

The Issue of Uncertainty Propagation in Spatial Decision Making

Rahim Ali. Abbaspour, Mahmoud R. Delavar, Reihaneh Batouli

Department of Surveying and Geomatic Eng., Engineering Faculty, University of Tehran,
Tehran, Iran

Tel : +98 21 8008841, Fax : +98 21 8008837

Email: abbaspour@geomatics.ut.ac.ir,
mdelavar@chamran.ut.ac.ir, batouli@ut.ac.ir

Abstract GISs give users facilities to integrate and analyze data from different sources with different scale, accuracy, resolution and quality of the original data which are the key aspects of GIS functionality, but it does raise the question as to what effects the combination of different levels of data uncertainty has on both the output maps and on the data derived from spatial query and analysis. In this paper, in addition to provide an overview of uncertainty propagation assessment in overlay analysis, an experiment using Monte Carlo simulation method has been performed and then the results were analyzed. Two polygons whose vertices have been perturbed by changing their coordinates randomly using Monte Carlo simulation method are overlaid so that their intersection defines the third polygons set which in turn were statistically analyzed using a developed program and some GPS data. Two mainly recommended indicators, i.e., area and perimeter of polygon, were used and ended up with consequence that the indices of these polygons whose vertices had error in position emerged less than those whose vertices were accurate.

1 Introduction

Geospatial information systems (GISs) permit a wide range of operations to be applied to spatial data in the production of both tabular and graphic output products. Too frequently, however, these operations are applied with little regard for the types and levels of error that may result. In numerous published articles detailing GIS applications, a critical examination of error sources is conspicuously absent and output products are presented without an associated estimate of their reliability. Unfortunately, in most cases these omissions do not imply that errors are of a sufficiently low magnitude that they may safely be ignored [12]. Moreover, the fact that input data are of relatively high quality is no guarantee that output products will be error free. The utility of GIS as a decision support system/science is dependent on the development and dissemination of formal models of error for GIS operations. The goal of users should at minimum be provided with a means of assessing the accuracy of the information upon which their decisions are based. Although error models have been developed for certain GIS operations, these models have not been widely

adopted in practice. One impediment to their more widespread adoption is that no single error model is applicable in all instances.

Uncertainty may arise from the very first step of conceptualizing the real world to the last operation which is the decision making process and it may take one or a combination of several forms as spatial, attribute, temporal, logical consistency and completeness. Several theories have been adopted to handle uncertainty in GIS [1,7]. This paper is concerned with developing methods able to estimate the confidence regions of GIS outputs by taking into account certain selected sources of uncertainty affecting spatial databases. A Monte Carlo simulation-based method is adopted as a general means of estimating the effects of input data uncertainty on the outputs after an arbitrary sequence of overlay analysis. The objective is to identify and handle the effects of data uncertainty in GIS by defining main uncertainty indices such as area and perimeter. This is considered the minimum need to allow a GIS to function in an uncertain data environment.

2 The Concept of Uncertainty in GIS

Uncertainty can be defined as a skepticism, mistrust, suspicion or lack of sureness about something. It can denote a lack of certainty or lack of definite knowledge about an outcome or result. Unlike the terms *accuracy* and *error*, in the context of spatial databases, there is a clear distinction between *error* and *uncertainty*, since the former implies some level of knowledge about the differences between the observations and the truth, while the latter conveys the fact that it is the lack of such knowledge which is responsible for hesitancy in accepting those same results and observations. In many cases, the term *error* is used when it would be more appropriate to use *uncertainty* [11].

Many theories have been developed for dealing with uncertainty in spatial databases. However, a comprehensive theoretical framework that can handle all the existing uncertainty forms in GIS has not been developed, yet. Existing theories such as spatial statistics, fuzzy set theory, interval theory, probability theory and mathematical theory of evidence can be adopted and applied to certain forms of inexactness in databases [1]. Every theory is based on specific assumptions and as such can only be effectively applied to those uncertainties that conform to the *a priori* assumptions. Among several theories that have been considered to handle uncertainty, geostatistics and probability theory can be used to manage those uncertainties caused by random components. Fuzzy set theory may be applied to handle vague concepts, e.g. linguistic variables, while evidential theory may be of use to manage the uncertainty due to information incompleteness [3]. It is argued that a single theory that can handle all types of uncertainty is yet to come.

