement. In contrast, question eight is a line example with agreement below the grand mean, but with significant agreement. It was plotted below the grand mean and significant because of its lower chance for random agreement. This situation can also be observed for questions two and twelve. The three questions with the least agreement and no statistical significance in agreement among users were three, four and nine, all polygon examples.

Figure 2.10 shows little correlation between percent agreement and effectiveness. For example, question one has the highest percentage agreement among users, but a very low effectiveness score. Question seven had the best overall statistics (high effectiveness and percent agreement). It is an example that is very similar to some methods used for georeferencing remote sensing data (Appendix A).

By transposing the y-axis to the x-axis and using the within equation (2), agreement and effectiveness within raters was also plotted (Figure 2.11). Agreement within user and effectiveness do not have similar patterns, however rater B, P and S clearly stand out as the best raters. Rater B and P considered themselves intermediate GIS users, while rater P was the only the only test subject to consider himself or herself an expert in GIS. A review of the statements from students for each question suggest that the interpreters relied heavily on their past experiences and knowledge. GPS knowledge, digitizing experience, remote sensing experience and spatial analysis experience were feedback responses most common among the better performing raters.
5. Discussion and Conclusions

The results on agreement provide important insights into how users of GIS perceive geospatial uncertainty, but pose an interesting problem for creating GIS tools for spatial confidence intervals. To some degree, although not statistically proven, experience and education explained a portion of the variance in responses. The only self-identified advanced GIS user was the best rater, making it important to assess a user’s experience in any tool produced from this research. A survey on only advanced users may result in better agreement. This should not discount the current significance of response agreement, instead that some linkage needs to be made between metadata and spatial confidence intervals. Training is likely to improve usage of this classification system and by improving the experimental design via making classes equiprobable, new research could improve use of this or a similar system.

It is most interesting that the point examples were of the highest percentage agreement and that four of five, point examples were statistically significant. One explanation, is that point based error modeling is more common to geospatial data. Conversely, the point questions had
fewer options for answers. Even though accounted for, it is hard to accept 31.82% agreement as better than chance. The debate over how much agreement is “good” is emphasized by the highest percent agreement (question 1) having an effectiveness score lower than random chance.

Nonetheless, this distinction between geometry needs further exploration. For instance, why would polygon features have lower percentage agreement than lines? A polygon is no more than a collection of lines that is closed. Are there conceptual differences between points, lines and polygons, or is it because they represent different types of geographic data?

One could speculate as to why agreement varies with geometry and why the test subjects agreed so little, but the greater importance of this framework is that it provides a systematic structure to model error in geospatial data. Individual analysts may find no universal answer for these twelve questions and for that matter any data, but applying some estimate for confidence in spatial operations is a step forward.

Specific situations may determine whether a GIS user creates a confidence interval using the provided metadata and frequentist methods, or chooses credible intervals, the Bayesian approach. The test subjects’ comments indicated that they did not really know how to apply meta-data to this framework and that they relied heavily on their past experiences. Therefore, it seems critical that the storage of geospatial data accuracy be advanced beyond the overall statistics used in metadata and be attributed at both the feature and segment level. These recommendations follow closely with the idea of measurement based GIS (Buyong and Kuhn 1992, Goodchild 1999, Leung et al. 2004a), but always having the original measurements is not a reality so there needs to be alternatives. Brown and Huevelink’s (2007) DUE is the closest working system for the issues addressed in this paper, but this research takes the DUE a step further by try to implement a methodical process for creating and visualizing confidence interval in a GIS. Only when we
can pair the human dimension with quantitative measurements will it be possible to generate meaningful and useful intervals for geospatial data.
6. References


Goodchild, M. F., 1999. Measurement-Based GIS. *In: International Symposium on Spatial Data Quality*, Hong Kong, Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University


Von der Dunk, F.G., 2008. Legal aspects of using space derived geospatial information for


Chapter 3:
GIS and statistical analysis of digitized data’s spatial confidence intervals for a forest clearing

Andrew S. Foy a, Laurence W. Carstensen a, Stephen P. Prisley b, James B. Campbell a, and Randel L. Dymond c

aDepartment of Geography, College of Natural Resources and Environment, Virginia Tech, Blacksburg, VA, USA; bDepartment of Forestry Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, USA; cDepartment of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, USA

Abstract: Perception affects geographic representations of the real world and people’s interactions with data because to model natural phenomena, it is necessary to abstract and generalize. These processes can lead to uncertainty in maps and in Geographic Information Systems. To account for this, error-bands are commonly used to model uncertainty of geographic objects, for example a buffer around a road or forest clearing. The bands are comparable to spatial confidence intervals, but there is not always a sample to use for statistics, and there are limited tools to calculate error-band widths associated with a probability. One method to generate a sample is to have people or instruments collect multiple realizations. Another is to use Monte Carlo simulation, but when these methods are not possible, there are few other options for modeling the uncertainty of a map or in a GIS. This paper seeks to expand these options.

The objectives of this paper are: 1) to determine if people digitizing a forest-clearing boundary can estimate uncertainty in their representation, and 2) to explore the geometry of a forest-clearing boundary’s error-band. The test subjects’ estimates of 90% confidence intervals, at individual transects along their digitized line, were significantly different from 90% confidence intervals from the sample. Different segments of the line had varying size error-bands. Uncertainty was overestimated at crisply defined boundaries between grass and forest, but underestimated at fuzzy boundaries. The mean estimated confidence interval width was the same as the mean confidence interval computed from the sample. This suggests that vegetative boundary uncertainties are adequately modeled and estimated with a digitizer using a uniform error-band, transects, and a systematic method to analyze the data. The test subjects mean estimated 90% confidence interval width for a forest clearing was 5.7154m. The mean width calculated from the sample was 5.0263m, which was not statistically different from the digitizers’ estimate. This indicated that people using a GIS might be able to estimate uncertainty in their data during digitization. This research provides insight into the logic needed to program GIS tools for uncertainty analysis, but there still needs to be better methods to capture and quantify belief. Bayesian statistics may be useful in future research.

Keywords: GIS, digitizing, error-band, spatial confidence interval, forest clearing
1. Introduction

Geographic Information Systems (GIS) usage is increasing due to rapid growth in information technology, software and hardware becoming less expensive, and data becoming readily available. The usefulness of GIS is without contest; but the rapid advances in geospatial technologies and the popularity of geo-enabled software such as GoogleEarth©, Open StreetMap© Twitter©, and Foursquare© have overshadowed critical issues about how humans model the world. Ultimately, perception affects both representations of the real world and people’s interactions with data because to model natural phenomena it is necessary to abstract and generalize (Burrough 1996). Monmonier (1991, pp.1) states this best, “It’s not easy to lie with maps, it’s essential...to present a useful and truthful picture an accurate map must tell white lies.” This paradox raises questions about the confidence one should have in GIS outputs because most objects on a map or in a GIS vary in scale, space, and time, thus forcing the analyst, cartographer, or person collecting the data to select specific aspects of phenomena to represent (Burrough 1996, Zhang and Goodchild 2002). Practical experience plays a large role in making such decisions (Frank 1996). To use such knowledge, there needs to be a conceptual and logical GIS model that can utilize human knowledge and belief to improve data quality and better understand uncertainty of geospatial data.

Many types of geospatial data represent natural phenomena (e.g. land cover, floodplains, habitat, glaciers, etc.) and have indeterminate boundaries, but man-made objects, which typically possess a lower degree of uncertainty in boundary locations, also present problems when depicted on a map or included in a GIS (Couclelis 1996, Frank 1996). For example, it may be easier to define the edge of a road than the edge of a forest, but when defining a beach and the
boundary of a beach, the question becomes both scale and context dependent (Lam and De Cola 1993). How can this information be recorded and useful to a GIS analyst?

Couclelis (1996) identified 120 different cases in which uncertain boundaries exist and categorized these in eight different classes based on how uncertainty may or may not realistically matter. She used a framework that considered: 1) the empirical nature of the measurable characteristics of an object, 2) the mode of observation, measurement, or representation, and 3) the purpose for which the representation will be used. Her paper was groundbreaking because it shifted research emphasis from an analysis of the uncertain boundaries themselves, to an examination of the conditions under which they develop. It focused on the process of constructing knowledge and quality of knowledge.

By meshing the concepts of Couclelis (1992, 1996, 2003, and 2010) with digital data entry uncertainty (Jenks 1981, Carstensen 1990), this paper explores human perception and knowledge of boundaries in efforts to take steps towards creating an ArcGIS© tool for uncertainty. Often, a measurement device’s accuracy and precision are so well known and accepted, that beliefs in the quality of results are unquestioned. For GIS to expand further in the scientific domain this paradigm must shift.

Most discussions of quality in a GIS focus on issues of data uncertainty, accuracy, and error (Goodchild and Gopal 1989, Beard et al. 1991, Goodchild 2010). However, these are only one aspect of a final product’s “quality.” Couclelis (1992) extended the quality debate by creating an outline for a mental “expert” system designed to interactively monitor the quality of a GIS product. By accounting for the human dimension, data quality became a product of the needs of the user combined with the precision and accuracy of the measurement device. Without a tool to utilize such a system, is it ethical to conduct research with unknown uncertainties, if there is a
strong belief in the result’s accuracy, usefulness, or meaningfulness? This question is open to
debate and leads to other interesting questions. For instance, can a person determine the
uncertainty or error of the data they use? Can they define error for data they create? On the
other hand, can they define an acceptable threshold for error based on the purpose of their GIS
analysis? All present interesting situations. However, the topic of interest in this paper is on
error in data people create.

The problems associated with assessing the quality and uncertainty of GIS outputs result
from the diverse methods for modeling geographic phenomena and the overlay process (Veregin
1995, Sae-Jung et al. 2008). Thus, it is critical to determine if the geospatial analysis methods
are appropriate given the data and problem, and to produce measures of confidence. To-date, no
tools for GIS software (such as ArcGIS©, MapInfo©, or IDRISI©) exist that specifically address
these issues. What the GIS community needs, is an uncertainty tool that can bridge the gap
between measurement-based observations, statistics, belief for both experts and the average user
of GIS, and an understanding of the beliefs of data entry personnel who create the data in the
first place.

There is an obvious need to link human cognition with geographic measurements, as noted
by Menniss et al. (2000) and more recently with ontological debates from Couclelis (2010).
These GIS papers parallel one of the biggest hurdles for artificial intelligence, which is
commonsense reasoning in the spatial domain (Brachman et al. 1992). Couclelis (2003) even
suggested the need for an “encyclopedia of ignorance to map the terra incognita of epistemic
impossibility.” It is obvious that GIScientists need to know the bounds of what can and cannot
be known. For GIS to advance and be considered a rigorous and valid scientific tool, the
methods of GIS analysis must evolve to capture and quantify uncertainty in all of its measurable
forms. If recording and using such data are possible, statistical inference in a GIS will be greatly enhanced.

The purpose of this research was to test techniques for recording information about the uncertainty of geospatial data in a GIS, to determine if digitizing personnel’s estimation of their digitizing uncertainty agreed with the true uncertainty in sample data. The findings of this paper contribute not only to the theory of uncertainty in GIS, but also to the development of GIS uncertainty tools. They also addresses the lack of ability to estimate spatial confidence intervals in a GIS, using Brown and Heuvelink’s (2007) Data Uncertainty Engine tool (DUE). Given the DUE tool’s limitations and need for GIS data to have statistics for uncertainty estimates, the goal of this research was to collect and analyze the error in multiple realizations of people’s digitized representations of forest-clearing boundaries and compare that sample to test the subjects’ believed or estimated uncertainty.

2. Uncertainty Basics

Stories of the unknown and infinitesimal, such as Zeno’s paradoxes (e.g. “Achilles and the Tortoise” and the “Grain of Sand,”) and other philosophical debates have particular relevance to the uncertainty of geographic representations. Literature on such debates in metaphysics and cartography are long researched topics, predating GIS. The widely scattered fields that have contributed to the advancement of GIS and uncertainty research include geodesy, cartography, photogrammetry, engineering, and computer sciences, but there are limited applied solutions to the problems that arise from the spatial fusion of data. An offshoot of these fields became GIScience, which considers uncertainty a central issue surrounded by society, the human, and the computer (Goodchild 2010). The discipline focuses on the big picture (i.e. the effects of GIS on
the scientific community). It also emphasizes that managing uncertainty is of vital importance in GIS, especially as the technology permeates a multitude of subjects. GIScience and geospatial analysis is expanding beyond traditional geography and influencing disciplines ranging from business, law, and engineering to archaeology.

With geospatial technology’s evolution, GIS professionals have become aware they may be held legally accountable for the accuracy and reliability of their information, especially if economic loss or harm results from its use (e.g. Indian Towing Co. v. United States [1955], Reminga v. United States [1978], and Zinn v. State [1983]). When important decisions or legal rulings are made, GIS estimates or results should come with the same statistics as a maternity or DNA test. It is critical to know the confidence of an estimate or the chance that it could be wrong.

