377

A General Framework For Error Analysis In Measurement-based GIS —— A Summary

Yee Leung

Department of Geography and Resource Management, Center for Environmental Policy and Resource Management, and Joint Laboratory for Geoinformation Science, The Chinese University of Hong Kong, Hong Kong
E-mail: yeeleung@cuhk.edu.hk

Jiang-Hong Ma

Faculty of Science, Xi'an Jiaotong University, Xi'an, and Chang'an University, Xi'an, P.R. China
E-mail: jhmath@pub.xaonline.com

Michael F. Goodchild

Department of Geography, University of California, Santa Barbara, California, U.S.A.
E-mail: good@ncgia.ucsb.edu

Abstract

This paper gives a summary of the main results obtained from a series of research on the development of a general framework for error analysis in measurement-based geographic information systems (MBGIS). The study provides a rigorous statistical approach to measurement error analysis and error propagations throughout GIS and spatial operations. It first constructs a basic measurement error model from which relevant concepts such as the approximate law of error propagation, covariance-based error band and maximal allowable limits for positional error are developed. The research then proceeds to the point-in-polygon analysis under measurement errors. Quadratic forms in the joint coordinate vectors are introduced in order to identify whether a point is inside a polygon. They are not only used in point-in-polygon analysis, but are also utilized as the identification of intersection points as well as the computation of length of a line and area of a polygon. An algebra-based probability model is constructed for such a purpose. It is simple but rigorous and can circumvent the complexities surrounding the geometric relations between points and polygons. As a consequence, simple analytic expression and an approximate law for error propagation of intersection points are established, and they are applied to polygon-on-polygon overlay and its error propagation. The general framework facilitates a formal and practical error analysis in MBGIS. It renders a consistent and effective treatment to error propagation in a variety of interrelated GIS and spatial operations

1 Introduction

Since the conception of GIS, study of errors in GIS has been rather extensive and diverse (Goodchild and Gopal, 1989; Heuvelink, 1998; Leung and Yan, 1998; Mowrer and Congalton, 2000; Shi et al, 1999; Stanislawski et al, 1996; Veregin, 1989; Wolf and Ghilani, 1997; Zhang and Goodchild, 2002). It is reckoned that different classes of spatial data exhibit different types of errors. Errors may be introduced during various stages of data manipulation and they may be propagated through GIS or spatial operations, ending up with an output consisting of certain level of inaccuracy.

Errors in spatial databases are complex and multivariate. They can generally be classified into the *inherent error* and the *operational error*. Inherent error is the error present in source documents, including the accumulated error in a map used as input to a GIS. Operational error, categorized as positional and identification errors, is produced through the data capture and manipulation functions of a GIS or is introduced during the process of data entry and occurs throughout data manipulation and spatial modeling. Since GIS databases, at the most general level, are based on a model of geographical data, errors can be classified into spatial, temporal and thematic error. From the modeling point of view, they can also be classified as either systematic or random. Systematic errors usually follow physical laws and they can be

predicted. It is however impossible to avoid entirely random errors in measurements (Wolf and Ghilani, 1997). Such random error is called measurement error (ME), which is one of the most important problems in the use of GIS data.

The epsilon band model of digitizing accuracy, for example, has been used to make estimates of the levels of positional uncertainty and ME that is due to digitizing polygon outlines (Dunn et al, 1990). Although the paper does not develop a comprehensive analytical approach to error for the conversion of map data into digital form, their empirical results on positional accuracy and ME suggest that the interaction of the scale and quality of source documents with the unique process of digitizing may introduce unexpectedly large amounts of error. The main form of this uncertainty is positional error, since it is the location of the polygon boundaries which is uncertain. But this in turn leads to ME, because the estimation of area will then be subjected to a large degree of uncertainty. There is thus an urgent need to develop methods for assessing the accuracy of vector-based GIS. Furthermore, these accuracy assessments need to be incorporated into the spatial analysis procedures, e.g., polygon overlay (the sliver polygon problem) and point-in-polygon operations, used in GIS. Compared to raster-based GIS, error analysis in vector-based GIS is much more complex.

It should be noted that a feature of conventional GIS designs is the representation of position by derived coordinates, rather than by original measurements. In such coordinate-based GIS, it is impossible to apply traditional error analysis or estimate uncertainties in derived products. It is because the process of derivation is so complex, and frequently involves human interpretation, and the original measurements are almost always lost during the process. Furthermore, the complexity of the process induces strong spatial autocorrelations between objects in the result. To overcome the difficulties in managing uncertainty intrinsic to conventional GIS, the concept of measurement-based GIS (MBGIS) has been proposed in Goodchild (1999). The basic idea is to retain details of measurements such that error analysis can be made possible, and corrections to positions can be appropriately propagated through the database.

This paper is a summary of a series of research on the rigorous development of a general framework for error analysis in MBGIS with a sound statistical foundation. Figure 1 depicts in brief the main problems of the research series around which our discussion is centered.

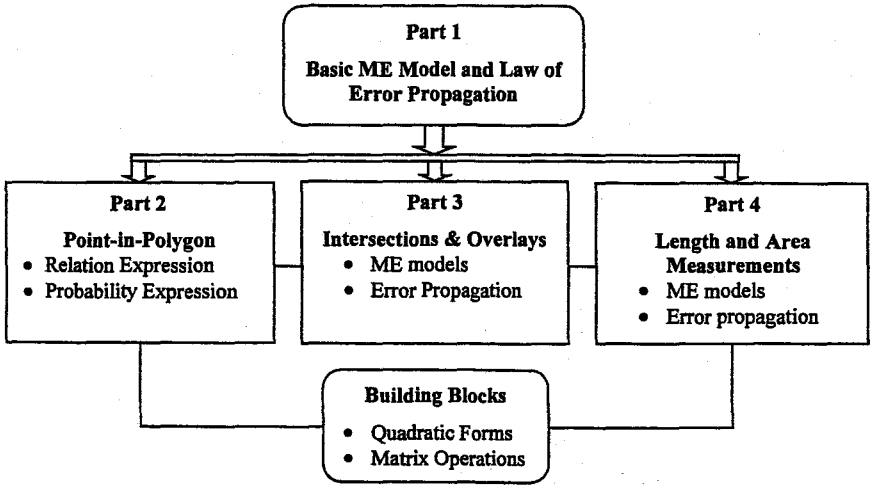

Figure 1. Project structure

2 Basic Measurement-Error Model and Related Concepts

This section summarizes the theoretical and empirical results in Leung et al (2003a).