3 Classification of Uncertainty in GIS

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

Although the terminology varies in the literature on this subject [14], it is generally agreed that there are several categories of uncertainties, which may contribute to the overall accuracy of products derived from GIS. What are usually unclear are the relationships among these errors and there is often confusion between their sources and the forms they may take. Classification would help to clarify this confusion.

While taxonomic study is a science in itself, it need not always be effective [13]. Calkins and Obermayer have elaborated on the application of taxonomies to describe the use and value of spatial databases and make it clear that though the classification process is not a simple task, it is nevertheless an achievable one.

Efforts to classify spatial database uncertainty have been made by a number of researchers [8,11]. Other authors have also tried to understand database errors through their classification. For instance, GIS errors have been categorized into three groups, (i) obvious sources of errors (such as the age of the data and the extent of area coverage), (ii) error resulting from the original measurement or through natural variation, and (iii) errors resulting from computer processing (such as rounding) [6]. Error classification can also be based on the GIS functionality [2]. The most widely used error taxonomy is embedded in the U.S. Spatial Data Transfer Standard (SDTS) [12]. Given the variations among respected researchers in organizing uncertainties, the question of uncertainty classification in this research is answered by the following three level taxonomy (Fig. 1).

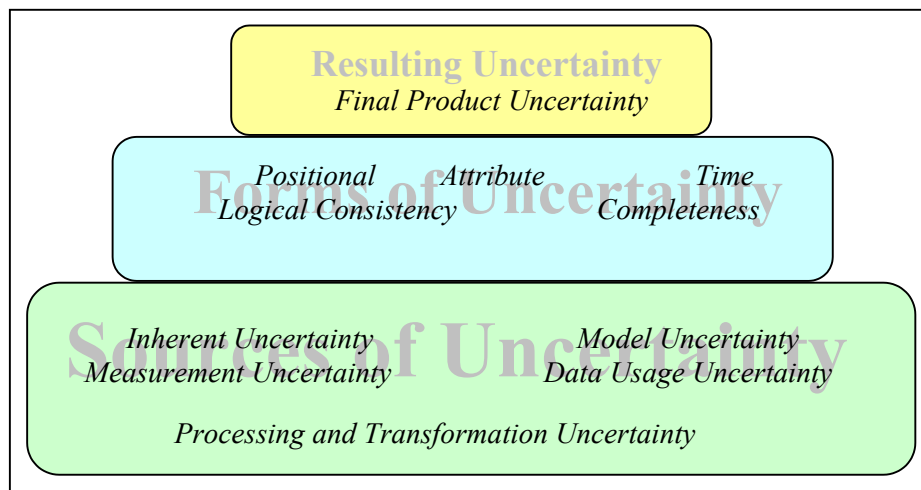


Fig. 1. A taxonomy of uncertainty in GIS [3]

As indicated in Fig. 1, the first level of the taxonomy deals with sources of uncertainty which have been classified into five categories: (i) the inherent

uncertainty of the phenomena being mapped, (ii) measurement uncertainty of spatial phenomena due to accuracy limitations of all observations, (iii) model uncertainty which arises due to the models that are used to communicate the measurements, (iv) processing and transformation uncertainty which refers to the secondary uncertainty caused during computer manipulation of data, following the data measurement, and (v) data usage uncertainty which has only recently received attention among researchers which is concerned with the manner in which spatial data are used.

The second level of the taxonomy classifies the forms of uncertainty into five aspects which are positional, attribute, time, logical consistency and completeness. This classification is mainly based on the SDTS.

The third level refers to resulting uncertainty. The separation of the final product uncertainty from the forms of them is done because of the manner in which they may occur in the product.

4 The Theory of Error Propagation in GIS

The error propagation problem can be formulated mathematically as follows. Let $X(.)$ be the output of GIS operation $g(.)$ on the n input attributes $A_i(.)$.

$$X(.) = g(A_1(.), A_2(.), \dots, A_n(.)) \quad (1)$$

The operation $g(.)$ may be one of various types, such as a standard filter operation to compute gradient from a gridded digital terrain model (DTM). The objective of the error propagation analysis is to determine the error in the output $X(.)$, given the operation $g(.)$ and errors in the input attributes $A_i(.)$. The output map $X(.)$ is also a random field, with mean $\mu(.)$ and variance $\tau^2(.)$. From an error propagation perspective, the main interest is in uncertainty of $X(.)$, as contained in its variance $\tau^2(.)$ [10].