Uncertainty in map and GIS research has progressed from the analysis of cartographic error, to thematic, positional, attribute, locational descriptive, and measurement error in manual digitizing, to creating error-bands that model positional uncertainty, to Monte Carlo simulation, to sensitivity analysis and using fusion tables for uncertainty analysis. It evident by the large amount and diversity of the literature, that there has been great progress in GIS uncertainty research, but there is still a need for a unifying conceptual data model for representing geographic data (Couclelis 1992, Goodchild et al. 2007, Voudouris 2010). Interestingly, progress has been slower at the conceptual and logical levels of abstraction (Voudouris 2010). For instance, it is still unknown whether people can estimate error and uncertainty in data they are using. Furthermore, uncertainty research is lacking in its application to generalized data, most often focusing on digitizing a sample line or polygon on a screen and not on digitizing an ill-defined boundary, such as ecological or vegetation boundaries.
The most distinctive type of human generalization is the creation and delineation of boundaries. These boundaries are the foundation of commercial GIS and many common GIS applications. By definition, boundaries are the outer limits of individual entities, but many times these boundaries are not recognizable in reality (Couclelis 1996). For instance, could you park your car on the border of your municipal boundary? Can you really stand on the US-Canadian border? Consider that boundaries, and how one models spatial phenomenon, vary based on inherent and operational error, as well as user needs. Certain features (e.g. buildings, roads, and parcels), are measured with high precision and accuracy, but a layer from a GIS analysis using multiple datasets (e.g. a map showing the habitat of a certain species or suitability map used for site selection), has few useable measures of uncertainty.

Quantifying geospatial data uncertainty is further complicated because a map or representation of a geographic phenomenon is usually a single observation, or if not, sample statistics are rarely provided. This equates to a sample size of one, and therefore it is less practical to make statistical inferences or imply cause and effect. A multiple realization approach to mapping, using more than one person or instrument can be used to increase observations, but this is not generally practical. One solution is to use computer generated multiple realizations from Monte Carlo simulation, but the design of such experiments requires either prior knowledge of the data’s error distribution or making assumptions and using arbitrary numbers in place of variance and mean. Guidelines or insights into how to apply estimates for Monte Carlo simulation are limited (Openshaw et al. 1991).

One potential alternative for uncertainty estimation with a small sample size is to use Bayesian statistics, in which belief is a component of probability. Bayesians logic allows one to say that there is a given probability that an estimate contains the mean, while frequentist logic
cannot. Bayesian probability is the degree of believability, whereas the frequentist probability references the number of occurrences. Bayesian logic lends itself well to using expert knowledge to evaluate the gains or losses expected from possible consequences of decisions, especially in ecological and biological studies (Wolfson et al. 1996 and Ellison 1996). Readers who are not familiar with the differences between Bayesian and traditional statistics may consult Ellison (1996).

2.1. GIS vs. Cartography

GIS is used in many fields extending beyond traditional geography, and influences disciplines from business to law, engineering, and archaeology. Even social network data from Foursquare© and Twitter© are spatially referenced and used in GIS. The growth of GIS and subsequent decline of paper maps can be attributed to maps having fixed scales, extents and symbology, while GIS is dynamic, can be applied to a wide array of scientific applications, and can handle many different types of data. GIS data can also be easily reproduced and distributed. For instance, the Federal Emergency Management Agency (FEMA) recently announced that they would only provide communities with one paper map along with a digital copy of the 2009 Flood Insurance Study (FEMA 2009).

Other advantages to GIS are that they utilize spatial data fusion in which multiple layers of spatially defined data that may vary in timeliness, source scale, origin, reliability and compilation approach are combined or overlaid. This computational power provides users with a rich set of geographic data, for a wide array of applications, but can also lead to high levels of uncertainty in analysis (Chrisman 1982, Burrough 1996, Zhang and Goodchild 2002). Therefore, spatial
data accuracy will always be a major issue within the GIS community (Goodchild and Gopal 1989, Goodchild 1991, Goodchild 2010).

Cartographers’ publications on paper map quality are arguably more useful than digital metadata, not because the information is different, but because it is harder to fuse graphic map data. For example, it is difficult to overlay a paper topographic map with an aerial photograph and subdivision plat because of spatial referencing and scale differences. Alternatively, if the data are georeferenced and in a digital format (as is the case in a GIS), the process is much easier. Data interoperability and the high precision of GIS have also expanded the domain of spatial analysis. Further advancements in printing, scanning, cameras, and remote sensing have made digital data more accessible and easy to create.

Digital data might seem more reliable, but it is arguable whether the quality of digital geospatial data is any better than paper maps. This is because a large proportion of today’s digital base data came from scanning and digitizing paper maps. Even with high quality data, it can be difficult interpreting an accuracy statement, especially when an analysis involves more than one dataset. The main limitation in applying cartographic research findings to GIS research on uncertainty is that data fusion is not nearly as prevalent.

Some critical research exploring GIS uncertainty comes from generalization studies in cartography. Puecker (1976) introduced the most widely used line generalization algorithm. It utilized recursive bands to remove unnecessary points from oversampled digital lines, similar to Ballard's ‘strip trees’ (Ballard 1981). Puecker’s paper has had long lasting effects on geospatial data analysis, but theory leads one to conclude that not all lines can be generalized with the same method. Mark and Csillag (1989) created a typology of cartographic line features and showed that lines of different types should be generalized differently. For example, a road should be
generalized differently than a climatic region because man-made linear features are constructed with smoothly flowing curves and straight lines segments. Natural linear features tend to be more complex, with larger and more frequent changes in direction. Information on how a line has been generalized, given there are multiple methods, could provide insight on how to better use geospatial data. For example, Figure 3.1 shows a surveyed municipal park boundary and generalized park boundary used for printing recreational maps. A conflict occurs over whether two structures are within the park boundaries or outside. The surveyed representation shows the structures outside the park, but the generalized version shows the structures inside. Generalized data are common in GIS because generalization can be inherent in the representation or mapping process, can result for processing the data, or both. It is imperative to know how and why data are generalized.

Figure 3.1. A typical point-in-polygon operation; displays the original and the generalized representation of a park boundary.

GIS data is much easier to manipulate and reproduce than a paper map. It is easier to process and overlay. Therefore, GIS data should have a unique data quality standard. There have been many efforts to define geospatial-data quality standards, but the efforts have struggled to keep
pace with technological innovation. The most notable published cartographic standard for spatial data was the National Map Accuracy Standard (USBB 1947), which applied to all U.S. federal agencies that produce maps. It predates digital GIS data, as it was intended for paper or graphic maps. Due to its continued reference in GIS, there are questions on how to apply that standard to digitally converted maps that adhere to NMAS. The standard sets a threshold for the error of large and small-scale maps. For example, the error of well-defined points can be no more than approximately 12.19m for a map at a scale of 1:24,000. Revisions to the NMAS were made several times by the American Society for Photogrammetry and Remote Sensing (ASPRS 1990), but these standards applied only to procedures used to produce the maps, not generally to a post-mapping assessment of those points, and there was a demand for a better standard that applied to graphic maps, digital data and data fusion.

To meet the growing needs of the GIS community and outdated map standards, in 1990 the Office of Management and Budget (OMB) established the Federal Geographic Data Committee (FGDC) for the coordination of geographic information and related spatial data activities. The FGDC published the United States Spatial Data Transfer Standard, defining five key components to spatial data quality: positional accuracy, attribute accuracy, lineage, logical consistency, and currency. This standard evolved into National Standard for Spatial Data Accuracy (NSSDA), a statistical and testing methodology for estimating the positional accuracy of points on maps and in digital geospatial data (FGDC 1998). This uncertainty standard emphasizes spatial confidence interval estimation in reference to measurements of better quality. This standard is fundamental to the concepts in this paper, but its use has progressed slowly because the design is more of a template for storing geospatial-data quality than a mandated procedure, and data sources of higher accuracy are not always available.
2.2. GIScience

Advancements in computer technology allowing small devices to process large datasets and complex algorithms have greatly expanded the development of GIS, and led to the evolution of new fields of research. For example, GIScience focuses more on how individuals use GIS rather than the individual programs that make up a GIS. Over the last 20 years, GIScience researchers have made great advances in defining, measuring, modeling, and visualizing uncertainty and data quality (refer to Fisher [2007] and Goodchild [2010] for a review of classic papers and key authors). An excellent, comprehensive synthesis on the nature of uncertainty and how it has been conceptualized and handled within GIScience is in Plewe (2002). As indicated by numerous conference sessions and journal articles devoted to the topic, uncertainty has become a central issue in GIScience research. Yet, basic questions remain unanswered. Does displaying uncertainty information help users by clarifying a map (Leitner and Buttenfield 2000; Edwards and Nelson 2001), or does it clutter the map (McGranaghan 1993)?

Other areas of research within the GIScience and GIS communities have looked at the application of “intelligent” systems to handle uncertainty. Burrough (1991) suggests intelligent GIS could help decision makers evaluate the consequences of using different combinations of data, technology, processes and products. The goal being to provide an estimate of the uncertainty expected in an analyses before they begin. Nijkamp and Scholten (1991) suggest using systems that evaluate proper uses of data for certain applications by asking analysts guiding questions.

Stoms et al. (1992) discussed knowledge-based approaches and predicted that GIS would be embedded in decision support systems. Others have tried to incorporate data quality modules in
a decision support system to assist users in the management of forest pest infestations (Elmes and Cai 1992). They tried to determine whether analysts' needs could be met by examining the data to be used, the spatial processes, and the nature of the outputs through an on-line question and answer tutorial.

Beard (1989) suggests that databases should be re-designed to help prevent misuse by structuring them so that validity of mathematical operations may be verified before processing. In that experiment, users were given explanatory warnings prior to analysis, in an effort to better inform them of the outcomes. While there are advantages to a warning system, Agumya and Hunter (1996) believe the uncertainty debate needs to advance from the effect of uncertainty in the information to considering the effect of uncertainty on the decisions which rely on it. Decision-makers need to be asking; “How good are my decisions?” as opposed to “How good is my information?” They proposed a risk-based assessment of fitness for use of geographic information, to determine acceptable use. While many methods have proved useful under certain circumstances and experimental conditions, there is still a need for more universal methods of uncertainty management.

2.3. Spatial Confidence Intervals, Errors, and Epsilon Bands

More specific to spatial confidence intervals is research on error/epsilon bands (Perkal 1956, Chrisman 1982, Blakemore 1984), s-bands (Shi and Ehler 1993, Shi 1994) g-bands (Shi and Lou 2000), and the g-donut (Dajun et al. 2003). The concept of using a buffer to show error developed out of research seeking methods to better define and represent the boundaries of the vector model (Perkal 1956). The Epsilon Generalization Model (EGM) uses circular bands to
represent the level of confidence in a boundary (Figure 3.2). However, the level of confidence is not associated with a probability or p-value.

Figure 3.2. Depiction of Perkal’s Epsilon Generalization Model

In the EGM the margin of error is equal across the length of a line, meaning that uncertainty is the same at the end and middle of the line. The EGM is conceptually easy to understand, but it is limited in its scope by a lack of suitable methods for calculating the width of the band, referred to as epsilon. Chrisman (1982) and Blakemore (1984) used the EGM to study cartographic error and more specifically categorized error generically, for example the error of a digitized line is 1.5m (refer to Jenks [1981] for a review of digitizing error). Research that is more recent suggests the shapes of such bands are not necessarily uniform (Dutton 1992, Shi and Liu 2000), and there is still a need for a unifying model, such as the error-band framework proposed in Chapter 2.

In a study similar to this paper, a method to assess the horizontal positional accuracy of class boundary location for large-scale categorical data in mountainous and protected areas was tested using a land cover dataset produced from color infrared orthophotographs (Bley and Haller 2006). They defined all possible pairs of neighboring habitat classes and classified the required accuracy by a confusion matrix (Table 3.1). The accuracy values are based on the guidelines for interpretation from a set of 200 areas and 981 sample points.
Table 3.1. Vegetation boundary error-band categories, \( r \) = the radius that contained \( \sim 95\% \) of the interpretations from Bley and Haller’s (2006) findings

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Accuracy Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Boundaries are defined sharply in the aerial image for example sealed roads and buildings</td>
<td>( r = 1 \text{m} )</td>
</tr>
<tr>
<td>2</td>
<td>Less sharply defined but nevertheless well recognizable as the transition of tree and bush vegetation to the other habitat types</td>
<td>( r = 2 \text{m} )</td>
</tr>
<tr>
<td>3</td>
<td>Indicated by gradual and frayed out transitions typical for boundaries between heath land with dwarf shrubs and vegetation poor areas</td>
<td>( r = 5 \text{m} )</td>
</tr>
<tr>
<td>4</td>
<td>Very blurredly definable boundaries which are typical for wooden areas</td>
<td>( r = 10 \text{m} )</td>
</tr>
</tbody>
</table>

In general, positional accuracy is defined through measures between the locations recorded in a database to locations determined with higher accuracy (Goodchild and Hunter, 1997). The problem with many types of environmental or ecological data is that the reference of higher accuracy is not available, so in Bley and Haller’s study, the true boundary was an interpretation. An interesting finding was that error was normally distributed. Prisley \textit{et al.} (1989), also found errors in a study on the mean and variance of GIS-based area estimates to be normally distributed, but suggest using caution because positional coordinates are not independent. More recently, Leung \textit{et al.} (2004) published a general framework for error analysis, but there is no GIS tool available to-date, and the methods focus on measurement based-GIS, limiting its application.