2.1 The basic measurement-error model

Let $\mathbf{x} \in R^p$ be a set of p measurements which can be directly observed, $\mathbf{y} \in R^q$ be the quantities to be measured, and f (a well-defined geographical procedure, called the *operation function*, or more generally, the *transformation function*) be the function linking measurements $\mathbf{x}$ to quantities $\mathbf{y}$, so that

$$\mathbf{y} = f(\mathbf{x}).$$

Once $\mathbf{x}$ is measured and f is known, y and the associated error distribution are uniquely and automatically determined. The basic ME model can simply be constructed as

$$(\mathrm{I}): \quad \begin{cases} \mathbf{Y} = f(\mathbf{X}), \\ \mathbf{X} = \boldsymbol{\mu}_x + \boldsymbol{\varepsilon}_x, \ \boldsymbol{\varepsilon}_x \sim (\mathbf{0}, \boldsymbol{\Sigma}_x), \end{cases} \tag{2.1}$$

where $\boldsymbol{\mu}_x$ is the true value vector, $\mathbf{X}$ the random measurement value vector, $\mathbf{Y}$ the indirect measurement value vector obtained through f, and $\boldsymbol{\varepsilon}_x$ the random ME vector with zero mean $\mathbf{0}$ and the variance-covariance matrix $\boldsymbol{\Sigma}_x$. According to (2.1), $\mathbf{Y}$ is random and its error variance-covariance matrix $\boldsymbol{\Sigma}_y$ is propagated by $\boldsymbol{\Sigma}_x$. To simplify, the variance-covariance matrix is called the covariance matrix henceforth.

Furthermore, if a new vector $\mathbf{z} \in R^r$ can be obtained through $\mathbf{y}$ and a known function $g(\cdot)$, model (I) can again be utilized for ME analysis. As a result, error propagates along the process of operations via the basic ME model taking the general form of (I) (Figure 2).

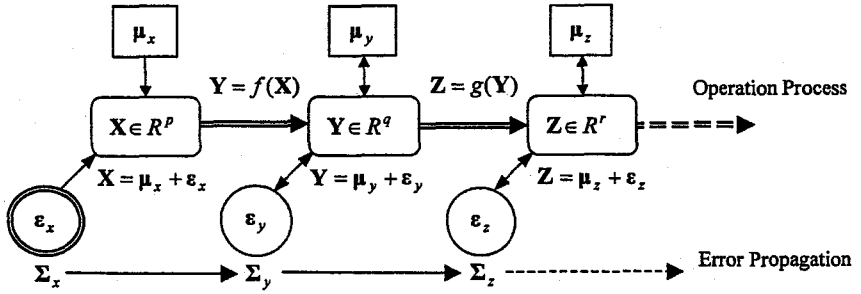

Figure 2. A flowchart for error propagation based on the basic ME model (I)

2.2 Approximate law of error propagation

In practice, the error distribution of $\mathbf{Y}$ can seldom be derived exactly from the relationship $\mathbf{Y} = f(\mathbf{X})$. When the error distribution cannot be exactly obtained or is too complex to be derived because of nonlinearity of f, a common method is to employ a linear (or second-order) approximation to $\mathbf{Y} = f(\mathbf{X})$. Then, for nonlinear f, we have the following *approximate law of error propagation*:

$$\boldsymbol{\Sigma}_y \approx \tilde{\boldsymbol{\Sigma}}_y \equiv \mathbf{B}_\mu \boldsymbol{\Sigma}_x \mathbf{B}_\mu^{\mathrm{T}}, \tag{2.2}$$

where $\mathbf{B}_\mu$ is a *Jacobian matrix* of f at $\boldsymbol{\mu}_x$. That is, $\boldsymbol{\Sigma}_y$ can be approximately expressed by $\boldsymbol{\Sigma}_x$ and $\mathbf{B}_\mu$. In general, we have the following law of error propagation:

$$\text{(II):}\quad \Sigma_y \begin{cases} \approx \widetilde{\Sigma}_y \equiv \mathbf{B}_\mu \Sigma_x \mathbf{B}_\mu^{\mathrm{T}}, & \text{if } f(\mathbf{x}) \text{ is nonlinear in } \mathbf{x}, \\ = \mathbf{B}\,\Sigma_x \mathbf{B}^{\mathrm{T}}, & \text{if } f(\mathbf{x}) \text{ is linear in } \mathbf{x}, \text{ i.e., } f(\mathbf{x}) = \mathbf{a} + \mathbf{B}\,\mathbf{x}, \\ & \quad \text{where } \mathbf{a} \text{ and } \mathbf{B} \text{ are constants,} \end{cases}$$

Although the term "law of error propagation" has been commonly used in the literature, it often lacks a rigorous mathematical expression and its usage is often unclear or sometimes incorrect. We especially stress in here the approximation and local property of this law for a nonlinear transformation function f. The emphasis on "approximation" is necessary in order to apply correctly the first-order Taylor approximation to derive (2.2). In addition, what should be pointed out is its local property as such an approximation is effective only in the local neighborhood of $\boldsymbol{\mu}_x$. Different from the notation adopted in Wolf and Ghilani (1997), Heuvelink (1998) and other publications, the Jacobian matrix $\mathbf{B}_\mu$ is appended with the subscript μ to reflect its dependence on the local point $\boldsymbol{\mu}$.

