5-2001

Strategies for Handling Spatial Uncertainty due to Discretization

Thomas Windholz

Strategies for Handling Spatial Uncertainty due to Discretization

By

Thomas K. Windholz

Dipl.-Ing. Technical University Vienna, Austria, 1997

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

May, 2001

Advisory Committee:

M. Kate Beard-Tisdale, Professor and Chair of Spatial Information Science and Engineering, Advisor

Peggy Agouris, Assistant Professor of Spatial Information Science and Engineering

Max J. Egenhofer, Professor of Spatial Information Science and Engineering

Andrew U. Frank, Professor and Chair of Geoinformation, Technical University Vienna

Gerard B. M. Heuvelink, Professor and Chair of Environmental Science, University of Amsterdam

Library Rights Statement

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for “fair use” copying of this thesis for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature: Thomas K. Windholz

Date: May, 2001

Strategies for Handling Spatial Uncertainty due to Discretization

By Thomas K. Windholz

Thesis Advisor: Dr. M. Kate Beard-Tisdale

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Spatial Information Science and Engineering)
May, 2001

Geographic information systems (GISs) allow users to analyze geographic phenomena within areas of interest that lead to an understanding of their relationships and thus provide a helpful tool in decision-making. Neglecting the inherent uncertainties in spatial representations may result in undesired misinterpretations. There are several sources of uncertainty contributing to the quality of spatial data within a GIS: imperfections (e.g., inaccuracy and imprecision) and effects of discretization. An example for discretization in the thematic domain is the chosen number of classes to represent a spatial phenomenon (e.g., air temperature). In order to improve the utility of a GIS an inclusion of a formal data quality model is essential. A data quality model stores, specifies, and handles the necessary data required to provide uncertainty information for GIS applications. This dissertation develops a data quality model that associates sources of uncertainty with units of information (e.g., measurement and coverage) in a GIS. The data quality model provides a basis to construct metrics dealing with different sources of uncertainty and to support tools for propagation and cross-propagation. Two specific metrics are developed that focus on two sources of uncertainty: inaccuracy and discretization. The first metric identifies a minimal

resolvable object size within a sampled field of a continuous variable. This metric, called detectability, is calculated as a spatially varying variable. The second metric, called reliability, investigates the effects of discretization on reliability. This metric estimates the variation of an underlying random variable and determines the reliability of a representation. It is also calculated as a spatially varying variable. Subsequently, this metric is used to assess the relationship between the influence of the number of sample points versus the influence of the degree of variation on the reliability of a representation. The results of this investigation show that the variation influences the reliability of a representation more than the number of sample points.

Acknowledgements

I gratefully acknowledge the financial support for this research by the National Imagery and Mapping Agency under grant NMA202-97-1-1021 and the National Science Foundation under grant SBR9700465.

I would like to thank my thesis advisor, Kate Beard, for her support, collaboration, and guidance. I would also like to thank all my committee members Peggy Agouris, Max Egenhofer, Andrew Frank, and Gerard Heuvelink for their valuable input and guidance. Additionally, I would like to thank Mike Goodchild for his collaboration.

A special thanks goes to my parents who always supported me in my decisions and enabled me to complete this program. I would also like to thank Craig Miller for his help improving my programming skills.

Finally, I would like to thank my wife Elizabeth for her moral support and patience.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

It is a challenge to capture an infinite universe in finite systems. Geographic information systems (GISs) represent aspects of our world in finite computer systems.

GISs model our reality in an immense variety of fields (Longley et al. 1999). GISs are used, for example, in utility (e.g., electric, water) management systems, models describing forest growth (from small regions to the entire world), and in medicine where GISs are utilized to map the human genome. GISs are vital to progress modeling, management, and the investigation of scientific as well as everyday phenomena. For all of these beneficiary applications we have to keep in mind that GISs are models of our world and are, therefore, constrained by our capability to model the complexity of processes and events.

The limitations of finite systems are reflected in several distinct yet interdependent aspects of uncertainty within a GIS. One source of uncertainty is commonly known as measurement error. Such errors occur whenever data are collected. Measurement errors are introduced by the limited resolution capability—what is referred to in this thesis as discretization—of any measurement system (Sinton 1978; Chrisman 1997). Models are to a certain degree copies of the original where the focus is on simulating its essential properties, not on duplicating an entirety. For example, when generating a road map the

emphasis should be on maintaining topology (Egenhofer and Herring 1990) rather than metric. Finite systems are further limited by storage and viewing devices, processor times, and the rates at which people can absorb and perceive information. These restrictions either reflect today's technology or simple logic that will persist in the future. Given these boundaries any spatial representations managed by GISs will include imperfections.

That errors exist in any form of representation is only one issue. The other issue is to what degree—if any—the errors or the imperfections are addressed and made explicit. In everyday human interactions, imperfections are implicit if not explicit. Common examples of information with associated imperfections are weather reports (implicit) that might include a percentage giving the probability of precipitation (explicit); polls that include an error margin (explicit); or a promise to your spouse that you will be home from work at *about* six (implicit).

The literature includes several discussions on the cognitive aspects of the human perception of maps (Kozlowski and Bryant 1977; Schone 1984; Kuipers and Levitt 1988; Dutta 1989). Early in development people build up models of the environment that they store in their brains. They gain the ability to estimate distances and have a feeling about how accurate these estimations are (Kuipers 1982). For example, if one drives from Orono, ME to Boston, MA it takes about four hours—give or take half an hour, depending on traffic. The extension and the geometry of the environment in combination with experience build the foundation of spatial knowledge and an assessment of associated imperfections. Conceptualization and understanding of space become increasingly important to improve computerized formalizations of spatial inferences in GISs (Egenhofer and Mark 1995).

This dissertation is based on the assumption that there is a need within the GIS community for models of imperfection. Currently the term community could include almost everybody. The GIS community extends from large governmental agencies

(e.g., environmental management organizations) to utility companies (e.g., phone, hydro, and electric) to the person on the street who uses a cellular phone to find the nearest Italian restaurant. Some applications are less susceptible to uncertainties than others. Nevertheless, the information “Bangor has two Italian restaurants” is too uncertain to locate any of them.

1.1 Terminology

This section gives a brief overview of some key terms used within the thesis. This explanation of terminology should be seen as a condensed and simplified clarification. Throughout the thesis each of the terms will be discussed in more detail.

We make the distinction between a representation of spatial data as the computer model (raster/vector model) and the presentation of data as the graphic display of the data (e.g., on screen).

Sources of uncertainty refer to all causes that contribute to the uncertainty of a spatial representation (e.g., inaccuracy and inconsistency).

Imperfection is a term that addresses a specific subset of the sources of uncertainty. The term imperfection implies that an error value of some kind is present. For example, a representation can be inaccurate, imprecise, invalid, incomplete, or inconsistent (as discussed in Chapter 3). Thus, inaccuracy, imprecision, inconsistency, incompleteness and invalidity can be seen as terms of imperfections.

Discretization is a source of uncertainty but not an imperfection. It is the conscious decision to subdivide a continuous spatial, thematic, or temporal domain. For example, the number of classes used to represent a continuous variable is thematic discretization. The distance between sample points is an example of spatial discretization and the time interval between samples is temporal discretization. A representation can be free of imperfections and yet bear a considerable amount of uncertainty due to discretization

choices. For example, a weather map showing one temperature class (representing minus 100F° to plus 200F°) can be perfect. However, it leaves a lot of room for speculation (i.e., uncertainty) as to what the temperature at a specific location is.

Units of information are an ordered grouping of elements where data are present. These units of information can exist inside (e.g., vector, coverage, or query) or outside of a GIS (e.g., a measurement or a print out).

In the context of this dissertation a *data quality model* is a conceptual model. It interprets, connects (e.g., data with data referred to in the lineage), and processes (e.g., detectability and reliability) data quality aspects within a GIS.

1.2 Motivation

This section uses examples to give an overview of the effects of spatial discretization on the uncertainty of a representation. The motivation for the research presented in this dissertation is to provide GIS users with tools to model and visualize these effects.

The following discussion uses primarily height measurements to illustrate effects of spatial discretization on uncertainty. In addition, we use the example to clarify the terms discretization and resolution. For the sake of simplicity we assume that all measurements mentioned in the following discussion are made without error. This assumption allows us to focus only on the effects of discretization.

Figure 1.1 shows an example of a spatial distribution of sample points observing the attribute value height. The sample points are not yet associated with a specific location in the real world. This association, however, is a decisive factor for our motivation.

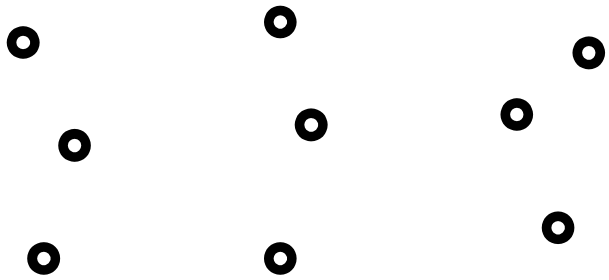


Figure 1.1—Spatial Distribution of Sample Points (Measured Variable: Height)

Thus, we would like to start with a simple environment as far as height measurements are concerned. The first environment represents flat terrain (Figure 1.2). The measured height values are fairly similar and the variation in height differences is negligible. By choosing nine sample points and their specific location we applied a certain level of discretization of space. In our example this space is the dessert region depicted in Figure 1.2. Each of these sample points is now a representative of the actual height in its neighboring region.

Figure 1.2—Spatial Distribution of Samples (Measured Variable: Height), Flat Terrain

The size of this region (Figure 1.3) is a direct result of the chosen level of discretization. The shape of these regions can be derived, for example, by Dirichlet tessellation (Green and Sibson 1978).

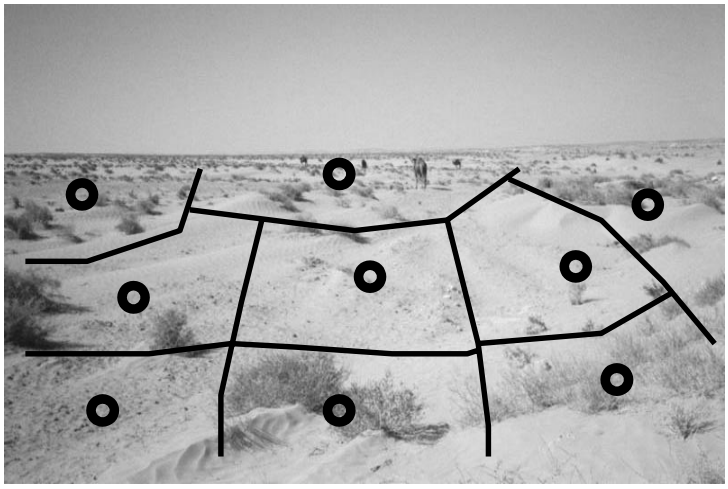


Figure 1.3—Resolution as a Direct Result of Discretization

Resolution is a term that is traditionally used in photogrammetry, where the size of a region associated with a measurement is equivalent to the pixel size of an image. The size of the pixel is dependent on the size of the sensor and distance between an object and camera. In the case of an image the distance between two pixel centers represents the chosen level of discretization and the area of a single pixel expresses its resolution. For an image we typically assume a constant level of discretization and resolution within the entire image. In our example of height measurements, however, discretization as well as resolution is a spatially varying entity.

We chose to represent the area by nine sample points Figure 1.2, resulting in a scenario where each of the sub regions has an uncertainty associated with it (Figure 1.3). To eliminate any uncertainties we would have to sample the whole region with an infinite number of sample points with an infinitely short distance between them. Since such a scenario is unrealistic, we have to accept the fact that uncertainties are present. The lack of a completely exhaustive sample set also implies that we lose information. We cannot recover this loss of information, however, we can estimate the resulting amount of uncertainty and communicate it to the GIS user.

Uncertainties are introduced because we lack the information on the actual heights between the sampled locations. As for the scenario in Figure 1.2 we can assume that the sampled attribute value height has little or no change over the entire region and, thus, a low uncertainty. Therefore, for this specific scenario the chosen level of discretization carries some redundancy with respect to uncertainties in the sub regions. Taking it one step further we can say that we can coarsen the level of discretization (i.e., lower the number of sample points) without any significant impact on the uncertainty of the representation. The green sample points represent the new sample design and the green line the division in the new neighborhoods (Figure 1.4). Each of the two newly generated regions has now a new level of uncertainty. Nevertheless, we can assume that when comparing the two levels of discretization (for the shown terrain) that there is no significant difference between the uncertainties of the two scenarios.

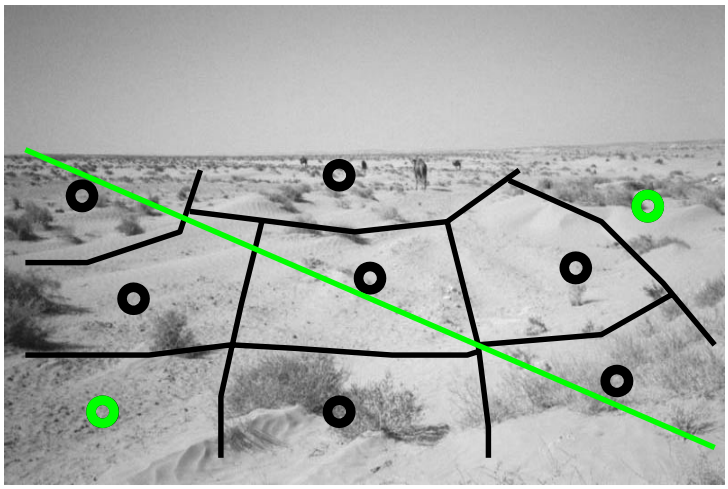


Figure 1.4—Decreasing the Level of Discretization

These statements, however, do not hold for the scenario shown in Figure 1.5, where an environment with a larger variation of the attribute value (i.e., height) is present. Here is a decisive difference between the uncertainties of the representations depending on the two levels of discretization. We can assume that the design with the nine sample points yields less uncertain results than the two sample points. This is an example of the loss of information due to discretization.

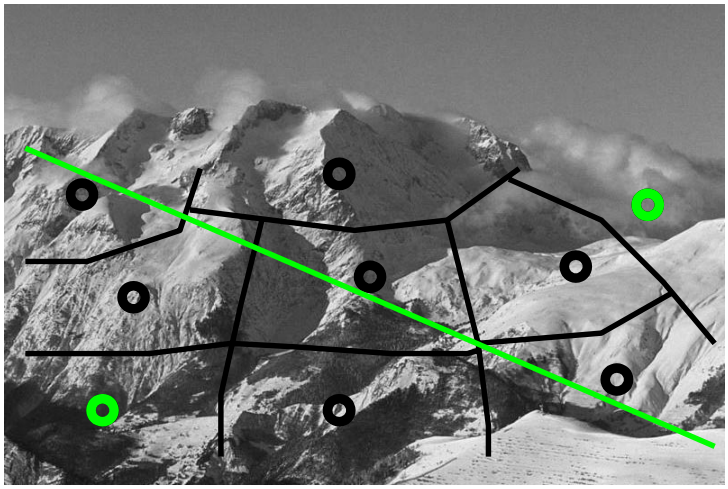


Figure 1.5—Height Measurements, Mountainous Terrain

The uncertainty of a representation can be made apparent in several different ways. In this dissertation we focus on two general types of GIS queries. For each we develop a specific metric to convey the uncertainties of a representation. The first type is the following scenario: a GIS user would like to identify a specific object in a representation. For example, this could be a ridge that the user would like to locate in Figure 1.5. It is of interest to estimate the uncertainties of the representation to advise the user if the representation is good enough for the desired task. We call this metric detectability. The detectability depends on the uncertainties as well as on the object itself. For example, it is easier to detect a broad and high ridge. For the second scenario description we use snowfall measurements instead of height measurements—assuming that the sample locations in Figure 1.5 are now weather stations. For an avalanche model it is of interest to know the exact tonnage of snowfall over a certain period of time. The model itself can achieve acceptable results only if certain quality constraints for the representation are met. Therefore, it is of interest to estimate the reliability with which we can calculate the tonnage of snow within a specified area. We call this metric reliability. The reliability of a representation depends again on the discretization underlying the representation.

1.3 Goal and Hypothesis

The goal of this research is to make the effects of spatial discretization within spatial databases explicit. To accomplish this goal, sources of uncertainty are identified, and managed through the information life cycle. The information life cycle begins with the measurement of the phenomenon and culminates in presentations to the user. Measurements are the initial source of all GIS datasets and substantial research has been devoted to measurement errors within a GIS context. Discretization, which is an essential aspect of a measurement framework, has received much less attention and is the central focus of this research.

This dissertation aims to answer the following key research questions:

- What are the elements of a data quality model?
- What are the relations among these elements?
- What are the effects of spatial discretization?
- How do discretization effects propagate?
- How are discretization effects related to other elements?
- How can we determine and track the reliability of a representation?
- Which properties (e.g., sample density, variation of attribute variable) of a representation influence its reliability?

The hypothesis addressed by this thesis is that:

The loss of information due to discretization is more strongly influenced by the variation of an attribute variable than the sample density.

1.4 Scope of the Thesis

Several factors contribute to the quality of spatial data within a GIS, including imperfections (i.e., inaccuracy and imprecision) as well as effects of discretization. To manage and document the quality of data within GIS, a data quality model (DQM) is necessary. A DQM is developed by incorporating information on the imperfections and on the resolution, discretization, or scale.

We are interested in the definitions of the terms of imperfection and units of information as the basis for a formal data quality model. Additional discussions focus on the role of resolution and discretization. We introduce two specific models that can be implemented for a better understanding and handling of the interaction between discretization and accuracy. The key effects of discretization are the detectability of objects and the reliability of spatial variation in the sub-regions of a representation.

This thesis is not concerned with a complete discussion on all possible cross-propagations among all terms of imperfections and discretization. The discussion of the interactions between accuracy and discretization is not exhaustive, rather two possible approaches are thoroughly investigated in this dissertation.

1.5 Approach

In order to coordinate the increasing interest in handling quality issues within geographic information systems we formally define imperfections and identify their association with units of information. In the past, terms used for units of information and imperfection have been ambiguous (e.g., accuracy, precision). A starting point is to clarify those ambiguities. The goal of the first section of the thesis is the determination and definition of occurrences of imperfection within a GIS and specification of their interactions.

This work provides a basis for generating necessary tools for propagation, cross-propagation, and specification of metrics for the identified imperfections (e.g.,

inaccuracy and imprecision) and their interaction with discretization. For example, one of the metrics of interest for inaccuracy is the root mean square error (RMSE). Elaborated relations help to consider if this metric is appropriate and how it propagates (cross-propagates) from one unit of information to the next. Furthermore, the impact of scale (level of geographic detail) changes is investigated using different approaches for object versus field models in order to define the necessary metrics.

Any measures of imperfection must consider that uncertainty varies through space and time and is context sensitive. Thus, it is important to capture the relations among the units of information and their relative importance in contributing to the overall uncertainty.

1.6 Major Results

Beginning with the determination and identification of imperfections, the terminology within the field of quality aspects is clarified. Contributions of this dissertation are formal specifications of data quality elements and their relationships to units of information and formal specifications of discretization effects as well as the identification of discretization propagation.

The contributions of this dissertation are aids to users of geographic information in enhancing their understanding of the reliability of their results. More specifically these contributions include:

1. A better understanding of types of imperfection (e.g., inaccuracy and imprecision) and discretization; not limited to the traditional concerns for positional accuracy, but examining all of the reasons why a user might be led to an incomplete understanding of the phenomena being represented.

2. Additional methods for measuring uncertainty, covering some forms of imperfection and providing a suite of readily computed and well-defined metrics.

3. Improved methods for communicating imperfections, to ensure that the user is informed about imperfections and their consequences by accessible, readily understood methods.

4. Concepts for controlling and modeling the propagation of imperfection and discretization, to ensure that the impacts of uncertainties on the user's decisions can be fully evaluated.

1.7 Intended Audience

The intended audience of this dissertation includes, but is not limited, to designers, developers, and users of GIS software. Especially addressed is the audience who has an interest in estimations of the reliability of any given data. This dissertation is also directed towards users who might be held liable for any decisions that were based on a geospatial database (e.g., emergency management).

1.8 Thesis Organization

The remainder of the thesis is organized into five chapters: Chapter 2 discusses the research and literature background relevant for the subsequent approaches. We elaborate on imperfection, discretization, propagation, and a data quality model.

In Chapter 3 we address components of a data quality model by providing unambiguous definitions of imperfections, linking these to specific units of information and investigating their potential to propagate across units of information.

Chapter 4 focuses on the first effect of discretization: the ability to resolve spatial objects. In this chapter we develop a specific metric that we call detectability. A case study is included.

Chapter 5 develops an explicit metric for the loss of information due to spatial discretization. We propose this metric as a spatially random field that provides an estimate of reliability at any given location. A case study is included.

In Chapter 6 we use the metric for reliability to investigate its dependencies. We compare the influence of the variation of a given attribute variable to the influence of the sample density on the reliability.

Chapter 7 concludes the dissertation with a summery of the major results. This chapter discusses future research questions that are based on the findings of this dissertation.

Chapter 2

Data Quality and Uncertainty

GIS users want to make decisions based on the geographic data stored in a GIS. The combination of geographic information leads to an understanding of relations among geographic phenomena and provides a helpful tool in complex decision-making. Internally, data variables of representation can be stored using any adequate data type (e.g., integer or real). The internal precision of such representations, however, must not be adapted to the accuracy of their represented topics. For example, a location of a feature is stored internally as having an x-coordinate of 123.26439 meters does not mean that the accuracy of this location is known to the hundredth of a millimeter. Thus, it is a problem to assess information about the reliability of the results. If GIS is to gain more widespread adoption as a scientific tool it is necessary to know more about the nature and behavior of uncertainties occurring in such a system and thus, increase the reliable and meaning of results. Although we are able to define accuracy in numerous ways outside the GIS, little information is included in existing systems (e.g., FGDC 1994 compliant data) for accuracy assessment. It is not necessarily a requirement that the GIS user be well acquainted with models of uncertainty. The producer has primary responsibility for providing more detail about the quality of the information and this information needs to be managed by the system for delivery to users. This capability is based on a *data quality model* and essential metrics. This chapter reviews the literature on topics related to the development of this data quality model.

2.1 Uncertainty

The term uncertainty has gained recent popularity but suffers from inconsistent and ambiguous usage. A recent compilation of most frequent interpretations is given by Mowrer (1999). Geographic Information Science (Chrisman 1997; Clarke 1997; Burrough and McDonnell 1998) is relatively new and has emerged as a combination of several different scientific fields (e.g., computer science, geography, surveying, and photogrammetry). Each of these scientific fields has a different view of uncertainty. Sometimes the principles of fuzzy set theory (Zadeh 1965) and entropy (Shannon and Weaver 1962) are used to characterize uncertainty (Morrisey 1990). However, some claim that there is a difference between a situation of risk and one of uncertainty (Joslyn 1992). The distinction is that in a risky situation a random event comes from a known probability distribution, whereas in an uncertain situation the probability distribution is not known.

We define uncertainty as a state of knowledge about a relationship between the world and a statement about the world (Motro and Smets 1997). This dissertation focuses on sources of uncertainty and a particular subset of these, which we refer to as imperfections. Imperfections are deficiencies in data or information and a source of uncertainty. For example, there are, however, sources of uncertainty we would not describe as deficiencies. For example, interpretation, modeling concepts, and discretization contribute to uncertainty but are not necessarily deficiencies in the information. Chapter 3 gives a more detailed discussion on sources of uncertainty and the role they play in the data quality model.

Current GISs for the most part lack explicit information about imperfections in the data. Potential problems with undetected and undocumented imperfections involve the inappropriate or ineffective use of geospatial information, which ultimately undermines decision-making. Many applications that use geospatial information depend heavily on knowledge of the reliability of the information.

With any GIS product there is a level of uncertainty about the nature of its quality. It is important to provide the GIS user with the necessary awareness that these problems exist. Although there is a continuing interest in improving data quality standards (FGDC 1994; CEN/TC 287 1995; CEN/TC 287 1995; FGDC 1996), commercial GIS packages put little or no effort into calculating and communicating the inherent imperfections to the user (Frank 1998). In the literature (Chrisman 1983; Goodchild 1989; Goodchild et al. 1992; Heuvelink 1993; Hunter and Goodchild 1993; Carroll 1995; Beard 1996; Parsons 1996; Heuvelink 1999), however, we find several approaches to handling either a single imperfection (e.g., inaccuracy) or a conglomerate of imperfections (e.g., imprecision and inconsistency).

To improve the management of quality within geographic information systems it is essential to detect occurrences of imperfections and furthermore to clarify some frequently used terms. Steps in this direction have been made over the last several years. Beginning with Chrisman (1983), in preparation for development of a Spatial Data Transfer Standard, and continuing with NCGIA (Goodchild and Gopal 1989), GISData (Burrough and Frank 1996), and other national and international efforts (Nijkamp and Scholten 1991; Guptill and Morrison 1995; Hunter and Goodchild 1997), there has been on going research to understand spatial data uncertainty. One problem, which is a result of the many disciplines involved, is the ambiguity and inconsistency in the use and definition of terms. Many terms that are used to describe imperfections in spatial data are used interchangeably and sometimes inappropriately (Goodchild et al. 1992).

Previous research (FGDC 1994; CEN/TC 287 1995) has identified several parameters (i.e., positional accuracy, thematic accuracy, temporal accuracy, logical consistency, completeness, and lineage) as encompassing the quality aspects of geographic information. The unit of information that has been the focus of most of this research has been the map or the digital map and its digital subcomponents (points,

lines, polygons or pixels). Restricting the focus to the map and its subcomponents limits our view and understanding of uncertainty.

There are a variety of approaches to the management of uncertainty (Bedard 1987). “Learning to live with errors in spatial databases” is essential (Openshaw 1989). Agumya and Hunter (1996) discuss possibilities for the assessment of the fitness for use of spatial information as one form of uncertainty measure. Beard (1989) offers a slightly different approach emphasizing the design of a GIS to avoid misuse of spatial information. Similarly, Burrough (1991) pushes the development of an intelligent GIS. Elmes and Cai (1992) focus on data quality issues with regard to a user interface design. Stoms et al. (1992) follow a more specific approach, investigating the influence of uncertainty on a specific wildlife habitat model. Blakemore (1985) discusses the relationship between the advantages of high resolution and the disadvantages of the accompanying high costs in GISs. To estimate the influence of resolution in a GIS representation specific metrics are included. It is important to further elaborate on the concepts of resolution in the spatial, thematic and temporal domains—or, as Sinton (1978) called it control.