3. Methods

3.1. Study Area

A forest clearing in the Fishburn Experimental Forest, Blacksburg, VA (Figure 3.3) was the study area for GIS digitizing exercises, designed to assess user estimation of confidence
intervals, error, and understand how the error-band geometry framework proposed in Chapter 2 applies to vegetation boundaries. The forest is a 1,353-acre demonstration forest of the Appalachian hardwood and mixed pine-hardwood type. It is a unique research spot situated in the Ridge and Valley physiographic province and divided in 15 compartments. Each compartment serves a different purpose, such as, prescribed burns or studies on different silviculture practices.

This study area, originally cleared as a wildlife food plot, has remained relatively unmanaged, although used for educational labs. The orthophotograph from 1998 shows forest-clearing boundaries in different stages of ecological transition. Some areas of the boundary
appear easily defined, others fuzzy. It is rare to observe succession in the classic sense. This study area is a product of contemporary succession where disturbance, most often human induced, is too frequent for a true climax community (Woodward 2009 pp. 60-66). This process is also noticeable on the ground. For example, Figure 3.4A (near transect E) is clearly a grassy area, likely from people walking and fauna grazing, whereas Figure 3.4B (near transect G) shows an area where there is more woody vegetation. This one clearing provided sufficient area to study digitizing uncertainty under various boundary conditions.

![Figure 3.4. Pictures of different boundary conditions in the field from the study area; A is near transect E and B near transect G (used with permission from Stephen Prisley, 2011)](image)

3.2. **Experimental Design**

Students (n=27) from a GIS for natural resources class were given a lab exercise asking them to digitize the boundary of a forest clearing from a color infrared orthophotograph (1m resolution; UTM, Zone 17N, NAD 1983) (Figure 3.3). After digitizing, they recorded uncertainty at each transect (Figure 3.3) by estimating a 90% spatial confidence interval for their interpretation of the boundary. As defined in the lab, the spatial confidence interval is a zone, in which it may be uncertain whether a point is in the forest or clearing. Think of it as representing the area in which the boundary lies with 90% probability. This experimental design gathers estimated confidence intervals, provides a sample for comparison, has potential for applications using
FDGC standards, and uses methods that can be replicated in research on other boundaries. The students used ArcGIS 9.3.1© for the exercise.

The first step involved students digitizing the boundary of the forest clearing using the Editor tool and only the orthophotograph as a base-map. To reduce any statistical bias the students had to wait until after they finished digitizing to add the transect layer to their map. The layer contained eight lines, labeled A-H. At the intersection of each transect with the test subject’s forest clearing boundary, they estimated the width of a 90% confidence interval perpendicular to their boundary edge, in both directions (Figure 3.5).

![Figure 3.5. An example of using the measure tool to estimate the width of a confidence interval at transect and clearing boundary intersection.](image)

The transects provided references for specific point locations to estimate uncertainty. The location and orientation of transects were designed to provide a complete sample of boundary conditions along the clearing. One concern was that a line may cross a specific transect more than once, leading to an unbalanced sample design. Transects were placed in locations to minimize the occurrence of multiple line crossings and have good coverage for the different boundary conditions (see Figure 3.3). Each student was identified with a random identification number to ensure anonymity.
3.3. Data Analysis

In some cases the estimated widths inside and outside of the boundary were not the same. As a result, the test subjects’ mean estimated confidence interval widths (m) at each transect for their line (EstBand) was used for comparison and statistical analyses with confidence intervals and the median digitized line computed from the sample. This was calculated using measurements of the test subjects’ line deviation from the “true” line (median line), at transects. In theory, the “true” line could be any measure of central tendency, but there are no good algorithms for computing a mean line or polygon. Therefore, it was more practical to compute the median boundary. The process involved converting each polygon feature to a raster. Each realization of the clearing was reclassified to a value of one and forest to zero. The rasters were added together using an additive overlay. The output of the raster additions was reclassified to select the 14th value, the median boundary edge. This was converted back to a polygon feature. Transect lines were split at points of intersection, from which the end-point at the intersection was selected and used to compute the distance from the median. This process produced measures of deviation (Deviation) for each realization of the forest boundary, from which it was possible to create confidence intervals widths (MeasBand) and summary statistics for the data.

Statistical analysis of the data was conducted with JMP 9© for summary statistics (Table 3.2) and to test several hypotheses. Out of the 27 students 21 correctly completed the exercise by answer all the questions and not having a digitized line crossing a transect more than once. All 27 representations were used for finding the median line, but for statistical analysis of deviation, the sample size varied at transects. Fourteen duplicate transect crossings were removed: six at H, five at F, two at E and one at G, creating an unbalanced design (Figure 3.6). Discarding these results reduced the power of the analysis slightly, but is better than guessing which intersection

57
was used by the test subject in his/her estimate. 201 transect crossings were used in the final analysis.

Table 3.2. Summary statistics for the transect and whole boundary

<table>
<thead>
<tr>
<th>Transect</th>
<th>N</th>
<th>EstBand Mean</th>
<th>Deviation Mean</th>
<th>MeasBand Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>27</td>
<td>3.168</td>
<td>0.7469</td>
<td>0.9541</td>
</tr>
<tr>
<td>B</td>
<td>27</td>
<td>3.2157</td>
<td>0.9813</td>
<td>1.2995</td>
</tr>
<tr>
<td>C</td>
<td>27</td>
<td>3.4683</td>
<td>0.9597</td>
<td>1.2556</td>
</tr>
<tr>
<td>D</td>
<td>27</td>
<td>4.2417</td>
<td>1.2881</td>
<td>1.7809</td>
</tr>
<tr>
<td>E</td>
<td>24</td>
<td>5.5292</td>
<td>2.8254</td>
<td>3.7738</td>
</tr>
<tr>
<td>F</td>
<td>22</td>
<td>7.4673</td>
<td>5.2389</td>
<td>7.4809</td>
</tr>
<tr>
<td>G</td>
<td>26</td>
<td>11.9785</td>
<td>10.8785</td>
<td>14.0875</td>
</tr>
<tr>
<td>H</td>
<td>21</td>
<td>6.6548</td>
<td>6.3260</td>
<td>9.5779</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td><strong>201</strong></td>
<td><strong>5.7154</strong></td>
<td><strong>3.6556</strong></td>
<td><strong>5.0263</strong></td>
</tr>
</tbody>
</table>

Figure 3.6. Individual realizations of the forest clearing and the point of intersection between each person’s line and transects (orthophotograph [online] http://geoserve.asp.radford.edu/doqq/B/blacksburg_1_01.htm, Photo 8, [accessed April 2 2009]).
To reiterate, the objectives of this research were to identify the type of error-band geometry and if the test subjects could accurately estimate a 90% confidence interval. The hypothesized questions were: 1) Are the group means of EstBand different at each transect (A-H)? 2) Are the group means of Deviation different at transects (A-H)? 3) Is the group mean of EstBand different from the group mean of MeasBand at transects (A-H)? 4) Is the grand mean of EstBand different from the grand mean of MeasBand?

3.3.1. Questions 1 and 2

The group means of EstBand and Deviation were compared to identify the error-band geometry of a forest clearing as either uniform or segmented (refer to Chapter 2); a single feature cannot be classified as non-uniform. Using ANOVA to compare the means and using Transects as a grouping variable, it was possible determine that the test subjects, on average, had different measures of deviation in their digitized line, and that they estimated 90% confidence interval widths differently at transects A-H. The ANOVA result on transect group means for Deviation (F Ratio = 13.3051, Prob>F < 0.0001) and EstBand (F Ratio = 9.6010, Prob>F < 0.0001) were significantly different (Figure 3.7). ANOVA was not run on Deviation and Estband using Transects as a blocking variable because of the unbalanced sample design and because EstBand is a 90% confidence interval, while Deviation is the digitizer’s distance from the median line.
Figure 3.7. Graph of ANOVA results on Deviation and EstBand and Transects as the grouping variable. The top and bottom of each diamond represent the 95% confidence interval for each group. The mean line across the middle of each diamond represents the group mean. Overlapping marks indicate that the group means are not significantly different.

In addition to the ANOVA, an ANOM was run to specifically identify which transects were significantly different from the grand mean of Deviation and EstBand. ANOM compares the group mean to the overall mean, where ANOVA compares the group mean to the other group means. ANOVA tells you if there is a significant difference overall, while ANOM indicates which group mean is significantly different. In Figure 3.8, the centerline indicates the grand mean, UDL is the upper dimensional limits, and LDL is the lower dimensional limits. If a group mean falls outside of the decision limits, then that indicates a significantly different group mean from the overall mean. Transects A and G were the only two that were significantly different in Deviation and Estband. Transects D, E, F, and H were not significantly different from the overall mean in both groups. Note that the averages differ slightly from Table 3.2 because the ANOM accounts for the different group sample sizes.
3.3.2. **Question 3**

The above questions answer whether uncertainty differs around the boundary of a forest clearing, but not if people can estimate the width of these varying confidence intervals. This question was tested by taking the mean Deviation and approximating a 90% confidence interval by using the Analyze tool in JMP 9®. This approximation (*MeasBand*) was compared to the estimated confidence interval from the test subjects, to determine if they could accurately estimate error and uncertainty in GIS data. The data were not normally distributed and the unbalanced sample design when grouping with transects lead to using the Wilcoxon Signed Rank test; a nonparametric version of the paired t-test that compares the sizes of the positive differences to the sizes of the negative differences (Conover 1999, pp. 350).

There was statistically significant evidence that an estimated 90% confidence interval (*Estband*) is different from the actual size of a confidence interval calculated from the sample (*Measband*) at all transects except for F (Table 3.3). There were both overestimations and underestimations. Overestimation happened at transects along crisply defined boundaries (A-D), and underestimation at E, G, and H.
Table 3.3. Wilcoxon Signed Rank test results identifying transect F as the only estimate that was not significantly different from the sample transect F 90% confidence interval.

| Transect | DF  | Signed Rank | Prob > |t| |
|----------|-----|-------------|--------|---|
| A        | 26  | 189         | <0.0001* |
| B        | 26  | 157         | <0.0001* |
| C        | 26  | 181         | <0.0001* |
| D        | 26  | 173         | <0.0001* |
| E        | 23  | 103         | 0.0015*  |
| F        | 21  | -39.5       | 0.2066  |
| G        | 25  | -78.5       | 0.0437*  |
| H        | 20  | -68.5       | 0.0131*  |

3.3.3. Question 4

Transects have different levels of uncertainty, and either overestimation or underestimation were more frequent at different transects, but what if the error geometry of the forest clearing is assumed to be uniform (i.e. the same width for the whole clearing). Do the test subjects better estimate a 90% confidence interval when the mean of transects are compared using ANOVA? The group transect means of EstBand and MeasBand are not significantly different from each other (F Ratio = 0.1162, Prob> F = 0.7382). The mean of MeasBand (5.0263m) was slightly below the grand mean (5.3708m), whereas, the mean of EstBand (5.7154m) was slightly greater.

When transects were used to estimate uncertainty, the transect means or group averages were no different from the true uncertainty identified from the data. There was a slight tendency to overestimate the width of the band. The map in Figure 3.9 showing the EstBand and MeasBand bands at transects overlaid with a uniform error-band uses the mean of EstBand and supports using an averaged error-band width estimated from transects. The estimated and true size of a
90% confidence interval is so close at transect F at the map scale, that only one buffer is visible. Areas of over and under estimation are clearly visible and related to crisp and fuzzy boundaries.

![Visualization of EstBand, MeasBand, and the Uniform Estimated Error Band](image)

Figure 3.9. Final map showing the different spatial confidence interval widths and a uniform error-band (orthophotograph [online] http://geoserve.asp.radford.edu/doqq/B/blacksburg_1_01.htm, Photo 8, [accessed April 2 2009]).