Based on the basic model (I) and the error propagation law (II), we have made three simple applications in geodesy: (1) measurement with a distance and a direction, (2) measurement with two distances, and (3) measurement with two angles, and derived their approximate laws for error propagation (see Leung et al (2003a) for details).

2.3 Covariance-based error bands

Uncertainty of a point can be derived from the covariance matrix associated with it and can be presented by an error ellipse. There are many uncertainty models for a line, such as the most commonly used epsilon band model (Perkal, 1956, 1966; Blakemore, 1984), the error band model (Shi, 1994, 1999), the ε_σ and ε_m error band models (Tong et al, 1999), and the positional uncertainty model (Alesheikh and Li, 1996; Alesheikh et al, 1999). Some of these models are either simple with no statistical rigor, or they are too complex to be lucid and effective. We formulate in this study a simple but statistically strict and unified error band model, named the "covariance-based error band", which contains error structures of the endpoints of a line segment. This turns out to be a natural and effective error band concept.

Since a line consists of points, uncertainty (i.e. the covariance matrix) of any point on the line should naturally reflect that of the line. For $i=1,2$, let $\mathbf{X}_i \equiv (X_{i1}, X_{i2})^{\mathrm{T}}$, $\boldsymbol{\mu}_i \equiv (\mu_{i1}, \mu_{i2})^{\mathrm{T}}$, and $\boldsymbol{\varepsilon}_i \equiv (\varepsilon_{i1}, \varepsilon_{i2})^{\mathrm{T}}$ be respectively the random, true, and ME vectors of the endpoint coordinates of the line segment V_1V_2. To generalize our discussion, we consider the joint ME vector $\boldsymbol{\varepsilon}_{(2)} \equiv (\boldsymbol{\varepsilon}_1^{\mathrm{T}}, \boldsymbol{\varepsilon}_2^{\mathrm{T}})^{\mathrm{T}}$ and let its covariance matrix be $\Sigma_{(2)}$, i.e.,

$$\Sigma_{(2)} \equiv \mathrm{cov}(\boldsymbol{\varepsilon}_{(2)}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Sigma_{ij} \equiv \mathrm{cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j),\ i,j=1,2,$$

where the subscript (2) indicates that we have two endpoints.

Thus for any point V' in the line segment V_1V_2, the covariance matrix of its coordinate vector $\mathbf{X}' = t\,\mathbf{X}_1 + (1-t)\mathbf{X}_2$ is derived as

$$\Sigma_t \equiv \mathrm{cov}(\mathbf{X}') = t^2\Sigma_{11} + (1-t)^2\Sigma_{22} + t(1-t)(\Sigma_{12} + \Sigma_{21}),\quad 0 \le t \le 1.$$

If we assume that $\boldsymbol{\varepsilon}_{(2)}$ is distributed as a normal distribution, then $P[\mathbf{X}_i \in R_i^{(\alpha)}] = 1-\alpha$, with

$$R_i^{(\alpha)} \equiv \{\mathbf{x}:\ (\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \le \chi^2_{2,\alpha}\},\ i=1,2,$$

where $\chi^2_{2,\alpha}$ is the upper α-quantile of the chi-squares distribution with 2 degrees of freedom. Based on $R_1^{(\alpha)}$ and $R_2^{(\alpha)}$, we can construct a larger region:

$$R_{\{1,2\}}^{(\alpha)} \equiv \{\mathbf{x}:\ \text{there is a real number } t\ (0 \le t \le 1) \text{ such that } (\mathbf{x}-\boldsymbol{\mu}_t)^{\mathrm{T}}\Sigma_t^{-1}(\mathbf{x}-\boldsymbol{\mu}_t) \le \chi^2_{2,\alpha}\},$$

where $\boldsymbol{\mu}_t \equiv t\,\boldsymbol{\mu}_1 + (1-t)\,\boldsymbol{\mu}_2$. It is obvious that $R_{\{1,2\}}^{(\alpha)}$ is the union of the confidence regions (ellipses) of all points on the line segment and is called the *covariance-based error band* for the line segment. It can be observed that when $\Sigma_{11} = \Sigma_{22} = \Sigma_{12} = \Sigma_{21} = \varepsilon^2\mathbf{I}_2$, $R_{\{1,2\}}^{(\alpha)}$ becomes the epsilon band model; when $\Sigma_{11} = \Sigma_{22} = \sigma^2\mathbf{I}_2$ and $\Sigma_{12} = \Sigma_{21} = \mathbf{0}$, $R_{\{1,2\}}^{(\alpha)}$ becomes the error band model. For different $\Sigma_{(2)}$, the covariance-based error band take on different shapes. Detailed theoretical analysis and simulation experiments can be found in Leung et al (2003a).

2.4 Maximal allowable limits for positional error

It is well known that one of the most important characteristics of spatial objects is their topological (or geometrical) property. Although we can tolerate a certain degree of ME in locations, ME must be restricted to a certain range so that it will not distort the original topology. That is, it is necessary to give a ME limit in order to guarantee that the original topology of a spatial object or topological relationship among spatial objects is invariant under ME. We call such a limit the *maximal allowable limit* (MAL). Large ME may change the topology and geometrical property of a spatial object and result in logical inconsistency or reality distortion. A simple original polygon, for example, can be distorted into a complex one if error goes above the MAL.

Let σ_{1i}^2 and σ_{2i}^2 be respectively the variances of the ME ε_{i1} and ε_{i2} of a vertex V_i, $i=1,\ldots,n$. Within our framework, the maximal allowable limits $\breve{\sigma}_{1i}$ and $\breve{\sigma}_{2i}$ for V_i are defined as the maximal allowable values of σ_{1i} and σ_{2i} so that the confidence ellipse of V_i does not intersect the covariance-based error bands of its disjoint edges. Since there is a confidence level attached to the covariance-based error band, it may also be treated as the confidence level of the MAL. Thus, as long as the variances σ_{i1}^2 and σ_{i2}^2 of ME in V_i are not greater than $\breve{\sigma}_{i1}^2$ and $\breve{\sigma}_{i2}^2$ respectively, the observed polygon in the presence of ME will not change the topological property of the true polygon with the confidence level, and their difference results only from errors in position of the vertices.