It must first be observed that the error propagation problem is relatively easy when $g(.)$ is a linear function [9]. In that case, the mean and variance of $X(.)$ can be directly and analytically derived. The theory on functions of random variables also provides several analytical approaches to the problem for nonlinear $g(.)$, but few of these can be resolved by simple calculation. Thus for the general situation, analytical methods are not very suitable.

5 Methods of Uncertainty Modeling

An error model refers to a stochastic process capable of simulating the range of possibilities known to exist for spatial data. These possibilities may exist because measuring instruments are known to be of limited accuracy, or because vital information, such as datum or map projection is missing [5].

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

The basic geometric components in 2D object-based GIS are points, lines and polygons. They are considered as different objects since their spatial properties and their uncertainty behavior are different [15]. Consequently, the uncertainty models that describe the behavior of objects should be different. It is elaborated that the analytical method is just an approximation technique. The simulation method requires several interactive operations and hence, the method may not be efficient in practice. Empirical method is based on comparing the object with its 'true' value, making it time and cost consuming. Error descriptors cannot simulate the possible locations of the spatial objects; therefore, they lack the conditions of being statistical error models. The selection of the best method depends on the application, time and cost.

6 Monte Carlo Simulation Method

In principle the problem of uncertainty transmission in GIS operations can be handled by using covariance propagation [6]. The lack of single continuous differentiable function renders the use of explicit equations for error propagation. Instead, it is simpler and more general to use a universal solution based on Monte Carlo simulation approach.

The idea of the method is to compute the result of (1) repeatedly, with input values $A_i(.)$ that are randomly sampled from their joint distribution. The model results from a random sample from the distribution $F_X(.)$ of X , so that parameters of $F_X(.)$ such as the mean, μ , and the variance, τ^2 , can be estimated from the sample. Properties of these estimators are well known from classical sampling theory. The method thus consists of the following steps:

1. Decide what levels and types of error characterize each data set as input to a GIS.
2. Replace the observed data by a set of n random variables drawn from appropriate probability distributions assumed to represent the uncertainty in the data inputs.
3. Apply a sequence of GIS operations to the step (2) data.
4. For this set of realizations l_i , store the results, x_i .
5. Compute summary statistics.

7. Accuracy of Monte Carlo Method

Let the outcomes of n times running operation $g(.)$ with the error-perturbed inputs be $X = [x_1, x_2, \dots, x_i, \dots, x_n]$. Then μ and τ^2 can be estimated by the sample mean, m_X , and sample variance, s_X^2 , as follows:

$$m_X = \frac{1}{n} \sum_{i=1}^n X_i \quad (2)$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - m_X)^2 \quad (3)$$

The sample mean and variance are both unbiased and consistent estimators of μ and τ^2 . Their variances are [10]:

$$\text{var}(m_X) = \tau^2 / n \quad (4)$$

$$\text{var}(s_X^2) = \frac{1}{n} [\mu_4 - \tau^4 \left(\frac{n-3}{n-1}\right)] \approx \frac{\mu_4 - \tau^4}{n} \quad (5)$$

where μ_4 is the fourth central moment of $X(\cdot)$. The sample skewness and kurtosis can also be computed, yielding unbiased estimates of the true skewness and kurtosis of $X(\cdot)$. The variances of these estimators can also be derived, although the resulting expressions are somewhat complicated [10].

When $X(\cdot)$ is normally distributed, Equation 5 reduces to Equation 6:

$$\text{var}(s_X^2) = \frac{2\tau^4}{n-1} \quad (6)$$

An interesting consequence from (5) is that the coefficient of variation of s_X^2 , $\text{cv}(s_X^2)$, is independent of τ^2 :

$$\text{cv}(s_X^2) = \sqrt{\frac{\gamma_2}{n} + \frac{2}{n-1}} \quad (7)$$

where $\gamma_2 = \mu_4 / \tau^4 - 3$ is the coefficient of kurtosis, which is zero for the normal distribution. If the coefficient of variation of the resulting sample variance is taken as a criterion for establishing its accuracy, then (7) enables one to tell in advance how large the number of Monte Carlo runs should be.