2.2 Errors in GIS

Before developing models for handling imperfection it is essential to know what kind of errors can occur in a GIS and the granules of information they are associated with. To produce a map—either paper or digital—we need data (and their spatial dependencies) that are collected or measured in the field. No matter how this material is obtained, there will be errors. For GIS one of the earliest approaches in error analysis can be found in Taylor (1982) who adopted the term “error analysis” for computer simulation models.

Errors are introduced through measurement (Goodchild 1993) and processing (Perkal 1956; Keefer et al. 1988) and can either be systematic or random. Systematic or random measurement errors have been discussed in other disciplines as well, for

example, in surveying (Reissmann 1976; Wolf and Ghiliani 1997) which provide us with the necessary statistical background. Examples of more GIS-specific errors are errors of orientation. These include errors due to the transformations that are used while digitizing a paper map or due to the orientation of a GIS raster. During the process of conversion of the data into a raster map the level of granularity changes, which is an additional error source when measuring for example the area or the perimeter of a polygon. Another GIS-specific error surfaces when generating map overlays due to sliver polygons (Goodchild 1979; Veregin 1989).

Two examples illustrate previous error metrics developed for GIS. The first one is a metric dealing with the propagation of thematic error through GIS overlay operations (Veregin 1989). The model is based on the number of occurrences of errors of omission and commission in input data. Another example of error propagation modeling deals with methods for visualization of the accuracy of geometrical data (Kraus and Kager 1993). Areas are represented by their boundaries. The vertices of these polygons are treated as stochastic information. The mathematical principle is based on the probability of the location of an arbitrary point within a closed polygon. This model can be used to determine the accuracy of an area segment by overlaying two areas with a map overlay operation. The latter quality model combines variances as well as correlation and systematic errors based on proven theoretical methods.

2.3 Principles of Data Quality Models

A data quality model (DQM) is one way of integrating and presenting uncertainty information to a GIS user. The DQM is a subschema in the concept of metadata (FGDC 1996 1997). It provides essential additional information to assess the decisions made with the help of a GIS. A model of the real world requires transformations of the data to reduce the information to the essential quantity. During this process we get discrete data from continuous reality, which introduces errors. Modeling data we follow three steps:

1. *observation* of the real world

2. *processing* to transform the data

3. *representation* of the data in a GIS

There are different data types such as continuous versus discrete data (although GIS data are all discrete)—and different possibilities of representations like vector or raster models. The DQM is a general model.

"The need to share and integrate spatial data has spurred an interest in metadata. Metadata is designed to tell users what they have and what it can be used for." (Timpf et al. 1996) It is data describing data and business aspects of it (CEN/TC 287 1995). Metadata serves several different roles one of which is to describe the quality of the data. Metadata covers several different kinds of extra information—such as facts on:

- the identification and ownership
- the data content and structure

  including: currency of dataset, *quality parameters*, reference system
- the availability and delivery (administrative metadata)
- the source
- the validity
- the processing

⋮

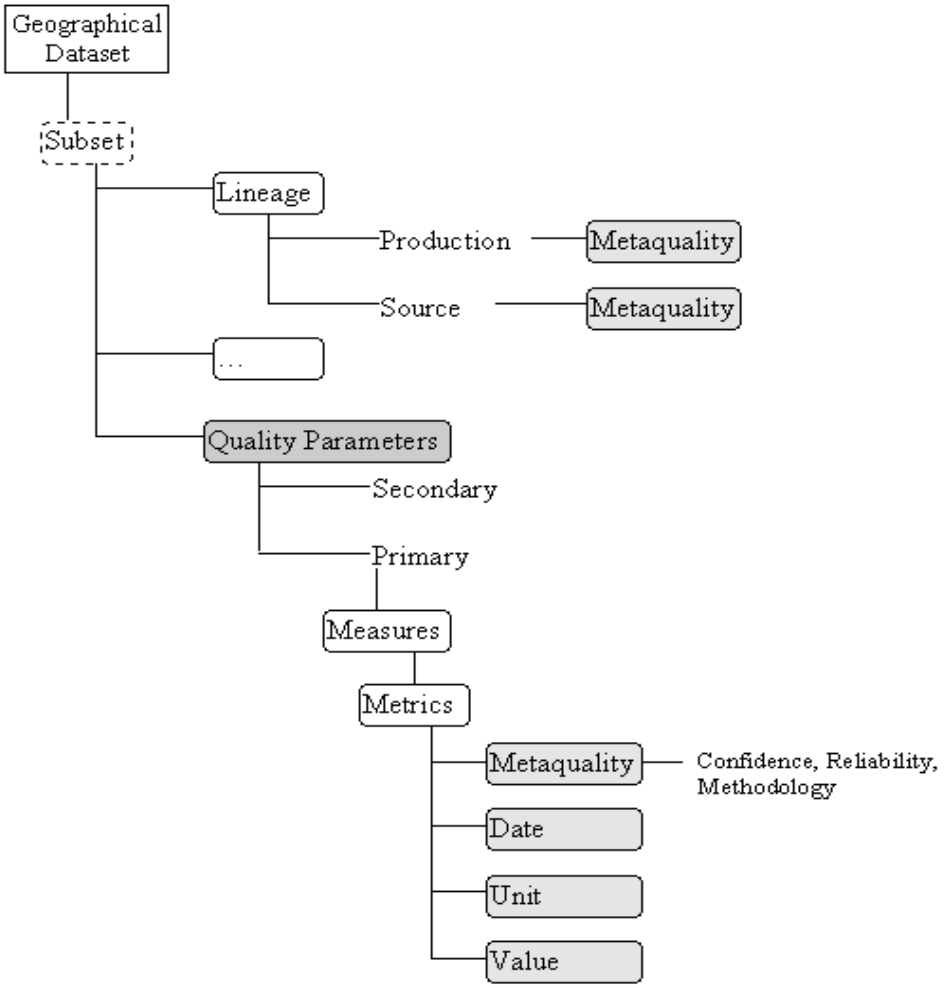


Figure 2.1—Integration of Additional Information in a Dataset (CEN/TC 287 1995; Timpf et al. 1996)

The providers of GIS datasets should be aware of the importance of offering this supplementary information as well (e.g.: for an increase in value of their GIS product or for the question of liability of the results). The metadata could be maintained in a meta-database and thus accessible via the World Wide Web for users around the world. There are numerous advantages for both sides. One of them is that the user has the ability to decide whether the product fulfills application requirements for accuracy and discretization or not.

A possible way to integrate the additional information in the datasets can be seen in Figure 2.1 that serves as an overall view for the model concept. The Lineage and its subsets of Metaquality give background information on the organization that was responsible for the data collection and acquisition. Furthermore it should give the source of the data - thus the user knows for what purpose the dataset was generated and if it satisfies application demands as well. But for this question another helpful topic ‘the Usage’ could give even more information, where for example the organization, the kind of usage, and its constraints or limitations are listed.

2.4 Propagation

Within a GIS, geographic information is stored in a database via representations that originate from discrete measurements of the real world. Usually, any interaction between a GIS user and the GIS itself is solely based on the representations and not on the measurements, since in common GISs the original measurements are not accessible. However, in the literature one can find an increasing interest in measurement based GISs (Buyong et al. 1991).

Inherent in a representation are deficiencies that accumulate and propagate during the process of generating a representation from measurements (e.g., from a few sample points one can generate a continuous representation). In order to handle those imperfections one can follow two different approaches. One could make inferences about the imperfections within a representation which result in vague approximations that at some point might even be wrong since inferences are based on imperfect values that one assumes a certain representation should have. On the other hand one could identify and measure the imperfections and derive more specific knowledge about the quality of the data. In some cases this is a decisive advantage and can increase the value of GIS products. For the latter approach it is necessary to include the original measurements and the applied transformations—and it is of interest to identify when

and where different components of imperfections are *initially* introduced in the process of generating the representations and finally a presentation.

The literature documents a wide variety of approaches beginning with Taylor (1982) and Veregin (1989). A general discussion of the topic can be found in Heuvelink (1993) as well as in Heuvelink (1998). Some approaches (Stanislawski et al. 1996; Kiiveri 1997) are focused on a specific subset of uncertainty (e.g., positional uncertainty). Frank (1998) uses error propagation to demonstrate that a simple quality measure cannot describe the effects on the results. Specific approaches, however, can be seen in almost all possible variations of implementation of GIS principles. Lanter and Veregin (1992), for example, investigate the aspects of error propagation for a layer-based GIS with an emphasis on raster representations whereas Kraus and Kager (1993) focus on vector-based GIS approaches. On the other hand some models for error propagation are geared towards a specific data type. For example, Goodchild et al. (1992) focus on an error model for categorical data. Yet other error propagation models involve a particular mathematical model—such as Monte Carlo simulations (Hammersley and Handscomb 1979). Others (Forier and Canters 1996) put more emphasis on the user-friendliness of tools for error modeling. Furthermore, in the literature one can also find very specific approaches such as in Carroll (1995) and Hunter and Goodchild (1997). And last but not least one can also find several implementations of developed methods. Some examples here would be: Wingle et al. (1994) and Pebesma and Wesseling (1997).

2.5 Scale, Resolution and Discretization

As pointed out by Sinton (1978), Chrisman (1983), and others, when making measurements, resolution is imposed across the three dimensions of space, theme, and time in the form of discretization. Control is a discretization along one or more dimensions so another dimension can be measured. The imposition of discretization results in a loss of information that contributes to the uncertainty about the variable or phenomena being described. In terms of uncertainty, the effects of discretization are

likely to be more substantial than measurement error. Work on uncertainty has tended to focus on measurement errors and yet the effects of discretization may be more substantial. In other words the imperfections in the measurements are less cause for concern than that which is not measured.

The imposition of discretization along a single dimension (e.g., the discretization on the spatial dimension imposed by satellite sensors) is not too difficult to track and account for. In most geographic representations, however, multiple levels of discretization are interacting. A representation may have heterogeneous levels of discretization (multiple levels of discretization along one dimension) and compound (combined spatial, temporal, and thematic) discretization as an outcome of multiple discretization processes, as a result of a sequence of operations on a representation (e.g., resampling, classification, interpolation) or as a consequence of the integration (e.g., overlay) of two or more representations. Attempts to monitor and measure the reliability of geographic representations need to track this interplay of discretization.

Tracking the interaction of multiple levels of discretization becomes particularly complex in the integration of several geographic representations. In determining effects of multiple discretizations within a composite map one can identify several dependencies that originate either with the input maps or the model used for the overlay. Formulation of a composite resolution may depend on the purpose of the composite map representation, for example, the integration of a vegetation representation and a cadastral representation for the purpose of planning an optimal route for a new highway. The vegetation representation can have a coarser discretization than the cadastre and still provide meaningful information for the composite representation. When calculating the reliability of the compound resolution the bias can be handled by weighing the importance of the input maps according to the requirements of an application. As another example, one may want to generate a vegetation coverage for a large area for which part of the data exists at a resolution of a single tree whereas other data obtained from satellite imagery have a resolution of 1km

by 1km. Information on the reliability of the resulting resolution requires the inclusion of the effects of the attribute discretization imposed by generic classes as well as the effects of the spatial discretization.

Several researchers discuss the effects of resolution or scale (Goodchild and Proctor 1997) in a broad variety of approaches. Watzek and Ellsworth (1992) for example, focus on an empirical approach to determine the perceived scale accuracy of computer visual simulations. Bruegger (1994) proposes spatial theory models for integrating datasets of different levels of resolution in GISs. Cushnie (1987) discusses the interactive effect of spatial resolution and degree of internal variability within land-cover types on classification accuracies. A different approach is taken in Canters et al. (1999) and Moody and Woodcock (1994) who focus on the errors introduced in land-cover proportions due to varying scale. Comparable methodologies were investigated by Burrough (1983) and Oliver and Webster (1986) where they concentrate on the influence of variations in a continuous field. Turner et al. (1989) investigate the effect of different scales on different landscape indices (e.g., contagious). An application specific approach (i.e., road density estimates) of scale dependent accuracies can be found in Wade et al. (1999). On a global scale Townshend and Justice (1988) elaborate on the effects of resolution in conjunction with a specific application—global monitoring of land transformations. Similar effects such as aggregation and support are discussed in Heuvelink (1999). Csillag et al. (1992) come close to articulating the problem of reliability but from a different perspective. In Prisley and Smith (1991) the effects of the underlying variation in the attribute variable on the decisions that were based on the GIS are investigated.

The influence of discretization on the quality of spatial representations has not been addressed in any systematic way. Van Groenigen and Stein (2000) as well as Burrough and McDonnell (1998) address a similar problem in a slightly different way. They are interested in optimizing the layout of a sample field. However, in their approach the underlying variation of the attribute does not play a central role. Their approach is

based on an a priori optimization whereas we are interested in estimating the loss of information a posteriori. Their approach is geared towards data producers whereas our approach concentrates on providing the user with helpful information on the inherent uncertainty. In general, the overall reliability of a spatial representation is less influenced by the accuracy or precision of a measurement than by the number, density, and spacing interval of the measurements. Accuracy measures are most often associated with well-defined points, which have little to say about unmeasured locations. Discretization is an implicit measure of what is not known or what might be missing as a result of the discretization.

2.6 Remarks

This chapter is an overview of important concepts related to data quality in GIS. There are distinctive differences between error perceptions when looking at a paper map versus a raster presentation versus a vector presentation. There are also distinctive differences in the models dealing with the inherent uncertainty. These distinctions are mainly based in the differences found in data collection, data representation and data storage.

There are several ways of dealing with uncertainty. Some models put the user in charge and some suggest dealing with uncertainties internally. However, all of the approaches are aiming at a better understanding and communication of uncertainty. In our opinion this increases the value of a product and decreases instances of misuse of a data set. The user should have the ability to judge the uncertainty of a conclusion that was based on GIS representations and analysis.

Discretization is a very important aspect in addressing the uncertainties for a given representation. In any given scenario, all sample points could be measured accurately—however the resulting presentation could show an overall uncertainty that is unacceptable due to an insufficient sample point density (i.e., spatial domain). On the other hand the same data set could be measured once every year—no conclusions could

be drawn about a certain day (i.e., temporal domain). Last but not least choosing a single class to represent the whole data set would make no sense at all (i.e., thematic domain).

Any sources of uncertainty such as discretization propagate through all phases of information management. This includes, for example, geo-referencing, generating a continuous representation, any transformation as well as map overlay operations. Thus, the implementation of metrics that can calculate or estimate the effects of the propagation of uncertainties is essential.

For the ability to implement the mentioned concepts it is essential to develop an adequate data quality model. Metrics for calculating uncertainties require certain input parameters (e.g., time of measurement, sample point distribution for a continuous representation) that can be stored or indexed in a data quality model.

Chapter 3
Data Quality Model

Data quality models are an essential part of any formal error analysis. Error metrics explore the effects of different sources of uncertainty on GIS applications or products. Increasingly complex error models require a more complex data quality model that allows a more focused access to necessary information on inherent imperfections. Progress on managing uncertainty in geospatial information will benefit from identifying sources of uncertainty and understanding how they affect units of information.

This chapter builds the foundation for a data quality model by defining what we call terms of imperfection where imperfections are a subset of sources of uncertainty (Figure 3.1). The chapter formally defines units of information (measurements, measurement vectors, spatial measurement fields, values, coverages, databases, queries, query results, and presentations) and relations among these units. Imperfections are present in all units of information but different types of imperfections show patterns of association with particular units of information. Specifically, certain types of imperfection originate in or apply to certain units of information. In addition propagation behaviors differ with the type of imperfection. The framework presented in this chapter provides a foundation for implementations in which metrics for specific

types of imperfection can be attached to units of information and appropriately propagated.

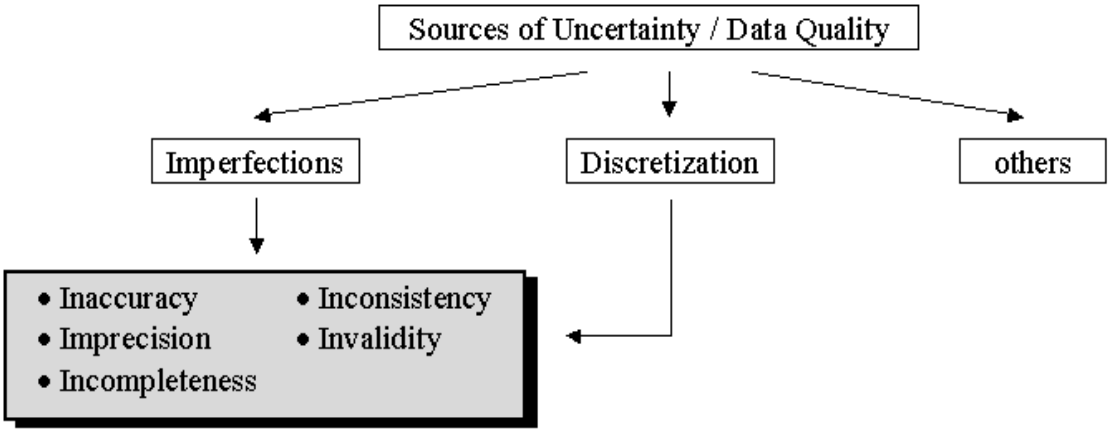


Figure 3.1—Sources of Uncertainty

The first section of this chapter defines units of information and follows with definitions for the terms of imperfection as well as a more detailed discussion of discretization. The last section develops a framework that describes how different types of imperfections attach to different units of information and whether the imperfections propagate from unit to unit.

3.1 Definitions

3.1.1 Units of Information

Units of information are logical units of information commonly encountered in an information system (Figure 3.2). A more detailed discussion of their relationships is given in section 3.2. These units of information range from raw observations to processed information and include both atomic and aggregate units. The units are organized into three categories: (1) units which are most closely associated with data acquisition, (2) units most closely associated with data or information management

(storage and processing) within a computer system, and (3) those associated with data retrieval and display.

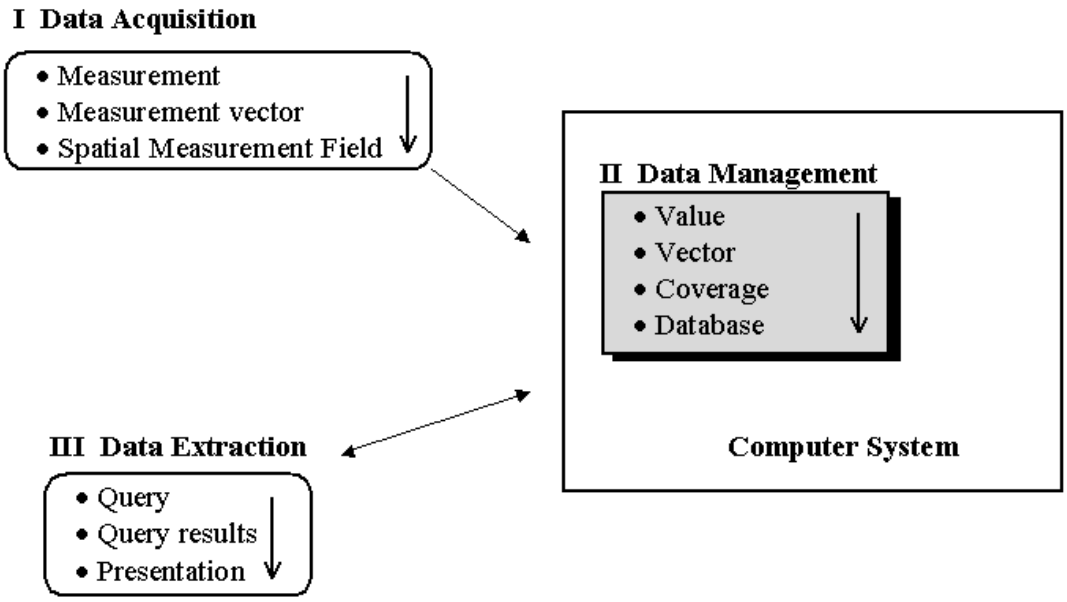


Figure 3.2—Units of Information

Each of the units of information has certain attributes and operations that can be performed on them. Thus, for an implementation of the data quality model we propose an object-oriented approach. Similar ideas of ontology from a philosophical position can be found in Frank (2001, to appear). The discussion in the following section describes possible attributes and operations for specific units.

3.1.1.1 Data Acquisition Units

Within the data acquisition level we identify three units of information: an individual measurement, a measurement vector, and a spatial measurement field.

Measurement: A measurement is any raw observation obtained by using some field instrument, laboratory procedure or survey questionnaire. A measurement can be an atomic or compound unit. Table 3.1 shows an outline of the object Measurement with a partial list of attributes and operations. Additional attributes could be attached to the

proposed Measurement object. We represent a measurement by M and its attributes by a_i where i = 1 … n attributes. Examples of a measurement include a single temperature observation (e.g., M = {c}), a GPS time observation (e.g., M = {t}), or a response to a single question on a survey questionnaire such as a question from the census (e.g., M = {q}).

Object	Measurement
Attributes	ID
	Variable
	Attribute value
	Level
	Units
	Support
	Instrument
Operations	Change units
	Change level of measurement

Table 3.1—A Measurement Object with its Attributes and Operations

An example of a measurement vector is M = {x,y} where x,y form a coordinate pair measured, for example using a digitizing tablet. An example of a measurement field is a raw satellite image (e.g., M = $\{s_i\}$ where i = 1 … n pixels in the image), or the set of responses to all questions on a survey questionnaire (e.g., M = $\{q_i\}$ where i = 1 … n questions). The attributes for each measurement are important metadata (FGDC 1994). Metadata for a measurement include the measurement units, measurement instrument, measurement procedure, measurement operator, the variable measured, its level of measurement, and its support (Beard 1996). Support is an aggregation level of the measurement (Heuvelink 1999).

Measurement Vector: A measurement vector is composed of measurement objects. It is a compound measurement for which a spatial and temporal measurement are made simultaneous with one or more attribute measures. Table 3.2 shows possible attributes and operations associated with a measurement vector.

Object	Measurement vector
Attributes	ID
	x-coordinate measurement_ID
	y-coordinate measurement_ID
	z-coordinate measurement_ID
	Time measurement_ID
	Attribute measurement_ID 1
	⋮
	Attribute measurement_ID n
Operations	Coordinate transformation

Table 3.2—A Measurement Vector Object with its Attributes and Operations

A measurement vector is represented by MV = {x,y,z,t,a_1,a_2,a_3,...,a_n} and can contain multiple attributes at the indicated spatial location at a given time. We distinguish different types of measurement vectors. A measurement vector can be, for example, spatial, temporal, or spatio-temporal. The interdependence among values forming a vector results from the measurement procedure used to generate the values and reflects a measurement framework (Sinton 1978; Chrisman 1997). An example for a spatio-temporal measurement vector is a compound of location measurements {x,y,z}, a temperature measurement, and a time stamp. An example of a spatial measurement vector could include x and y measurements along with a temperature measurement.

Spatial Measurement Field: The measurement field is a collection of spatial measurement vectors that form a logical unit based on some commonly shared attributes or metadata. The field is assumed to be composed of measurement vectors made by the same procedures and instruments. As specified in Table 3.3 the spatial measurement field is represented by the set SMF = {MV | P(MV)} where P(MV) indicates a property of MV.

Object	Spatial measurement field
Attributes	ID
	Measurement vector_ID 1
	⋮
	Measurement vector_ID n
	Discretization
Operations	Weighted average of m measurement vectors
	Adjustment on network of m measurement vectors
	Change support (punctual vs. block average)

Table 3.3—A Spatial Measurement Field Object with its Attributes and Operations

For example, a field could be the set of climatic observations made at several climate stations by similar instruments, a set of all questionnaires from one census period, or a set of GPS observations made over the period of a week using the same type of receivers and base station. Of particular interest in the context of this chapter are spatial measurement fields that share the same instruments and procedures and hence will have similar imperfections, which allows us to define the attribute discretization. For example, for a regular sample point distribution of SO_4 measurements we can record a constant spatial discretization value based on the distance between two sample locations. However, if an irregular (e.g., clustered) sample point distribution were encountered then the attribute discretization would spatially vary.

3.1.1.2 Data Management Units

Within the data management level we identify four units of information: values, vectors, coverages and databases.

Value: A value is an atomic unit of information stored in a computer system. We represent values (VAL) by *x, y, z, t,* or *a* depending on whether the value represents a spatial, temporal or thematic quantity or class. A value is typically derived from one or

more measurements through a transformation. A transformation may simply involve conversion to a digital form or conversion from one digital form to another. Such transformations can involve a change of measurement unit (e.g. Celsius to Fahrenheit), level of measurement (e.g. interval to ordinal), measurement framework (Chrisman 1997), or support. The Value object is derived from a Measurement Object and thus has similar properties to the Measurement Object. The distinctive difference here is that the value has a specific data type assigned to it (e.g., integer or double) for representation within a computer. Examples of Values are temperature values, an elevation value, an x, y or z coordinate value transformed from a GPS time observations, a pH value obtained from a bulked soil sample, or a population value for a census enumeration unit. In this last example the value results from a transformation (summation) of census measurements over a geographic area. In the transformation, the support has changed from a household to an enumeration unit. Values will have metadata which include the values' units, level of measurement, the measurement(s) from which the value was derived, the type of transformation used to derive the value where applicable, and any transformation parameters.

Vector: A vector is the counterpart of the measurement vector. It is a set of interdependent values that include spatial, thematic and/or temporal dimensions. A spatial vector might consist of an *x, y* and *a* value. A vector may also be a single value. The Vector object —similar to the Value object —is derived from a measurement vector (again with the specification of data type for each component). We represent a vector by V such that we might have a vector $V_1 = \{x,y,z,t,a_1,a_2,a_3,...,a_n\}$ which includes x,y,z values which describe a position in three dimensional space, t which indicates a time stamp value, and $a_1,a_2,a_3,...,a_n$ which indicate a set of thematic values associated with the specified location and time. Another vector $V_2 = \{x,y,z,t_i\}$ might be a three-dimensional coordinate with an associated time stamp. A vector can be a polygon $p = \{V_k\}$ with $k = 1 \ldots n$ or a grid cell with one or more associated attributes. A spatial measurement field gathered from one observation campaign can result in a set of related vectors. For example, the stereo compilation from one aero triangulation

would be a set of related vectors sharing a common lineage. Metadata for a vector should include the measurements used to construct the vector, its data type, the transformation applied to a vector where appropriate, and its parameters.