3 An Algebra-based Probability Model for Point-in-Polygon Analysis

Point-in-polygon, line-in-polygon, and polygon-on-polygon analyses are fundamental but important problems in vector-based GIS. They can arise in spatial queries or overlay operations. The basic starting point is point-in-polygon analysis (Leung and Yan, 1997, 1998; Rigaux et al, 2002). Because of the uncertainty about spatial objects such as points and polygons, the analysis becomes rather difficult and complex, especially from the geometric point of view.

How to perform point-in-polygon analysis under ME is thus a basic problem in MBGIS. Except for the conceptual discussion in Leung and Yan (1997), in-depth theoretical analysis of the issue when points and polygons are both in error has not been made. We construct in Leung et al (2003b) an algebra-based probability model for point-in-polygon analysis under ME and provide a concise and unified expression for computing the probability that a point is inside a polygon (concave or convex). Our basic idea is to first convert a point-in-polygon analysis into several point-in-triangle problems. For each triangle, the corresponding probability can then be concisely computed. The algebraic approach is not only simple and effective, it can also circumvent complications and difficulties encountered in geometric solutions of the problem. This section summarizes the theoretical and empirical results in Leung et al (2003b).

3.1 The triangle model

Under the effect of ME, let three vertex positions of the underlying triangle $\Delta V_1V_2V_3$ be $V_i(\mathbf{X}_i)$, $i=1,2,3$, the corresponding coordinate column vectors are $\mathbf{X}_i \equiv (X_{i1}, X_{i2})^{\mathrm{T}} \in R^2$, $i=1,2,3$, and $\mathbf{X}_v \equiv (X_{v1}, X_{v2})^{\mathrm{T}} \in R^2$ is the coordinate column vector of any point V. Define $\mathbf{X}_{(4)}^{\mathrm{T}} \equiv (\mathbf{X}_1^{\mathrm{T}}\ \mathbf{X}_2^{\mathrm{T}}\ \mathbf{X}_3^{\mathrm{T}}\ \mathbf{X}_v^{\mathrm{T}})$, and $\mathbf{H}_i \equiv \Delta_i \otimes \mathbf{H}_0$, $\mathbf{H}_0 \equiv \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, $i=1,2,3,4$,

$$Z_i \equiv \mathbf{X}_{(4)}^{\mathrm{T}} \mathbf{H}_i \mathbf{X}_{(4)},$$

$$\Delta_1 \equiv \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 \end{pmatrix}, \Delta_2 \equiv \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 1 & 0 \end{pmatrix}, \Delta_3 \equiv \begin{pmatrix} 0 & 1 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}, \Delta_4 \equiv \begin{pmatrix} 0 & 1 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The probability that the point V is inside the triangle region $R(V_1V_2V_3)$ is obtained as

$$P[V \in R(V_1V_2V_3)] = P[Z_i, i=1,2,3,4, \text{ have the same sign}], \tag{3.1}$$

or

$$P[V \in R(V_1V_2V_3)] = P[Z_\Delta \geq 0],$$

where $Z_\Delta \equiv Z_{\min} Z_{\max}^{-1} \geq 0$, $Z_{\min} \equiv \min_{1\leq i\leq 4}\{Z_i\}$, and $Z_{\max} \equiv \max_{1\leq i\leq 4}\{Z_i\}$.

3.2 The polygon model

Since triangles are the most essential and simple geometric objects and any polygon (whether it is convex or concave) can always be decomposed into triangles (Berg et al, 2000), we can transform the polygon problem into a multiple triangle problem on the basis of the triangle model (3.1). Thus, unlike any geometric method, the algebraic model makes the convexity or non-convexity of a polygon an non-issue.

Let V be an uncertain point with ME, $R(V_1V_2...V_n)$ be a point set representing the region of a polygon with vertices $V_1, V_2, ..., V_n$. Let a partition of a simple polygon region $R(V_1V_2...V_n)$ be

$$R(V_1V_2...V_n) = R(V_{11}V_{12}V_{13}) + R(V_{21}V_{22}V_{23}) + \cdots + R(V_{m1}V_{m2}V_{m3}),$$

where V_{i1}, V_{i2}, V_{i3} are the vertices of the triangle corresponding to $\Delta V_{i1}V_{i2}V_{i3}$. Then we can derive the following probability formula for point-in polygon analysis:

$$\begin{aligned} P[V \in R(V_1V_2...V_n)] &= P[V \in R(V_{11}V_{12}V_{13})] + P[V \in R(V_{21}V_{22}V_{23})] + \cdots + P[V \in R(V_{m1}V_{m2}V_{m3})] \\ &= P[Z_{\Delta 1} \geq 0] + P[Z_{\Delta 2} \geq 0] + \cdots + P[Z_{\Delta m} \geq 0], \end{aligned} \tag{3.2}$$

where $Z_{\Delta i} \equiv Z_{\min,i} Z_{\max,i}^{-1}$ is the random variable corresponding to $\Delta V_{i1}V_{i2}V_{i3}$.

The results can be applied to polygons with holes or objects consisting of several non-connected polygons. It should be noted that although the number of triangles resulted from triangulation may be large, the number of triangles necessary to the point-in-polygon analysis can usually be reduced since triangles outside the allowable maximal range of a point do not contribute to the probability in (3.2).

4 Error Analysis for Intersections and Overlays

Besides point-in-polygon analysis, line-in-polygon and polygon-on-polygon operations are also important in GIS. When a polygon is convex, line-in-polygon operation can be reduced to point-in-polygon operation since a line segment is inside a polygon as long as its endpoints are inside the polygon. When a polygon is non-convex, however, it involves the problem of knowing whether a line segment intersects

with the polygon. Moreover, polygon-on-polygon operations are also closely related to the intersection problems. Therefore, error analysis for intersection points is a key problem in many GIS operations. This section summarizes the theoretical and empirical results in Leung et al (2003c).