The results of a bivariate analysis on EstBand and MeasBand from the data in Table 3.2 help to visualize the relationship between the two estimates of error-band width and identify associations (Figure 3.10). The regression line and 95% ellipse clearly show that the test subjects mean estimate of a 90% confidence interval is similar to the 90% confidence interval computed from the sample. The figure helps support that the digitizers could identify whether
they were confident or not confident in a line location. There is an $R^2=0.92$ for the regression line and a correlation coefficient ($R = 0.9629$), which indicates there is a strong association between the two values, which is what would be expected if the test subjects were good estimators. It should be noted that the regression equation from that bivariate analysis is not good for prediction because it does not follow all the assumptions for predictive inference using parametric statistics, but it does offer insight into peoples’ intuitive abilities to estimate uncertainty (e.g., they know whether the error-band is wide or narrow).

![Figure 3.10](image)

Figure 3.10. Bivariate analysis on $EstBand$ and $MeasBand$ with a regression line and normal correlation ellipse; the horizontally stretched band, points’ close proximity to the regression line, and upward trend show a strong positive association between $EstBand$ and $MeasBand$

4. Discussion and Conclusions

This paper sketches out a template for spatial confidence interval estimation and error analysis under various boundary conditions. Several findings from this research have relevance to GIS uncertainty modeling and theory. Foremost, it supports the adoption of a unified conceptual
error-band framework and methods, but raises some theoretical questions on its proper use. For example, if you know a polygon boundary’s confidence interval has segment-based error-band geometry, is it acceptable to use an averaged interval width and treat the error-band as uniform? Does it matter which model is applied, especially since subjects did not estimate the individual transects particularly accurately? There is certainly a demand for a simple, unified conceptual data model that manages GIS uncertainty, but despite all the literature, it is hard to apply any of the current published models (Chrisman 1982, Blakemore 1984, Dutton 1992, Shi and Liu 2000, Dajun et al. 2003, Leung et al. 2004, Voudouris 2010). The best attempt at an error-band modeling tool was in Gürçük’s (2007) thesis, but it was limited in scope and availability. The main difficulty in using error-band models is there are no tools readily available that implement them, whereas all the necessary geoprocessing tools and the model for assessing spatial uncertainty are available in this paper.

Evidence supports incorporating beliefs or estimates of uncertainty into GIS models as a feasible alternative to generating multiple realizations of data, but at a simplified level and with a systematic approach. This finding supports exploring and potentially using Bayesian methods for confidence interval estimation. In general, the estimates of uncertainty in this research were conservative; the test subjects overestimated the width of error-bands, perceiving greater uncertainty than actually existed. These findings will greatly assist in the development of a tool for spatial confidence interval estimation in a GIS, but many other types of boundary and positional uncertainties need to be included in a flexible GIS uncertainty tool.

The major theoretical debates involving the results from this paper relate to the shape and formulation of the error-bands. Take for instance the results from Question 1, that the error and uncertainty of a digitized forest-clearing boundary are different when measured at transects. It is
likely that the shape and size of a spatial confidence interval has as much to do with the purpose of the data as the collection method and sample statistics. This suggests there is some flexibility in how one models uncertainty. This logic is consistent with Couclelis (1996) boundary class framework, and the line generalization typology of Mark and Csillag (1989). There are also parallels with Bley and Haller’s (2006) study, which grouped vegetation boundaries into four categories. Interestingly, the mean confidence interval for this forest clearing is similar to their third category (5m error-band), but far narrower than Jenks’ (1981), 1.5 m digitizing error. It is worth mentioning that much of the literature is inconsistent in the use and interpretation of the level of confidence associated with the error-bands, and this makes comparisons difficult.

It seems that there may more than one method to estimate a spatial confidence interval (e.g. segmented vs. uniform), but there are tradeoffs with the method chosen. For example, if the forest-clearing’s spatial confidence interval was used for area calculations to estimate the number of seeds to plant, the choice of segmented or uniform error-band could have significant impacts on the area estimates. Alternatively, a point-in-polygon analysis could yield the same results, regardless of error-band geometry. Further research is needed in respect to the variable methods and the consequences of using one method over another.

This study supports the theory that natural boundaries are likely to have varied levels of uncertainty and that the shapes of their spatial confidence intervals are segmented-based across the length of boundaries. This finding has particular relevance when considering the widespread use of landscape metrics (Arnot et al. 2004), and for detecting change in vague interpretations of the landscape (Fisher et al. 2006). Calculating acres of deforestation, carbon sequestration, and even ecosystem services rely on area measurements. Arnot et al. (2004) offers limited reassurance in area estimates at fuzzy boundaries; but by utilizing the methods presented in this
paper, it is possible to provide real measures of confidence in spatial estimates. However, it is unlikely that a U.S. Forest Service GIS technician would GPS the perimeter of a clearing from a wildfire multiple times to generate a sample.

Due to the logistics of sampling spatial data, the test subjects were asked to estimate the width of error-bands as 90% spatial confidence intervals. Such estimates could greatly enhance our understanding of landscape change. For example, there is direct application to Resler et al. (2004) and Malanson et al. 2007 research on alpine tree line mapping in response to climate change. Resler et al. (2004) mapped vegetative boundaries and reported accuracy as the number of correctly classified pixels divided by the total number of reference points. A classical accuracy assessment does not translate easily into an error-band model. Findings presented herein could aid future research on landscape change. The simple fact that test subjects had a tendency to overestimate the size of a confidence interval at crisply defined boundaries and underestimate at fuzzy boundaries could be a significant factor in change detection. However, they know whether an error-band is wide or narrow.

It is perplexing that the actual and perceived error and uncertainty were different along the length of the boundary, and that people could not accurately estimate 90% confidence intervals at transects. However, the average estimate of a 90% confidence interval, uniformly distributed across the perimeter of a forest clearing, is the same as an approximated 90% confidence interval from the sample. This is promising, but not optimal. It could be that the concept of segmented error-band geometry is too complicated, that it is too hard to estimate, or just too new a concept at this point in data development. Alternatively, the uniform error-band geometry may be more easily estimated and appropriate given the application or type of data. Even with the variances in estimation, the estimates appear to be strongly correlated. A larger, balanced sample and more
transects along boundaries could offer more insight into how people conceptualize and model error-bands.

The experimental design of this study is flexible enough that one could study many different types of boundaries. Such studies could include both natural and manmade features. It is likely with further research that some linkages or suggestions for different types of data could be offered via a GIS “wizard-based” tool. The information collected through this experiment can easily be merged with attribute based, semantic and ontological driven geographic data models. The problem with comparing this study to other research is that the test subjects were asked to estimate a confidence interval and not deviation. It might be that confidence intervals are conceptually difficult to estimate, especially when 90, 95, 99% intervals are frequently used. Using error or deviation in another study may provide better information because it is easier to compare the students’ deviation to an estimate of deviation.

It would also be interesting to incorporate credible intervals using Bayesian statistics for comparison with the estimated and realized confidence intervals. The estimates in this study are a measure of belief and might fit well in to such a model. Further research needs to include test subject experience, knowledge on the subject matter, scale, time, etc. Such variables or relationships between knowledge and the uncertainty of a geographic representation could have major impacts within the GIScience and artificial intelligence fields.
5. References


Bley, D. and Haller, R., 2006. Checking the spatial accuracy of class boundaries with a varying range of accuracy requirements. In: 7th International Symposium on Spatial Accuracy in Natural Resources and Environmental Sciences, 2006 Portugal, 249-263.


Chapter 4:

Applying GIS uncertainty tools to floodplain mapping and decision making

Andrew S. Foy a, Laurence W. Carstensen a, Randel L. Dymond c, Stephen P. Prisley b, James B. Campbell a

aDepartment of Geography, College of Natural Resources and Environment, Virginia Tech, Blacksburg, VA, USA; bDepartment of Forestry Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, USA; cDepartment of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, USA

Abstract: Floods are the most common and expensive hazard in the United States. To communicate flood hazard risk FEMA publishes DFIRMs, which are digital maps of Special Flood Hazard zones. These maps require continuous maintenance and revision, and are subject to uncertainty and dispute. In addition to DFIRM uncertainty, building location uncertainty can also alter the results of GIS overlays used to identify buildings in the 100-year floodplain. While there is extensive research on floodplain uncertainty, there is very little literature that incorporates the uncertainty from spatial data fusion. To address this problem, a simple tool generates overlays that incorporate spatial confidence intervals. The tool allows a GIS analyst to interpret how variable GIS results could be and therefore make better decisions. The objectives of the research were: 1) to test the usefulness of this tool by determining how the results of a GIS analysis differ when uncertainty of the two sets of data is accounted for in the overlay process, and 2) to analyze the spatial pattern and characteristics of positional uncertainty in the 100-year floodplain boundary.

There was statistically significant evidence that the results of 100-year floodplain and building overlays are different when the uncertainty of the two layers is incorporated into the overlay process. The maximum percent-change between crisp and uncertain overlays for building counts was 66% and lowest 11%. The impacts of uncertainty on the overlays varied based on municipality and other geographic variables. The uncertainty of the floodplain had a larger impact on the overlay results than the uncertainty of the buildings, but it was very difficult to utilize the metadata in its current published format. As a result, some assumptions had to be made on the size of the confidence interval for the error-band geometry and the size of a 95% spatial confidence interval. These assumptions were estimates based off the best available metadata and current FEMA and NSSDA standards. The need to make assumptions highlights the inadequacy of metadata for a large portion of GIS data. The results from this paper clearly show the utility and need for a tool that incorporates spatial confidence intervals. The tool presented in this paper could be a tremendous resource for many GIS analysts because it clearly displays the high degree of uncertainty that can result when using traditional crisp overlays.

Keywords: Uncertainty tool, spatial data fusion, GIS, overlay, floodplain
1. Introduction

Floods are the most common and expensive hazard in the United States (FEMA 1992, Conrad et al. 1998), and have increased significantly in frequency and cost of damages caused over the last four decades. To address flood hazards, the National Flood Insurance Act was passed in 1968 to establish the National Flood Insurance Program (NFIP), designed to create national Flood Insurance Rate Maps (FIRMs). These flood maps identify the areas along streams that are subject to flood risk at the county level. The Federal Emergency Management Agency (FEMA) currently uses new digital FIRMs (DFIRMs), which for a majority of the U.S. have been modified and improved to set flood insurance rates, regulate floodplain development, and inform people living in the 100-year floodplain of potential hazards (GAO 2010).

Due to both natural and anthropogenic factors, these maps require continuous maintenance and revision, and are subject to uncertainty and dispute. For example, land development and natural changes to the landscape can alter hydrology, weather can be unpredictable, and uncertainty is inherent in both science and in mapping. Flood maps are typically created using a base-map of elevation data and hydrologic models to compute flood inundation extent, water depth and flow velocity, but the elevation base-data and hydrologic methods vary. While great progress has been made in improving accuracy of the mapping and in solving the flow equations efficiently and reliably (Hunter et al. 2007, Bates et al. 2010), uncertainty reduction persists as a major issue, especially in the vertical measurements of elevation data and the resulting horizontal accuracy of floodplain boundaries—they are dependent on each other. Uncertainty is inherent in all geospatial datasets and “not just a flaw to be excised” (Couclelis 2003 pp. 166, Roth 2009).

In response to stakeholders’ concerns about the quality of flood data used to develop new flood maps during the Flood Map Modernization Program (1997), FEMA issued the Floodplain
Boundary Standard (FEMA 2007). The purpose of the FBS was to ensure that locations of the predicted floodplain boundary and base flood elevation lines drawn on flood maps are comparable to the topographic data selected for the study area. This standard was a step towards reducing and better understanding uncertainty, but it does not take into account uncertainty that occurs from the fusion of flood maps with structure and building maps. Such overlays are used in loss-estimate analyses and for identifying structures at risk of being flooded.

Misuse or a poor interpretation of floodplain maps carries real world consequences. For example, uncertain data contributed to the underestimation of the worst-case flooding scenario of the Upper Mississippi in the Great Midwest Flood of 1993, causing many flood levee failures, which led to 12-16 billion dollars in damages, the dislocation of 50,000 people, and significant loss of life (FEMA 2003). A similar scenario occurred along the Gulf Coast after hurricane Katrina in 2005, and in April and May of 2011, the Mississippi River experienced a 500-year flood. The flooding was unexpected and resulted from above average rainfall and snowmelt runoff. Flooding and flood mapping are uncertain phenomena.

The research reported here contributes to the knowledge base on the uncertainties in floodplain mapping by addressing the uncertainty in a typical GIS analysis used to determine what structures fall within the 100-year floodplain using a new GIS uncertainty tool for spatial data fusion. The tool allows a GIS analyst to generate spatial confidence intervals and perform overlays that incorporate spatial uncertainty. The objectives of the research were: 1) to test the usefulness of this tool by determining how the results of a GIS analysis are different when the uncertainty of the data is accounted for in the overlay process, and 2) to analyze the spatial pattern and characteristics of positional uncertainty in the 100-year floodplain boundary. This paper expands on FEMA’s Standard Floodplain Boundary Audit Procedure (2007) to better
inform landowners, residents, insurance companies and mortgage lenders of the likelihood a structure is in the floodplain.