Coverage: A coverage is an assemblage of vectors with a common dimension, such as a set of vectors $V_i = \{x,y,t,a_j | j = 3\}$, i = 1 … n, (e.g., related across the common attribute a_3). For example, a raster representation of soil pH generated by kriging (Cressie 1991) a set of spatial vectors $SV_i = \{x,y,a_j\}$ i = 1 … n is a coverage related by a common thematic value a_j where a_j equals pH. A processed satellite image is another example of a coverage that has a common attribute as well as a common time stamp. Coverages are represented by $R = \{SV_i\}$, i = 1 … n.

Database: A spatial database is an organized collection of coverages designated by D (e.g., $D = \{VAL_i, SV_i, Ri\}$ i = 1 … n). The database is a derived object with the spatial measurement field as its parent object. A database may be homogeneous as in a set of satellite images from one type of sensor (storing multiple coverages of the same type), or heterogeneous as in a set of satellite images plus a set of vectors of water quality observations plus kriged maps of water quality variables (storing different types of coverages that can either be exhaustive in their representation or consist of sample points).

3.1.1.3 Data Extraction Units

The data extraction level includes three units of information: queries, query results, and presentations.

Query: A query is a unit of information constructed by a user for the purpose of retrieving information from an information system. This unit will typically be an expression formulated from some combination of the proceeding units of information. We assume for purposes of this dissertation that a query is expressed in some controlled language or formal query language such as SQL. At a minimum a query

specifies a database and conditions for coverages (R), vectors (V) or values (VAL) residing in the database. For example, a query Q might be a request to a database for a set of vectors containing a similar time value (e.g., Q = {D, V | V(t) = 1997}), a request for coverages with a common thematic value (e.g., Q = {D, R(a) | a = soil pH}), or a request for coverages depicting a specific geographic area (e.g., Q = {D, R(x,y) | n1 ≤ x ≤ n2, n3 ≤ y ≤ n4}).

Query Result: A query result is the unit of information generated by an information system in response to a query (i.e., Q ⇒ QR). This unit will be a set of values, vectors, or coverages with one or more attributes matching one or more attributes specified within the initiating query. A query result may consist of a null value. Metadata for a query result should include a count of the total units returned along with the generating query.

Presentation: A presentation is defined here as a query result which has been transformed for communication to a user, P = *f*(QR). The presentation can be in a textual, graphic, or even auditory format (e.g., oral driving instructions). One transformation, for example, may be a symbolization or graphic encoding of individual components of a query result such that a map is created. In generating a map, typically individual values or value ranges will be assigned specific visual variables (e.g. color, size, shape). Metadata for a presentation should include encoding rules, scale, projection, etc.

3.1.1.4 Running Examples

This section contains an example that illustrates the various units of information described in this chapter. The example describes a particular scenario and gives a description of the units of information contained within the example.

Description: SO_4 contamination of the soil within a certain area. This example introduces a combination of two requirements: not only the position but also the

attribute (i.e., the level of contamination) are to be measured. For the sake of simplicity we assume that the spatial locations have been assessed directly in the field (without considering time measurements of GPS or directions and distances of a tacheometer).

Measurement: the contamination of the soil (at a certain location – where the information concerning the location has to be handled as a separate measurement). Table 3.4 shows a possible scenario for a measurement of SO_4 concentration recorded as an attribute value (VAL). Additional Objects would be needed to record the coordinates (e.g., COOR) of the location as well as multiple measurements at the same location.

Object	Measurement	
Attributes	ID	VAL-212
	Variable	SO_4
	Attribute value	100.2
	Level	ordinal
	Units	ppm
	Support	$1cm^3$
	Instrument	S-318
Operations	Change units	
	Change level of measurement	

Table 3.4—A Measurement Object Showing an SO_4 Measurement

Measurement Vector: Table 3.5 shows one of several Measurement Vector objects that handles a spatial coordinate (e.g., x,y, and z) and the measured SO_4 level of several SO_4 measurements.

Object	Measurement vector	
Attributes	ID	SV-14
	x-coordinate measurement_ID	COOR-X310
	y-coordinate measurement_ID	COOR-Y310
	z-coordinate measurement_ID	COOR-Z310
	Attribute measurement_ID 1	VAL-212
	⋮	
	Attribute measurement_ID n	VAL-215
Operations	Coordinate transformation	

Table 3.5—A Measurement Vector Showing an Aggregated Vector for SO_4 Contaminations

Spatial Measurement Field: is a set of measurement vectors showing contaminations and their spatial location. Table 3.6 shows an example of a spatial measurement field. In the given example we can also see that we are now able to indicate the inherent spatial discretization.

Object	Spatial measurement field	
Attributes	ID	SMF
	Measurement vector_ID 1	SV-1
	⋮	
	Measurement vector_ID n	SV-187
	Discretization	Regular at 50x50m
Operations	Weighted average of m measurement vectors	
	Adjustment on network of m measurement vectors	
	Change support (punctual vs. block average)	

Table 3.6—A Spatial Measurement Field Object as an Aggregation of Measurement Vectors

Value: the degree of contamination in the required unit (e.g., percentage) with specification for a computer representation (e.g., integer or real). For example, the

previously indicated measurement of 100.2 might now be stored as “100” as a result of choosing integers to store SO_4 measurements.

Vector: Each value has a specified computer representation (i.e., data type) similar to value.

Coverage: Assuming that the goal is a coverage of the SO_4 contamination within a certain area, one solution is to represent the data using a raster-representation. The xy-parameters of a vector are used to assign a contamination value to a certain pixel that can be seen as the initial coverage where only a few pixels actually have assignments of contamination values. Another coverage is the result of the application of a spatial process like kriging—where we generate levels of SO_4 contamination across the whole area of interest. This method generates new vectors one for each pixel at a spatial discretization specified by the process which collectively form the coverage.

Database: the assembly of the above coverages – or a combination with more coverages based on different attributes of the same area or a wider ground coverage.

Query: A request for areas that have a percentage of contamination that is higher than 20% within the stored coverage *a*: SQL> select * from *coverage a* where *c > 20*.

Query result: all vectors from coverage *a* that have a percentage of contamination (i.e., c) > than 20.

Presentation: A map depicting the vectors extracted by the query plus possible contextual information (e.g., roads) for the same geographic area.

3.1.1.5 Relations Among Units of Information

Figure 3.3 summarizes relations among units of information in the form of an entity-relationship diagram. From this diagram we can identify the relationships that exist

between one unit and any other. Several hierarchical relationships occur and are designated by aggregation “is member of” relations in Figure 3.3 (e.g., value is member of a vector, which is member of a coverage, which is member of a database). There are also several relationships (e.g., measurements transform to values), which we call transformation relations, and are designated by “transforms to” in the diagram. Moreover, Figure 3.3 illustrates that a value may be transformed to a new value, a vector to a new vector, and a coverage to a new coverage. The differences between aggregation and transformation relations become particularly important in modeling propagation of the imperfections.

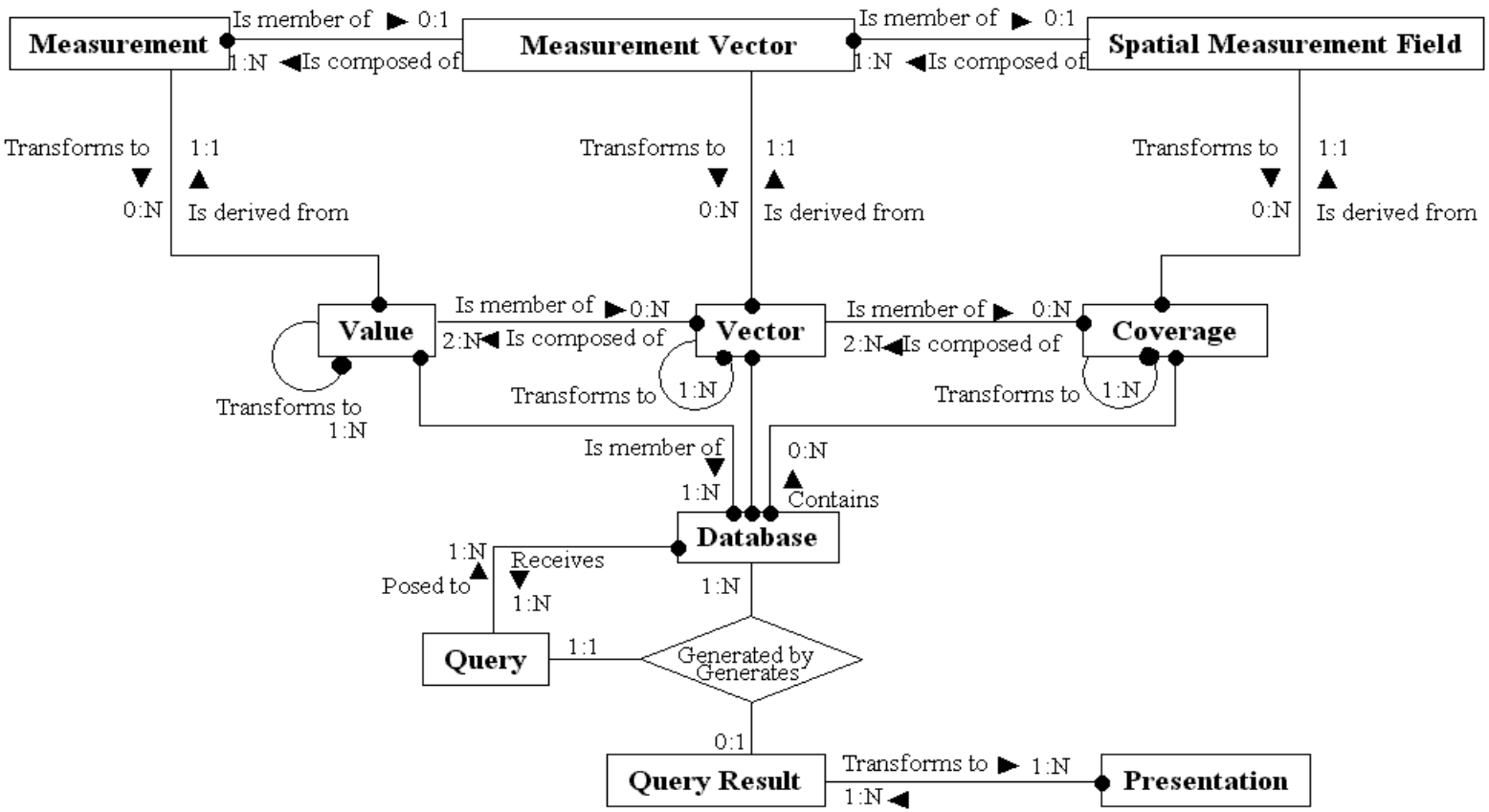


Figure 3.3—Relationship Among the Units of Information

The units of information represent the information life cycle beginning with a single measurement to the final stage the presentation. The identified units are exhaustive with regards to possible intermediate forms information can take. In everyday life, however, some of these units can also be combined. For example, a weather station can be composed of a digital temperature sensor and a database. In this scenario the unit measurement and the unit value are combined. Nevertheless, to handle all data quality issues adequately it is advantageous to separate all entities (i.e., units) in a conceptual model—such as the discussed data quality model.

3.1.2 Sources of Uncertainty

In this subsection we formally discuss two sources of uncertainty. The first part focuses on imperfections, whereas, the second discusses discretization effects.

3.1.2.1 Terms of Imperfection

An imperfection is the broad term used to cover all defects found in units of information. A unit of information that is imperfect may have one or more of the following defects: it may be inaccurate, imprecise, incomplete, inconsistent or invalid. This section defines the types of imperfections that potentially apply to the units of information described above. These closely relate to previously identified data quality components (CEN/TC 287 1995) but we clearly distinguish each term and indicate measures of these terms.

Inaccurate: A unit of information is considered to be inaccurate if it deviates from the true value or a value accepted as the true value. Thus, probably every measurement is—to a certain degree—inaccurate. Inaccuracy is measured as the difference between a unit of information and another unit of information specifically identified as ground truth (quantity accepted to be true). It is distinguished from imprecision by the following example. The elevation of a point can be measured by two different methods;

one method uses a stereoplotter and the other GPS. With the first one we can achieve a precision (see below) at the meter level, whereas with the latter one precision is achieved at the centimeter level. To measure the inaccuracy of the stereoplotter data we may accept the GPS measurements to be true, and thus, in this example, the inaccuracy is the difference between the two quantities.

Imprecise: Imprecision has a number of different meanings. A unit of information is considered precise if it is obtained with a high precision instrument (instrument precision) and represented or stored with high precision (numerical precision). Instrument precision is associated with repeatability. A precise instrument (or laboratory process) is one that can generate very similar measurements over many trials. Numerical precision relates to the number of significant digits used to represent a value. Numerical precision is often entirely independent of instrument precision. In the first case, imprecision in a measurement is inherited from the instrument. Instrument precision is usually available from instrument (process) specifications or calibration tests. With numerical precision precision is inherited from a value, vector, or coverage specification (i.e. single precision, double precision, number of decimal places). Precision can also be associated with an information processing operation. In this case, the precision is typically inherited from a parameter of the process (e.g., a tolerance value). For example, a line smoothing operation changes the precision of a polyline by changes in a smoothing parameter or tolerance value.

Imprecision as used in the information science community (Morrisey 1990; Parsons 1996; Smets 1997) refers to the case where units of information are ranges or sets of values (e.g. $[a \leq u_I \leq b]$ or $\{u_{a,}, u_{b,}, u_{c,}, u_d\}$). For example, rather than assign age a single value (e.g. John is 30) it may be assigned a range (e.g. John is between 30 and 40). The statement that John is between 30 and 40 is imprecise but accurate if John is 31. This type of imprecision can be independent of instrument, numerical, or process precision. For example, a survey question may ask a person's age and give ranges as choices. The choices are imprecise and hence the measurement will be imprecise since

it is multi-valued. A person may make the exact same choice of age range over many trials and hence the measurement will not be imprecise in the sense of instrument precision. We will refer to this type of imprecision as multi-value imprecision.

Inconsistent: Units of information are inconsistent if they are in contradiction with a determined set of rules, commonly expected relationships (Kaintz 1995), or with other units. For example, a database may contain information on a person whose age is stored as "10", and marital status is stored as "Married". In a spatial example, two polylines may be considered inconsistent in a topological vector-representation if they cross. A unit of information may exhibit inconsistencies among its component parts if it is a compound unit. We will refer to this as internal inconsistency. When an atomic unit or a compound unit is inconsistent with some other units of information, it is referred to as external inconsistency (e.g., one coverage inconsistent with another coverage). There are two approaches to addressing inconsistency. One approach is to simply flag the inconsistency when detected, the other is to correct it if possible. If the inconsistencies are flagged, one measure of inconsistency is the number of occurrences.

Incomplete: An incomplete unit of information is one lacking some part. This term will not apply to a unit with a single part (i.e., an atomic unit such as a value). A vector representing a 3D spatial coordinate is incomplete if the z value is missing; e.g., V = [x,y,,*,t,a]. A satellite image is incomplete if a line drops out due to a transmission problem. An incomplete unit of information can use a null value as a placeholder for the missing information (Codd 1979). A measure of incompleteness would, therefore, be the number of nulls. One interpretation that takes this a step further is that a null value can be a value that is either "undefined", "inapplicable", or "nonexistent" (Parsons 1996). In this case, a measure of incompleteness would be the counts of each of the specified types of nulls.

Invalid: The term invalid is another type of imperfection with several meanings. By one definition, a unit of information is invalid when one or more of its component parts

are outside a set of possible or allowed quantities or classes (e.g., a soil class outside the set defined for a county). Another definition of invalid applies to a measurement that is not in fact a measure of the intended phenomena. A measurement of a complex concept (i.e., intelligence, biodiversity) for which there is no commonly accepted measurement procedure is more likely to be found invalid, but this definition is not measurable and we do not consider it further. A designation of invalid can apply to an impossible relation as well as impossible values. For example, within vector V = {x,y,a}, a value such as the thematic value may not be valid for the spatial component (i.e., population at a point, angle of inclination for an area). A unit of information can also be invalid as a result of an invalid operation. For example, there may be two coverages (temperature and population density) that contain valid ordinal values. If these coverages are added, the resulting values (coverage) may be considered invalid. We also recognize the potential for temporal invalidity (CEN/TC 287 1995; Guptill and Morrison 1995). A coverage may be invalid as a response to a query if its timestamp is not current. For example, queries to a transportation information system during a large sporting event may give invalid results if the database has not been updated to reflect temporarily modified traffic patterns. Specifically we define invalid as being any value or relation that is outside a specified domain. The determination of invalidity thus comes form specification of a domain. For example assume air temperature has a domain of -70°C to +70°C, any value outside this range will be designated as invalid. Invalidity implies that the unit of information must also be inaccurate (since it is not within the attribute-domain it cannot be identical with ground truth).

3.1.2.2 Discretization

In photogrammetry the term resolution is often used to describe the resolving power of an image. In digital photogrammetry, for example, it indicates the ground area represented by a pixel. Within this dissertation we use the term discretization. Thus, the photogrammetric term resolution translates to discretization in the spatial domain (i.e., spatial discretization).

GIS databases contain multiple examples of discretization across thematic, spatial, and temporal dimensions. Examples of thematic discretization are the number of classes used to present a continuous variable or the symbol chosen to depict a specific feature in a presentation (e.g., single houses versus outline of a city). Instances for spatial discretization are, for example, the spacing of sample points or the size of a pixel in a raster based GIS representation. Instances of temporal discretization are, for example, a temporal sample interval or the time between two updates of a database.

3.2 Framework

This section links occurrences of imperfections and discretization effects to specific units of information. Table 3.7 is a cross-tabulation of units of information against types of imperfection. Table 3.8 shows the cross-tabulation of units of information against discretization. The tables are designed to show in which information units the various imperfections first occur. A checkmark appears in the tables if the indicated imperfection or discretization can be determined for the specified unit of information. By determined we mean that the imperfection can be identified and measured as an attribute of the indicated unit of information or that a new level of imperfection or discretization can be determined beyond that propagated from the generative unit. For example, Table 3.7 indicates that inaccuracy can be determined for spatial measurement field and query result. No checkmark indicates that either the imperfection does not apply to a unit or occurs in the unit by propagation (Heuvelink and Burrough 1993; Heuvelink 1998) only.

Within this framework we are only interested in initial occurrences of uncertainty. The advantage of detecting the initial occurrence of any source of uncertainty is that it allows for identifying and handling the uncertainty at its origin. Subsequent propagation of a specific uncertainty through the information life cycle captures the entirety of the associated effects. For example, a GIS user determines that a certain coverage stored in his database has been adequate for a previous task, but, carries too many uncertainties for a second task he intends to perform. A closer look at the

coverage reveals that the initial measurements were sufficiently precise. Further investigations expose that the largest contributor to the uncertainty of the coverage is an affine-transformation to a geo-referenced coordinate system. Since, in our example, the user is not interested in solving the given task in an absolute coordinate system he can use a previous stage of the information where the data is provided in a relative coordinate system.

	inaccurate	imprecise	inconsistent	incomplete	invalid
Measurement		✓	✓	✓	✓
Measurement Vector	✓			✓	✓
Spatial Measurement Field	✓		✓	✓	
Value		✓			✓
Vector			✓	✓	
Coverage			✓	✓	
Database			✓	✓	
Query		✓	✓	✓	✓
Query result	✓				
Presentation		✓			

Table 3.7—Units of Information - Terms of Imperfection: Initial Occurrences

	spatial discretization	thematic discretization	temporal discretization
Measurement		✓	
Measurement Vector			
Spatial Measurement Field	✓		✓
Value		✓	
Vector			
Coverage			
Database			
Query	✓	✓	✓
Query result			
Presentation[1]	(✓)	(✓)	(✓)

Table 3.8—Units of Information - Discretization: Initial occurrences

Since there is at least one contributor to imperfection for each unit of information, Table 3.7 indicates that all units of information as imperfect. Furthermore, we have to deal with the fact that the imperfections propagate among units of information through both transformation and aggregation relations as shown in Figure 3.3. Propagation of imperfection within the system occurs as shown in Figure 3.4. Each type of imperfection propagates in a different manner and varies with the type of relations between information units. For example, if there is a transformation relation between units of information, inaccuracy must be recalculated by comparing the transformed unit of information to the ground truth. On the other hand, the imprecision of the same transformed unit of information does not need to be recalculated because functional propagation models can handle the propagation. Other imperfections can be associated with a unit of information by inheritance.

[1] dependent on the medium of presentation (e.g., a single map cannot depict a time series)

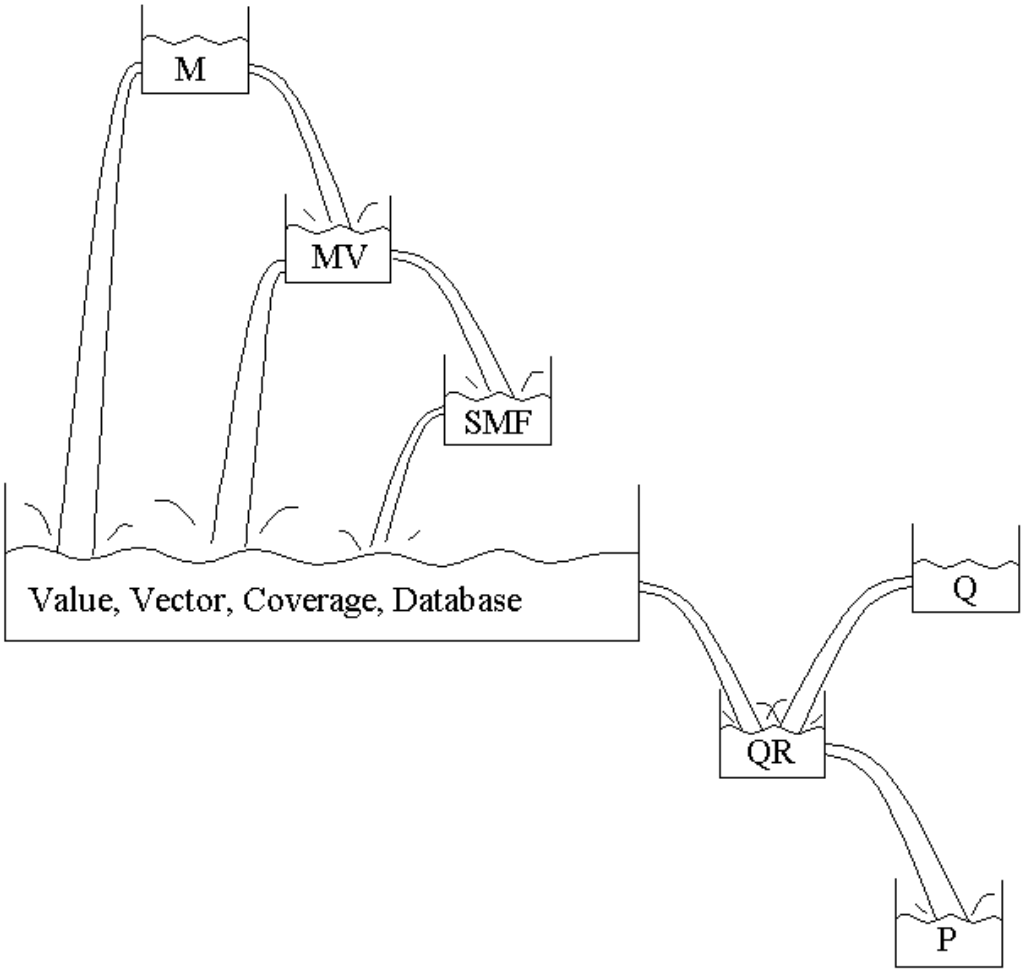
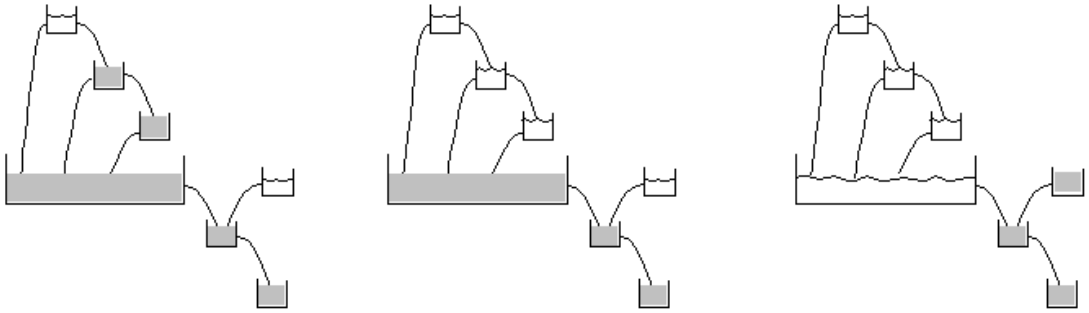


Figure 3.4—Buckets and Pools Representing the Propagation of Imperfections

In Figure 3.4 the buckets and pools represent the different units of information and the water represents the possibility to transport any imperfection into the next unit(s). Figure 3.5a-c illustrate the propagation of any introduced sources of uncertainty. In Figure 3.5a the uncertainty is introduced with the measurement vector (could be in the form incompleteness or invalidity; see Table 3.7) and propagated into the next pool of value, vector, coverage, and database followed by the propagation to the query result and finally to the presentation. Figure 3.5b depicts the situation where any of the units:

value, vector, coverage, or database introduces a form of uncertainty. This example shows the reason why the units: value, vector, coverage, and database are within one pool. Let us assume that a transformation of a coverage (e.g., smoothing of a set of polylines) introduces an imperfection (e.g., imprecision) then this imperfection would affect all the values and vectors that belong to this coverage. Furthermore, Figure 3.5b shows the rest of the propagation of the imperfection similar to Figure 3.5a. Figure 3.5c shows the propagation of any source of uncertainty introduced with the unit query. In contrast to Figure 3.3 there is no connection between the query and the database. On the other hand there is a new connection that shows the propagation to the query result. The following section describes the measurement and propagation of each imperfection type and discretization in greater detail.



a b c

Figure 3.5a-c—Examples of Propagation of Imperfections; a: Measurement Vector to Presentation, b: Value, Vector, Coverage, and Database to Presentation, c: Query to Presentation

3.2.1 Measurement and Propagation of Inaccuracy

Determination of accuracy requires a “ground truth” for comparison. A “ground truth” can be generated from redundant measurements, by independently collected measurements, by stochastic simulation (Ripley 1987) (where one can actually simulate the amount of inaccuracy and not ground truth per se), or by expert opinion. Calculating the mean (e.g., true value) and its standard deviation can be considered as a

ground truth. No measure of accuracy is possible without one of these forms of “ground truth”.