4.1 Approximate law for error propagation in intersection coordinates

Under ME, denote by $\mathbf{X}_{(4)}^{\mathrm{T}} \equiv (\mathbf{X}_1^{\mathrm{T}}\ \mathbf{X}_2^{\mathrm{T}}\ \mathbf{X}_3^{\mathrm{T}}\ \mathbf{X}_4^{\mathrm{T}})$ the joint coordinate vector of the endpoints for two line segments $V_1(\mathbf{X}_1)\ V_2(\mathbf{X}_2)$ and $V_3(\mathbf{X}_3)\ V_4(\mathbf{X}_4)$, $\overline{\mathbf{H}} \equiv \overline{\Delta} \otimes \mathbf{H}_0$, $\overline{\mathbf{H}}_i \equiv \overline{\Delta}_i \otimes \mathbf{H}_0$, $\overline{\Delta}_i \equiv \Delta_i$, $i=1,2,4$, $\overline{\Delta}_3 \equiv -\Delta_3$, and

$$\overline{\Delta} \equiv \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \end{pmatrix}, \quad \lambda_i(\mathbf{X}_{(4)}) \equiv \frac{\mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}}_i \mathbf{X}_{(4)}}{\mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}} \mathbf{X}_{(4)}}, \quad i=1,2,3,4.$$

Then the ME model for the position vector $\mathbf{X}_c$ of the intersection point V_c of two line segments $\overline{V_1V_2}$ and $\overline{V_3V_4}$ is

$$\begin{cases} \mathbf{X}_c = f(\mathbf{X}_{(4)}) = [\lambda_1(\mathbf{X}_{(4)})\mathbf{D}_1 + \lambda_2(\mathbf{X}_{(4)})\mathbf{D}_2]\mathbf{X}_{(4)}, \\ \mathbf{X}_{(4)} = \boldsymbol{\mu}_{(4)} + \boldsymbol{\varepsilon}_{(4)}, \quad \boldsymbol{\varepsilon}_{(4)} \sim (\mathbf{0}, \boldsymbol{\Sigma}_x), \end{cases} \tag{4.1}$$

where $\mathbf{D}_i \equiv \mathbf{e}_{4,i}^{\mathrm{T}} \otimes \mathbf{I}_2$ is a 2×8 matrix, $i=1,2$, $\mathbf{e}_{n,i}$ is a $n\times 1$ unit column vector whose the ith component is 1 and 0 elsewhere, $\boldsymbol{\mu}_{(4)}^{\mathrm{T}} \equiv (\boldsymbol{\mu}_1^{\mathrm{T}}\ \boldsymbol{\mu}_2^{\mathrm{T}}\ \boldsymbol{\mu}_3^{\mathrm{T}}\ \boldsymbol{\mu}_4^{\mathrm{T}})$ and $\boldsymbol{\varepsilon}_{(4)}^{\mathrm{T}} \equiv (\boldsymbol{\varepsilon}_1^{\mathrm{T}}\ \boldsymbol{\varepsilon}_2^{\mathrm{T}}\ \boldsymbol{\varepsilon}_3^{\mathrm{T}}\ \boldsymbol{\varepsilon}_4^{\mathrm{T}})$ are respectively the corresponding true and ME vectors.

Since the transformation function f in (4.1) is nonlinear in $\mathbf{X}_{(4)}$, an approximate law of error propagation for $\mathbf{X}_c$ is obtained as

$$\boldsymbol{\Sigma}_c \approx \widetilde{\boldsymbol{\Sigma}}_c = \mathbf{B}_\mu \boldsymbol{\Sigma}_x \mathbf{B}_\mu^{\mathrm{T}},$$

where $\mathbf{B}_\mu$ is a 2×8 matrix given by

$$\begin{aligned} \mathbf{B}_\mu = {} & \frac{2}{\mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}} \mathbf{X}_{(4)}} [\mathbf{D}_1 \mathbf{X}_{(4)} \mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}}_1 + \mathbf{D}_2 \mathbf{X}_{(4)} \mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}}_2] \\ & + [\lambda_1(\mathbf{X}_{(4)})\mathbf{D}_1 + \lambda_2(\mathbf{X}_{(4)})\mathbf{D}_2] \left(\mathbf{I}_8 - \frac{2}{\mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}} \mathbf{X}_{(4)}} \mathbf{X}_{(4)} \mathbf{X}_{(4)}^{\mathrm{T}} \overline{\mathbf{H}} \right). \end{aligned}$$

4.2 Approximate law for error propagation in polygon-on-polygon analysis

Due to its methodological and technical complexities, developing error propagation models for overlay operations, especially on vector-based data, has seldom been attempted. Error propagation in polygon-on-polygon operation is a typical example. A formal model is constructed in Leung et al (2003c) and is briefly described as follows:

Consider directed polygons P_i with n_i vertices $V_j^{(i)}$ listed in a counter-clockwise manner: $V_1^{(i)}$, $\cdots, V_j^{(i)}, V_{j+1}^{(i)}, \cdots, V_{n_i}^{(i)}, V_{n_i+1}^{(i)} = V_1^{(i)}$, $i=1,2$, where in the superscript (i) denotes the ith polygon and the subscript j denotes the jth vertex. Assume that vertices $V_j^{(i)}$ have the coordinates vectors $\mathbf{X}_j^{(i)}$ under ME, their true coordinates vectors are $\boldsymbol{\mu}_j^{(i)}$, and the corresponding ME vectors are $\boldsymbol{\varepsilon}_j^{(i)}$, $j=1,\ldots,n_i$,

$i=1,2$. Let $\varepsilon_{\{1-2\}}$, $\varepsilon_{\{2-1\}}$, $\varepsilon_{\{1,2\}}$, $\varepsilon_{\{1\}}$, and $\varepsilon_{\{2\}}$ be respectively the joint ME vectors of vertices coordinates counter-clockwisely listed, of the overlaid polygons $P_1 - P_2$, $P_2 - P_1$, $P_1 \cap P_2$, and the original input polygons P_1 and P_2. Denote the joint input ME vector by $\varepsilon = (\varepsilon_{\{1\}}^{\mathrm{T}}, \varepsilon_{\{2\}}^{\mathrm{T}})^{\mathrm{T}}$.