Flood hazards are on the rise, making it reasonable that some people may err on the side of caution and insure property for flooding, even if it is outside the floodplain. Alternatively, there are people who might have difficulty affording flood insurance or they may not think they need coverage, when indeed they do. In such situations, it would be helpful to know if there is a chance a building is in the floodplain or if there is a chance a person has been unnecessarily paying for flood insurance. Maps that show only floodplain boundaries imply that every building in a designated flood zone may flood and every building outside the zone is safe; history has proven this is not the case. Baldassarre et al. (2010) suggest that a deterministic approach to floodplain mapping is misleading and a probabilistic approach produces a more correct representation of a floodplain. This paper builds on similar deterministic vs. probabilistic debates by applying probabilistic measures of uncertainty to discrete geographic representations of floodplain boundaries.

In addition to better informing the public, the findings from this research have relevance to FEMA by presenting a method for identifying places where there may be greater uncertainty in floodplains, whether it is due to land development patterns or the physical geography, and not necessarily the map. Improving flood hazard map analysis has the potential to improve hazard mitigation and emergency response. Furthermore, there is a demand in the GIS community for a tool to use for uncertainty analysis in spatial data fusion. Such a tool as presented in this paper has universal application, but a major question remains before releasing such a tool. Are the results from GIS overlays, such as polygon-intersects-polygon, statistically different when uncertainty is included in the analysis?
2. Flood Mapping History

Despite over 60 years of flood management guided by federal agencies, policy, and a large amount of respected literature, flood mapping and prediction remain uncertain and are not properly communicated to the public (Downton et al. 2005). Accurately mapping and predicting flood events are important to land management, water usage, agriculture, and disaster relief. The U.S Geological Survey (USGS), U.S. Department of Agriculture (USDA), National Oceanic and Atmospheric Administration (NOAA), FEMA, and several other agencies play a large role in flood mapping and management, but FEMA is the lead agency. Their purpose is to build, sustain, and improve the U.S. response, resilience, and mitigation for all types of hazards.

Flooding is the most frequent and widespread hazard, therefore the most damaging (GAO 2010). The high risk of floods caused FEMA to produce flood maps that have been in evolution since 1968, when NFIP was introduced. Since its inception, numerous revisions have been made to policy, methods, and maps, which have caused considerable unrest amongst stakeholders. The main purpose of NFIP was to provide flood insurance, reduce flood damages through floodplain management regulations, and identify and map the nation’s floodplains (FEMA 2002).

In 1997, FEMA designed the Map Modernization (Map Mod) plan to update paper FIRMs to digital and raise public awareness of the importance of the maps and of community involvement in map revisions. Map Mod was a critical program and technological advance, considering people use flood maps an estimated 15 million times annually for state and community floodplain management regulations, for calculating flood insurance premiums, and for determining whether specific property owners are required by law to have flood insurance (FEMA 2002). The initial plan was to convert approximately 80% of existing paper maps to
digital format, update 20% with new flood risk information and convert to digital, and to add 13,700 new digital maps for unmapped communities. There was also a planned maintenance program for updates based on new engineering data (FEMA 2002).

A mid-program evaluation of the program in 2006 resulted in an adjustment of the program’s goals and objectives. The objectives became to: 1) produce digital products, 2) provide new, updated, and validated engineering analysis, and 3) implement the Floodplain Boundary Standard (GAO 2010 pp. 16). Other revisions included the addition of a flood risk ranking system to prioritize and schedule mapping projects. The county scale ranking system used factors such as population, growth trends in housing units, flood insurance claims, repetitive loss properties, and flood disaster history. Map Mod ended in 2008, producing digital flood maps for 92% of the continental U.S. population. Many of these areas had outdated maps or no maps at all, making the program a success to some, but after a $1 billion investment, only 21% of the population had maps that met all of FEMA’s data quality standards (NRC 2009).

After Map Mod, FEMA and NOAA asked the National Research Council (NRC) to examine the factors that affect flood map accuracy, assess the costs and benefits of producing more accurate maps, and recommend ways to improve mapping, communication, and management of flood-related data. An innovative study from the North Carolina Floodplain Mapping Program (NCFMP) that used high-accuracy topographic data (such as LIDAR) and maps for floodplain analysis in mountains, rolling hills, and coastal plains, allowed the NRC to compare new and traditional mapping methods under various geographic conditions (NCFMP, 2008, GAO 2010). That study found significant differences in elevation values between LIDAR data and the USGS National Elevation Dataset (NED) used to delineate the 100-year floodplain for the three study regions. Topography and elevation errors have the greatest impact on floodplain map quality.
NED datasets do not always meet the requirements for DFIRMs for acceptance as FEMA compliant. Therefore improving elevation data will improve floodplain mapping, but it is expensive. The committee concluded that benefits outweighed the costs of more accurate flood maps and that improving the accuracy of elevation data would significantly improve prediction of flood extents.

The foundation of Map Mod and findings from the NRC evaluation of the program led to the Risk MAP multi-year plan, a five-year effort envisioned to improve data quality and increase public awareness. It implemented three standards for ensuring the quality of data in developing flood maps. The new ‘FEMA’s Guidelines and Specifications’ required using the Floodplain Boundary Standard (FBS) and two of three elements of the New, Validated and Updated Engineering (NVUE) standard (GAO 2010 pp. 16-17). The Mapping Information Platform (MIP) information system was established to provide resources to monitor the quality and process of flood mapping under Risk MAP. The goal of Risk MAP was to raise flood map standard compliance to 80%. As of January 2010, FEMA reported a 52% compliance rate. However, many have raised issues with FEMA’s metadata management process, and it has been suggested that FEMA establish separate measures of compliance from detailed and approximated study areas so that the FBS can be better used as a measure of map accuracy (GAO 2010 pp. 44).

Flood mapping is the process of defining an inundation or flood hazard zone on a map that graphically represents the land area covered during a flood event. In concept, the process compares the surface of water elevations with the surface of ground elevations to identify places where the ground elevation is below the water surface. There are various programs used to model flood hazards and risk, such as HEC-RAS©, HAZUS©, and ArcGIS©; and there are several different methods for delineation. For example, FEMA has specific guidelines for conducting
detailed and approximate study areas. Approximate study areas result in the delineation of floodplain boundaries for the 1% annual chance flood, but do not include base flood elevations (BFEs) or flood depths and rely on interpolation. The 1% annual chance flood zone is considered to reach the BFE boundary or be the 100-year floodplain. Detailed studies, at a minimum, result in the determination of floodplain boundaries for the 1% annual chance flood, include BFEs and use detailed engineering methods. Those types of studies typically involve using hydraulic modeling with HEC-RAS©. The data from detailed study areas are held to a higher quality standard, and it has been proven that including BFE data in DFIRMs improves map use; but the current metadata standards do not differentiate between the two types of studies (GAO 2010). FEMA does however, have a detailed flood-zone classification system.

FEMA calls flood hazard zones on FIRMs special flood hazard areas (SFHAs), defined as the area that will be inundated by a flood event having a 1% chance of being equaled or exceeded in any given year (Figure 4.1). SFHAs are labeled on FIRMs as A, AO, AH, A1-A30, AE, A99, AR, AR/AE, AR/OA, AR/A1-A30, AR/A, V, VE, and V1-V30. Moderate flood hazard areas are labeled as zone B or X (with shading on paper FIRMS). Those are the areas between the limits of the BFE boundary and the 0.2% annual chance, or 500-year floodplain. Minimal flood hazards are the areas outside the SFHA, higher than the elevation of the 500-year flood and labeled zone C or X (not shaded on paper FIRMs). FEMA’s numerous codes for the 100-year floodplain exist to communicate the type of flood study and take into consideration the geography of the area. Despite the detailed zoning typology, zones in similar geographic regions are often modeled using different methods. For example, zone A is an approximated floodplain, usually in sparsely populated areas, or in places where there has been little previous floodplain study done. Zone AO implies a detailed hydraulic analysis in an area with high flood velocities,
such as alluvial fans and washes. The AE zone on DFIRMs implies a detailed study area. It replaced A1-A30 on FIRMs. Although reclassifying AE zones may seem like generalizing, AE zone standard compliance is at a minimum, equal to that under Map Mod.

Figure 4.1. Depiction of FEMA’s Special Flood Hazard Area and normal stream bank.

The process for producing flood maps involves three main phases:

- **Scoping** - identifying flood risk, assessing immediate and future needs, and determining what type of flood study is possible with available resources. FEMA, state and local governments are all involved.

- **Development** - collecting technical data, modeling, creating a preliminary map, and performing quality control and quality assurance. Modeling and map production are carried out by a FEMA mapping partner, such as a contractor, state or local government employee.

- **Adoption** - period for public comment and appeal. FEMA, contractors, and state and local government agencies involved in the process must respond to comments made within the appeal period.
Four main approaches are used to study flood hazard along rivers and streams: 1) detailed studies, 2) limited detailed studies, 3) approximate studies, and 4) redelineation. Most of the time they are just referred to as detailed or approximate study areas. Each approach produces different information, yet all are used for the same purpose. The type of study depends on the type of flood hazard, the resources available, and the risk of flood damage.

The most expensive approach is the detailed study, which provides the most information about flood hazards by establishing base flood elevations, special and moderate flood hazard areas, and floodways. FEMA guidelines specify that the dimensions and elevations of all hydraulic structures and underwater sections adjacent to the structures must be obtained from available sources or field surveys (FEMA 2003). Limited-detailed studies provide similar information, but structures such as bridges or culverts represented in the models do not need to be verified in the field. The bridges and hydraulic structures data are typically extracted using field measurements or as-built data, rather than precise survey measurements.

Approximate studies provide an approximate outline of the floodplain or best guess given the available data, but no base flood elevations, floodways, moderate hazard areas, or other details. However, comparison of the floodplain boundaries to a topographic map can be used to get an estimate of the BFE, but the estimates are not suited for regulatory purposes. FEMA provides written guidance (FEMA 1995, FEMA 2009) and software for calculating approximate water surface elevations on open channels based on specified field measurements and methods. The bridges and hydrology data can be estimates from photographs, orthophotos, or existing topographic mapping, thus do not require field surveys (FEMA, 2003).

Two floodplain mapping methods may be used for approximate study areas. *Contour interpolation* plots limits on a contour map. The edges of the floodplain at any given location
can then be matched to a contour line on each side, or to an area between contour lines. The other, simplified method is *data extrapolation*. When a site is within 500 feet of a flood zone, BFEs have been determined and drawn as a profile and the ground surface over that area slopes at approximately the same slope as the studied area the profile line of the flood elevation can be extended and used as the flood elevation for the project.

Redelineation studies are no longer conducted by FEMA due to data quality and compliance standards, but they still are useful. These types of studies produced digital representations of flood maps as part of the national digital flood layer initiated under Map Mod prior to the mid-program adjustment. This method was considered superior to simply scanning the older FIRMs to convert to digital format. Redelineation uses existing flood elevation information to redraw flood boundaries using updated topographic data.

The detailed methods of flood mapping involve the use of georeferenced data, a hydrologic model and a one-dimensional or two-dimensional hydraulic model. FEMA DFIRMS have approximate and detailed study zones that may be constructed by quite different methods. The simplest form uses a one-dimensional hydraulic model consisting of the following steps (IACWD 1982, Noman *et al.* 2001, FEMA 2003, Merwade *et al.* 2008):

1. Estimate flow (e.g., 100-year flood) with a calibrated hydrologic model and precipitation input, or through statistical analysis.
2. Develop cross sections for the flood-prone study area, either by surveying river transects in the field or by extracting elevations along river transects from a digital terrain model, or more recently, from LIDAR.
3. Estimate water surface elevations with a hydraulic model using flow calculated in Step 1, the cross sections developed in Step 2, and any data on any structures (e.g., bridges, dams and culverts) along the waterway or other hydraulic parameters.

4. Georeference the water surface elevations from the hydraulic model on the digital terrain model, and a water surface, which is usually a triangular irregular network (TIN).

5. Subtract the values of the digital terrain model from the water surface values to create a map of water-depth. Positive values identify areas in the inundation zone.

One-dimensional models are computationally efficient and are considered to produce ‘accurate’ surface water profiles (Büchele et al. 2006). The one-dimensional approach assumes flow velocity varies only in the direction of the longitudinal channel slope. It is averaged over the depth and the width of the flow at cross sections. The depth of the water over all points in a cross section is determined by extending a horizontal water surface elevation line across the channel to create a single water surface elevation value. The floodplain boundary is delineated at the location where the water surface elevation line intersects the topographic surface representing land surface elevation (NRC 2009).