From (4) and (5) it follows that the standard deviations of m_X and s_X^2 are approximately inversely related to the square root of the number of Monte Carlo runs (Fig. 2). This is a general result that holds for other parameters of the distribution as well. It means that to double the accuracy, four times as many Monte Carlo runs are needed. The accuracy thus slowly progresses as n increases.

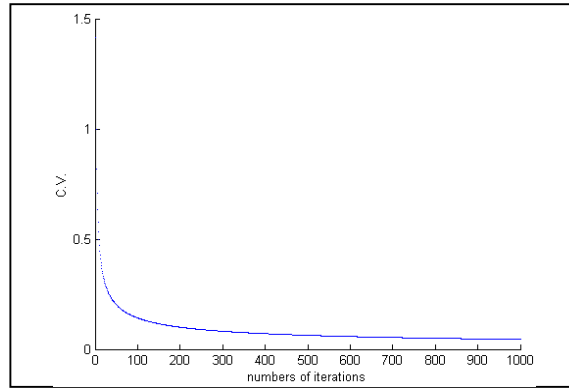


Fig. 2. Relationship between Monte Carlo runs and coefficient of variation [1]

8 Methodology

Much of the functionality of GIS lies with their ability to overlay one or more digital maps for the purpose of Boolean or network analyses. In this study, Monte Carlo simulation method has been used to estimate the levels of uncertainty in the output of an overlay analysis, so a simulation is made by generating random numbers from their probability distributions (which assumed to be normal) defined by covariance matrix of each polygon and adding them to the coordinates of vertices of two measured polygons (see Fig. 3). This process is repeated 100 times and two groups of polygons were generated. Each of the randomized input polygons is then overlaid on another one. The final results were then saved.

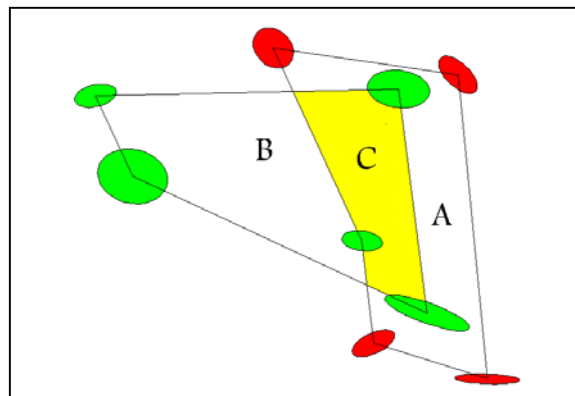


Fig. 3. Polygons, their overlays and vertices with error ellipses (ellipses are not in scale)

Fig. 4 shows the histograms, which illustrate the frequency of 100 overlaid polygons with respect to their area and perimeter. They approximately follow the normal distribution. The entire process is repeated 100 times. Then the total set of output

maps were used to calculate the area and perimeter of confidence regions that were overlaid on the deterministic result. Table 1 shows the results where the mean area and perimeter of 100 generated overlaid polygons are less than the area and perimeter of overlaid polygon when its spatial uncertainty is ignored respectively.

The GIS software used in this case study was Arc/Info 7.2. A program was developed in MS Visual Basic, using the ODE of Arc/Info.

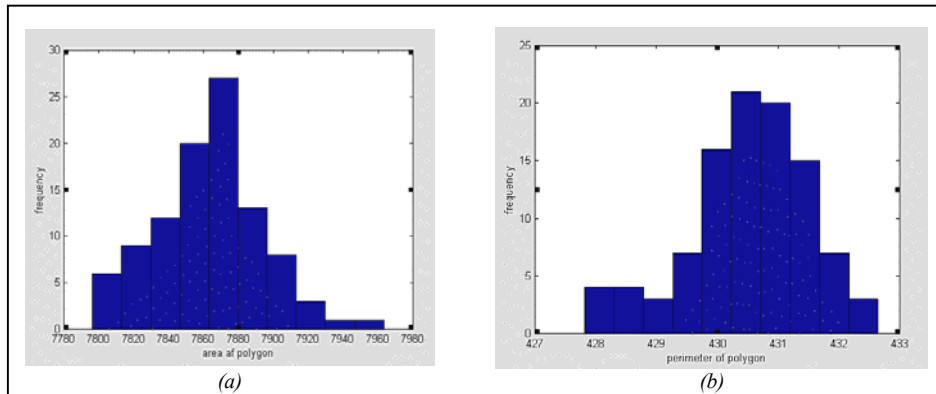


Fig. 4. Frequency of the overlaid polygons with respect to their area (in square meters) (a) and perimeter (in meters) (b)