Employing the results in the approximate law for error propagation in intersection points, we can express approximately the joint output ME vectors $\varepsilon_{\{1-2\}}$, $\varepsilon_{\{2-1\}}$, and $\varepsilon_{\{1,2\}}$ as

$$\tilde{\varepsilon}_{\{1-2\}} = \mathbf{D}_{\{1-2\}}\varepsilon, \quad \tilde{\varepsilon}_{\{1,2\}}^{\mathrm{T}} = \mathbf{D}_{\{1,2\}}\varepsilon, \quad \tilde{\varepsilon}_{\{2-1\}}^{\mathrm{T}} = \mathbf{D}_{\{2-1\}}\varepsilon,$$

where the determinations of $\mathbf{D}_{\{1-2\}}$, $\mathbf{D}_{\{1,2\}}$ and $\mathbf{D}_{\{2-1\}}$ depend on the intersection points. Thus, we can obtain the following approximate laws of error propagation in polygon-on-polygon operation:

$$\tilde{\Sigma}_{\{1-2\}} = \mathbf{D}_{\{1-2\}}\Sigma\mathbf{D}_{\{1-2\}}^{\mathrm{T}}, \quad \tilde{\Sigma}_{\{1,2\}} = \mathbf{D}_{\{1,2\}}\Sigma\mathbf{D}_{\{1,2\}}^{\mathrm{T}}, \quad \tilde{\Sigma}_{\{2-1\}} = \mathbf{D}_{\{2-1\}}\Sigma\mathbf{D}_{\{2-1\}}^{\mathrm{T}}.$$

5 Error Analysis in Length and Area Measurements

In addition to the study of absolute errors for a single location, we, on the basis of the discussions in Kiiveri (1997) and Hunter and Goodchild (1996), and in light of the locational error models advanced in Leung and Yan (1998), have also performed detailed analysis of relative errors in length and area measurements in MBGIS. This section summarizes the theoretical and empirical results in Leung et al (2003d).

5.1 Error analysis in length measurements

Let $V_1(\mathbf{X}_1)$ and $V_2(\mathbf{X}_2)$ be the endpoints of a line segment under ME, $\mathbf{X}_{(2)}^{\mathrm{T}} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})$, and L the length of the line segment. Let $\mathbf{G}_{(2)} \equiv \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \mathbf{I}_2$. Then the ME model for length measurement becomes

$$\begin{cases} L = f(\mathbf{X}_{(2)}) = (\mathbf{X}_{(2)}^{\mathrm{T}}\mathbf{G}_{(2)}\mathbf{X}_{(2)})^{\frac{1}{2}} \\ \mathbf{X}_{(2)} = \boldsymbol{\mu}_{(2)} + \varepsilon_{(2)}, \ \varepsilon_{(2)} \sim (\mathbf{0}, \Sigma_{(2)}). \end{cases}$$

It is shown in Leung et al (2003d) that L^2 can be represented as

$$L^2 = \sum_{i=1}^{m} \lambda_i \chi^2_{p_i,\delta_i^2},$$

where λ_i are the distinct non-zero eigenvalues of $\mathbf{G}_{(2)}\Sigma_{(2)}$, p_i their respective orders of multiplicity, and δ_i^2 the quantities determined by $\Sigma_{(2)}$ and $\boldsymbol{\mu}_{(2)}$. Theoretically, the exact law for error propagation in L can be obtained by the statistical methods. For simplicity, we can give its approximate law of error propagation as:

$$\tilde{\sigma}_l^2 = \mathbf{B}_{\mu_{(2)}}\Sigma_{(2)}\mathbf{B}_{\mu_{(2)}}^{\mathrm{T}}, \text{ where } \mathbf{B}_{\mu_{(2)}} = (\boldsymbol{\mu}_{(2)}^{\mathrm{T}}\mathbf{G}_{(2)}\boldsymbol{\mu}_{(2)})^{-1/2}\boldsymbol{\mu}_{(2)}^{\mathrm{T}}\mathbf{G}_{(2)}.$$

We can also analyze the error of the perimeter of a polygon propagated from the ME of the vertices coordinates. Assume that the vertices of a simple n-sided polygon are $V_i(\mathbf{X}_i)$, $i=1,\ldots,n$, $V_{n+1} = V_1$. Let $\mathbf{X}_{(n)}^{\mathrm{T}} = (\mathbf{X}_1^{\mathrm{T}}, \cdots, \mathbf{X}_n^{\mathrm{T}})$ and $\mathbf{G}_{(n),i} \equiv [(\mathbf{e}_{n,i} - \mathbf{e}_{n,i+1})^{\mathrm{T}}(\mathbf{e}_{n,i} - \mathbf{e}_{n,i+1})] \otimes \mathbf{I}_2$.