Cross-section geometry is determined at regular intervals along the centerline and at structures, river bends, and major points of change. Accurate representation of structures and river bends is important for identifying flow constrictions and areas where water can pool. Information about surface roughness, also known as flow resistance, is required for each cross section. Many different equations relate surface roughness to flow characteristics, but the most popular in open-channel flow computation is the Manning’s equation (Straatsma and Huthoff 2011).
Unlike the one-dimensional model, two-dimensional models are computationally demanding and require expertise to prepare and execute. Flow velocity is averaged over the flow depth, and velocity components are computed in directions both parallel and perpendicular to the longitudinal channel slope, giving velocity magnitude and direction. The models can interactively provide solutions that advance in space and time, but FEMA only requires a single discharge value for peak flow of the 100-year flood event, so most two-dimensional models assume steady flow (i.e., the water surface elevation is constant over time). The Transportation Research Board (2006) found that the choice of model dimension can significantly affect the determination of the floodplain because one-dimensional models cannot capture complex features, such as braided streams, multiple openings, and bridge crossings. DFIRMs do not provide any metadata to reference the type of model or parameters set.

3. Uncertainty in Flood Maps

The uncertainty in flood mapping is a major concern for GIScientists, engineers, regulatory agencies and the public because the misuse or inaccurate comprehension of floodplain maps carries real world consequences (for example, human loss of life and billions of dollars in damage). The uncertainty exists because of the complex and diverse methods for modeling floodplains (Anderson 2000, Robayo et al. 2004, Knebl et al. 2005, Merwade et al. 2008). Numerous software programs, levels of detail for a study, and scales of analysis are used to fuse geospatial data in floodplain boundary mapping. Furthermore, datasets for model parameters, such as stream flow (Parodi and Ferraris 2004), flood frequency (IACWD 1982, Griffis and Stedinger 2007), and elevations (Whang and Zheng 2005), are inherently uncertain. Data
commonly vary in accuracy, precision, and timeliness, all of which significantly affect the horizontal extent on flood maps (Merwade et al. 2008).

Traditional GIS uncertainty results from six factors: lineage, positional uncertainty, attribute-uncertainty, logical consistencies, completeness, and temporal uncertainty. These same factors also exist in flood mapping, specifically in the stream flow records, flood frequency analysis, regional flood frequency analysis, hydrologic models, terrain datasets, hydraulic models, and level of detail for the study area (Smemoe et al. 2007, Merwade et al. 2008). How does such drastic variance influence a representation of a floodplain boundary?

Stream flow records are used to model flood frequency and as parameters in hydrologic models. Flow rates are derived using stage-discharge rating curves and are rarely computed for low-frequency events. Furthermore, they are subject to inherent uncertainty (Fread 1982, Freeman et al. 1996, Schmidt 2002, Parodi and Ferraris 2004, Merwade et al. 2008). Guidelines for estimating flows are published by the Interagency Advisory Committee on Water Data (IACWD 1982), which recommends using the Log-Pearson Type III probability distribution with a historical weighting procedure to model flow. However, a recent study by Griffis and Stedinger (2007) found that incorporating additional historical or even paleoflood information into the models can significantly improve parameter estimation.

Flow is usually estimated with water gauge data, though when these are not available, regional regression equations, developed by the USGS, can be used. These equations incorporate drainage area and shape, channel slope, impervious surface area and precipitation to model the hydrologic processes of a region. Merwade et al. (2008) cite abundant literature on uncertainty in the hydrologic modeling process, but much of their research focuses on the uncertainty of the model parameters and not on the outcomes that result from the fusion of
geospatial data. Some have computed confidence and prediction limits in stochastic floodplain models, but an increase in the understanding of the uncertainty in the data is needed to improve map interpretation. For example, the error of prediction associated with some flow equations currently ranges from 15 to 100 % (Sauer et al. 1983, Ries and Crouse 2002), yet little is known about how those errors propagate through the flood mapping process.

Hydraulic and hydrologic models are used to create a water surface to compare with the land topography. Errors in either surface can dramatically alter the extent of a floodplain. Merwade et al. (2008) used the flow equation from Pope et al. (2001) and found that a one-meter rise in water surface caused the inundation extent to increase 47.9 m. They found that the uncertainty bounds of the water surface elevation are positively correlated with drainage area (i.e., larger drainage areas have greater uncertainty).

It is well known that terrain datasets contribute to uncertainty in flood inundation mapping in three ways:

1. they affect the discharge values estimated from hydrologic models (Brasington and Richards 1998, Valeo and Moin 2000, Hancock 2005, Chaubey et al. 2005),
2. they affect the water surface elevation derived from hydraulic models (Marks and Bates 2000, Werner 2001, Vazquez et al. 2002), and

As technology improves, the impacts of data quality on floodplain mapping are becoming more apparent. There are also differences based on resolution, data type (e.g. TIN vs. raster), algorithms, and methods. For instance, ArcHydro® is a terrain processing tool that fills sinks (artificially low areas) in the terrain model, but WISE® does not.
Lastly, hydraulic models are sensitive to the number of cross sections, spacing between cross sections, slope, roughness, etc. Straatsma and Huthoff (2010) found that the various sources of error in hydraulic models lead to large variations in the floodplain roughness or Manning’s coefficient. Errors as simple as the misclassification of land cover can ultimately affect the outcome of a floodplain model.

The vast number of model parameters and sources of uncertainty have led to alternative approaches to flood mapping. Instead of a discrete approach to mapping, some scientists have used probabilistic approaches (Smemoe et al. 2007, Merdwade et al. 2008). The probabilistic approaches either used multiple sources of data or artificially introduced error into the model via Monte Carlo simulation (Merz et al. 2004). Monte Carlo simulations are used frequently to create confidence limits for loss estimates. Others have looked past the actual process of floodplain mapping and have tried to determine how people interpret flood risk using maps under uncertain conditions. Roth (2009), specifically looked at the impact of user expertise in the interpretation of flood risk assessment, perceived assessment difficulty, and assessment confidence using an online survey that displayed three delineations of a floodplain. The users were asked to rate certain locations’ risk of flooding given the different interpretations of the floodplains. He found that user expertise plays a significant role in the decision process.

From a practical standpoint, the ever-increasing use of geospatial data demands a better understanding of its limitations, because the user and the use of information are changing. The FBS standard sets an accepted level of confidence for floodplain boundaries, but no GIS tools incorporate that standard. A detailed audit procedure (FEMA 2007) and a tool called WISE© can be used to evaluate the accuracy of a floodplain for a given region; however, that process does
not include building uncertainty. Sensitivity analyses, also widely conducted, again do not include buildings.

4. Methods

4.1. Study Area

The study area includes all counties and municipalities in Virginia’s New River Valley Planning District (NRVPD), which includes the counties of Floyd, Giles, Montgomery, Pulaski, and the independent City of Radford (Figure 4.2). There are ten towns, two state universities, and a major federal facility. The 2010 U.S. census reported approximately 78,340 housing units in the region. Floods have the highest risk rating of all natural hazards in the NRVPD. The region experiences riverine flooding and flash floods and has several dams that make it susceptible to flood events resulting from dam failure (NRVPDC 2010). Pulaski County has the greatest area covered by a 100-year floodplain, but Montgomery County has reported more repetitive loss properties; ones that have suffered flood damage at least twice within a ten-year period. Montgomery County has 15 of the NRV’s 28 repetitive loss properties (NRVPDC 2010).

The NRV extends into two physiographic provinces: the Blue Ridge Province (Floyd), and the Valley and Ridge Province (Pulaski, Montgomery, Giles, and Radford). The region sits on the Eastern Continental Divide; part of the land drains into streams that flow to the Atlantic Ocean, while other parts drain into the Gulf of Mexico.

The average elevation of the NRV is approximately 762m. The elevation ranges from 448m above mean sea level at Glen Lyn to 1326m at Bald Knob in Giles County. The New River is the largest river in the region. It is dammed at Claytor Lake, producing the region’s largest lake. There are 15 federally regulated dams in the NRVPD, but the Claytor Lake dam is
the largest. Four of the dams have high or significant hazard ranking, making floods resulting from dam failure a major hazard.

![NRVPD Study Area for Floodplain Uncertainty Analysis](image)

Figure 4.2. The NRVPD, the region used for the floodplain uncertainty analysis.

### 4.2. Data Analysis

The purpose of this research was to test a practical application of a GIS tool for uncertainty analysis in spatial data fusion. After using the tool, the results of numerous GIS overlays were statistically analyzed to identify significant differences when uncertainty of the data was accounted for in the overlay process. The tool was specifically tested through intersect overlays to determine whether buildings are in or outside of the floodplain. This research also looked at the differences between deterministic and probabilistic floodplain mapping by exploring the spatial patterns and characteristics of horizontal positional uncertainty in the 100-year floodplain
boundary. All GIS analyses used ArcGIS10©. JMP9© was used for statistics. The tool uses VBA©, ArcObjects©, GeoProcessingObjects, Python, and the Model Builder. It was originally built for ArcGIS 9.3©, but is being migrated to the Microsoft VB .NET© framework to ensure that it is compatible with ArcGIS 10© and any future releases. VBA is no longer supported by ESRI© and will not be available after version 10.0.

4.2.1. Base Data

The data for the overlays included DFIRMs for the NRVPD, downloaded from FEMA’s Map Data Center4 and buildings layers for the NRVPD, collected from individual localities. All of the DFIRMs for the region have been recently updated and meet the standards set under FEMA’s “Guidelines and Specifications”, but the metadata only describes the DFIRMs at the dataset level (GAO 2010). This made it difficult to identify a unique spatial confidence interval for different study areas, although in reality they are likely different. The individual DFIRMs were combined and modeled using uniform error-band geometry (reference Chapter 2). The 100-year floodplain boundary was dissolved into a single-part feature class. Therefore, it was assumed the entire NRVPD DFIRM’s spatial confidence interval met the FBS because there was no other metadata. That standard states the data should be compiled and tested to meet a 38ft 95% confidence interval for horizontal positions of DFIRMs (FEMA 2007).

The Virginia Information Technology Agency (VITA) originally published the buildings data in 2007 as part of the 2006 and 2007 orthophotography update cycle of the Virginia Geographic Information Network's (VGIN) Virginia Base Mapping Program (VBMP). The buildings data

4 http://www.msc.fema.gov/webapp/wcs/stores/servlet/FemaWelcomeView?storeId=10001&catalogId=10001&langId=-1
did not come with metadata; however, VITA publishes the metadata for products they distribute\textsuperscript{5}.

The buildings GIS data was updated on an as needed basis. They were in an ESRI\textsuperscript{©} personal geodatabase format as feature classes. The data was produced at 1:2400 scale source mapping. All structures with roofs over 200 square feet or greater were delineated by on-screen digitizing using orthophotography as a base-map. The digitizing was done in a 3D environment and includes elevation data for rooflines. The metadata stated NSSDA compliance\textsuperscript{6}, but that does not make sense because the NSSDA is a procedural standard rather than a threshold-based standard. However, the NSSDA does reference the ASPRS standards for planimetric data (FGDC 1998). For this study, it was assumed the buildings met the Class I limiting RMSE of two feet because of the detailed procedure for digitizing. In reality, however, individual localities updated the buildings, but there was not enough metadata to choose a variable class or non-uniform value. The limiting RMSE was converted to a NSSDA 95\% confidence interval using NSSDA methods, thus the data was compiled to meet 3.416ft horizontal accuracy at 95\% confidence level. Using NSSDA procedure, the width for the buffers representing a 95\% confidence interval for the horizontal positional accuracy of the 100-year floodplain boundary was $\pm$ 38ft, and for buildings, $\pm$ 3.416ft. Given the available information, both were modeled with uniform error-bands.

\subsection*{4.2.2. \textit{The Processing Toolset}}

The tool includes three modules and is conceptually simple. There are also shortcuts for running overlays with previously processed data and a tool that will randomize point locations based on a

\textsuperscript{5} http://gisdata.virginia.gov/Portal/DiscoveryServlet
\textsuperscript{6} http://gisdata.virginia.gov/Portal/DiscoveryServlet
mouse click (Figure 4.3). The process involves identifying the type error-band geometry, generating error-bands and running overlay operations. The toolset records information in the attribute table at the feature level and reports the counts for the selected features that result from the overlays.

The first module is *FrameWorkID*, used to define the type of error-band geometry (Chapter 2) and to indicate if metadata or other information about uncertainty is available (Figure 4.4). If metadata is present then it is the reference for the input parameters. If metadata is not present or not detailed enough then one can make a best guess or estimate the size of an error-band using a wizard (Figure 4.5). In this experiment, the input values for the error-bands were 38ft for floodplains and 3.416ft for buildings, as defined in the metadata and computed using NSSDA methods.