Table 1: Results of the practical work

Number of generated polygons	Mean		Standard deviation		Overlaid polygon (ignoring spatial uncertainty)	
	Area	Perimeter	Area	Perimeter	Area	Perimeter
100	7858.610	430.519	34.487	3.028	7869.997	438.271

9 Conclusion

Handling uncertainty in GIS is of vital importance as it plays a considerable role in decision-making. One of the significant aspects of uncertainty study is its propagation through the GIS analysis and modeling. Different methods exist in GIS uncertainty modeling. Four classes of methods have been identified and formulated. They are analytical, simulation, empirical and error descriptors where each of them has its advantages and disadvantages. The Monte Carlo method uses a simulation approach to analyze the propagation of error in GIS operations. It repeatedly draws realizations from the joint probability distribution of the input attributes, each time substituting these realizations into the operation, computing the result and storing it. In this way a random sample from the output distribution is obtained which is analyzed by using techniques from classical sampling theory. In this study, this method was used to model the spatial uncertainty results due to inherent uncertainty of input data and its propagation along with the overlay analysis. To achieve this aim, two data sets of

GPS were overlaid on each other. Each vertex in the polygons was perturbed by selecting random numbers from a normal distribution, based on the given covariance matrix. The results achieved by Monte Carlo showed the difference while considering and ignoring spatial uncertainty of input data.

References

1. Ali Abbaspour, R.: Error Propagation in Overlay Analysis in GIS. , MSc. Thesis, Surveying and Geomatic Engineering Dept., University of Tehran, Iran (2002)
2. Alai, J.: Spatial Uncertainty in a GIS. , MSc. Thesis, Department of Geomatic Engineering, University of Calgary, Canada (1993)
3. Alesheikh, A.A.: Modeling and Managing Uncertainty in Object-Based Geospatial Information System., PhD Thesis, University of Calgary, Canada (1998)
4. Aronoff, S.: Geographical Information Systems: A Management Perspective. , WDL Publication, Ottawa (1989)
5. Blais, J.A.R. and Boulianne M.: Comparative analysis of information measures for digital image processing, Int. Archive of Photogrammetry and Remote Sensing, Vol. 27, Part B8, Commission 3, Kyoto (1998) 34-44.
6. Burrough, P.A.: Principles of Geographic Information Systems for Land Resources Assessment. , Clarendon Press, Oxford (1986)
7. Delavar, M.R.: Development of Probability Maps to Assess the Accuracy and Reliability of Information in the Output of a GIS System, Ph.D. Thesis, The University of New South Wales, Australia (1997)
8. Goodchild, M.F., Guoqing, S. and Shiren, Y.: Development and Test of an Error Model for Categorical data., Int. J.GIS, **6** (1992) 87-104.
9. Heuvelink, G.: P.A. Burrough and A. Stein: Propagation of error in spatial modeling with GIS. , Int. J. GIS. , **3** (1989) 303-332
10. Heuvelink, G.: Error Propagation in Quantitative Spatial Modeling. , Ph.D. Thesis, Universiteit Utrecht (1993)
11. Hunter G.J.: Handling Uncertainty in Spatial Database., Ph.D. Thesis, Department of Surveying and Land Information, University of Melbourne, Australia (1993)
12. NIST: Spatial Data Transfer Standard (FIPS 173), National Institute of Standards and Technology, US Department of Commerce, Washington DC (1992)
13. Obermeyer, N.J.: A systematic approach to the taxonomy of geographic information use, Proceedings of the GIS/LIS '89 Conference, Orlando, **2** (1989) 421-429
14. Veregin, H: Error modeling for the map overlay operation. In: Goodchild M .and Gopal S. (eds): Accuracy of Spatial Databases., Taylor & Francis Inc., London (1989)
15. Zhou, F.: Uncertainty Management for Object-Based Geographic Information Systems. (1995)