Measurement: A single measurement does not allow the calculation of an accuracy measure. Although it might be inaccurate, the accuracy cannot be determined from a single measure. While comparison with a ground truth (i.e., an accepted true value for the measurement), will result in an accuracy value, in this chapter we exclude any external information that is not either generally accepted or common knowledge (e.g., for a right triangle, Pythagorean Theorem; America is larger than Australia; Europe is not an island). The reason for this restriction is that “ground truth” are measurements themselves (including those calculated, for example, by using Monte Carlo simulations). “Ground truth” generated by different methods will produce different accuracy values and, therefore, introduce new imperfections beyond the scope of this dissertation. Under this restriction, an individual measurement cannot have a measure of inaccuracy (and thus does not receive a check mark in Table 3.7).

Measurement Vector: A measurement vector can be a result of redundant measures of the same phenomenon (see Example) at the same location. When such redundancies are present we can calculate inaccuracy measurements by computing the mean and standard deviation and hence measurement vector receives a check mark in Table 3.7.

Spatial Measurement Field: A spatial measurement field can introduce inaccuracies if the spatial measurement field is a geodetic network. For this scenario we assume that the measurement vectors consist of redundant direction and distance measurements. The redundancy of the system allows the calculation (via an adjustment) of locations that are accepted as being ground truth and their associated inaccuracy values. Thus we suggest that the spatial measurement field receives a check mark in Table 3.7. In the case of a heterogeneous (i.e., measurements of different variables collected over the same geographic area), the redundancies are not present to compute an inaccuracy measure.

Value, Vector, Coverage: Each of these units of information can be inaccurate but the inaccuracy cannot be determined without the presence of some "ground truth". A single value, vector, or coverage like a single measurement does not in itself contain the information to determine inaccuracy. Assuming the presence of a "ground truth" such that inaccuracy can be measured we can examine whether it propagates. In the case of a transformation relation, the inaccuracy of the transformed unit of information must be recomputed by comparing it to the ground truth once again. For example, the inaccuracy of an attribute value *a* is calculated *a* - GT $\Rightarrow$ inaccuracy measure, where GT is ground truth. After a transformation is performed, the inaccuracy of the transformed value *a'* is then recalculated as *a'* – GT' $\Rightarrow$ inaccuracy measure. In the case of an aggregation relation (e.g., values assembled for a coverage) the inaccuracies may be summed or averaged over common values but this is not particularly useful for spatial coverages where the variations in accuracy over space are of interest (Kyriakidis et al. 1999).

Database: No single measure of inaccuracy is applicable to a database only to sub units of the database.

Query: For our definition the term inaccuracy is not associated with a query.

Query Result: A query to the spatial database might require a process in which a new measure of inaccuracy applies. For example, assume the spatial database contains information on the spatial location of a coastline observed during several days with one coverage generated for each hour. In order to eliminate effects due to high and low tide, the query demands an operation for extraction of a mean coastline. Due to present redundancies, this operation could also give the calculated standard deviation for the requested result.

Presentation: Since the presentation is a transformation of the query result, all possible inaccuracies are propagated as a function of the transformation.

3.2.2 Measurement and Propagation of Imprecision

As mentioned in the previous section there are various types of imprecision. Numerical precision is present in all units that are stored within a computer system. The only units of information that may not incur numerical precision are measurement (e.g., tape or thermometer measurements recorded manually) and presentation (e.g., a printout of an area of interest). Any instrument or numeric imprecision found in units of information is inherited from the generative unit. A measure of instrument imprecision can be recorded as the precision of the instrument (e.g., ± centimeter). Process imprecision must be considered any time a transformation is applied to a unit of information. Multi-valued imprecision (see section 3.1.2.1) can be present in any unit and propagated. The metadata associated with a unit should document each of these types of imprecision. By tracking these independently, instrument and process imprecision can be used to govern excess numeric imprecision (i.e., limit the number of significant digits reported). Table 3.9 shows the break down of the identified types of imprecision and the units of information to which they apply.

	numeric imprecision	instrument imprecision	multi-valued imprecision	process imprecision
Measurement		✓	✓	
Measurement Vector				
Spatial Measurement Field				
Value	✓			✓
Vector				
Coverage				
Database				
Query	✓		✓	
Query result				
Presentation				✓

Table 3.9—Units of Information - Break Down of Imprecision

Measurement: A measurement introduces the classical imprecision, instrument precision, which is derived from the ability of the instrument to repeatedly measure a spatial phenomenon with the same exactness (i.e., the measurement results vary within the given interval of precision). A measurement inherits imprecision from an instrument or measurement process. Precision is independent of the ground truth and, thus, must not be confused with the measure of inaccuracy. Since measurements are the origin of other units of information, instrument imprecision in a measurement propagates through all subsequent units of information (except the query). Multi-valued imprecision can occur in measurements as in response to a survey question whose options are multiple classes or ranges as discussed above.

Measurement Vector: Since the measurement vector is purely a grouping of existing measurements no new imprecision is introduced. The imprecision existing within the unit measurement vector is purely propagated from each associated measurement.

Spatial Measurement Field: No new imprecision is introduced in a spatial measurement field, since it simply involves the assembly of measurements (hence no checkmark in Table 3.7 or Table 3.9). If imprecision is present in a measurement it will

be present in the spatial measurement field. The measure of imprecision of a field as a whole however must be modified if the field is made up of measurements with different levels of precision (e.g., water quality measures carried out by different labs then assembled as a field). In such a case there are two options: to record each level of imprecision separately for each subset of the field or to indicate some function of imprecision such as minimum, mean, or maximum for the field. If the field is homogeneous with respect to instrument and procedures the measure of imprecision will be the same for the field as for an individual measurement.

Value: Process imprecision can be introduced with the transformation of measurement to value. A transformation can be a spatial process like kriging where interpolation may generate additional imprecision. Two aspects might not be obvious: a) why imprecision is introduced at the value level and b) why the term precision and not accuracy? Regarding the first question, when a transformation is applied to a coverage it is in fact applied to the values. The argument here is that the process precision has to be assigned to the lowest unit of information where it can be determined—and this is the value. From the value the imprecision propagates to the vector and the coverage (and later on to the database, query result, and presentation). Regarding the second question, let us first take a look at the similarity of a process and a measurement device. Kriging generates values for unmeasured locations. Thus, the process can be seen as a measurement process that generates a value at a spatial location. The process, however, is an interpolation or extrapolation and not a result of redundant measurements (the interpolation/extrapolation tool is the variogram in combination with a functional model). Since kriging does not generate redundancies the attached imperfection has to be considered as process imprecision.

Vector, Coverage, and Database: No new imprecision is introduced with these units of information. If they are imprecise, it is a result of the imprecision being propagated from generative units through transformation or aggregation. Applying a transformation (e.g., coverage $\Rightarrow$ new coverage) may introduce process imprecision

that affects values as discussed above (see Figure 3.4—pool representing the units: value, vector, coverage, and database).

Query: A query introduces imprecision when it is not formulated precisely. Looking at the following example, this statement becomes clearer. A user wants to view the highest mountain peak within a certain region, but is not sure how high that mountain is. The only thing the user is sure about is that the highest mountain is above 5000 meters. The user formulates the query in a way that the system shows all mountains that are higher than 5000 meters. Let us assume that the result shows three mountains. The query was imprecise because not only is the requested spatial phenomenon returned, but others are returned as well. This is an example of multi-value imprecision as described above (section 3.1.2.1). In comparison, "Find the location of the mountain peak with an altitude of 5746" is a more precise query.

Query result: The query result can be imprecise due to propagation of any imprecision that occurs in a previous unit. As shown in the example above, in addition to the required result, two other mountains were given. Another possibility is due to cross-propagation (propagation of one imperfection type to another) for example, an incomplete query can result in an imprecise query result. If one wants to know all parcels that are co-owned by Horace & Daniel but in formulating the query forgets to include the name Horace, the result shows all parcels owned solely by Daniel in addition to those that are co-owned by Horace & Daniel. Thus, the query result is imprecise (multi-valued imprecision). This can be interpreted as an introduction of a new component of imprecision within the query result as a consequence of the incomplete query. Since cross-propagation is not considered in this chapter it is not indicated as a source of imprecision (hence no checkmark in Table 3.7 or Table 3.9).

Presentation: In a presentation, imprecision is introduced if one applies any form of reclassification of an attribute. Figure 3.6 illustrates the reclassification of an attribute from classification scheme 1 to classification scheme 2. A similar concept is also

shown in Frank et al. (1997). The gray areas denote the imprecision introduced as a result of the reclassification. For example, assume a soil scientist gathers information at different sample sites and then produces a map coverage where the thematic classes are dependent on the actual grain size (e.g., increments in fractions of inches using a logarithmic scale - A1 to A9). On the user side a geologist is interested in the soil type (Sand - 0.0025 in. and Silt - 0.00015 in.(B1), pebble - 5/32 in. and granule - 5/64 in. (B2), cobble 2 ½ in. (B3), and boulder - 10in. (B4)), so a reclassification has to be conducted. The provided increments do not match the required increments and the result of the reclassification to 'granule and pebble' (B2) is the sum of A3, A4, and A5, where the gray shading in Figure 3.6 stands for the introduced imprecision. This can be viewed as another example of multi-value imprecision.

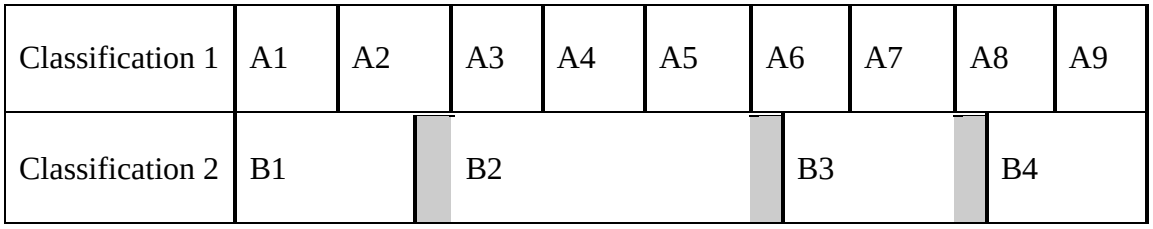


Figure 3.6—Imprecision in Reclassification

A textual presentation may use the numeric precision specified for a value. If excess precision has been specified to avoid rounding errors the presentation may show excess numeric precision (a form of imprecision) that is misleading. If we allow instrument, process, or multi-value imprecision to govern presentation of numeric precision we could get more faithful view of the pertinent imprecision.

3.2.3 Measurement and Propagation of Inconsistency

As noted above, units of information are inconsistent if they are in contradiction with a set of rules or commonly expected relationships. In order for an inconsistency to

be present an implicit or explicit relationship must be present. Atomic units of information (single measurement or value) can thus not be inconsistent. For example a measured z value for a single building corner cannot be inconsistent. If the set of measured z values for the four corners of a building are 350, 352, 349 and 373 meters the set is inconsistent since an expected relationship among heights of building corners (assuming that the rule that one can expect that if 3 corners of a building are within three meters the fourth corner does not vary significantly from the other three) has been violated. The measurement 373 meters may be inaccurate but without another measure to establish this we can only say it is inconsistent. Measures of inconsistency can be computed as counts of occurrences of identified inconsistencies or as probabilities.

Measurement: A measurement can be internally or externally inconsistent. Internal inconsistencies can be introduced in a measurement only if it is a compound unit. For example, answers on a survey instrument might be contradictory (e.g., answer 4: age is 8 – answer 16: plan to do the driver's license within the next 6 months). For an atomic unit (e.g., an individual temperature measurement), external information must be present to determine inconsistency.

Measurement Vector: A measurement vector cannot introduce any inconsistencies without referring to external knowledge. The measurement vector combines two or more measurements of different types and thus, one lacks the ability to compare it internally with measurements of the same type and draw conclusions on whether a singular measurement vector bears inconsistencies or not.

Spatial Measurement Field: Inconsistencies can also be introduced in a spatial measurement field. Internal inconsistencies in a spatial measurement field mean one or more measurements are inconsistent with the majority of measurements or measurements within a spatial neighborhood. As an example a sequence of measurements from a climate station taken an hour apart might look as follows: 9:00 am 47°F, 10:00 am 75°F, 11:00 am 52°F. This sequence of measurements is clearly

inconsistent with an expected pattern. The measurements cannot be claimed to be inaccurate because there is no truth-value available for a comparison. The values are also not invalid since they all are within the possible range of degree Fahrenheit. Answers to questions on survey instruments are prime examples of measurement collections prone to inconsistencies. An example would be an individual reporting an age as 4 years old, and the same individual reporting an income of over $100,000. This is an example in which the inconsistency might be measured as a probability of the relation being true.

Values: In general since values are atomic by themselves they cannot be inconsistent. Inconsistencies found in values occur by propagation from measurements or measurement collections. They are not introduced here as a source of imperfection.

Vector: As defined in the previous section a vector is a set of interdependent values. These interdependencies imply a relationship and so a vector may be internally inconsistent. Inconsistencies can be introduced at this level if a vector is assembled incorrectly. For example, an inconsistency could occur in a vector consisting of spatial coordinates along with some timestamp and attribute value if the attribute value was inconsistent with the location and timestamp (e.g., a temperature of 82°F for Alaska in January).

Coverage: Inconsistency at the coverage level can be introduced through aggregation or transformation. For an example of an aggregation problem, suppose spatial coordinates V = (x,y,z) of a GPS campaign were to be transformed and stored in meters. Another set of vectors was transformed and stored as decimal degrees. An inconsistency is introduced if the two sets of vectors are assembled to form the coverage. Transformations can also generate inconsistent coverages. For example, the Douglas algorithm can cause polylines to cross, creating topological inconsistencies (Douglas 1972).

Database: Inconsistencies can be introduced in a database. For example, if a parcel owner sells her property, the database will need to be updated to indicate the transfer of property. The database may have two coverages R_1 and R_2 where parcel owner information is stored. If the update is only made in one of the two coverages, the database will be inconsistent.

Query: A query is always formulated by a user. Therefore, inconsistencies in this unit of information are directly related to the user's knowledge of the database. This knowledge serves as a rule to establish inconsistency. For example, a query is inconsistent if a user queries the database for a value or vector that it does not contain (i.e., requesting the capital of Maine from a database of Alaska). This example is not necessarily invalid since no specific range of possible values is associated.

Query Result: Inconsistencies found in the query result are propagated from the previous units of information. They are not introduced in this unit of information. The query result for an inconsistent query will be the null set.

Presentation: Inconsistencies found in a presentation are propagated from the previous units of information. They are not introduced in the presentation. As seen in the above description of the query result, a null set may be the result. If this is the case, there is no presentation.

3.2.4 Measurement and Propagation of Incompleteness

A unit of information is incomplete if a part is missing. The missing information can range from the absence of a specific attribute, to an unanswered question on a survey, to an area in a sampling scheme where information was not collected. Causes for missing values can be the result of a measurement not made, a value dropped in assembly, or values lost in a transformation. Determination and measurement of

incompleteness requires knowledge of the whole (entire questionnaire, sampling scheme, etc).

Measurement: Incompleteness can only apply to compound units of information. If a single measurement is an atomic unit it cannot be incomplete. Typically measurements will be sufficiently structured such that missing components will be trivial to detect. A compound measure such as a measured coordinate pair will be clearly incomplete if one of the coordinate pair is missing (e.g. M = {x, *}. Similarly in the example of a measurement being the set of responses to all questions on a survey questionnaire described in the previous section the measurement is incomplete if one or more questions are not answered.

Measurement Vector: A measurement vector can introduce incompleteness. A measurement vector is incomplete if one or more components of the vector are excluded. For example, within the assembly process the location measurements (x, y, and z) were compiled but not the attribute measurement.

Spatial Measurement Field: A collection can clearly be incomplete if the incompleteness is due to propagation from an incomplete measurement. For example a spatial measurement field could be considered incomplete if in a collection of x,y coordinate pairs one of the coordinates is missing or if in the case of a set of questionnaires one of the questionnaires is not entirely filled out. In these examples, the incompleteness of the single measurement propagates to the spatial measurement field through the aggregation relation. An example of incompleteness being introduced in this unit of information is if a sampling scheme has been specified, but not all points have been sampled (e.g. a sampling scheme is developed for a soil survey, but not all the samples are measured). In this case the incompleteness is detectable because there is specification of the whole.

A problem arises in detecting incompleteness in a spatial measurement field where the entirety of the collection cannot be clearly specified. For example a spatial measurement field could consist of the building footprints extracted from an aero triangulation model. Unless an independent inventory of buildings exists incompleteness in such a collection cannot be detected.

Value: A value is atomic and thus cannot be incomplete.

Vector: Incompleteness can be introduced in a vector or propagated from an incomplete measurement or spatial measurement field. Incompleteness is easily detected within a vector because a vector must have a specified structure. For example, in a vector with a specified structure of V = {x,y,z}) for coordinates derived from GPS time observations, a missing z-coordinate (i.e., V = {x,y,*} is easily detected.

Coverage: Incompleteness can be introduced in a coverage or propagated from the generative units described above. Incompleteness can be introduced by the loss of entire vectors through assembly or transformation relations.. If the incompleteness is a result of propagation from incomplete vectors, the measure of incompleteness can be straightforward. The number and type of missing components from the assembled vectors can be tallied. (e.g. 3 missing timestamps, 4 missing attribute values). Using the soils example from the spatial measurement field unit described above, if the sampling scheme developed for a soil survey contains the vectors $V_i = (x,y,z,a_1)$ i = 1,2 … n, where x,y,z are the spatial coordinates and a_1 is the soil type, an incomplete coverage is produced if not all the sample sites are measured. Incompleteness due to loss of vectors may occur through transformation operations such as through a generalization operation which remove polygons or polylines. The measure of incompleteness in this case is the number of lost vectors. Undetected missing components in a spatial measurement field will result in missing vectors in a coverage but the loss remains undetectable.

Database: A database can be incomplete by any of the propagation scenarios described above. For example, if a coverage R_1 in the database is incomplete, the database D_1 will therefore be incomplete. Incompleteness can be introduced at the database level if entire coverages are missing. For example, if the required coverages R_i i = 1,2 … n, are not present in the database, then the database is incomplete. However if there is no specification of which coverages must be included in a database this will not be measurable.

Query: A query is incomplete if a user does not specify all of the required components. For structured queries, missing components should be easy to detect and correct. Using the example found under imprecise query result, if a user wants to know all parcels that are co-owned by Horace & Daniel, but only queries the database for parcels owned by Daniel, the query is incomplete. When a query is incomplete, it can lead to null, imprecise, or invalid results.

Query Result: Incompleteness in a query result is the result of incompleteness in the database or databases to which the query was posed. It is not introduced in the query result. For example, if a population database does not contain the population values for all of the states, a query requesting all of the states with a population over 100,000 may result in an incomplete listing of states.

Presentation: Incompleteness in a presentation is propagated from the previous units of information. It is not introduced in the presentation. In comparison to the previous units of information, a graphic presentation however can be incomplete due to the lack of important ancillary information (e.g. scale, legend, north arrow) that is essential for interpreting the presentation. These aspects of incompleteness are beyond the scope of this dissertation.

3.2.5 Measurement and Propagation of Invalidity

As described in the previous section there are a number of possible definitions for invalidity. However, the aspect of invalidity we focus on is the case of values being outside the range of a specific domain or values that are not possible. The definition also includes relations that are not possible. Under this definition the measure of an invalid unit is a binary measure: it is invalid or not. In contrast to inconsistency, invalidity documents the impossibility of a value or relation as opposed to contradictions among possible values

Measurement: The determination of an invalid measurement will depend on a specified domain range of a spatial, temporal or thematic variable. For example air temperature has a valid range, the spatial extent of the state of Connecticut has a valid range in some coordinate system say UTM, as does the temporal extent of activities or events such as a hunting season. Measured values that exceed the given range are flagged as invalid.

Measurement Vector: For a measurement vector one can consider the possibility of invalidity. An invalid vector can be introduced if a data file is corrupt. For example, if a measurement vector is defined as having the components of x, y, and an attribute value but was actually compiled as x, attribute value, y. Then any subsequent operations performed on this vector leads to deficient results. The source of these deficiencies should be identified as an invalid measurement vector.

Spatial Measurement Field: Invalidity is not introduced in this unit of information. A spatial measurement field can only be invalid through propagation from an invalid measurement. Specifically the presence of an invalid measure in a field will necessitate that the field is invalid.

Value: A value can be invalid even if the original measurement is valid. For example an incorrect transformation may cause a measurement to be converted to an

impossible value. A value can also be invalid through propagation of an invalid measurement.

Vector: A vector is invalid solely due to the propagation of invalid measurements and values.

Coverage: A coverage is only invalid due to the propagation of invalid units of information discussed above.

Database: A database is only invalid through propagation of invalid units of information as discussed above.

Query: A query is invalid when the user requests a value that is outside a given domain. For example, a query requesting all states with a population of -50 is invalid (e.g. Q = {D, R | R(a) = -50).

Query Result: The query result is invalid solely through propagation from the previous units of information. This type of imperfection is not introduced here. It is possible that the query result from an invalid query is the null set. This is the case in the previous example when requesting all states with a population of -50.

Presentation: The presentation is invalid solely through propagation from the previous units of information. This type of imperfection is not introduced here.

3.2.6 Measurement and Propagation of Discretization

The initial occurrences of discretization are indicated in Table 3.8 above. As discussed in a previous chapter discretization itself is not seen as an imperfection of the data. The following discussion highlights the units of information where one can initially determine a discretization effect.

Measurement: Neither the concept of spatial or a temporal discretization is associated with a single measurement. Both require the distance or time interval to the neighboring measurements. On the other hand thematic discretization might be introduced at this level. One has the option to measure an attribute at an interval or an ordinal level. For example, if a classification is previously determined than the measured attribute value is assigned to a specific class that bears a certain thematic discretization.

Measurement Vector: For the unit measurement vector we cannot determine any new levels of discretization. Here the concept of a spatial and temporal discretization is still not applicable. Since the measurement vector is a compound unit the thematic discretization is purely propagated.

Spatial Measurement Field: For the spatial measurement field the spatial distances between neighboring measurement vectors determine the level of spatial discretization. Similarly, for the temporal discretization the time interval between two measurement vectors determines the level of temporal discretization. Thematic discretization within a spatial measurement field is purely propagated from the previous units.

Value: The concepts of spatial and temporal discretization are not applicable. Thematic discretization might be introduced with the unit value. For example, if the measurement was conducted at an ordinal level or when switching to an internal data type one chooses to represent the data by a different thematic classification. The new classification would then introduce a new thematic discretization that is not necessarily introduced through propagation.

Vector, Coverage, Database: Within these units the level of discretization is purely propagated.

Query: The query can be posed at a certain level of discretization, which would introduce a new level of discretization. For example, the posed query might be aimed to

retrieve every tenth sample point stored in a coverage. Then the spatial discretization would be different from the level of discretization given in the coverage. It would be adapted to the level of discretization specified in the query.

Query Result: The query result does not introduce any new levels of discretization.

Presentation: Within the presentation one might find a new level of discretization. Here the level of discretization is dependent on the medium of the presentation. For example, a single map cannot represent a time series. The spatial discretization might also be different when compared to the stored coverage due to the extent of a given area. For example, one could store a coverage depicting temperature values of the whole world at a resolution (spatial discretization) of 1km x 1km. Subsequently, the discretization of the presentation would have to be changed if one would like to print this map on a single sheet of paper (letter or A4).

3.3 Remarks

There are several different approaches to handling uncertainty within a GIS. One current method is to produce results that provide evidence of uncertainty but undifferentiated by source (e.g.: inconsistency versus inaccuracy), where one could apply techniques such as Monte Carlo simulations. This is a powerful approach in the current environment since there is often limited ability to discriminate sources of uncertainty. Assuming we make progress towards better managing all units of information, starting with measurements, we should be better able to distinguish sources of uncertainty, and measure, and track them.

This chapter has discussed an object oriented data quality model. We also provide unambiguous definitions of imperfections, link these to specific units of information and demonstrate whether they propagate among these units. The specified relations among units provide the conceptual foundation for the development of mathematical tools (i.e. metrics, propagation algorithms) to handle sources of uncertainty.

One advantage of the suggested model is the potential foundation it provides for distinguishing sources of uncertainty. The simulation approach is robust but shows only the aggregate of all imperfections. With this approach the user lacks the ability to determine an individual source of uncertainty. Using the defined terms of imperfection and discretization and tracking them early on in the information lifecycle will allow presentation of the individual sources. The advantage of dealing with each source of uncertainty explicitly lies in the ability to identify a specific uncertainty as an unacceptably large contributor. Suppose a GIS user wants to take a look at a special area of interest in order to calculate distances as "accurately" as possible. A simulation may indicate that the overall uncertainty evaluation is unsatisfactory. Assume the main contributor to uncertainty is an imprecise transformation process from a local to a global coordinate system. Using the simulation approach the user is not able to determine the source of the unacceptable uncertainty. The latter model could provide the necessary information on the imprecision of the transformation process.

The approach presented in this chapter does not cover all aspects of propagation of imperfection existing in a GIS. For future work it is important to focus on the influence of cross propagation of terms of imperfection. These occur but are not fully described in this chapter. Some examples are the influence of incomplete data on imprecise results; imprecision on invalidity (the imprecision of a measure increases such that the resulting values lie outside the predefined interval); incompleteness on imprecision—where an incomplete census results in imprecise census values; or spatial discretization on inaccuracy (wide spacing between sample points diminish the accuracy of an interpolated coverage).

It is important to find a common terminology so that everyone can express and understand the aspects of uncertainty inherent in geospatial information. Adoption of common terminology enhance the users understanding of the data quality, which allow for more informed decisions. The following chapters make the assumption that one has

access to the data quality model—as described in this chapter and thus, access to the initial sample locations and attribute values.

Chapter 4
A Model for Detectable Objects

The chapter elaborates on the possibilities of determining the effects of discretization within spatial datasets. In the discussion we use the term resolution as an indicator for the ability to identify a certain object within a given GIS representation. This interpretation of the term resolution combines some properties of the photographic heritage related to the degree of discernable detail, and some of the properties inferred by the scale of a paper map—the users expectation to identify specific features at a certain scale. Resolution is a source of uncertainty as it constrains both what we can observe and represent. Without a model and measures of resolution we cannot formulate a measure of what may be missing from a spatial representation.

The model developed in this chapter considers the combined effect of spatial and thematic dimensions. The objective is a metric to resolve “objects” in “fields”. From a three-dimensional representation of the residuals (stored representation vs. higher accuracy) we obtain a relief map showing the minimal determinable variations—which can be used to detect the minimal size of a resolvable object. Thus, the resolvability of a spatial object can be determined by a function of the spatial extension of an object, its attribute value, and the three-dimensional relief of the inherent accuracy of the thematic representation. "Objects" in the context of this chapter are considered to be patches of

higher concentration, density, etc. A large patch may not be resolvable if its attribute value is weak compared to the accuracy of the “field” representation.