The ME model for the perimeter measurement is

$$\begin{cases} L_{p(n)} = f(\mathbf{X}_{(n)}) = \sum_{i=1}^{n} (\mathbf{X}_{(n)}^{\mathrm{T}}\mathbf{G}_{(2),i}\mathbf{X}_{(n)})^{\frac{1}{2}} \\ \mathbf{X}_{(n)} = \boldsymbol{\mu}_{(n)} + \varepsilon_{(n)}, \ \varepsilon_{(n)} \sim (\mathbf{0}, \Sigma_{(n)}), \end{cases}$$

and the corresponding approximate law of error propagation is

$$\tilde{\sigma}_{l_{p(n)}}^2 = \mathbf{B}_{\mu_{(n)}}\Sigma_{(n)}\mathbf{B}_{\mu_{(n)}}^{\mathrm{T}}, \text{ where } \mathbf{B}_{\mu_{(n)}} = \sum_{i=1}^{n} [\boldsymbol{\mu}_{(n)}^{\mathrm{T}}\mathbf{G}_{(n),i}\boldsymbol{\mu}_{(n)}]^{-\frac{1}{2}}\boldsymbol{\mu}_{(n)}^{\mathrm{T}}\mathbf{G}_{(n),i}.$$

5.2 Error analysis in area measurements

For area measurement, most of the research in the literature is on raster-based data (Lloyd, 1976). Under the assumption that error at each vertex is independently and identically distributed around the local mean, effects of point error on area calculations has been reported in Chrisman and Yandell (1988). For more results, see Goodchild and Gopal (1989).

We consider in here an n-sided polygon with the vertices $V_i(\mathbf{X}_i)$ under the ME vectors $\varepsilon_i \equiv (\varepsilon_{i1}, \varepsilon_{i2})^{\mathrm{T}}$, where $\mathbf{X}_i \equiv (X_{i1}, X_{i2})^{\mathrm{T}}$, $i=1,\ldots,n$, $V_{n+1} = V_1$. The signed area of this polygon is given by

$$A_{(n)} = \frac{1}{2}\sum_{i=1}^{n} (X_{i1}X_{i+1,2} - X_{i2}X_{i+1,1}).$$

Let $\mathbf{X}_{(n)} \equiv (\mathbf{X}_1^{\mathrm{T}}, \cdots, \mathbf{X}_n^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\mu}_{(n)} \equiv (\boldsymbol{\mu}_1^{\mathrm{T}}, \cdots, \boldsymbol{\mu}_n^{\mathrm{T}})^{\mathrm{T}}$, $\varepsilon_{(n)} \equiv (\varepsilon_1^{\mathrm{T}}, \cdots, \varepsilon_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{H}_{(n)} \equiv \Delta_{(n)} \otimes \mathbf{H}_0$, and

$$\Delta_{(n)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & -1 \\ -1 & 0 & 1 & \ddots & \vdots & 0 \\ 0 & -1 & 0 & \ddots & 0 & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \vdots & \ddots & \ddots & 0 & 1 \\ 1 & 0 & \cdots & 0 & -1 & 0 \end{pmatrix}_{n\times n}, \quad n \geq 3$$

Accordingly, the ME model for the area measurement of a polygon becomes

$$\begin{cases} A_{(n)} = f(\mathbf{X}_{(n)}) = \frac{1}{4}\mathbf{X}_{(n)}^{\mathrm{T}}\mathbf{H}_{(n)}\mathbf{X}_{(n)}, \\ \mathbf{X}_{(n)} = \boldsymbol{\mu}_{(n)} + \varepsilon_{(n)}, \ \varepsilon_{(n)} \sim (\mathbf{0}, \Sigma_{(n)}). \end{cases} \quad (5.1)$$

Assume that $\varepsilon_{(n)} \sim N_{2n}(\mathbf{0}, \Sigma_{(n)})$. Then the area of the polygon $A_{(n)}$ in (5.1) can be represented as

$$A_{(n)} = \frac{1}{4}\sum_{i=1}^{m} \lambda_i \chi^2_{p_i,\delta_i^2},$$

where $\lambda_i \equiv \lambda_i(\mathbf{H}_{(n)}\Sigma_{(n)})$ are the distinct non-zero eigenvalues of $\mathbf{H}_{(n)}\Sigma_{(n)}$, p_i their respective orders of multiplicity, and $\delta_i^2 \equiv \delta_i^2(\Sigma_{(n)}, \boldsymbol{\mu}_{(n)})$ the quantities determined by $\Sigma_{(n)}$ and $\boldsymbol{\mu}_{(n)}$. The resulting exact law of error propagation is obtained as

$$\sigma^2_{A_{(n)}} \equiv \mathrm{var}(A_{(n)}) = F_n(\Sigma_{(n)}; \boldsymbol{\mu}_{(n)}) \equiv \frac{1}{8}\sum_{i=1}^{m} \lambda_i^2(\mathbf{H}_{(n)}\Sigma_{(n)})[p_i + 2\delta_i^2(\Sigma_{(n)}, \boldsymbol{\mu}_{(n)})].$$

This equation indicates that the variance $\sigma^2_{A_{(n)}}$ of area measurement for polygons is not only related to the covariance matrix $\Sigma_{(n)}$ of the locational error of the vertices, but is also related to the true locations $\boldsymbol{\mu}_{(n)}$ of the vertices. This conclusion may give a valuable insight on area measurements in general.

6 Conclusion

Corresponding to Figure 1, the main results in Leung et al (2003a-d) are summarized in Figure 3. The theoretical underpinnings and empirical analyses of the present study provide a sound and formal basis for error analysis in MBGIS. It will play an even more important role when future GIS are no longer digitized maps but maps constructed directly with multi-source data with varying measurement accuracies.