Module 2 is the error-band generator (*ErrorBandGenerator*), which computes the size of an error-band based on mean, standard deviation and alpha, a defined probability density function, or a distance. The best-case scenario occurs if the metadata includes a 95% spatial confidence interval reported using NSSDA. The third module (*DataFusion*), runs multiple iterations of intersect overlays and stores the counts for the number of features that intersect each other and the layers involved in the overlay (Figure 4.6). The *DataFusion* module uses the variables identified and displayed in the dialog box to implement a script designed in model builder (Figure 4.6). The uncertainty info is stored as feature-level attribute data. Distance and alpha are stored in the attribute table and used as references for the buffer operations that generate the error-bands.

This is just one application of the toolset. It can be used for point-in-polygon, polygon-intersects-polygon, line-intersects-polygon, and line-intersects-line. For example, it could also
be used an overlay using crime data and census tracts, or an oil spill perimeter and critical habitat. The toolset can also simulate error and run Monte Carlo simulation using Python.

Figure 4.3. The toolbar for the uncertainty tools.

Figure 4.4. The dialog box for Module 1, used to select feature class error-band geometry.

Figure 4.5. The dialog box for Module 2 used to create fields and for storing and calculating error-band widths
Figure 4.6. The DataFusion module (Module 3), defines the type of overlay, layer involved and displays the parameter that should have been already set or available in the attribute table.

The tool is currently only capable of running overlays that involve two layers. Nonetheless, that results in nine iterations of overlays. First, a traditional crisp overlay is run with no error-bands. Next, the inner extent of layer one (100-year flood boundary) is overlaid with the original boundary of the second layer. The first inner boundary layer is then overlaid with the inner and outer extent of layer two (buildings). The process is then repeated with the outer extent of the first layer. The resulting overlay groups and IDs that identify the type or treatment are as follows: InF-RegB, InF-OutB, InB-InF, OutF-RegB, OutF-OutB, OutF-InB, RegF-RegB, RegF-OutB, and RegF-InB. For example, the inner buffer extent of the floodplain is InF, and the regular buffer extent of the buildings is RegB. Figure 4.7 illustrates the buffer overlay process and graphically displays the IDs that reference the overlay treatment. The number of features that intersect each other in the numerous overlays is stored in an array with an ID that references the original feature ID. The array can be exported to a database or text file.
and then imported into JMP© for statistical analysis. All overlays that include error-bands become a polygon-intersect-polygon topologic operation; therefore use the same model.

![Buffer overlay example and visualization of buffer IDs.](image)

Figure 4.7. Buffer overlay example and visualization of buffer IDs.

4.2.3. Results

For statistical analysis, the data was grouped based on the locality. The county/city was used as a blocking/stratifying variable to account for differences in building counts that may result from the municipality’s size, population density, or development patterns. For example, Floyd and Radford have significantly fewer buildings in the 100-year floodplain than the other municipalities, but they are also different in size and in physiographic province; therefore, these differences need to be accounted for in the analysis. Table 4.1 shows the results from the nine treatments of overlays grouped by county. The data from Table 4.1 was used in an ANOVA to test whether the treatment means were statistically different from one another. There was statistically significant evidence that the number of structures in the 100-year floodplain is
different when uncertainty is accounted for in the overlay process (F Ratio = 7.0191, Prob>F < 0.0001).

The percent decrease between the crisp overlay (RegF-RegB) and the inner buffer overlays (InB-InF) and percent increase from the RegF-RegB and OutF-OutB overlays was calculated (Table 4.2). The largest differences in percent increase and decrease were in Floyd County, which is the only municipality in the Blue Ridge physiographic province. The smallest were in Radford City, which has the smallest land area and is downstream from a large dam. Floyd County also has the largest percent difference between the OutF-OutB and InB-InF overlay. The percent increase and decrease from the crisp overlay were not symmetrical. The inner buffers have a greater impact on the number of building in the 100-year floodplain than the outer buffers.

Table 4.1. The table displays the number of buildings that intersect the 100-year floodplain per municipality by treatment.

<table>
<thead>
<tr>
<th>Municipality</th>
<th>InF-RegB</th>
<th>InF-OutB</th>
<th>InB-InF</th>
<th>OutF-RegB</th>
<th>OutF-OutB</th>
<th>OutF-InB</th>
<th>RegF-RegB</th>
<th>RegF-OutB</th>
<th>RegF-InB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floyd</td>
<td>34</td>
<td>36</td>
<td>32</td>
<td>84</td>
<td>90</td>
<td>83</td>
<td>54</td>
<td>55</td>
<td>53</td>
</tr>
<tr>
<td>Radford</td>
<td>39</td>
<td>41</td>
<td>39</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>53</td>
<td>55</td>
<td>49</td>
</tr>
<tr>
<td>Pulaski</td>
<td>604</td>
<td>626</td>
<td>587</td>
<td>1180</td>
<td>1199</td>
<td>1105</td>
<td>826</td>
<td>849</td>
<td>795</td>
</tr>
<tr>
<td>Giles</td>
<td>607</td>
<td>623</td>
<td>588</td>
<td>960</td>
<td>978</td>
<td>940</td>
<td>756</td>
<td>776</td>
<td>737</td>
</tr>
<tr>
<td>Montgomery</td>
<td>677</td>
<td>693</td>
<td>656</td>
<td>1080</td>
<td>1117</td>
<td>1070</td>
<td>859</td>
<td>878</td>
<td>837</td>
</tr>
<tr>
<td>NRVPD</td>
<td>1961</td>
<td>2019</td>
<td>1902</td>
<td>3363</td>
<td>3443</td>
<td>3257</td>
<td>2548</td>
<td>2613</td>
<td>2471</td>
</tr>
</tbody>
</table>

Table 4.2. The percent increase and decrease in building counts between a crisp overlay and ones that involved in the inner and outer buffer extents, and the percent difference.

<table>
<thead>
<tr>
<th>Municipality</th>
<th>Percent Decrease</th>
<th>Percent Increase</th>
<th>Percent Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floyd</td>
<td>66.67</td>
<td>40.74</td>
<td>54.72</td>
</tr>
<tr>
<td>Pulaski</td>
<td>45.16</td>
<td>28.93</td>
<td>41.01</td>
</tr>
<tr>
<td>Montgomery</td>
<td>30.03</td>
<td>23.63</td>
<td>31.90</td>
</tr>
<tr>
<td>Giles</td>
<td>29.37</td>
<td>22.22</td>
<td>30.66</td>
</tr>
<tr>
<td>Radford</td>
<td>11.32</td>
<td>26.42</td>
<td>25.48</td>
</tr>
<tr>
<td>NRVPD</td>
<td>35.13</td>
<td>25.35</td>
<td>35.07</td>
</tr>
</tbody>
</table>
The OutF-OutB and InB-InF layers were overlaid with a layer that contained buildings in every single iteration of the overlays. This made it possible to visualize buildings that had a 90.25% (0.95 x 0.95 = 0.9025) chance of being in or outside of the floodplain (Figure 4.8). Blakemore’s (1984) terms that defined possibly in, possibly out, and definitely in geographic objects were used as building classes in the map. The buildings that were possibly in or outside of the floodplain were combined and converted to point features. The point data was then used with the Kernel Density tool to identify clusters of uncertain buildings (Figure 4.9). The tool calculates a magnitude per unit area from point features using a kernel function to fit a smoothly tapered raster surface to each point. The raster data was reclassified into three classes (High, Medium, and Low), using Jenks natural breaks. Narrows, Pembroke, the Town of Pulaski, Shawsville and Elliston-Lafayette had the highest amount of clustering for buildings that may possibly be in or out of the 100-year floodplain. The uncertainty toolset does not currently run such an operation, but a kernel density tool might be a useful addition for identifying areas of high uncertainty in an overlay.
Figure 4.8. The map identifies buildings always in the 100-year floodplain and ones that may or may not be in for the Town of Pulaski.
5. Discussion and Conclusions

The results from this paper have major implications for floodplain mapping and highlight the utility of a GIS tool that incorporates spatial confidence intervals into the overlay process. While some have looked at floodplain boundary uncertainty with similar methods, this paper was the first to incorporate a second layer’s uncertainty, building uncertainty, into the flood mapping process. It is also the first tool of its kind. The fusion of uncertainties produces some very interesting data that could drastically alter the number of insurable houses, loss estimates, and
floodplain management. This tool has potential to be used in conjunction with the FEMA audit procedures, and to help identify places where detailed studies could improve floodplain management. In addition, there are many locations where insurance requirements could be reevaluated, such as the places identified as hot spots in Figure 4.9.

The spatial pattern of the outputs from the tool raises some important questions. For instance, there are two distinct groups of percent difference in Table 4.2 (municipalities over 32% and municipalities under 32% percent-difference). The difference could be related to floodplain management and hazard mitigation plans, related to the physical geography of a place, or a combination of the above. For instance, Radford City has a long history of restricting development in the floodplain, converting a large amount of public land into parks along the river, but it is also directly downstream of Claytor Lake dam. In contrast, the Town of Pulaski has extensive development in the floodplain and is above Claytor Lake dam. It is not likely coincidence that the town of Pulaski has suffered at least eleven 100-year floods and one 500-year flood in the past 90 years, while the last classified 100-year flood in Radford was in 1940, before the dam was built.

The municipality with the most repetitive losses was Montgomery County, but it did not have the highest percent-change, percent-difference, or largest number of buildings in the 100-year floodplain. This is somewhat confusing, as it was expected that Montgomery County would also have the most uncertainty in floodplain mapping. Unfortunately, data was not available for determining which properties were repetitive losses, therefore not possible to look at the variables that may contribute to having more repetitive losses. Research related to hazard mitigation plans and floodplain management may provide more insight into why certain places have more uncertainty when trying to determine which building are in the 100-year floodplain.
One struggle was finding an acceptable value to use as the spatial confidence interval. The metadata was not easily interpreted for use in the uncertainty tool. This was not an oversight in the tool; rather it was due to the inconsistent, incomplete, and inaccurate metadata. It is documented that FEMA was unable to manage metadata for detailed study areas (GAO 2010). Though metadata does reference other metadata, FEMA does not provide it when contractors conducted a floodplain study, making for a labor-intensive process of tracking down contractors to acquire metadata. As a result, the confidence interval for horizontal positions under the FBS was used as an input parameter in the tool. The same problem existed with the building data because it only cited NSSDA compliance and did not define a value. The solution was to assume uniform error-band geometry and apply a best guess or belief using the available metadata. While this might not have been the ideal method, the results from Chapter 3 support the notion that a GIS analyst can often make a reasonable estimate of the uncertainty in data with which they are working. An alternative approach could have involved field tests that compared the data used in the overlay analyses to data of higher accuracy, but there is no data of higher accuracy for comparison in regards to floodplain maps.

In conclusion, it is clear that the uncertainty of floodplain boundaries and of building and structure footprint data significantly affects the number of structures computed to be both inside and outside of floodplains. The tool greatly enhanced the overlay operations used to identify buildings at risk of being flooded. However, the uncertainty from the two layers did not contribute equally to the results. The floodplain boundary uncertainty has a far greater impact on the results than the buildings (38ft vs. 3.416ft), which makes sense due to the large range between the buffer values. If only building uncertainty was included in the overlay, it would not have created significantly different results, but that does not mean building uncertainty is not
important in other overlays. There is some evidence to support the idea that GIS analysts do not need to worry about the uncertainty of high quality planimetric data and that natural features are more uncertainty than man-made features. More research is clearly needed to understand fully how different types of geographic data’s positional uncertainty affect overlays. The large amount of variation should raise concerns for anyone who owns property near the 100-year floodplain. The results from this paper are also a red flag and question the reliability of all simple overlays. The tool presented in the paper has the potential to alter how GIS analysts interpret and use results from overlays. Many other situations exist in which the tool has application, making this tool and the methods important to a wide variety of disciplines that use spatial analysis and GIS.
6. References


Chapter 5: Conclusion

1. Conclusions

Geographic Information Systems (GIS) technology has revolutionized how scientists study the earth. It has become the most influential tool for managing geographic data in today's geospatial and environmental science disciplines because the technology allows for the fusion of maps with other geographic data and provides tools for analysis and visualization. The results from GIS and other geospatial technologies are inherently subject to uncertainty, and such geospatial data uncertainty will remain a major concern within the disciplines that utilize spatial analysis (Goodchild 2010).

This dissertation addressed the above concern by filling three broad research gaps related to geospatial data uncertainty: 1) the lack of a unifying conceptual data model from classifying geospatial uncertainty and modeling error-bands, 2) the absence of suitable methods to estimate epsilon for error-bands, and 3) the lack of tools for incorporating spatial confidence intervals into overlay operations. The research presented here reflects a progressive and cohesive study that thoroughly reviewed geospatial uncertainty literature, defined a simplified error-band framework for uncertainty classification, and developed GIS tools to incorporate spatial confidence intervals into overlay operations.