This chapter includes a case study of a sea surface temperature dataset collected off the coast of Maine. The approach for the case study is focused on—but not limited to—the investigation of the effects of discretization on kriged maps (i.e., a continuous raster-based representation) from a given sample dataset. We investigate the differences in the ability to detect a certain object by reducing the number of sample points. Based on the residuals from a comparison of a kriged map versus a representation that is accepted as being ground truth (which is also generated by applying simulation algorithms), the result provides a visualization of the inherent uncertainties due to discretization. Furthermore, the model provides the user with the possibility to analyze a stored representation for its ability to reveal an object of a certain spatial extension (i.e., x,y-coordinates) and a given attribute value.

4.1 General Considerations

There are several methods of generating a GIS map. One of them is to generate a raster representation of a continuous variable (e.g., sea surface temperature). For example, we could sample the variable of interest and generate a kriged map. Then we could ask whether if the resulting map is “good” enough for the purpose of finding an “object” of a certain spatial extension within the field representation (e.g., an eddy of warmer water with an extension of one square mile). Thus, this chapter investigates a model that provides the GIS user with the necessary tools to judge the quality of a stored map with respect to its ability to identify a certain object in a continuous field representation.

Terminology—From Scale and Resolution to Detectability: In general one can say that a representation stored within a GIS models the real world at some scale and resolution. This representation cannot be identical with the real world and thus introduces imperfections (e.g., inaccuracies). To avoid confusion over terminology we clarify the terms discretization and detectability.

The stated problem of deciding whether one is able to detect a specific object within a given field representation is dependent on the combined imperfections within the representation—one of which is the discretization imposed on the field representation. On the other hand we could also use the term “level of geographic detail”, which is discussed by Goodchild and Proctor (1997) as a possible augmentation of the term “scale” in the digital geographic world.

Since some of the terminology is used differently in different disciplines we do not want to reuse terms like resolution or scale for the model introduced in this dissertation. Thus, we introduce another term: detectability, which combines properties of the field (i.e., discretization) and properties of the object. Their distinct dependencies (e.g., sample size or object size) are explained in more detail in the following section.

4.2 Dependencies of Detectability

This section discusses the parameters that influence the outcome of the question where (within the field representation) one can identify objects. We refer to this as the dependencies of detectability. An intuitive approach to this question suggests that there are two main components influencing the results. On the one hand there is the field representation and on the other hand there is the object. Figure 4.1 presents a more detailed list of factors.

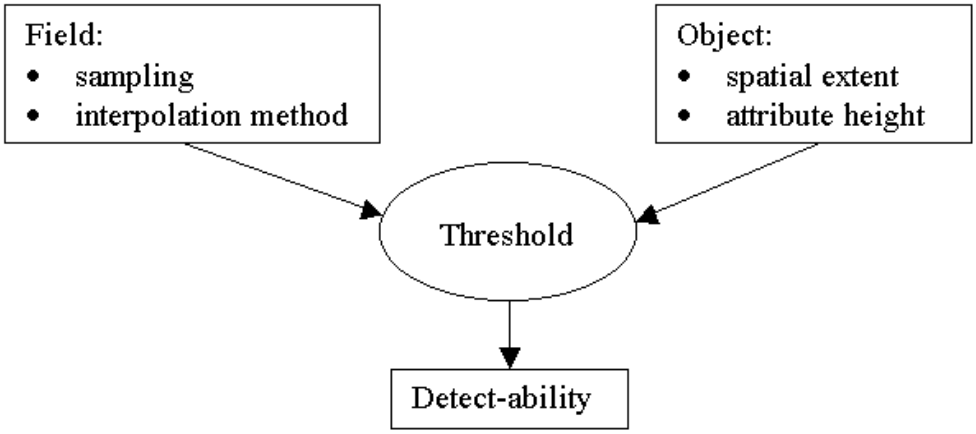


Figure 4.1—Dependencies of Detectability

First let us take a look at the parameters of the field (Peuquet et al. 1999) representation. We assume that we start with sampling the underlying variable. A sampling scheme involves two aspects: a) the number of sample points and b) their distribution. The next step in generating a continuous coverage is deciding which interpolation process (e.g., universal kriging) to choose. The final field representation will differ if any of these three components vary. Some of the will be more accurate than others (e.g., more sample points) and some of them will be smoother compared to others—depending on the interpolation method. Moreover, the accuracy of the representation determines the ability to detect an object or not.

Second, we consider some properties of the object itself. There are two components that are of interest when formulating its detectability within a field representation: a) the spatial extent of the object and b) the attribute height (or strength) of the object. Assuming that an object within a field exhibits a compact outline, its spatial extensions

can be given by a single value, namely by its area in square units. The attribute height of the object is in the same units as the field representation and is a relative comparison to its neighborhood. The influence on its detectability involves a

consideration of the following facts. For an object with a small spatial extent it will be more easily detected with a larger attribute height. On the other hand an object with a small attribute height will be more detectable with a large spatial extent.

A threshold forms the third dependency. The threshold determines the percentage of the object that has to be visible for its detection. A GIS user can specify this parameter up to a certain degree of freedom. The determination of object visibility and thus, detectability are discussed in the following section.

4.3 The Model—How to Determine Detectability

The method for generating a representation (e.g., sampling followed by kriging) introduces some constraints on the level of detail that one is able to provide within the GIS. In this section we discuss a model that results in a binary map that identifies areas where a certain object can be determined and where it cannot.

4.3.1 Approach

The model is based on the residuals calculated by subtracting an interpolated field representation from a ground truth. For an implementation we can substitute ground truth with any layer that we accept as being true. This could either be a comparable representation of higher accuracy (if available) or multiple (e.g., n = 100) realizations generated by conditional simulation (e.g., Gaussian Simulation). The size of the residuals can be seen as a result of a) the sample method and b) the model effects inherent in the interpolation method used to generate the field representation (e.g., kriging). The residuals represent an indicator of how well the representation matches reality—or one could say that this is the accuracy of the map. This is one way of interpreting these residuals. Here we are looking beyond the numeric information, to the spatial distribution of the values of the residuals. These residuals are used to determine the detectability of an object in a given representation.

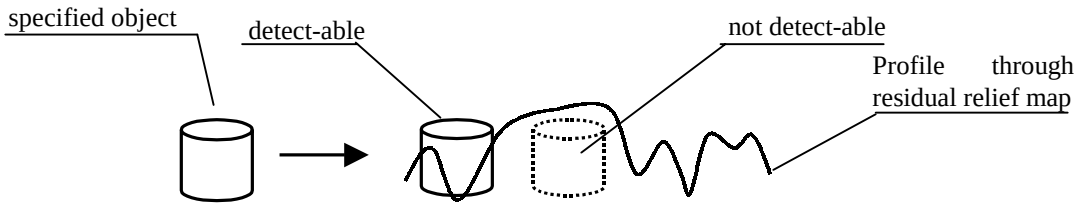


Figure 4.2—Object Representation and Relief Map of the Residuals

Within the field representation—at any given location—one cannot determine a feature occurring in the real world if the spatial and attribute dimensions are smaller than the residuals. Let us take this idea a step further: if we generate a three-dimensional representation of the residuals we obtain a relief map (similar to a DEM) of the minimal determinable variations. Looking at this relief map we can now specify a representative object and compare it to the outlines of the relief map (Figure 4.2). If the object is hidden by the relief map then we say it cannot be detected from the kriged map representation. On the other hand if the object is fully visible on the outside of the relief map than we would be able to determine the object within the kriged map.

Next we discuss the generation of a representative object for the comparison mentioned above. We suggest a representative object in the form of a cylinder. This is a result of the fact that a circle is the most compact form and that the height is a parallel movement of the object's outline. The radius of the circle is determined by the spatial extent of the object (e.g., we want to identify an object that has an area of π square units than the radius of the cylinder would equal 1 unit). The height of the cylinder represents the attribute value.

Finally we could combine the relief map with the cylinder. In order to determine the areas of possible detectability the cylinder is moved over the relief map. At each new location occupied by the cylinder (i.e., representation of the object) we now have to determine whether the top of the cylinder extends beyond the relief map (i.e., inherent inaccuracies/noise) or not. If the top of the cylinder is visible we can infer that an object located at this position is not overwhelmed by the inaccuracies and thus, is detect-able.

However, we can say that if the spatial extent of the cylinder is represented by, for example, 100 pixels it is still sufficient enough to see 99 pixels in order to detect the cylinder. Thus, the introduction of a threshold for the detectability allows a percentage (e.g., 5%) of the cylinder to be obscured by the relief map.

Figure 4.3 shows a schematic representation of the calculation of the detectability from the relief map and a representative object. The result is a binary map, where areas of positive detectability (i.e., the object can be detected) are marked white and areas of negative detectability (i.e., the object cannot be detected) are marked black. The areas refer to the center of the object. Thus, if parts of the object are within a black area, but it is centered within a white area, we would still be able to detect the object. Regarding the visualization of the resulting binary map it might be better to represent areas of positive detectability as green and areas of negative detectability as red. These color settings might improve the communication of the inherent imperfections to the GIS user.

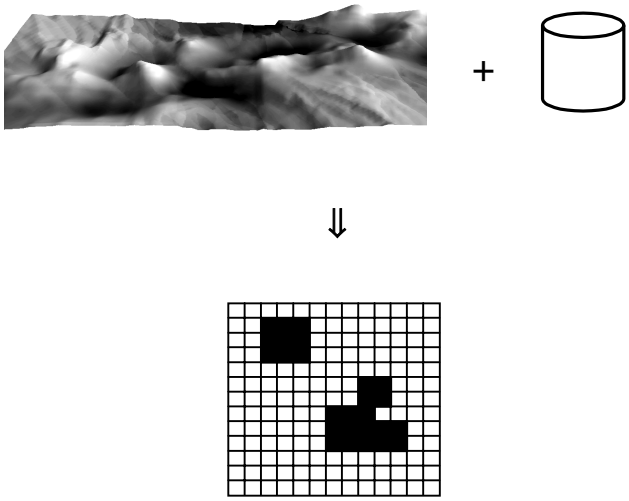


Figure 4.3—Schematic Representation of the Moving Object and the Resulting Binary Map

	Certain about statement	Uncertain about statement
Object is in observed area	white pixel 	black pixel 
Object is *not* in observed area	white pixel 	black pixel 

Table 4.1—Implied Inferences

The resulting binary map needs some more discussion in order to clarify the inferences we can make about the areas of positive (white) and negative (black) detectability. For the white areas we can say that whether an object is present or not, the field representation is “good” enough to state that we are certain about the represented facts (i.e., there is an object or not). For the black areas we have to state that the field representation does not allow us to make any inferences about the existence or non-existence of the defined object. Thus, all inferences made about objects within a black area introduce uncertainty in any derivations made from these field representations. This relationship is shown in Table 4.1.

4.3.2 Applications

In this section we take a closer look at some applications of the discussed approach. In general one can divide the applications into two major categories. On the one hand there are those applications where the whole area of interest is already sampled or where—in addition to sampling—the kriged map is already generated. Here the model would be able to tell the user if the quality of the representation is sufficient to derive conclusions with a desired certainty. The model could also be used to determine the appropriate sample size for a specific purpose (i.e., detecting objects of a certain size).

First, we would like to discuss examples where the whole study area has been sampled. Applications could include the identification of, for example, warm core rings (i.e., warmer water pools), which would lead to a different ecological system within a cold-water area. This phenomenon occurs in the Gulf of Maine when warm core rings get separated from the Gulf Stream. The sizes of these separations have to fulfill minimum requirements regarding their spatial extent in order to have an impact on the ecological system. The issue is to prove that the change in an ecological system was initialized by one of these pools. Thus, it is of interest to have the ability to say—with certainty—that there was no such object (i.e., pool) within a given field representation (i.e., map of sea surface temperature generated from sample points). Another application could be the detection of patches of high concentration of soil pollution in a rural area. This case introduces another interesting aspect, where operators of a chemical plant might have an interest to establish—with certainty—that there are no high concentrations of soil pollution in a specific sub-area. Thus, here we deal with a legal issue to prove that a map is fit for the specific purpose.

Second, a slight modification of the discussed model could be used to determine whether a proposed sample size is efficient for detecting a certain object prior to sampling the whole area of interest. Here the problem is more focused on the determination of whether the combination of the applied methods (i.e., sampling and interpolation method) will yield a sufficiently accurate field representation. The first step would require collecting sample points within a predefined sub-area, where objects do not necessarily have to be located. Then, at arbitrary locations within the sub-area, perturbations of the size of the given object are introduced. Finally an application of the suggested model to determine detectability would clarify if the applied methods (i.e., sampling and interpolation method) were sufficiently accurate. If there are any black areas in the resulting binary map, changes are necessary (e.g., increasing the sample size). This method requires the implementation of conditional simulations—as discussed earlier.

4.4 Case Study

In this case study we want to determine whether we can detect pools (with a radius of about 10km) of different water temperature (e.g., ±2°C and ±5°C). Here we discuss two different approaches. First, we generate the accuracy information by subtracting a kriged map (generated with an isotropic variogram model and punctual kriging) from the satellite image (i.e., ground truth). The second approach uses 50 conditional simulations (Gaussian) to generate the required accuracy information. In the latter approach we used the difference between the lowest/highest simulated attribute value and the kriged map for each location within the study area.

4.4.1 The Used Data

We use a satellite image showing the sea surface temperature (Figure 4.4) in the Gulf of Maine and a set of 231 sample points taken within the area represented in Figure 4.4. The sample points follow a regular distribution with a spacing of about 20km between them.

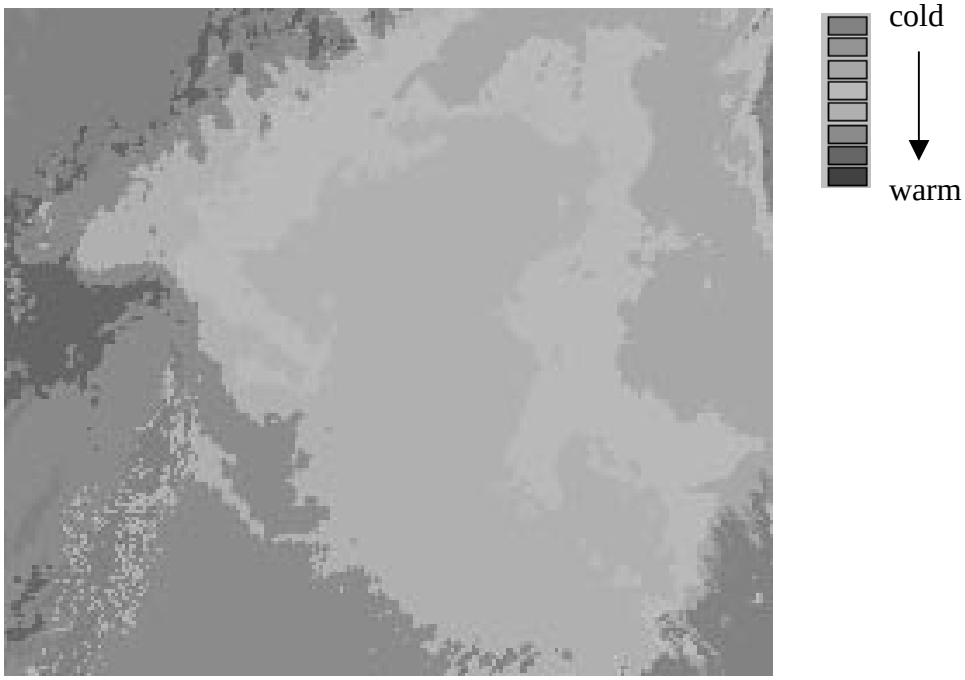


Figure 4.4—Satellite Image, Showing Sea Surface Temperature

In the satellite image (Figure 4.4) the northeast corner of the image is represented by the coolest temperature class due to the fact that it is actually land (coast of Maine) and not the sea surface. Furthermore, in the southeast region and in the southwest corner extensive cloud coverage is evident.

4.4.2 Results Using the Satellite Image

Processing the data:

- Removed the linear trend surface inherent in the sample data. This step is necessary since GS+ only supports ordinary kriging.

- Determination of the semi-variogram (used software: GS+).

- Execution of punctual kriging, which results in an interpolated continuous representation (used software: GS+).

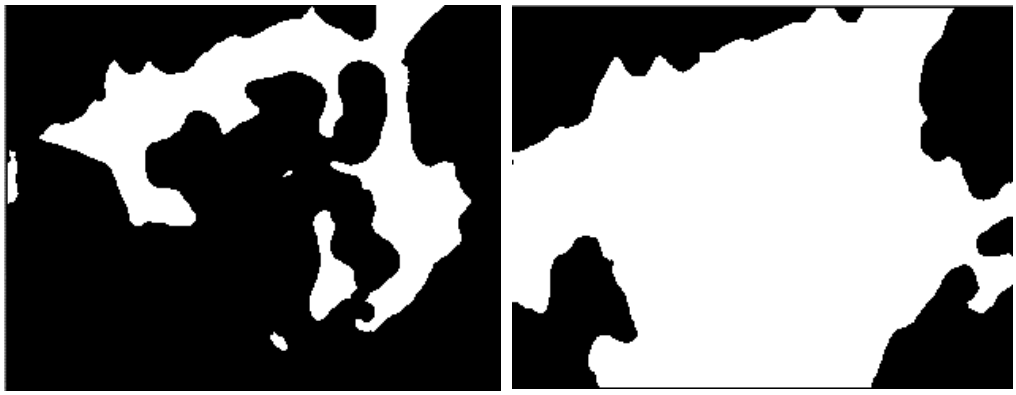
- Addition of the trend surface (using a short c++ program), which results in an interpolated continuous representation of the sea surface temperature in the surveyed area.

- Generation of the relief map of the residuals. This is accomplished by simply subtracting the interpolated surface from the satellite image (i.e., ground truth) (used software: ARC/INFO).

- Creation of a binary image of result. We use an AML in ARC/INFO to calculate the resulting binary maps—with the definition: `cylinder.aml <input grid> <output grid> <attribute height> <threshold> <radius>`. A defined cylinder is centered over each pixel within the relief map. At each location we can now calculate the number of pixels where the relief map exceeds the cylinder. A threshold

decides whether the center pixel results in a white (i.e., detect-able) or in a black (i.e., not detect-able) output pixel.

When applying the discussed model we investigate the detectability for two different objects. One of them with an attribute height of 2°C and the other one with an attribute height of 5°C, where the remaining dependencies (e.g., radius = 10km, threshold = 85%) of detectability are kept constant. The results can be seen in Figure 4.5a—for a 2°C object—and in Figure 4.5b—for a 5°C object.



a b

Figure 4.5—Resulting Binary Maps a: for the 2°C Object and b: for the 5°C Object

A comparison of the two results (shown in Figure 4.5) confirms the assumption that the areas where inferences about an object of an attribute height of 2°C can be made with certainty are clearly smaller than the areas where inferences about an object of an attribute height of 5°C can be made with certainty. These results lead to the following conclusions:

- If objects of 2°C attribute height need to be detected the representation and applied method (e.g., sample spacing) are not adequate.

- If objects of 5°C attribute height need to be detected the representation and the applied methods are sufficient.

4.4.3 Results Using Conditional Simulations

For the second approach we used the same data set of 231 sample points extracted from the satellite image to generate a continuous representation. Furthermore, we also used the same objects (of 2°C and 5°C, respectively). The remaining dependencies were also kept constant at a radius of 10km and a threshold of 85%. The difference for this approach however, is the generation of the accuracy information and the subsequent calculation of the binary results.

To obtain the accuracy information required for this approach we generated 50 Gaussian simulations (conditional upon data) of the given area (software used: gstat). In order to increase the quality of the results we added 100 additional sample points. The addition of 100 sample points might seem a lot at this point. However, within the scope of the dissertation the goal is to prove the general approach discussed in this chapter. The resulting maps where then compared to the kriged map (using 231 sample points) consequentially generating 50 binary images showing the given detectability. In a final step all binary images were added. The pixels were assigned a “black” value if one or more of the 50 generations indicated a “black” value. On the other hand they were assigned a “white” value only if all 50 generations resulted in a “white” value. If we would generate 100 simulations we would allow 2 generations to show a black pixel and still assign a white pixel to our final result. This approach is taken to gain independence of the number of generated simulation.

Figure 4.6 shows a few simulation results. The area shown in Figure 4.6 is about one fourth of the complete study area representing the northeastern region (compare with Figure 4.4). The simulation results were viewed in a different program (since they were in a different file format)—with the effect that different color schemes and scaling is applied. Here the strong edge effects where no data points were available are apparent (top in Figure 4.6).

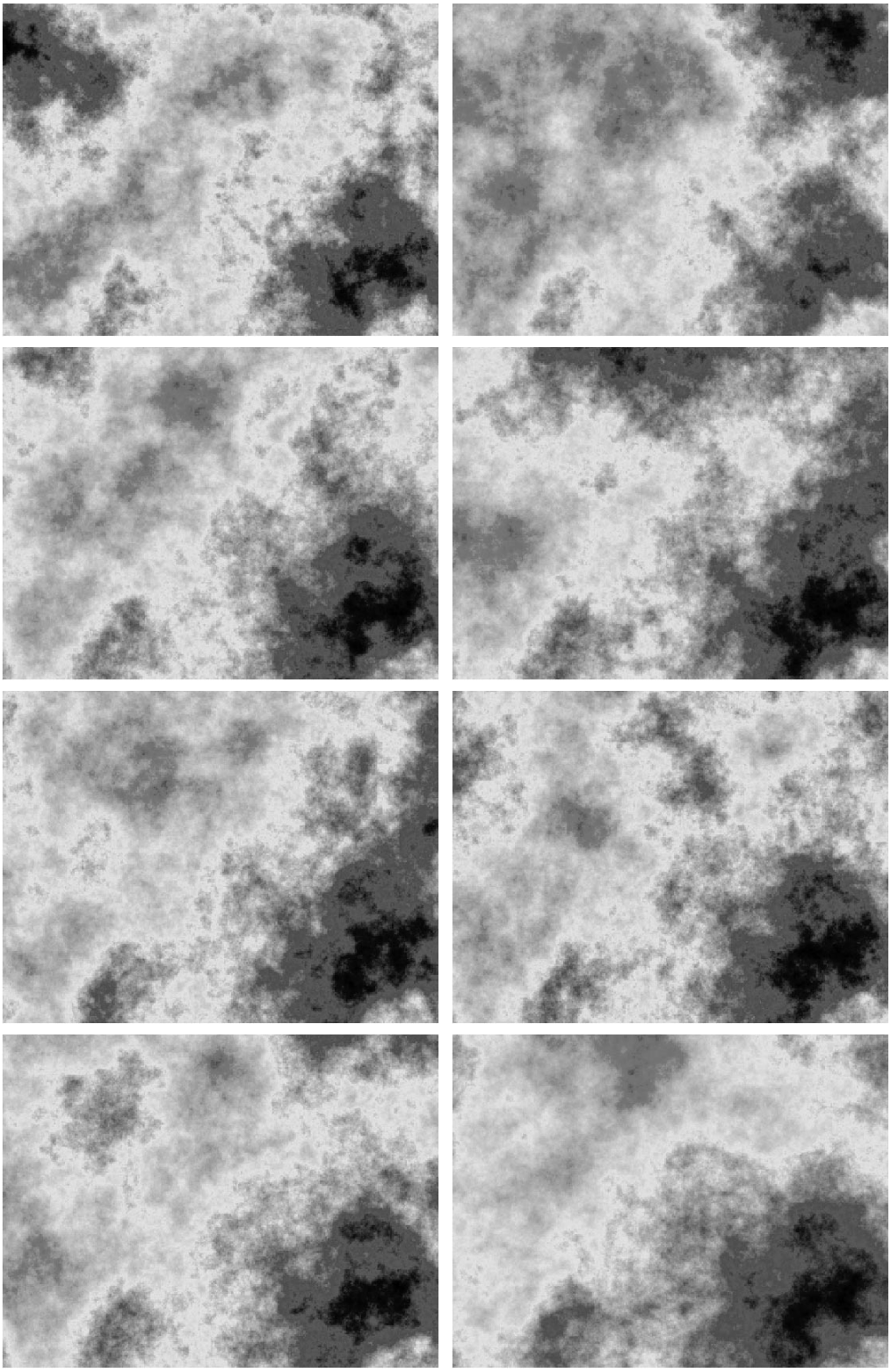
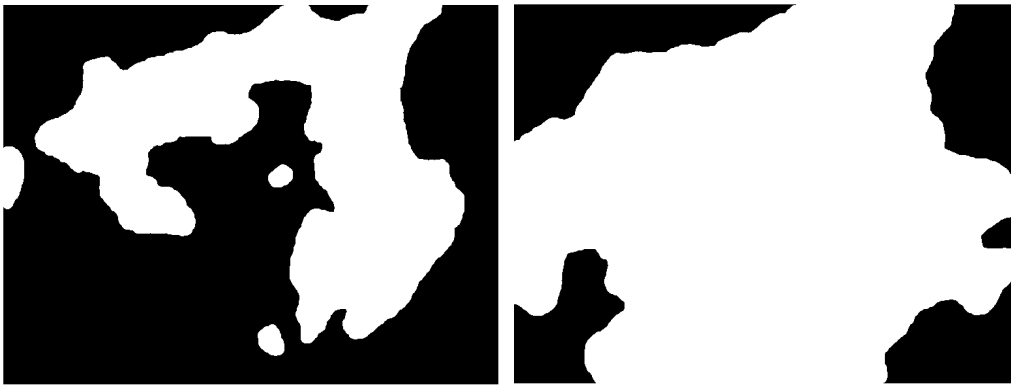


Figure 4.6—Simulation Results

From the results shown in Figure 4.7a—for the 2°C object—and in Figure 4.7b—for the 5°C object one can see that similar conclusions can be derived. However, as expected one can see an obvious smoothing effect when compared to the previous results.



a **b**

Figure 4.7—Resulting Binary Maps a: for the 2°C Object and b: for the 5°C Object

4.5 Remarks

The detectability metric enables the GIS user to determine whether the quality of a given field representation is sufficient to detect a representative object. The result is presented via a binary raster representation in which the user can identify areas of positive and negative detectability. The user has to provide the spatial extent and the attribute height of the object. Furthermore, if required, the user should have the ability to vary—up to a certain degree (e.g., 0% to 20%)—the threshold for the determination of detectability.

In our case study we used two different approaches. First, a satellite image was used as a reference (i.e., ground truth) to calculate the necessary residuals for the relief map. Subsequently we focused on including the model of conditional simulations to gain independence of a ground truth reference. In our approach we included additional sample points for the simulations to increase the quality of our results.