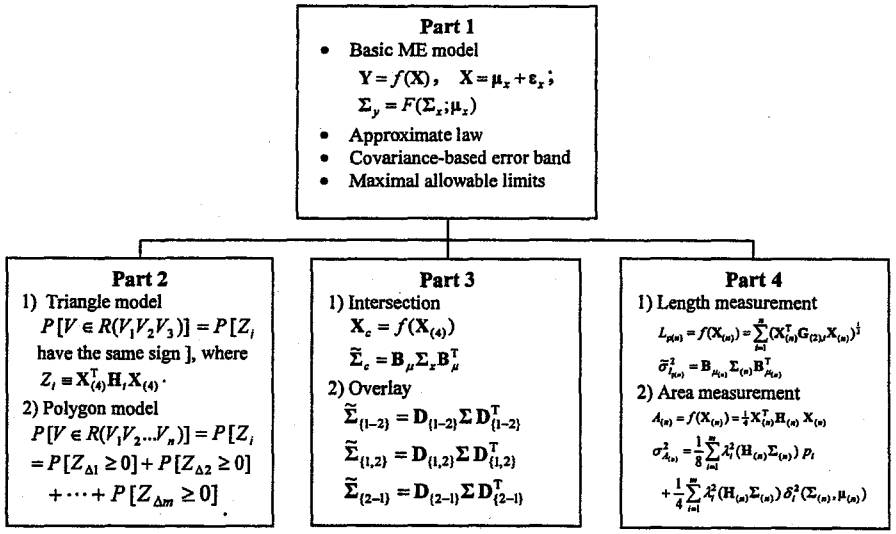

Figure 3. Basic results

Acknowledgement

This project was supported by the earmarked grant CUHK 4362/00H of The Hong Kong Research Grants Council.

References

Alesheikh, A.A., J.A.R. Blais, M.A. Chapman, and H. Karimi. 1999. Rigorous geospatial data uncertainty models for GISs. In *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*, K. Lowell and A. Jaton (eds), pp. 195-202, Chelsea, Michigan: Ann Arbor Press.

Alesheikh, A.A., and R. Li. 1996. Rigorous uncertainty models of line and polygon objects in GIS, *Proceedings of GIS/LIS '96*, Denver, CO, pp. 906-920.

Berg, M. de, M. van Kreveld, M. Overmars, and O. Schwarzkopf. 2000. *Computational Geometry: Algorithms and Applications* (2nd ed.), Berlin: Springer-Verlag.

Blakemore, M. 1984. Generalization and error in spatial data bases, *Cartographica*, 21,131-139.

Chrisman, N.R., and B.S. Yandell. 1988. Effects of point error on area calculations: a statistical model. *Surveying and Mapping*, 48(4): 241-246.

Cressie, N.A.C. 1993. *Statistics for Spatial Data, Revised Edition.* John Wiley & Sons, New York.

Dunn, R., A.R. Harrison, and J.C. White. 1990. Positional accuracy and measurement error in digital databases of land use: an empirical study. *Int. J. Geographical Information Systems*, 4(4), 385-398.

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

Goodchild, M.F. 1999. Measurement-based GIS, In Shi, W.Z., Goodchild, M.F. and Fisher, P.F. (Eds), *Proceedings of the International Symosium on Spatial Data Quality* '99, Hong Kong: Hong Kong Polytechnic University, 1-9.

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

Hunter, G.J. and M.F. Goodchild. 1996. A new model for handling vector data uncertainty in geographical information systems. *Journal of the Urban and Regional Information Systems Association*, 8(1), 51-57.

Kiiveri, H.T. 1997. Assessing, representing and transmitting positional uncertainty in maps. *Int. J. Geographical Information Science*, 11(1), 33-52

Leung, Y., and J.P. Yan. 1999. Point-in-polygon analysis under certainty and uncertainty. *GeoInformatica*, 1, 93-114.

Leung, Y., and J.P. Yan. 1998. A locational error model for spatial features. *Int. J. Geographical Information Science*, 12, 607-620.

Leung, Y., J. H. Ma, and M.F. Goodchild. 2003a. A general framework for error analysis in measurement-based GIS---Part 1: the basic measurement-error model and related concepts. (unpublished paper)

Leung, Y., J. H. Ma, and M.F. Goodchild. 2003b. A general framework for error analysis in measurement-based GIS---Part 2: the algebra-based probability model for point-in-polygon analysis. (unpublished paper)

Leung, Y., J. H. Ma, and M.F. Goodchild. 2003c. A general framework for error analysis in measurement-based GIS---Part 3: error analysis for intersections and overlays. (unpublished paper)

Leung, Y., J. H. Ma, and M.F. Goodchild. 2003d. A general framework for error analysis in measurement-based GIS---Part 4: error analysis for length and area measurements. (unpublished paper)

Lloyd, P.R. 1976. Quantisation error in area measurement. *The Cartographic Journal*, 13(1), 22-25.

Mowrer, H. T., and R. G. Congalton. (eds). 2000. *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*. Chelsea (Michigan): Ann Arbor Press.

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

Perkal, J. 1966. On the length of empirical curves, Discussion Paper Number 10, Michigan Inter-University Community of Mathematical Geography.

Rigaux, P., M. Scholl, and A.Voisard. 2002. *Spatial Databases with Application to GIS*. San Francisco : Morgan Kaufmann Publishers.

Shi, W. 1994. *Modeling Positional and Thematic Uncertainties in Intergration of Remote Sensing and GIS*, Ph.D. Thesis, ITC Publication No. 22. ITC, The Netherlands.

Shi, W., M. Ehlers, and K. Tempfli. 1999. Analytical modeling positional and thematic uncertainties in the intergration of remote sensing and geographical information systems, *Transactions in GIS*, 3(2): 119-136.

Stanislawski, L.V., B.A. Dewitt, and R. S. Shrestha 1996. Estimating positional accuracy of data layers within a GIS through error propagation. *Photogrammetric Engineering and Remote Sensing*, 62(4), 429-433.

Tong, X., W. Shi, and D. Liu. 1999. Development of error models for transition curves in GIS. In Shi, W.Z., Goodchild, M.F. and Fisher, P.F. (Eds), *Proceedings of the International Symosium on Spatial Data Quality* '99, Hong Kong: Hong Kong Polytechnic University, 299-307.

Veregin, H. 1989. A taxonomy of error in spatial databases. *Technical Paper 89-12, National Center for Geographic Information & Analysis*, Geography Department, University of California, Santa Barbara, California.

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

Zhang, J., and M.F. Goodchild. 2002. *Uncertainty in Geographical Information.* New York: Taylor and Francis.