Paper1, An evaluation of a GIS framework for generating spatial confidence intervals to enhance geospatial data fusion (Chapter 2) summarizes current literature on error-bands and introduces a new simplified framework for generating spatial confidence intervals in a GIS. Prior to this work, many publications proposed various models for positional uncertainty, but most of the examples and applications of the models, while complex, used examples that were over-simplified. They involved detailed mathematical procedures, but did not take into account
the realities of geospatial data, such as the fact that many data are generalizations or have indeterminate boundaries (Burrough and Frank 1996). The models typically used only a line segment, a line, or a polygon and therefore did not adequately reflect the conditions under which one models error-bands in a real GIS analysis. My analysis of GIS uncertainty literature identified key components in multiple models that were incorporated into a new error-band classification system flexible enough—to generate spatial confidence intervals or estimate them using measurement tools. Different methods still must be explored to determine the ideal analysis method for specific projects and classes of data. There are some obvious parallels to conventional statistics’ need to choose between using pooled and unpooled statistical tests or parametric and non-parametric statistics. No single, unified method exists for conducting overlays with spatial confidence intervals; but the tools to explore various outcomes are now available and one day may be as diverse as conventional statistical procedures.

In the error-band classification framework’s experimental phase of the research, great diversity in the response results was found, yet there was correlation between user agreement and the geometry of data. The evidence suggests point-based data are better understood in terms of uncertainty. Paper 1 concludes that common metadata and methods for conveying data quality are hard to incorporate into a model for determining spatial confidence intervals and that GIS experience is important for assessing uncertainty. For a tool to be practical and useful, users need guidance and training, whether via a wizard, series of questions, or measurement devices in a GIS tool that help them assess data-quality. A framework would be much easier to use if the GIS community made efforts to provide better metadata. It was obvious throughout the study that metadata are hard to interpret, inconsistent, and sometimes simply erroneous.

Chapter 3, GIS and statistical analysis of digitized data’s spatial confidence intervals for a
forest clearing, addresses that lack of ability to use metadata to estimate confidence intervals that was identified in Paper 1. I explored peoples’ ability to estimate confidence intervals for use in the error-band classification framework given the absence of metadata or with inadequate metadata. Error-band literature makes it very apparent that there is a major need for methods to estimate epsilon. The findings of Paper 2 show that people can produce reasonable estimates for widths of spatial confidence intervals following the methods from the paper, although they do have a tendency to overestimate them.

Another interesting finding that also supports the framework from Paper 1 is that the error and uncertainty of a digitized forest-clearing boundary was different when measured at transects, calling for the use of segmented error-band geometry. Yet, the mean spatial confidence interval from transects was the best representation of the uncertainty in the forest-clearing boundary. It is likely that the shape and size of a spatial confidence interval have as much to do with the purpose of the data and analysis, as the collection method and sample statistics. This suggests there is some flexibility in how one models uncertainty. This logic is consistent with Couclelis’s (1996) boundary class framework and the line generalization typology of Mark and Csillag (1989). However, their findings in conjunction with the results from Paper 2 raise questions about the correct application of newer error-band models such as the s-band and g-band. Not all data come in their original state or are measurement-based, which limits models’ real-world application. People have a natural ability to interpret geographic objects; perhaps this intuitive ability translates to geographic representations.

The final part of this dissertation, Applying uncertainty GIS tools to floodplain mapping and decision making (Chapter 4) discusses the testing of a toolset I developed for incorporating spatial confidence intervals into the overlay process. The goal of the test was to determine if the
number of buildings in a 100-year floodplain was significantly different from a crisp overlay compared to overlays that involved spatial confidence intervals. The results were highly significant. Outcomes were different up to 66%, raising major concerns with all overlay operations, especially if results may affect a person’s life. These tests gave important insight into GIS analysts’ need to worry about the outputs of overlays and demonstrated that the tool developed has real world application. Furthermore, the major limitations of Brown and Huevelink’s (2007) Data Uncertainty Engine (DUE) were overcome, making this the first tool to be able to compute spatial confidence intervals in a GIS universally and use those statistics in the overlay process. The results clearly indicated that spatial uncertainty could have major impacts on geospatial data fusion.

It is hard for people to agree on a universal interpretation of metadata for use in statistical operations, but people are good estimators of data quality and uncertainty if they follow a systematic approach and use their average estimate to define spatial confidence intervals. Given these findings, I believe the toolset and the error-band classification framework to be useful resources for interpreting overlay results. It is scheduled for release in Fall 2011 and will allow a GIS user to incorporate geospatial uncertainty into the overlay process. The framework and tools presented have the potential to alter how people interpret and use geospatial data. The hope is that this paper will prompt inquiry and questions about the reliability of all simple overlays and that by following the framework and methods presented here, geospatial data fusion, and more importantly, decisions based on its results will be enhanced. Many situations exist in which this research has relevance, making the framework, the tool, and the methods important to a wide variety of disciplines that use spatial analysis and GIS.
2. Future Research

Many future research possibilities, across multiple disciplines, stem from the results reported in this paper. The most urgently needed next step is to finish the migration of the toolset from ArcGIS 9.3.1© to VBA© to VB .NET© for ArcGIS10©. Once completed, the tools will be released to the public in hopes that their use becomes commonplace in overlay operations. The validity of overlay operations needs testing in countless situations. A comprehensive application of the toolset on Couclelis’s (1996) boundary class framework and the line generalization typology of Mark and Csillag (1989) could provide useful information for improving the use of the tools by utilizing that information in a wizard that helps guide a user of the tool through the classification and fusion process.

Despite the usefulness of the toolset, it was limited to overlays that involve two layers, but many GIS operations include more than one overlay, so work is needed in order to manage more layers. Another major area for future research is the use of the framework and tools in 3D GIS. With advances in LIDAR and other terrain measurement devices, 3D data are becoming more frequent and easier to acquire. It is only a matter of time before the need for spatial confidence intervals in the 3D domain becomes a major topic of inquiry. Already, in the medical field, Butler (1999) concluded that confidence intervals had different shapes dependent upon the vertical or horizontal planes of stereotactic brain imagery. No similar research appears in GIS or GIScience literature. Among other improvements to the framework and toolset could be the inclusion of a module that propagates error-bands in which the variances are different in the X and Y coordinates.

Research endeavors in the uncertainty of geospatial data fusion are no doubt vast and multi-disciplinary. Disciplines ranging from natural resource management to engineering and artificial
intelligence to business are potential research landscapes as our world becomes increasingly reliant on geospatial technology. The increased complexity of models and computers’ ever-increasing ability to process data demand the development of mathematical and statistical procedures for communicating confidence and quality.
3. References


Appendix A:
Test questions from the experiment in Chapter 1

Question 1:
These data represent fire hydrants collected in the field (2004-2009) with a sub-meter GPS. The data are updated as needed. Data are post processed with base station information to improve accuracy. Positional accuracy information for individual fire hydrants is recorded in a table as attribute information. How would you classify the data?

<table>
<thead>
<tr>
<th>The data's uncertainty knowledge is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Known</td>
</tr>
<tr>
<td>B. Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The data's error band geometry is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Uniform</td>
</tr>
<tr>
<td>B. Non-uniform</td>
</tr>
</tbody>
</table>

Question 2:
This data represents the "best available" seamless source of the federal surface management agency boundaries (U.S. Federal Lands). It contains the most current data sets acquired from the state or regional offices of the Bureau of Land Management, United States Forest Service, National Park Service, and other federal agencies. The data were processed to form a national seamless dataset. Data for the Eastern U.S. comes from the U.S. Geological Survey's National Atlas. The source scale of the data is 1:2,000,000 as reported by the U.S. Geological Survey's (USGS) National Atlas. The USGS adheres to national map accuracy standards

<table>
<thead>
<tr>
<th>The data's uncertainty knowledge is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Known</td>
</tr>
<tr>
<td>B. Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The data's error band geometry is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Uniform</td>
</tr>
<tr>
<td>B. Non-uniform</td>
</tr>
<tr>
<td>C. Segmented</td>
</tr>
</tbody>
</table>
Question 3:
If you were to digitize or draw a forest stand based on the forest boundary as perceived from a ortho-photo, what would be the best way to classify the data using the suggested framework? One pixel in the photo represents one meter on the ground (i.e. 1 meter resolution).

The data's uncertainty knowledge is__________
A. Known
B. Unknown

The data's error band geometry is__________
A. Uniform
B. Non-uniform
C. Segmented

Question 4:
A local engineering firm created a map depicting flood risk polygons from a digital elevation model (DEM) for a local insurance company. You were asked to use this map to find homes that are at high risk for flooding. You were provided with a GIS layer containing polygons that represent the risk of flooding at a given elevation. How would you classify the data?

The data's uncertainty knowledge is__________
A. Known
B. Unknown

The data's error band geometry is__________
A. Uniform
B. Non-uniform
C. Segmented
**Question 5:**

This data represents the locations of crime produced from geo-coded addresses. (i.e. each street segment is attributed with address ranges). Geocoding takes an address, matches it to a street section or line feature and then interpolates the position of the address, within the range along that line as a point). No other metadata information available. How would you classify the data?

**The data's uncertainty knowledge is__________**
A. Known
B. Unknown

**The data's error band geometry is__________**
A. Uniform
B. Non-uniform

---

**Question 6:**

This data represents water distribution lines for a municipality. Data is a composite from old paper maps, GPS, surveys and expert knowledge. No accuracy or scale information for data collected using the old maps or from expert knowledge. Only line data collected with a GPS include spatial accuracy as an attribute. How would you classify the data?

**The data's uncertainty knowledge is__________**
A. Known
B. Unknown

**The data's error band geometry is__________**
A. Uniform
B. Non-uniform
C. Segmented
**Question 7:**
Underground water valve locations were marked by spray painting an 8x8 ft. cross at the center of each valve for a fly-over. If you were to create GIS data for the water valves by digitizing their locations using 1 ft resolution ortho-photography showing the spray painted valves; how would you classify the data?

<table>
<thead>
<tr>
<th>The data's uncertainty knowledge is</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Known</td>
<td>B. Unknown</td>
</tr>
</tbody>
</table>

The data's error band geometry is

| A. Uniform | B. Non-uniform |

---

**Question 8:**
This data represents recent wildfires in the US, published by the GEOMAC (Geospatial Multi-Agency Coordination Group). This data is a composite from multiple field offices. Perimeters are collected in the field by infrared flights for moderate accuracy or GPS for high accuracy. GPS perimeters contain horizontal accuracy information. How would you classify the data?

<table>
<thead>
<tr>
<th>The data's uncertainty knowledge is</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Known</td>
<td>B. Unknown</td>
</tr>
</tbody>
</table>

The data's error band geometry is

| A. Uniform | B. Non-uniform | C. Segmented |
Question 9:
As part of Gypsy Moth distribution research you were asked to map areas where they were visible evidence of Gypsy Moth damage. First you tried to identify the areas on a USGS 7.5 x 7.5 minute topographic map, but because the region has experienced major land cover changes, and it is large, you needed to observe the area from a helicopter to get a better perspective of the infected areas. While in flight, you draw polygons around the affected areas using the topographic map. It was last revised in 1983, is at a scale of 1:24,000 and adhered to national map accuracy standards at the time of publication. How would you classify the data you create?

<table>
<thead>
<tr>
<th>The data's uncertainty knowledge is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Known</td>
</tr>
<tr>
<td>B. Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The data's error band geometry is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Uniform</td>
</tr>
<tr>
<td>B. Non-uniform</td>
</tr>
<tr>
<td>C. Segmented</td>
</tr>
</tbody>
</table>

Question 10:
If you were asked to digitize or draw a recreational trail from an orthophotograph because the current recreational trail was not accurate, how would you classify the data? The old trail is overlaid with the orthophotograph to provide you with a reference. The image has a one meter pixel resolution.

<table>
<thead>
<tr>
<th>The data's uncertainty knowledge is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Known</td>
</tr>
<tr>
<td>B. Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The data's error band geometry is__________</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Uniform</td>
</tr>
<tr>
<td>B. Non-uniform</td>
</tr>
<tr>
<td>C. Segmented</td>
</tr>
</tbody>
</table>
Question 11:
You need to create a GIS layer and map of gate valves on a water distribution system. The only information available is a paper map of the water distribution system from 1990. Valve locations were derived from expert knowledge. There is a graphic scale and a statement that the map has never been field verified. How would you classify the data?

The data's uncertainty knowledge is __________
A. Known
B. Unknown

The data's error band geometry is __________
A. Uniform
B. Non-uniform

Question 12:
You work for a local government and for an emergency response plan you were asked to determine the homes within 1000 meters of the railroad. This is important because these homes are at greater risk for damage if there is a railroad crash, or other emergency. The building data is good quality, but the engineers only gave you a railroad digital map they had created using a paper map from 1980, which is no longer readable. How would you classify the railroad data?

The data's uncertainty knowledge is __________
A. Known
B. Unknown

The data's error band geometry is __________
A. Uniform
B. Non-uniform
C. Segmented