Another promising research area using this model is the investigation of the influence on the binary result map of varying the dependencies of detectability. For example, we could reduce or increase the number of sample points and then analyze the relation between the number of sample points and the area of positive detectability. It would also be of interest to investigate variations in representative object shapes. Here we would like to look into the outcomes of replacing the cylinder by less compact shapes such as a line.

This chapter investigates a simple approach to communicate aspects of inaccuracy and discretization effects of a field representation to the GIS user. Future work will show aspects of an exploration of the effects of the dependencies of detectability on the results.

Chapter 5

The Effect of Discretization on Reliability

The objective of this chapter is to develop an explicit metric for the loss of information due to discretization. The goal of the metric is to capture the discretization effect in the spatial dimensions and the propagation of the discretization effect through various operations such as overlay. We propose this metric as a spatially random field that provides an estimate of reliability at any given location.

For the remainder of this chapter we define reliability of a representation as the users expectation of the level of fitness of a representation for a specific purpose. This corresponds to a definition of data quality (Chrisman 1983) but we use it here with respect to a specific metric and in this case as a specific metric for reliability with respect to discretization. Units of reliability are measured in percent—where 100% indicates that the reliability is perfect and the data is fit for use within a given scenario. A more detailed explanation of the calculation of the percentage follows in the case study.

5.1 General Considerations

The result of the proposed metric is geared towards an indicator that establishes a relation between the level of discretization and the requirements of a specific application. For example, given a DEM (Digital Elevation Model) with discretization

in the spatial domain of 90m, for most areas, the level of discretization would be sufficient for planning a logging road. Conversely, a more slope sensitive application such as planning a railroad track would be more problematic, unless the planning area is located in flat terrain where the variation of the underlying variable (i.e., height) between two measurement points is relatively small. At this point we would also like to note that the accuracy, lets say ±2m or ±7m, of the measurement points is rather insignificant in comparison to the influence of the spatial discretization. The spatial variation, described in more detail in the next section is part of an interaction between discretization and model specific requirements in our approach.

The reliability due to discretization is a function of the size (C) of the discretization unit and a measure of variation (θ).

$$R_c = f(C, \theta)$$

In general, we assume that the following statement holds true (in regard to one specific purpose): wide spacing of sample points in an area of low variation is equally reliable to short spacing of sample points in an area of high variation. For our example this implies that there might be areas of low variation (flat terrain) where a 90m DEM is sufficient to plan a railroad track. However, there might also be areas of high variation (mountain peaks) where a 90m DEM is insufficient to plan a logging road.

The variation is unknown and must be estimated. We compute this on a spatial varying estimate for local neighborhoods. The definition of variation is dependent on whether the underlying variable is continuous or discrete. For a continuous variable we assume a field of sample locations and define the variation as the standard deviation of the slope at an arbitrary location, where for the calculation of the slope the closest 6

neighbors are taken into account (see discussion on number of neighbors below—5.3). For a discrete variable we propose to use the diversity index as an indication for the underlying variation in the variable (O'Neil et al. 1988). The diversity index is based on entropy (Shannon and Weaver 1962) and is used as a controlling factor in Csillag et al. (1992) to determine an optimal resolution.

When talking about reliability, we refer to a specific application that a given dataset or representation is used for. The question arises of how to incorporate the specification of an application into a reliability metric. Our approach is to incorporate the maximum allowed variation for a unit of spatial extent. The unit could either be of one, two, or possibly three dimensions. An example for the one-dimensional case would be the maximum slope of a railroad track at a given distance unit. A two dimensional scenario would be precipitation within a given area unit. Within the last scenario we also have the possibility of an extension to a third dimension by focusing on accumulation of precipitation over a given period of time in an area unit (which would also introduce an effect of a temporal discretization dimension).

5.2 Approach

This section gives a detailed derivation of the proposed reliability measure for the two dimensional case. The following is based on the assumption of a continuous field and estimation of variation as a standard deviation of the slope. The required inputs are:

- Sample points (i.e., the locations and attribute measurements—e.g., weather stations measuring precipitation)

- Object size (its spatial extent—e.g., km^2 of mountain slope)

- Acceptable error in the object value (dependent on object size) (e.g., liters of water)

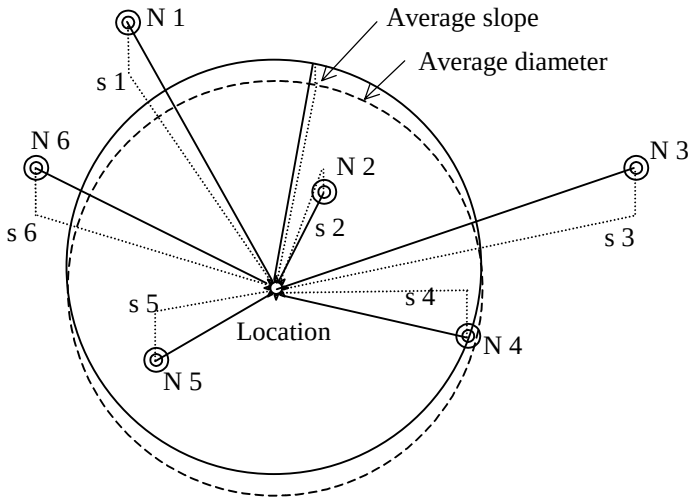


Figure 5.1—Average Slope at a Location Based on its Six Nearest Neighbors (N1-N6).

First we discuss the possible error of an object at a given location based on the variation of the underlying phenomenon. In order to calculate the variation we calculate the error in the average slope of the attribute value for a local neighborhood. As shown in Figure 5.1 and Eq.(1), we define the average slope at a location as the mean of the slopes between the location and its six nearest neighbors. As one can see, in Eq.(1) we also assign the weight of 1/distance to each of the slopes. Given the average slope one can now calculate its error m_M with Eq.(2).

$$\text{average slope} = \frac{\sum_{n=1}^{6}\left(s_n \cdot \frac{1}{d_n}\right)}{6 \cdot \sum_{n=1}^{6}\left(\frac{1}{d_n}\right)} \qquad (1)$$

with s_n … slope between location and neighbour n

d_n … distance between location and neighbour n

For the sake of simplicity Eq.(2) is based on the assumption that the value in each of the points (i.e., the location and its six neighbors) is given accurately. Consequently, this allows us to focus on the error introduced by the variation in the variable and neglects the influence of measurement errors in the value completely. Furthermore, we state that the resulting standard deviation (i.e., m_M) can be used as a measure of variation, or more exactly as a measure for the mean variation of a circular area with the average distance as the radius.

$$m_M = \sqrt{\frac{\sum_{n=1}^{6}\left[\left(average\ slope - s_n\right)^2 \cdot \frac{1}{d_n}\right]}{\sum_{n=1}^{6}\left(\frac{1}{d_n}\right) \cdot 5}} \qquad (2)$$

Next, we use this measure of variation to calculate an accuracy value of an object (Windholz et al. 2001) that has the dimensions of a cylinder, where the base is a circle of radius r = average distance and the height is the attribute value of the sampled variable. However, we do not interpolate the actual value of the underlying attribute for

the whole circular base. Instead we take the slope of the variable and derive the mean attribute value within the circular base (Figure 5.2 and Eq.(3)). Note that we are only interested in the relative attribute value and not in an absolute one.

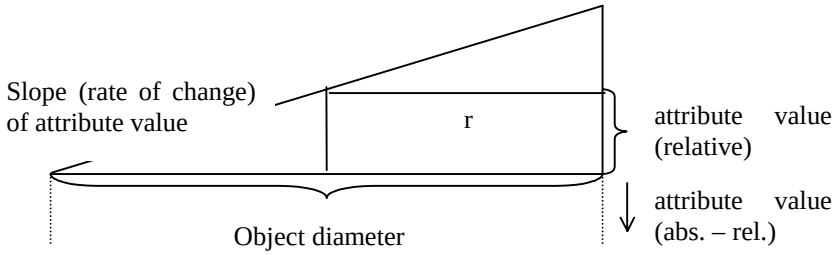


Figure 5.2—Calculation of the Mean Attribute Value for a Given Circular Area.

$$object\ value = r^2 \cdot \pi \cdot (mean\ slope \cdot r) \qquad \text{Eq.(3)}$$

The resulting error is now based on the propagation of the mean slope error m_M. Eq.(4) shows the calculation of the mean error of the object value. Or in other words the accuracy of the object value (e.g., amount of water in an area, where the attribute values in the sample points are liter per square unit) based on the underlying variation.

$$\frac{\delta(object\ value)}{\delta((mean\ slope)} = r^3 \cdot \pi \quad \Rightarrow \quad m_{ov} = \sqrt{\left(r^3 \cdot \pi \cdot m_M\right)^2} \quad \Rightarrow$$

$$m_{ov} = r^3 \cdot \pi \cdot m_M \qquad \text{Eq.(4)}$$

Next we are interested in the rate of change of the variation—or in other words in the dependency of the mean error of the slope (i.e., m_M) over distance—based on the distance from the location. Figure 5.3 illustrates a possible way of interpreting the rate of change at an arbitrary distance from the location. The model used in Figure 5.3 is a linear model and assumes that the error in the variation increases with distance from the location (with zero at the location and a maximum at the average distance from Figure 5.1). The rate of change m_{SM} of the variation can now be calculated by Eq.(5).

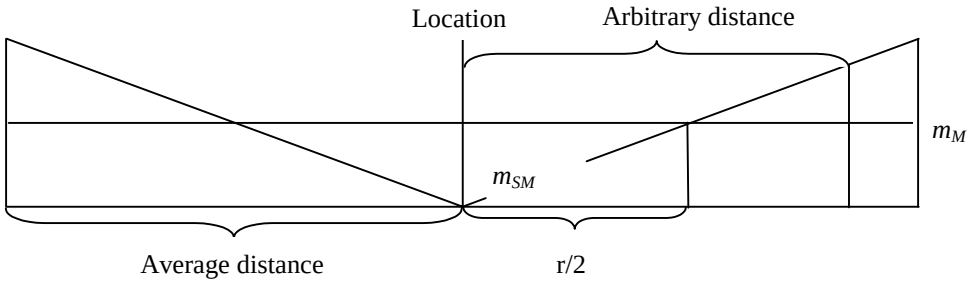


Figure 5.3—mSM the Rate of Change of Variation (i.e., mM).

$$m_{SM} = \frac{2}{r} m_M \qquad \text{Eq.(5)}$$

The rate of change m_{SM} allows us to calculate the variation at any arbitrary distance from a sample point and is an essential issue in our approach. Thus, the first step of our model requires the calculation of the rate of change in each sample location based on the six nearest neighboring sample points.

At this point we can estimate the variation (i.e., slope error—m_{SL}) at any given location. As shown in Eq.(6), we do this by taking the variations of the six nearest sample points into account using m_{SM}.

$$m_{SL} = \sqrt{\sum_{n=1}^{6}\left(\frac{m_{SM_n} \cdot d_n}{6}\right)^2} \qquad \text{Eq.(6)}$$

The next logical step would be to take the double integral over a continuous layer representing m_{SL} with the bounds of any desired object area to calculate the error in the object volume. However, since we only have finite computing systems we have to rely on Riemann sums. Thus, we generate a sufficiently fine raster where the unit size of a pixel in the raster serves as an increment in the Riemann sum. After estimating a m_{SL} value for each pixel within the raster based on Eq.(6) one can also calculate the unit error of volume for each pixel based on Eq.(7)—where we substitute the squared pixel area with a circle of equal area.

$$unit\ error = \frac{(pixel\ width)^3 \cdot m_{SL}}{\sqrt{\pi}} \qquad \text{Eq.(7)}$$

The next step is to sum the unit errors of all pixels congruent with any given object resulting in the error of this object's volume. The information about the error of the object volume can now be associated with the object's center (pixel). Subsequently, we can compare the given value with the acceptable error in the object volume—initially defined by the user. Finally, based on this comparison we can assign a reliability value to the center pixel. In order to get a continuous representation of reliability we move the object center to each location of our generated raster.

The resulting representation gives information on the error of the object volume based on the object size and more importantly based on the variation of the underlying variable at any given location.

5.3 Case Study

In the case study we investigate plankton concentrations (objects) within the Gulf of Maine. The first part of the input consists of 230 measured sample points that are evenly spaced within the area of interest (315km by 240km). For the second part of the input, we define an object size (area of plankton concentration) and an acceptable error in the object value in order to calculate the reliability of the sample point distribution and density under consideration of the inherent variations. The results are percentages of reliability indicating that the estimated average plankton concentration within the defined area is within the required error margins. For the percentage calculations we assume that if the calculated error is greater or equal to the user specified margins, the reliability is set to "0%". On the other hand if the calculated error is less than the user specified margins the reliability is calculated as a percentage based on those two values.

At this point we would like to show that the chosen number of 6 neighbors could be justified by the following Figure 5.4a-g. These figures show the resulting reliability with 2 to 8 neighboring sample points. As we can see there is a drastic change from two to five neighbors (Figure 5.4a-d). Followed by relatively stable reliability results for five to seven neighbors (Figure 5.4d-f). The inclusion of more than 7 neighbors on the

other hand dilutes the picture and over-smoothens any reliability in the given data set (Figure 5.4f-g). We would also like to point out that the inclusion of 6 neighbors is fitting for the given data set. Additional investigations might have to be conducted for different data sets. Moreover, one should keep in mind that the metric of reliability is an estimate. For the following investigations in the subsequent chapter the number of neighbors is irrelevant as long as it is kept constant for the entire case study.

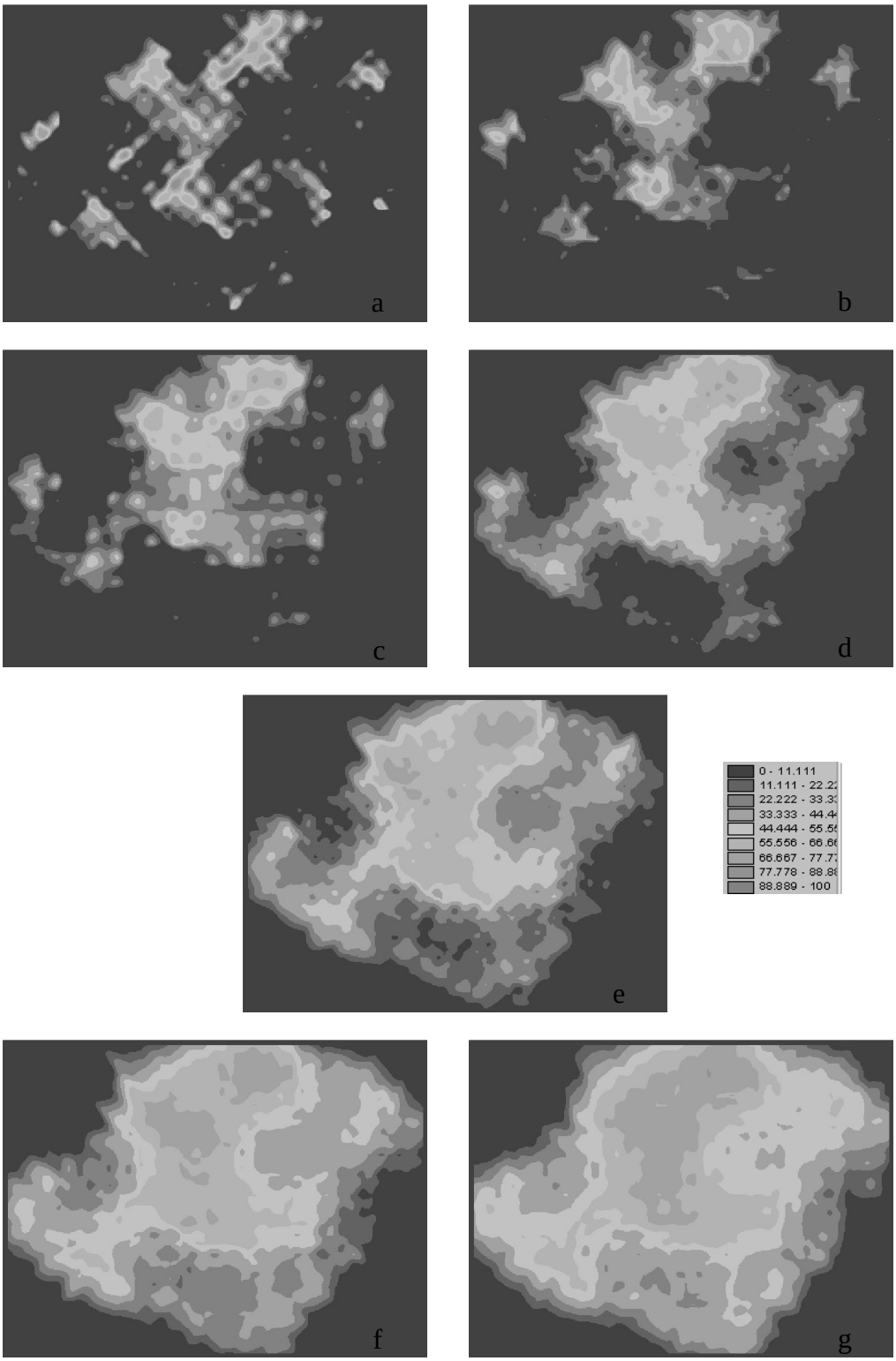


Figure 5.4a-g—Estimated Reliability with 2-8 Neighbors—the Legend Indicates Percentages of Reliability

Figure 5.5 - Figure 5.12 show the results of the reliability calculation for different scenarios described within the following discussion.

For Figure 5.5 - Figure 5.7 the object area is set to about 55km^2 whereas for Figure 5.8 - Figure 5.12 the object area is set to about 165km^2, as the larger “no data” margins (in black) at the edges of the presentations indicate. In Figure 5.5 - Figure 5.7 we increased the acceptable error margins.

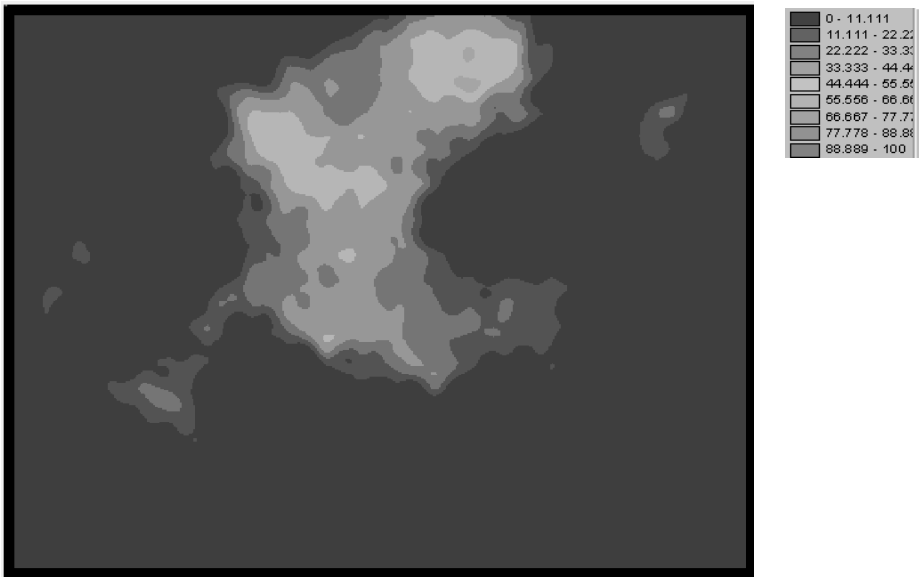


Figure 5.5—Object: 55km^2, Object Error 1/2 of Figure 5.6 and 1/3 of Figure 5.7—the Legend Indicates Percentages of Reliability

Figure 5.6 has twice the acceptable error margin of Figure 5.5 and Figure 5.7 has three times the error margin of Figure 5.5. As expected the areas of higher reliability increase from Figure 5.5 to Figure 5.7.

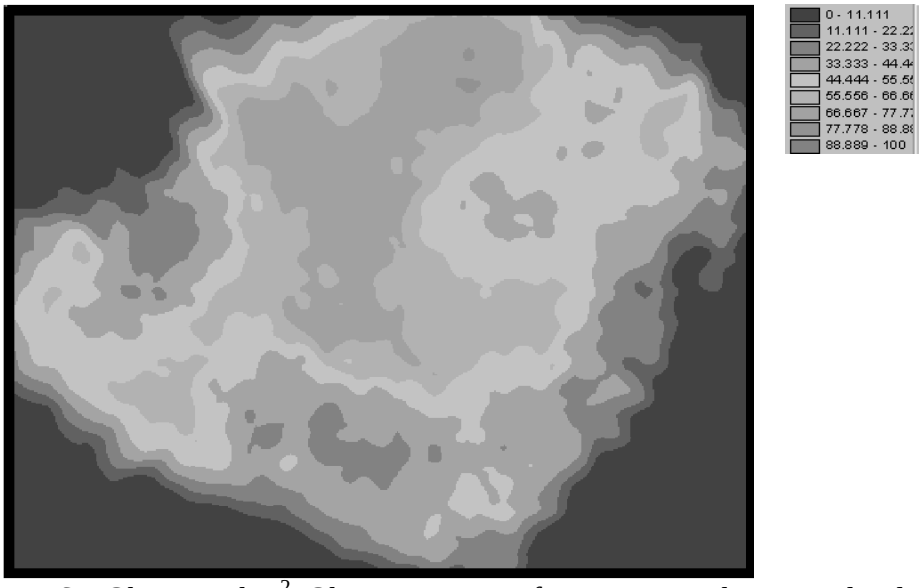


Figure 5.6—Object: 55km^2, Object Error 2x of Figure 5.5—the Legend Indicates Percentages of Reliability

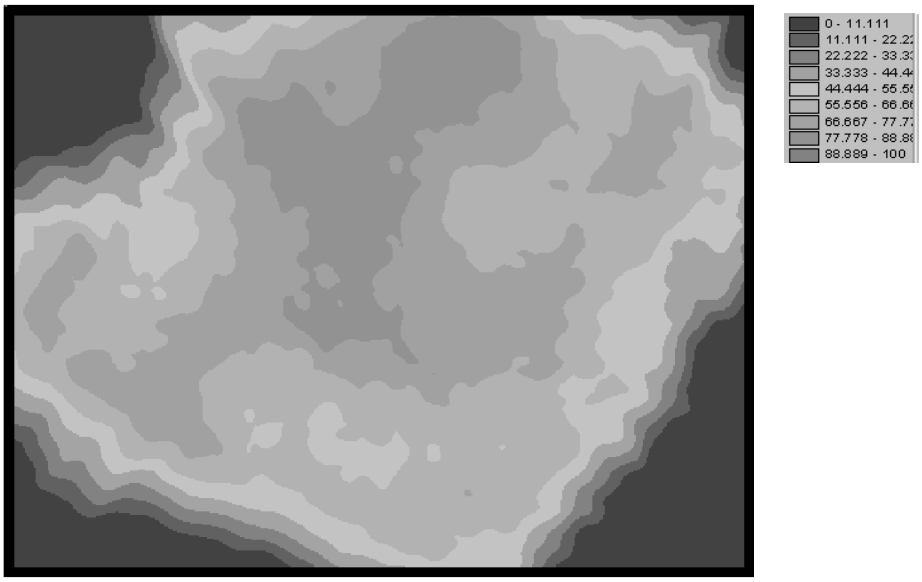


Figure 5.7—Object: 55km^2, Object Error 3x of Figure 5.5—the Legend Indicates Percentages of Reliability

In the following step (Figure 5.8) the error margin remains constant and the area of the object is increased (as mentioned above) by a multiplication of 3. The results show clearly that due to the underlying variation, the error in calculating the accumulated plankton concentration increased and consequently, the reliability decreased.

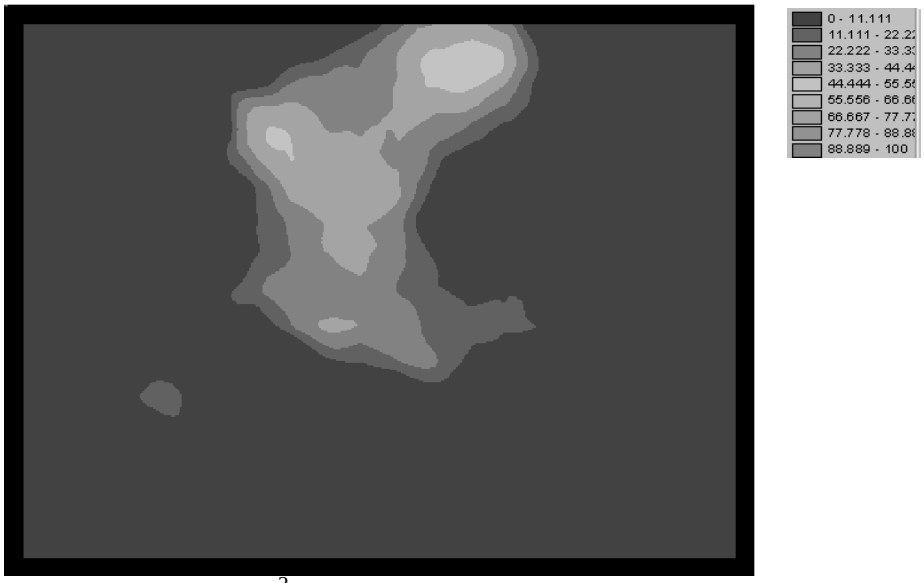


Figure 5.8—Object: 165km^2, Error the Same as in Figure 5.7—the Legend Indicates Percentages of Reliability

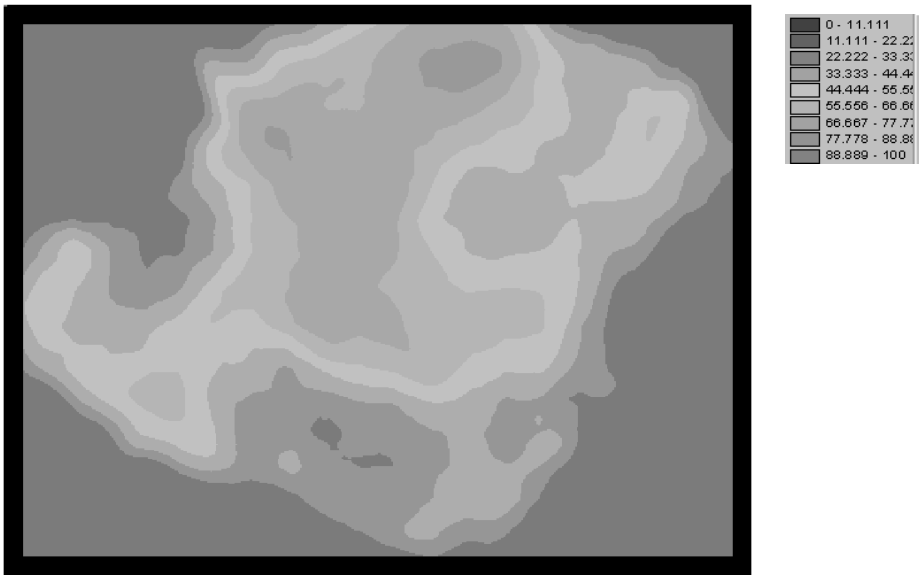


Figure 5.9—Object: 165km^2 with Increased Error Margins—the Legend Indicates Percentages of Reliability

Next we increased the acceptable error (Figure 5.9 - Figure 5.10) to a point where most of the area of interest shows a “good” reliability.

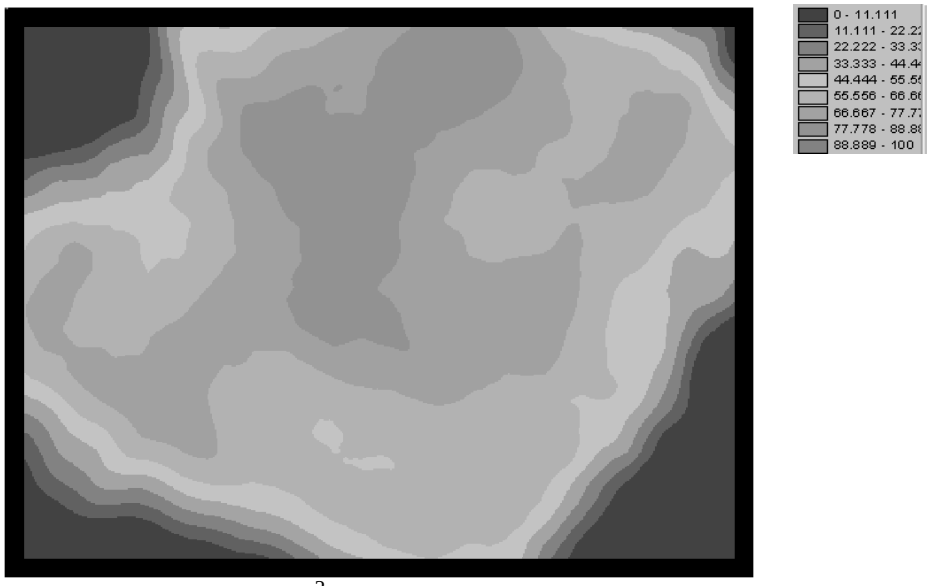


Figure 5.10—Object: 165km^2 with Further Increased Error Margins—the Legend Indicates Percentages of Reliability

In the final two runs we hold the object size and its acceptable error constant and perturb the underlying sample points. For the presentation shown in Figure 5.11 we randomly eliminated one third of the sample points, which results in a lower reliability overall—especially in those areas where sample points were removed.

For the presentation shown in Figure 5.12 we multiplied the measured attribute value (of the complete sample set) by a factor of ten. The reliability distribution is identical to the presentation shown in Figure 5.10—with one essential exception: the acceptable error limits were multiplied by ten (in comparison to Figure 5.10).

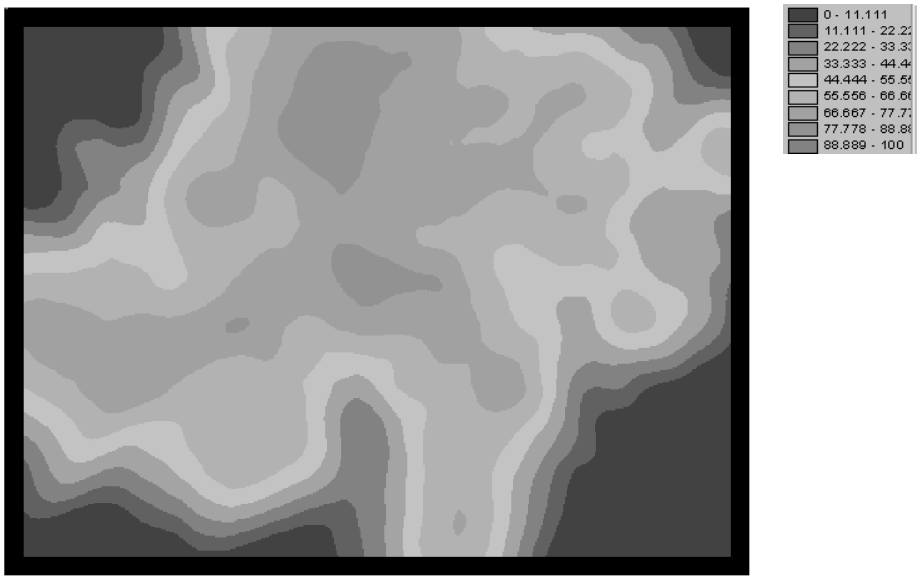


Figure 5.11—Object 165km^2 with 1/3 Less Sample Points Compared to Figure 5.10—the Legend Indicates Percentages of Reliability

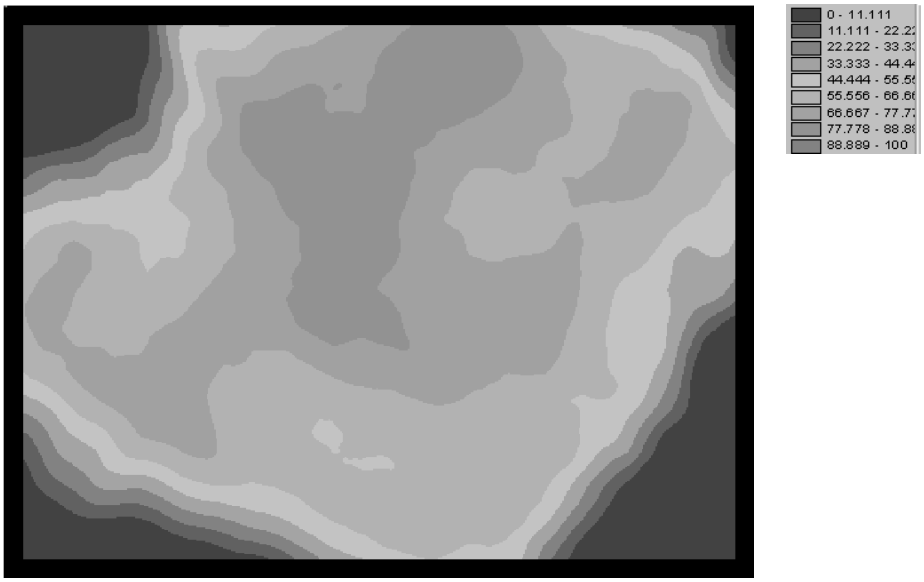


Figure 5.12—Object 165km^2, 10x Attribute Values in Sample Points & 10x Error Margins Compared to Figure 5.10—the Legend Indicates Percentages of Reliability

As we can see from a comparison between Figure 5.10 and Figure 5.12 the variation of a sample field and the acceptable error of the specified object are in a direct relation. If the amplitude of the attribute field increases by a factor *x* the acceptable

error of the object has also to be multiplied by the same factor *x* to achieve the same reliability. In the following chapter the discussion focuses on the more complex interaction between the number of sample points and the variation.

5.4 Remarks

The calculated reliability values are a helpful indicator for the usefulness of a dataset for a specific purpose. The metric isolates the effect of the discretization and presents it as a spatial variable. This is a useful extension to the reliability due to measurement error. The approach can be applied on discrete as well as continuous data. Also the approach should be generalizable to the temporal and thematic domains to account for effects of temporal or thematic discretization. Further research will investigate the propagation of the discretization effect through various GIS operations.

The subsequent chapter discusses a comparison of the dependencies of reliability in more detail.

Chapter 6

A Comparison of the Dependencies of Reliability

In this chapter the discussion is focused on the application of the previously developed metric of reliability. Perturbations are introduced to the given dataset to investigate the effects on the resulting reliability. The goal is to prove the hypothesis that we can make inferences about the influence of the amplitude of the variation and the influence of the sample density in respect to the reliability of the representation. The following discussion is based on an empirical approach. In general, a more thorough mathematical approach is preferable, which is, however, beyond the scope of this dissertation. Nevertheless, to ensure the validity of the empirical approach taken in this case study we used three different datasets. Two of them (sea surface temperature and plankton concentration) were highly correlated and thus, not presented separately in this chapter. Conversely, we were able to achieve more satisfying results (regarding the ability to state a generalized relation) with a third dataset (height measurements) that is included in the case study below.

6.1 Approach

To answer a specific question using a GIS (or to be more precise one or more representations) with a desired certainty requires that the data meet a certain level of reliability. The reliability of a map depends on several different circumstances. If it would be possible to copy the real world at a scale of 1:1 and sample at indefinitely

small intervals the accuracy of the map would approach 100% and consequently, also the reliability. Since this procedure is not feasible the sample density as well as its distribution influences the accuracy and the reliability of a representation.

The sample design, however, is not the sole factor that determines the reliability of a representation. As indicated in the previous chapter the variation of the attribute variable within a region of interest is also an important contributor to the reliability of a representation. For example, when measuring the attribute value ‘elevation’ the terrain dictates the number of sample points needed to describe the variation satisfactorily. Intuitively, one can assume that in a flat terrain (e.g., Salt Lake) less sample points are required than in a mountainous region (e.g., Rocky Mountains) to capture the variation of the elevation (i.e., height differences of neighboring areas). Thus, we can state that the variation of the attribute value under consideration might also be as important as the sample distribution.

Given the two factors, the question of interest is if we can make any assumptions of the amount each of the factors contributes to the reliability of a representation. Another interesting approach is to determine whether there is a relationship between the two factors. Within the scope of this dissertation we develop an experimental configuration to empirically demonstrate that the two factors are comparable.

The approach is based on the following scenario: If we decrease the number of sample points by a certain percentage—by what percentage do we have to decrease the amount of variation of the sample field to regain at least the same reliability as before the sample reduction?

Prior to the investigation the reliability of a set of sample points is calculated (as presented in the previous chapter). The first step of the approach requires reducing the number of sample points. This should be done based on a stratified selection

In the second step the reliability of the reduced sample set is calculated. The spatially varying reliability shows areas of lower reliability where the missing sample points were located. The decrease of reliability is dependent on the interaction of the local sample distribution and the local variation. Thus, some areas (e.g., higher local variation or a locally greater distance between sample locations) might be more affected by the exclusion of samples then others.

The third step is an iterative process beginning with decreasing the amplitude of the attribute variable by multiplying the measurements by a given factor. For example, we can say that multiplying the measurements by a factor of 0.9 decreases the amplitude of the attribute variable by 10% and thus, the variation of the attribute field. The next step in the iteration is the recalculation of the reliability of the representation. Subsequently, we can compare this newly computed reliability to the reliability of the original data set. Since the decrease of reliability is not the same for the whole study region (as mentioned above) the third step evolves around the constraint that all areas (using a pixel by pixel reliability comparison) must show at least the same reliability when compared to the full sample set to state that the reliabilities are equal. The iteration is now repeated until the correct factor for decreasing the amplitude is chosen.

The results of the approach allow a comparison of the influence of the sample point distribution to the influence of the amplitude of the attribute variable. These results also allow users to estimate changes to a sample point distribution if changes in the variation within the field are observed.

6.2 Case Study

This case study uses the metric of reliability (discussed in chapter 5) and follows the above-mentioned approach. For this case study we used two different datasets—the same dataset as in chapter 5 and a DEM (Digital Elevation Model). The accuracy requirements for the object are set constant for the subsequent case study. For the initial representations of reliability in both data sets the accuracy value was set to generate a

minimum reliability of at least 33% within the entire study region. Setting the starting values at a responsive level (as compared to setting the entire study region to 99 or 0 percent) allows us to enhance making changes of reliability (caused by perturbations in the samples) evident.

First, we use the data set from chapter 5 showing plankton concentrations in the Gulf of Maine. Figure 6.1 shows the reliability of the complete dataset of 230 sample points. The chosen object size is 165km^2.

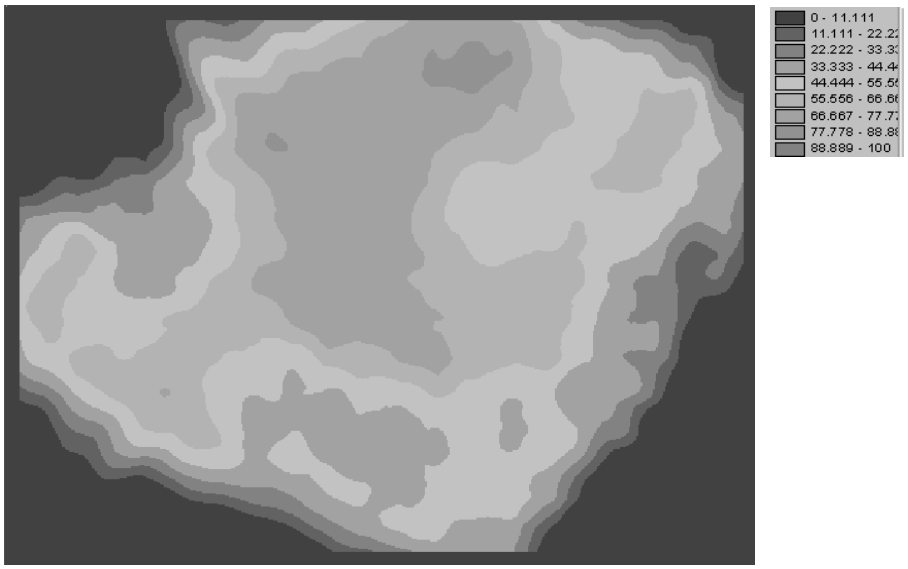


Figure 6.1—Full Set of Sample Points (230)—the Legend Indicates Percentages of Reliability

In comparison Figure 6.2 shows the estimated reliability of a subset of sample points. For the estimation of the reliability depicted in Figure 6.2 the original sample population was decreased by about 33% to 2/3 of the original population resulting in a total of 153 sample points. The methodology used to eliminate the 77 sample sites followed the previously discussed procedure. One can see that in some areas—especially in the south of the study region—the reliability decreased a few class intervals (i.e., ~ 40% of reliability). Also notable is the fact that some areas maintained a relatively high percentage of reliability. When comparing these results to the original

dataset one can see that the areas of lower variation are the areas where the reliability is relatively stable and vice versa where the variations are high (southern part of study region) the reliability is decreasing faster when excluding sample locations.

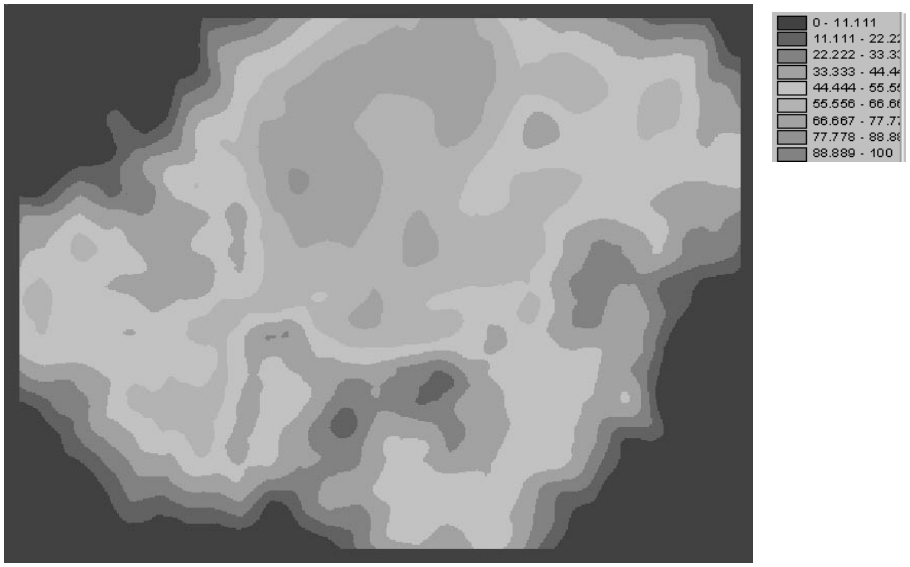


Figure 6.2—Subset of 153 Sample Points—the Legend Indicates Percentages of Reliability

Next, the variation is decreased incrementally. Figure 6.3 depicts the reliability estimates for the subset of 153 sample points with a decreased variation. The amplitude of the attribute field is reduced by a multiplication factor of 0.85—or 15% compared to the original. As one can see there are still some areas in the south of the region where the reliability is not as ‘good’ as in Figure 6.1. According to the constraints discussed in the approach a further decrease of the amplitude of the attribute field is required.

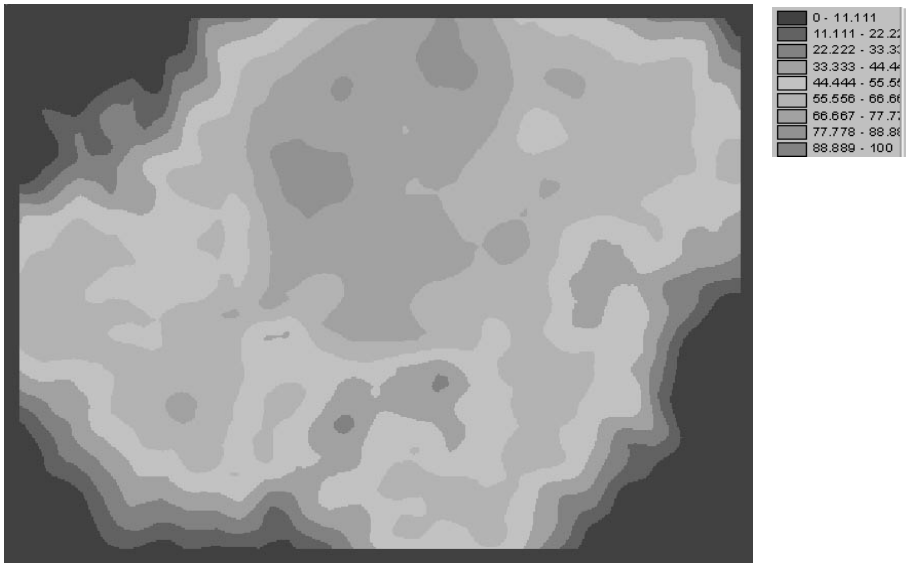


Figure 6.3—Subset of 153 Sample Points, Decreased Variation (-15%)—the Legend Indicates Percentages of Reliability

Figure 6.4 shows the final run of the iterative process. It is evident that the reliability is now at least as high as in our original dataset of 230 sample points in all areas of the study region. For Figure 6.4 the amplitude of the attribute field is set to 0.8 times the value of the original. Thus, showing a decrease of variation of –20%.

The results of this case study show that a decrease of 30% in sample points require a decrease of 20% of the variation in the attribute field to achieve similar reliability values. However, the results can also be interpreted the other way around. Namely, that for a decrease of 20% in the amplitude, or variation the number of sample points can be decreased by 30% without having a negative influence on the reliability of the representation of a continuous variable. On the other hand the results can also be interpreted as follows: Given a scenario where the amplitude of the attribute variable increases by about 20% a supplementary 30% of sample sites have to added to attain the same reliability values.

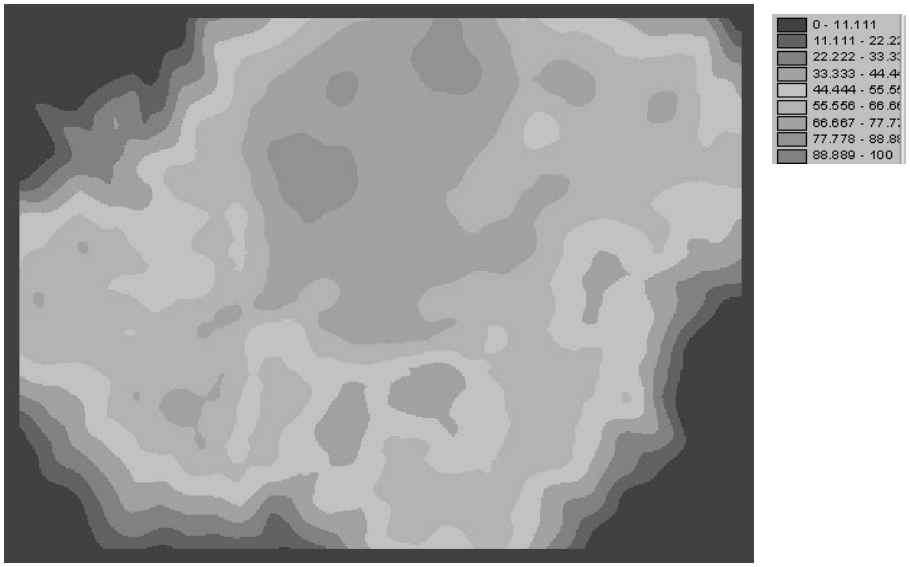


Figure 6.4—Subset of 153 Sample Points, Decreased Variation (-20%)—the Legend Indicates Percentages of Reliability

The same approach was followed using a second data set of height measurements (Figure 6.5). Figure 6.6 shows the reliability for a sample set of 255 sample points and a specific sub-region size, which is kept constant for the entire approach. The 255 sample points were taken randomly from a 30m by 30m DEM representation. Similar, to the investigation of the previous dataset, the parameters (sub-region size and acceptable inaccuracy) were chosen to yield a minimum reliability of 33% throughout the entire study region. This approach allows us to start with a representation of reliability that is susceptible to perturbations.

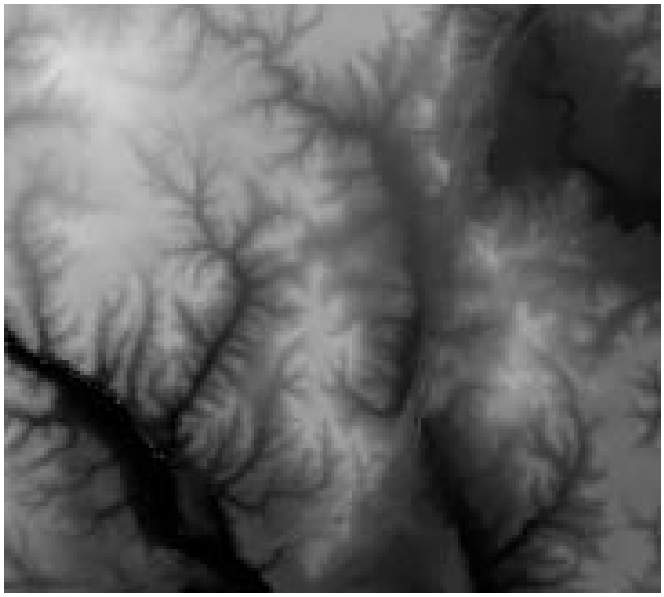


Figure 6.5—Presentation of the DEM Used in this Case Study

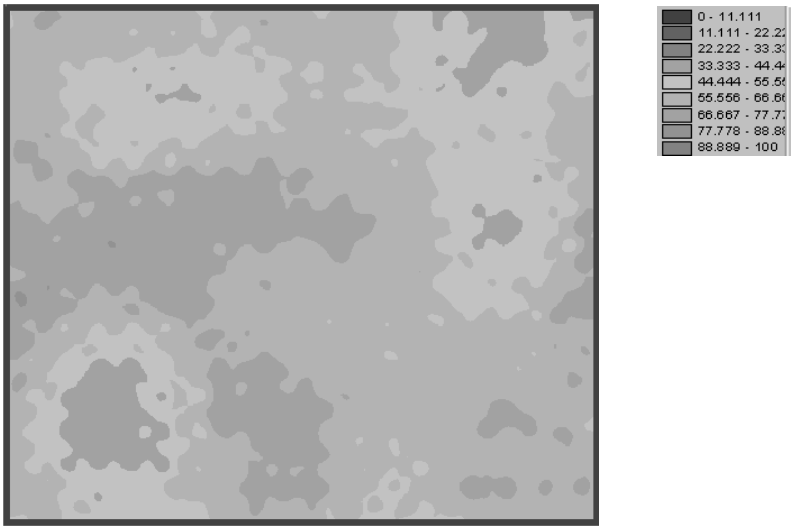


Figure 6.6— Full Set of Sample Points (255)—the Legend Indicates Percentages of Reliability

As in the previous data set we now decrease the number of sample points by approximately 33% to 172 locations. The resulting reliability estimates are shown in Figure 6.7—indicating the decreased reliability.

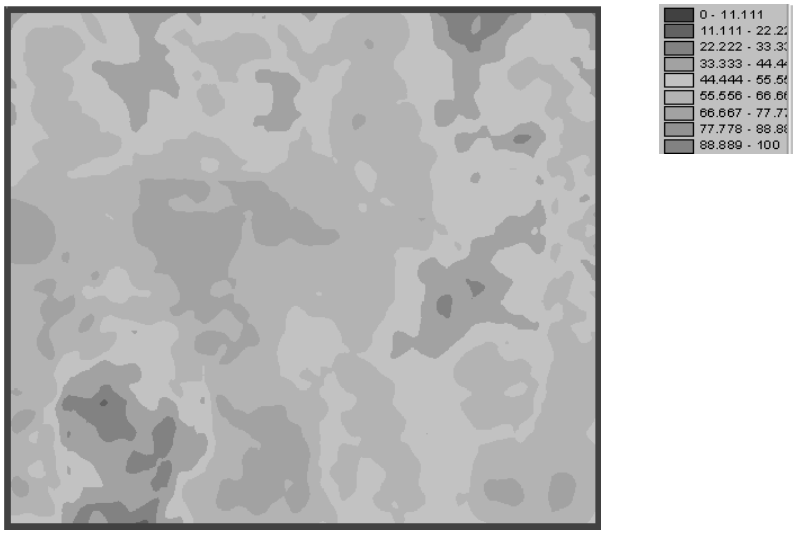


Figure 6.7—Subset of 172 Sample Points—the Legend Indicates Percentages of Reliability

The next step is to decrease the amplitude of the attribute value of the sample set to achieve at least 33% reliability (i.e., the initial reliability) throughout the entire study region. The results can be seen in Figure 6.8. Multiple runs with different multiplication factors resulted in an optimal value of 0.81. This value is equivalent to a decrease of variation in the attribute value of 19% compared to its original version. Using a factor of 0.815 (i.e., a decrease of 18.5%) yields in a reliability representation that shows 2 pixels within the 22 to 33-percentage class.

The difference between the two case studies is a 1% decrease of the variation of the attribute variable. Thus, we assume that there is no significant difference between the two results. Although, we did not apply any statistical tests to confirm this statement we are confident that for these two datasets our conclusions are correct.

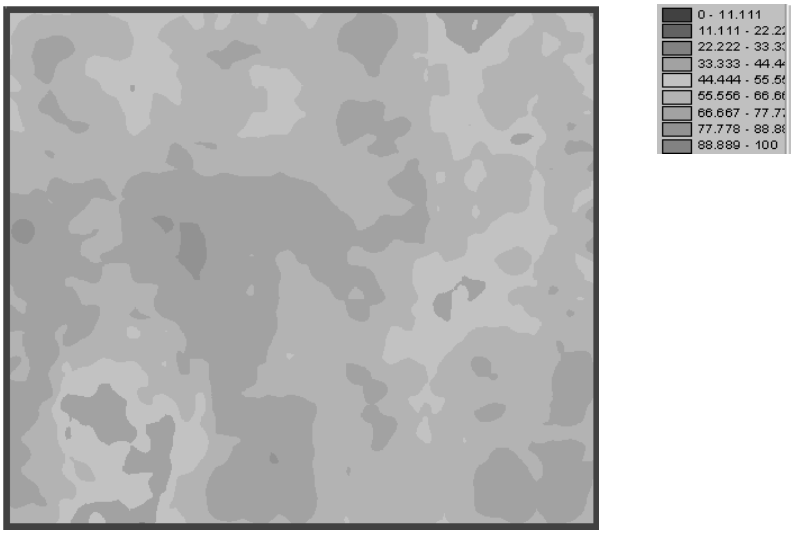


Figure 6.8—Subset of 172 Sample Points, Decreased Variation (-19%)—the Legend Indicates Percentages of Reliability

At this point we would also like to emphasize the fact that the case study shows that the influence of the sample density is comparable in size to the influence of the amplitude of the attribute field. Interestingly, the amount of influence of the amplitude seems to be larger than that of the sample density. These results were achieved with a data set showing plankton concentrations as well as with the second data set depicting height measurements. Both data sets indicate the possibility for generalization of the above-mentioned findings.

6.3 Remarks

In this chapter the focus of the discussion is based on the application of the reliability metric developed in the preceding chapter. Two influences on the reliability of a representation are empirically compared—the number of sample points (a representation of spatial discretization) and the amplitude, or variation of the attribute field.

The discussion begins with an outline of the applied approach. The approach is based on the fact that the reliability decreases with a decrease in the number of sample

points and increases with a decrease of variation. First, we give a description of the methodology to make the influences of the number of sample points evident. We suggest reducing the number of sample points by a certain percentage in a stratified pattern and applying the metric of reliability. Subsequently, we propose an iterative process to decrease the amplitude. The iteration process stops when the reliability of the entire representation reaches at least the amount of reliability generated with the complete sample set.

First, we set the accuracy requirement of the object to a constant. This constant is chosen to result in a reliability layer that is sensitive to perturbations. Subsequently, we follow the outlined approach. The number of sample points is reduced by 33% followed by a decrease of 20% or 19% of the variation to regain the original reliability. The results show that we can empirically compare the two influences. This case study also shows that the influence of the variation is slightly larger than the influence of the number of sample points.

Chapter 7
Conclusions and Future Work

7.1 Conclusions

This dissertation has addressed several aspects associated with various sources of uncertainty. The main focus is on modeling and visualizing effects of inaccuracy and discretization on the uncertainty of a representation. This is accomplished by developing an object oriented data quality model as well as specific metrics that require the implementation of such a DQM.

The first two chapters set the stage for the subsequent discussions. The first chapter emphasizes the importance of knowing the inherent uncertainties of a representation. Not knowing the uncertainty of a representation can lead to misinterpretations as pointed out in scientific, practical, and legal examples. The second chapter reviewed previous literature on the pertinent concepts used throughout the dissertation.

The proposed object oriented data quality model serves as the basis for more complex metrics that deal with uncertainties occurring in any GIS. The DQM links the units of information (e.g., measurement and coverage) to initial occurrences of specific sources of uncertainty (i.e., imperfections and discretization). The approach allows the more specific association of imperfection and discretization effects with identified units

of information and presents a more clearly identified path of propagation (or even cross-propagation) for specific sources of uncertainty. The essential part of this methodology is knowledge about the lineage and access to previous stages of a data set. The ability to access previous stages of the data (e.g., sample points that were used to interpolate a continuous representation) allows us to develop uncertainty metrics that depend on the original data set (i.e., the sample points).

Detectability is a metric that gives the user specific uncertainty information about the data at hand. The detectability metric estimates whether an object of a certain dimension can be detected within a field or not. The dependencies of this metric include properties of the sample field as well as the extent of the object. Thus, user input is required for this approach. The user can specify the size of the object she is looking for in the data field. The methodology is based on the comparison between an interpolated continuous surface and ground truth (i.e., some form of higher accuracy representation of the study region). From this comparison we can estimate the amount of noise that is introduced by the inaccuracy of the representation. This inaccuracy originates from the density (or the lack thereof) of the sample points and the interpolation process itself. The inaccuracy of the measurements might also contribute to the final inaccuracy of the representation but this effect is considered as small compared to the sources mentioned. Considering this fact we can conclude that it is not an absolute necessity for this approach to have knowledge of ground truth per se. We can estimate ground truth through simulations using the sample points—although they carry the same inaccuracies from the measurements. In the discussed case study we suggest to use additional sample points to increase the consistency of the results of the metric detectability.

Reliability is the second metric discussed in this dissertation. This metric also has the objective of making the influences of discretization on the uncertainty of a representation apparent. The reliability metric is also dependent on the variation within the sample field. Both of these components are vital for this metric that provides the

user with a reliability estimate of a representation. Again the reliability is dependent on user input, as the reliability of a representation depends on the user's expectations. A particular representation might be reliable enough to answer one query with sufficient degree of certainty, yet fails to perform as required for another task. The differences between the metric detectability and the metric reliability are mainly a) in the methodology to arrive at intermediate inaccuracy results and b) the way these intermediate inaccuracy results are interpreted. For the reliability metric, the local variations among neighboring sample points are investigated and subsequently inaccuracy values are derived. Since we have no knowledge about the variation of the reality (i.e., ground truth) the reliability metric results in estimations only. However, the achieved results were satisfactory. Subsequently, the interpretation of the inaccuracy estimates requires some information on the extent of the area the user is interested in. In the final step of the calculations for the reliability metric the comparison of the estimated inaccuracy values and the user defined accuracy requirements lead to the reliability estimates of a representation. The results are spatially referenced and help the user to judge the reliability of sub regions within a representation.

In our next step we explore the similarities and differences of the dependencies of the reliability metric in more detail. For this investigation we introduce perturbations to a specific dataset and observe the changes occurring in the resulting reliability. The perturbations are aimed at isolating the influence of the spatial discretization on the reliability from the influence of the underlying spatial variation. First, reducing the number of sample points by a given percentage and secondly, decreasing the variation in the sample field achieve this. The effect is accomplished by multiplying the observations by a constant factor representing the percentage of the variation after the multiplication (e.g., multiplying by 0.8 results in a variation that is 80% of its original). Finally, we can compare the two causes and conclude that the influence of the variation is similar in amount to the influence of the spatial discretization. For the specific datasets at hand a slightly larger influence of the variation is evident.

7.2 Future Work

In this section we discuss some possibilities for future research projects for each of the above-mentioned models (i.e., data quality model, detectability, reliability, and the comparison of the influences on reliability).

The proposed data quality model is rather complex in nature and possible cross-propagations need to be investigated and mapped in greater detail. So far formal approaches (i.e., metrics) have been developed only for a small portion of the sources of uncertainty (as discussed in the DQM). This statement is not necessarily restrictive to this dissertation but considers general approaches described throughout the pertinent literature.

The general model for the detectability metric is thoroughly discussed in this dissertation. Nevertheless, everyday applications would profit from more detailed case studies. Especially, when using conditional simulations, a rule of thumb for the number of additionally required sample points needs to be evaluated and optimized.

For the derivation itself we made somewhat arbitrary assumptions. Consideration for general applicability, however, would need more detailed investigations aiming at the arbitrariness of our choices. Here, specifically one should focus on determining the number of neighboring sample points for optimal results in a more systematic way. Within the text we give several guidelines to adapt the approach for discrete data structures. Future work might focus at an implementation of these ideas.

The comparison of the influences on reliability was done in an empirical way. It is doubtful that a relationship can be formally defined, however, investigations with additional datasets and regular sample point distributions might establish more rigid formalizations of the relationships among the different influences.

I believe that the future of GISs is a bright one. GISs are well established in an amazingly broad variety of applications. They can capture attributes of and relationships among galaxies in the universe as well as human genes on the DNA ladder. Yet, there is still room for improvement. In my opinion the current stage of geographic information science is an early one—when looking at its potential to grow. Future work is needed and if there is an interest to work on additions to the suggested approaches found in this dissertation I would be delighted to be part of it.

References

Agumya, A. and G. J. Hunter. 1996. Assessing Fitness for Use of Spatial Information: Information Utilisation and Decision Uncertainty. Proceedings of the GIS/LIS '96 Conference, Denver, CO: 359-370.

Beard, M. K. 1989. Designing GIS to Control Misuse of Spatial Information. Proceedings of the URISA '89 Conference, Boston, MA: 245-255.

Beard, M. K. 1996. A Structure for Organizing Metadata Collection. Proceedings 3rd International Conference on Integrating GIS and Environmental Modeling, Santa Fe, NM: 21-26.

Bedard, Y. 1987. Uncertainties in Land Information Systems Databases. Proceedings of Auto Carto 8, Falls Church, VA: 175-184.

Blakemore, M. 1985. High or Low Resolution? Conflicts of Accuracy, Cost, Quality, and Application in Computer Mapping. *Computers & Geosciences* 11 (3): 345-348.

Bruegger, B. P. 1994. *Spatial Theory for the Integration of Resolution-Limited Data*. Orono, ME, Dissertation.

Burrough, P. A. 1983. Multiscale sources of spatial variation in soir. II. A non-Brownian fractal model and its application in soil survey. *Journal of Soil Science* 34: 599-620.

Burrough, P. A. 1991. The Development on Intelligent Geographical Information Systems. Proceedings of the 2nd European Conference on GIS, Brussels, Belgium: 165-174.

Burrough, P. A. and A. Frank. 1996. *Geographic Objects with Indeterminate Boundaries*, London: Taylor & Francis.

Burrough, P. A. and R. A. McDonnell. 1998. *Principles of Geographical Information Systems*, Oxford: Oxford University Press.

Buyong, T. B., A. Frank and W. Kuhn. 1991. A Conceptual Model of Measurement-Based Multipurpose Cadastral. *Journal of the Urban and Regional Information Systems Association* 3 (2): 35-49.

Canters, F., H. Eerens and F. Veroustraete. 1999. Estimation of Land-Cover Proportions from Aggregated Medium-Resolution Satellite Data. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, Ann Arbor Press: 263-270.

Carroll, S. S. 1995. Modeling measurement errors when estimating snow water equivalent. *Journal of Hydrology* (172): 247-260.

CEN/TC 287, N. 1995. Geographic Information - Reference Model. *Draft for discussion - technical report.*

CEN/TC 287, W. 1995. Geographic Information - Data description-Quality. *Draft for discussion - technical report.*

CEN/TC 287, W. S. 1995. Working draft on Geographic information - Definitions. *Draft for discussion - technical report.*

Chrisman, N. 1983. The Role of Quality in Information in the Long Term Functioning of a Geographic Information System. Proceedings Auto Carto 6: 303-312.

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

Clarke, K. 1997. *Getting Started with Geographic Information Systems*, Prentice Hall.

Codd, E. F. 1979. Extending the Database Relational Model to Capture More Meaning. *ACM Transactions on Database Systems* 4: 397-434.

Cressie, N. A. C. 1991. *Statistics for spatial data*, John Wiley & Sons

Csillag, F., A. Kummert and M. Kertesz. 1992. Resolution, Accuracy and Attributes: Approaches for Environmental Geographical Information Systems. *Computers, Environment and Urban Systems* 16: 289-297.

Cushnie, J. L. 1987. The ineractive effect of spatial resolution and degree of internal variability within land-cover types on classification accuracies. *International Journal of Remote Sensing* 8 (1): 15-29.

Douglas, D. 1972. It makes me so CROSS. In *Introductory readings in GIS*. D. Mavble and D. Peuquet, eds. Taylor & Francis.

Dutta, S. 1989. Qualitative Spatial Reasoning: A Semi-quantitative Approach Using Fuzzy Logic. In *Design and Implementation of Large Spatial Databases*, First Symposium SSD, Santa Barbara, CA: 345-364.

Egenhofer, M. J. and J. Herring. 1990. A mathematical framework for the definition of topological relationships. Fourth International Symposium on Spatial Data Handling, Zurich, Switzerland: 803-813.

Egenhofer, M. J. and D. M. Mark. 1995. Naive Geography. In *Spatial Information Theory: A Theoretical Basis for GIS*. A. U. Frank and W. Kuhn, eds. Berlin, Springer-Verlag: 1-15.

Elmes, G. A. and C. Cai. 1992. Data Quality Issues in User Interface Design for a Knowledge-based Decision Support System. Proceedings of the 5th International Symposium on Spatial Data Handling, Charleston, SC: 303-312.

FGDC. 1994. *Content Standards for Geospatial Metadata*. Reston, VA: U.S. Geological Survey.

FGDC. 1996. *Draft Geospatial Positioning Accuracy Standards (Standards for Geodetic Networks)*, ftp://fgdc.er.usgs.gov.

Forier, F. and F. Canters. 1996. A User-friendly Tool for Error Modelling and Error Propagation in a GIS Environment. In *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. H. T. Mowrer, R. L. Czaplewski and R. H. Hamre, eds. Fort Collins: USDA Forest Services General Technical Report: 225-234.

Frank, A. U., G. S. Volta and M. McGranaghan. 1997. Formalization of Families of Categorical Coverages. *International Journal of Geographical Information Science* 11 (3): 215-231.

Frank, U. A. 1998. Metamodels for Data Quality Description. In *Data Quality in Geographic Information - From Error to Uncertainty*. R. Jeansoulin and M. Goodchild, eds. Paris, Editions Hermès: 15-29.

Frank, U. A. 2001, to appear. The rationality of epistemology and the rationality of ontology. Rationality and Irrrationality, Proceedings of the 23rd International Ludwig Wittgenstein Symposium, Kirchberg am Wechsel, Hölder-Pichler-Tempsky

Goodchild, M. F. 1979. Effects of generalization in geographical data encoding. Map Data Processing. H. Freeman and G. G. Pieroni. New York, Academic Press: 191-206.

Goodchild, M. F. 1989. Modeling error in objects and fields. In *Accuracy of Spatial Databases*. M. Goodchild and S. Gopal, eds. London, Taylor & Francis: 107-113.

Goodchild, M. F. 1993. Data Models and Data Quality: Problems and Prospects. In *Environmental Modeling with GIS*. M. F. Goodchild, B. O. Parks and L. T. Steyaert, eds. New York, Oxford University Press: 94-104.

Goodchild, M. F. and S. Gopal. 1989. *Accuracy of Spatial Databases*, London: Taylor & Francis.

Goodchild, M. F. and J. Proctor. 1997. Scale in a Digital Geographic World. *Geographical & Environmental Modelling* 1 (1): 5-23.

Goodchild, M. F., G. Sun and S. Yang. 1992. Development and Test of an Error Model for Categorical Data. *International Journal of Geographical Information Science* 6: 87-104.

Green, P. J. and R. Sibson. 1978. Computing dirichlet tessellations in the plane. *The Computer Journal* 21 : 168-173.

Guptill, S. and J. L. Morrison. 1995. *Elements of Spatial Data Quality*, Elsevier Science.

Hammersley, J. M. and D. C. Handscomb. 1979. *Monte Carlo Methods*, London: Chapman and Hall.

Heuvelink, G. B. M. 1993. *Error Propagation in Quantitative Spatial Modelling*. Utrecht, Drukkerij Elinkwijk.

Heuvelink, G. B. M. 1998. *Error Propagation in Environmental Modelling with GIS*, London: Taylor & Francis.

Heuvelink, G. B. M. 1999. Aggregation and Error Propagation in GIS. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, Ann Arbor Press: 219-225.

Heuvelink, G. B. M. and P. A. Burrough. 1993. Error Propagation in Cartographic Modelling using Boolean Logic and Continous Classification. *International Journal of Geographical Information Science* 7 (3): 231-246.

Hunter, G. J. and M. F. Goodchild. 1993. Managing Uncertainty in Spatial Databases: Putting Theorie into Practice. *Journal of the Urban and Regional Information Systems Association* 5 (2): 55-62.

Hunter, G. J. and M. F. Goodchild. 1997. Modeling the Uncertainty of Slope and Aspect Estimates Derived from Spatial Databases. *Geographical Analysis* 29: 35-49.

Joslyn, C. A. 1992. Web Dictionary of Cybernetics and Systems. http://pespmc1.vub.ac.be/:/ASC/uncertainty.html.

Kaintz, W. 1995. Logical Inconsistency. *In Elements of Spatial Data Quality*. S. Guptill and J. Morrison, eds. Elsevier Science: 109-137.

Keefer, B. J., J. L. Smith and T. G. Gregoire. 1988. Simulating manual digitizing error with statistical models. GIS/LIS '88, Falls Church, VA, American Society for Photogrammetry and Remote Sensing/American Congress on Surveying and Mapping: 475-483.

Kiiveri, H. T. 1997. Assessing, Representing, and Transmitting Positional Uncertainty in Maps. *International Journal of Geographical Information Science* 11: 33-52.

Kozlowski, L. T. and K. J. Bryant. 1977. Sense of Direction, Spatial Orientation and Cognitive Maps. *Journal of Experimental Psychology: Human Perception and Performance* 3: 590-598.

Kraus, K. and H. Kager. 1993. Genauigkeiten abgeleiteter Daten in einem Geo-Informationssystem. *Mitteilung der geodaetischen Institute der TU Graz* Festschrif Guenther Schelling zum 70. Geburtstag (Folge 78): 95-101.

Kuipers, B. J. 1982. The 'Map in the Head' Metaphor. *Environment and Behaviour* 14 (2): 202-220.

Kuipers, B. J. and T. S. Levitt. 1988. Navigation and Mapping in Large-Scale Space. *AI Magazine* 9 (2): 25-43.

Kyriakidis, P. C., A. M. Shortrige and M. F. Goodchild. 1999. Geostatistics for Conflation and Accuracy Assessment in Digital Elevation Models. *International Journal of Geographical Information Science* 13 (7): 677-708.

Lanter, D. P. and H. Veregin. 1992. A Research Paradigm for Propagating Error in Layer-Based GIS. *Photogrammetric Engineering & Remote Sensing* 58: 825-833.

Longley, P. A., M. F. Goodchild, D. J. Maguire and D. W. Rhind. 1999. Epilogue. In *Geographical Information Systems: Principles, Techniques, Applications and Management*. P. A. Longley, M. F. Goodchild, D. J. Maguire and D. W. Rhind, eds. New York, John Wiley & Sons: 1009–1021.

Moody, A. and C. E. Woodcock. 1994. Scale-Dependent Errors in the Estimation of Land-Cover Proportions: Implications for Global Land-Cover Datasets. *Photogrammetric Engineering & Remote Sensing* 60 (5): 585-594.

Morrisey, J. M. 1990. Imprecise Information and Uncertainty in Information Systems. *ACM Transactions on Information Systems* 8: 159-180.

Motro, A. and P. Smets. 1997. *Uncertainty Management in Information Systems: From Needs to Solutions*, Boston: Kluwer Academic Publishers.

Mowrer, H. T. 1999. Accuracy (Re)assurance: Selling Uncertainty Assessment to the Uncertain. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, Ann Arbor Press: 3-10.

Nijkamp, P. and H. J. Scholten. 1991. Information Systems: Caveats in Design and Use. Proceedings of the 2nd European Conference on GIS, Brussels, Belgium: 737-746.

Oliver, M. A. and R. Webster. 1986. Combining Nested and Linear Sampling for Determining the Scale and Form of Spatial Variation of Regionalized Variables. *Geographical Analysis* 18 (3): 227-242.

O'Neil, R. V., J. R. Krummel, R. H. Gardner, G. Sugihara, B. Jackson, D. L. DeAngelis, B. T. Milne, M. G. Turner, B. Zygmunt, S. W.Christensen, V. H. Dale and R. L.Graham, 1988. Indices of Landscape Pattern. *Landscape Ecology* 1: 153-162.

Openshaw, S. 1989. Learning to Live with Errors in Spatial Databases. In *Accuracy of Spatial Databases*. M. F. Goodchild and S. Gopal, eds. London: Taylor & Francis: 263-276.

Parsons, S. 1996. Current Approaches to Handling Imperfect Information in Data and Knowledge Bases. *IEEE Transactions on Knowledge and Data Engineering* 8 (3): 353-372.

Pebesma, E. J. and C. G. Wesseling. 1997. Gstat, a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences* 24 (1): 17-31.

Perkal, J. 1956. On epsilon length. *Bulletin de l'Academie Polonaise des Sciences* 4: 399-403.

Peuquet, D., B. Smith and B. Brogaard. 1999. The Ontology of Fields. Report of a Specialist Meeting Held Under the Auspices of the Varenius Project, Bar Harbor, Maine

Prisley, S. P. and J. L. Smith. 1991. The Effects of Spatial Data Variability on Decisions Reached in a GIS Environment. In *GIS Applications in Natural Resources*. M. Heit and A. Shortread, eds. 167-170.

Reissmann, G. 1976. *Die Ausgleichsrechnung: Grundlagen und Anwendungen in der Geodaesie*. Berlin, Verlag fuer Bauwesen.

Ripley, B. 1987. *Stochastic Simulation*. New York, John Wiley & Sons.

Schone, H. 1984. *Spatial Orientation - The Spatial Control of Behavior in Animals and Man*. Princeton, NJ, Princeton University Press.

Shannon, C. E. and W. Weaver. 1962. *The mathematical theory of communication*, Urbana: University of Illinois Press.

Sinton, D. F. 1978. The inherent structure of information as a constraint to analysis: mapped thematic data as a case study. *Harvard Papers on Geographic Information Systems* 6: 43-59.

Smets, P. 1997. Imperfect Information: Imprecision and Uncertainty. In *Uncertainty Management in Information Systems: From Needs to Solutions*. A. Motro and P. Smets, eds. Boston: Kluwer Academic Publishers: 225-254.

Stanislawski, L. V., B. A. Dewitt and R. L. Shrestha. 1996. Estimating Positional Accuracy of Data Layers within a GIS Through Error Propagation. *Photogrammetric Engineering & Remote Sensing* 62 (4): 429-433.

Stoms, D. M., F. W. Davis and C. B. Cogan. 1992. Sensitivity of Wildlife Habitat Models to Uncertainties in GIS Data. *Photogrammetric Engineering & Remote Sensing* 58 (6): 843-850.

Taylor, J. R. 1982. *An Introduction to Error Analysis*, Mill Valley: University Science Books.

Timpf, S., M. Raubal and W. Kuhn. 1996. Experiences with Metadata. Spatial Data Handling, Delft, Netherlands, TU Delft: 12B.31-12B.43.

Townshend, J. R. G. and C. O. Justice. 1988. Selecting the spatial resolution of satellite sensors required for global monitoring of land transformations. *International Journal of Remote Sensing* 9 (2): 187-236.

Turner, M. G., V. H. Dale and R. H. Gardner. 1989. Predicting across scales: Theory development and testing. *Landscape Ecology* 3: 245-252.

Turner, M. G., R. V. O'Neill, R. H. Gardner and B. T. Milne. 1989. Effects of changing spatial scale on the analysis of landscape pattern. *Landscape Ecology* 3 (3): 153-162.

van Groenigen, J. W. and A. Stein. 2000. Constrained Optimization of Spatial Sampling in a Model-based Setting using SANOS Software. Accuracy 2000, Amsterdam, Delft University Press: 679-685.

Veregin, H. 1989. Error modeling for the map overlay operation. In *Accuracy of Spatial Databases*. M. Goodchild and S. Gopal. London, eds. Taylor & Francis: 3-18.

Veregin, H. 1989. A taxonomy of error in spatial databases. Santa Barbara, National Center for Geographic Information and Analysis.

Wade, T. G., J. D. Wickham and D. F. Bradford. 1999. Accuracy of Road Density Estimates Derived from USGS DLG Data for Use in Environmental Applications. *Photogrammetric Engineering & Remote Sensing* 65 (12): 1419-1425.

Watzek, K. A. and J. C. Ellsworth. 1992. Perceived Scale Accuracy of Computer Visual Simulations. *Landscape Journal*: 21-36.

Windholz, T. K., K. Beard and M. Goodchild. 2001, to appear. Data Quality: A Model for Resolvable Objects. In *Advances in Spatial Data Quality*. W. Shi, M. F. Goodchild and P. F. Fisher, eds. Taylor & Francis.

Wingle, W. L., E. P. Poeter and S. A. McKenna. 1994. UNCERT: a geostatistical uncertainty analysis package applied to groundwater and contaminant transport modeling. http://uncert.mines.edu/

Wolf, P. R. and C. D. Ghiliani. 1997. *Adjustment Computations: Statistics and Least Squares in Surveying and GIS*. New York, John Wiley & Sons.

Zadeh, L. A. 1965. Fuzzy Sets. *Information and Control* 8: 338-353.

Biography of the Author

Thomas K. Windholz was born in Vienna, Austria on May 19th, 1971. He was raised in Biedermannsdorf—a small town with a population of about 2000 inhabitants located about 5 miles south of Vienna. Thomas graduated in 1989 from the scientific high school nw. Realgymnasium Mödling, Franz-Keim-Gasse, NÖ, Austria. After serving the mandatory time of 8 months in the Austrian Army at the PzB 33, Zwölfaxing, he continued his education at the Technical University of Vienna at the Department of Geoinformation (at that time this department was a sub-department within the Department of Surveying). While working on his Master's thesis he became a visiting student to the University of Maine for a period of six months, subsequently earning his Master's degree at the Technical University of Vienna (Dipl.-Ing. in Surveying) in 1997.

In January of 1998, Thomas entered the Ph.D. program in the Department for Spatial Information Science and Engineering at The University of Maine. He has been employed as a graduate research assistant and will continue work as a post-doctoral researcher in the Department of Spatial Information Science and Engineering at The University of Maine. Thomas is married to Elizabeth and has no children. Thomas is a candidate for the Doctor of Philosophy degree in Spatial Information Science and Engineering from The University of Maine in May, 2001.